Prédiction de la progression du myélome multiple par imagerie TEP : Adaptation des forêts de survie aléatoires et de réseaux de neurones convolutionnels

Ludivine Morvan

CENTRALE NANTES

Thèse de doctorat de

L'École Centrale de Nantes

École Doctorale N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
Ludivine MORVAN

Prédiction de la progression du myélome multiple par imagerie TEP : adaptation des forêts de survie aléatoires et de réseaux de neurones convolutionnels.

Thèse présentée et soutenue à l'Ecole Centrale de Nantes, le 7 décembre 2021
Unité de recherche : UMR 6004, Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Su RUAN	Professeure des universités, Université de Rouen-Normandie
Mathieu HATT	Directeur de recherche, Université de Bretagne Occidentale, Brest

Composition du Jury :

Présidente :	Carole LARTIZIEN	Directrice de recherche - CNRS, Université de Lyon
Examinateur :	Clovis TAUBER	Maître de conférences HDR, Université de Tours
Dir. de thèse :	Diana MATEUS LAMUS	Professeure des universités, Ecole Centrale de Nantes
Co-dir. de thèse :	Thomas CARLIER	Docteur - HDR, CHU de Nantes

Invitée :

Françoise KRAEBER-BODERE	Professeure, CHU de Nantes et ICO

Remerciements

Ce travail a été soutenu en partie par le Fonds européen de développement régional, la région des Pays de la Loire sur le programme Connect Talent MILCOM (Multi-modal Imaging and Learning for Computational-based Medicine), Nantes Métropole (Convention 2017-10470), et l'Agence nationale de la recherche appelée "Investissements d'Avenir" Labex IRON no ANR-11-LABX-0018-01 et INCa-DGOS-Inserm_12558 (SIRIC ILIAD). Ces travaux n'auraient pas étaient possibles sans les personnes ayant contribué à la production et à la labellisation des données utilisées lors de cette thèse, ainsi que celles m'ayant permis de les utiliser.

Je tiens avant tout à remercier mes directeurs de thèse, Pr. Diana Mateus et Thomas Carlier, pour avoir proposé ce projet, géré le financement et orienté le contenu scientifique de cette thèse. Je les remercie également de m'avoir fait confiance, et pour avoir passé des week-ends et des soirées à m'aider dans la rédaction des papiers et de ce manuscrit.
Mes remerciements vont également à mes collègues de l'équipe SIMS pour m'avoir accueilli chaleureusement lors de ces trois années, et les doctorants des sessions de "Deep Reading" pour ces échanges fort intéressants et instructifs.

Je remercie Pr. Su Ruan et Mathieu Hatt pour avoir accepté d'être rapporteurs. Je remercie également les membres du jury (et invités) Caroline Lartizian, Clovis Tauber et Pr. Françoise Kraeber-Bodéré.

Je souhaiterais exprimer ma gratitude aux professeurs de mon école d'ingénieur (ISBS), et notamment Pr. Hugues Talbot, qui m'a convaincu de me lancer dans ce domaine et dans la recherche. Il fut aussi d'une grande aide pour la recherche de cette thèse et des stages la précédant.

Inger Persson and the other members of AlgoDx allow me to approach the end of my thesis with peace of mind by giving me the opportunity to work with them next year on a wonderful project.

Mes amis ont eu une grande influence sur la réussite de cette thèse. Un grand merci à mes collègues de bureau, Diane et Vanessa pour ces discussions fort passionnantes. A elles mais aussi à Guillaume, Maël, Valentin et Samuel de m'avoir donné envie de venir au laboratoire et de continuer même lorsque les journées et semaines étaient difficiles. Merci particulièrement à Guillaume et Vanessa pour ces débats autour de la politique et du sens de la vie ! Merci à Adeline pour m'avoir soutenue pendant toutes ses années et bonne chance pour tes projets futures !

Je suis redevable à ma famille, et notamment mes parents, mon frère et mes grand-mères, pour leur soutien et pour la relecture du manuscrit. Enfin, merci à Julian pour sa patience et ses encouragement et pour m'avoir supporté pendant plus de deux ans.

Résumé

Contexte :

L'objectif de ces travaux est la conception de modèles par estimation statistique permettant la prédiction de la survie et l'identification de biomarqueurs dans le contexte du myélome multiple, à l'aide de l'imagerie TEP (Tomographie à émission de positons) et de données cliniques. Cette thèse est divisée en deux parties : la première fut dédiée à la création d'un modèle basé sur les forêts de survie aléatoires (RSF), et la seconde à l'adaptation de réseaux de neurones convolutionnels à la survie.

Méthodes :

Les RSF, bien que devenues la méthode de référence dans l'analyse de survie, n'ont jamais été étudiées dans le contexte du myélome multiple et des images TEP. Une première méthode basée sur les RSF est donc conçue, englobant la prédiction de la progression et la sélection de variables. Elle comporte trois étapes : une étape d'optimisation automatique des paramètres des arbres, une étape de prédiction de l'ordonnancement des variables les plus prédictives par VIMP (Variable Importance), et une étape de prédiction par RSF avec ces variables les plus prédictives. Ce modèle fournit la survie des patients, une classification en deux groupes de risque et une liste de biomarqueurs par ordre d'importance. En entrée du modèle sont utilisées les valeurs radiomiques et volumiques des images TEP sur la lésion focale la plus fixante et les données cliniques.

La deuxième méthode proposée, appelée M2P2, est un réseau de neurones convolutionnels (CNN) 3D qui apprend à extraire les valeurs radiomiques les plus pertinentes à l'aide d'un module d'attention sur les filtres. Pour entraîner le modèle à prédire des risques nous faisons une étude de fonctions de coût adaptées à la survie, en proposant deux nouvelles basées sur le contraste de risques entre les éléments d'un triplet. C'est nouvelles fonctions sont aussi efficaces pour apprendre un espace de caractéristiques discriminant pendant une étape de pré-entraînement du réseau. Après entraînement, M2P2 permet lui aussi de prédire la survie et un groupe de risque.

Deux bases de données prospectives multi-centriques, l'une française (IMAJEM) [1] et l'autre italienne (EMN02/HO95) [2], sont utilisées pour entraîner et valider nos modèles.

Résultats :

Nous avons montré que notre modèle basé sur les RSF est plus performant que les modèles classiques (erreur de prédiction de 0,36 au lieu de 0,56 pour Lasso-Cox et 0,48 pour Gradient-Boosting Cox). Nous avons aussi montré la prédictibilité et la stabilité de la sélection par VIMP comparée aux autres méthodes de sélection (0,47 et 0,43 pour "Minimal Depth" et "Variable-Hunting" respectivement). Parmi les modèles testés, notre modèle est le seul à permettre une séparation significative des patients en groupes pronostiques (p-value inférieure à 0,05). Finalement, nous avons montré l'intérêt d'utilisation des radiomiques en combinaison des données cliniques comme biomarqueurs d'intérêt dans un

contexte pronostique du myélome multiple.

Concernant l'apprentissage profond, nous avons créé un modèle prenant en compte les problématiques de petites lésions de tailles variables, du nombre limité de patients et de la survie. Nous avons montré l'intérêt de l'utilisation du modèle CNN (c-index de 0,57) et plus particulièrement le CNN 3D (c-index de 0,60), comparé aux méthodes de prédiction de survie classiques (c-index maximal de 0,51). Nous avons aussi prouvé l'intérêt de l'attention sur les filtres (c-index de 0,63 au lieu de 0,61) et des méthodes de pré-entraînement par classification binaire puis contrastif (c-index de 0,66 au lieu de 0,61), et testé d'autres comme l'attention spatiale et la SPP (Spatial Pyramidal Pooling).

Contributions :

Les contributions principales de la thèse sont les suivantes :

— Conception d'un modèle basé sur les RSF et les images TEP permettant la prédiction d'un groupe pronostique pour les patients atteints de myélome multiple.

— Détermination de biomarqueurs grâce à ce modèle.

— Démonstration de l'intérêt des caractéristiques radiomiques des images TEP.

— Production d'un modèle CNN 3D, appelé M2P2, permettant la prédiction de la survie et de groupes de risques des patients atteints de myélome multiple.

— Extension de l'état de l'art des méthodes d'adaptation de l'apprentissage profond à un nombre limité de patients et à de petites images.

— Étude des fonctions de coût utilisées en survie, et proposition de nouvelles.

De plus nous sommes, à notre connaissance, les premiers à avoir étudié l'utilisation des RSF dans le contexte du myélome multiple et des images TEP, à étudier la survie du myélome multiple grâce à des CNN, à utiliser du pré-entraînement auto-supervisé et contrastif avec des images TEP et avec une tâche de survie, et, à adapter la fonction de coût triplet à la survie.

Mot clés : Myélome multiple, Réseaux de neurones convolutifs, Analyse de survie, Forêts de survie aléatoires, Tomographie à Emission de Positons

Table des matières

Table des figures

Liste des tableaux

Notations

Générales

N	Le nombre d'individus
N_c	Le nombre de caractéristiques
$\mathcal{X}$	L'ensemble de données tel que $\mathcal{X} = \{\mathbb{X}, \mathbb{Y}\}$
$\mathbb{X}$	Matrice de données contenant N images $\mathbb{X}_i$ ou vecteurs de caractéristiques $\mathbf{x}_i$
$\mathbf{x}_i$	Vecteur de caractéristiques pour le patient i avec $\mathbf{x}_i = \{x_{i1}, ..., x_{ij}, ..., x_{iN_c}\}$
$\mathbb{X}_i$	Image du patient i
$\mathbb{X}_\text{t}$	Images du set d'entraînement et validation d'une validation croisée. $\mathbb{X}_\text{t} = \mathbb{X}_\text{train} + \mathbb{X}_\text{val}$
$\mathbb{X}_\text{train}$	Images du set d'entraînement
$\mathbb{X}_\text{val}$	Images du set de validation
$\mathbb{X}_\text{test}$	Images du set de test
$\mathbb{M}_i$	Masque pour le patient i
$\mathbf{x}_j$	j$^{\text{ème}}$ caractéristique
$\mathbb{Y}$	L'ensemble des données de survie tel que $\mathbb{Y} = \{t_i, \delta_i\}_{i=1}^N$
t_i	Temps de survie d'un individu i
δ_i	Censure d'un individu i
B	Taille du batch

Relatives à la survie

S_i	Fonction de survie d'un individu i
h	Fonction de hasard/risque
H	Fonction de risque cumulé
t_j	Les temps d'évènement avec $j \in \{1, ..., J\}$
Υ_j	Nombre d'individus à risque au temps t_j
d_j	Nombre d'évènements au temps t_j
$\hat{p_j}$	Probabilité d'avoir un évènement dans l'intervalle $]t_{j-1}$,$t_j]$ sachant que l'on était vivant en t_{j-1}
β	Paramètre d'intérêt dans le modèle de Cox qui représente l'effet des covariables sur le risque instantané

Relatives au calcul des radiomiques

Φ	Angle
X_v	Ensemble de N_v voxels inclus dans la région d'intérêt (ROI)
X_g	Ensemble des N_g intensités discrétisées (niveaux d'intensité) des N_v voxels de la ROI
$\mathcal{H}$	Histogramme des fréquences d'apparition n_i de chaque intensité discrétisée i dans X_g
$\mathcal{P}_i$	Probabilité d'occurrence de chaque intensité discrétisée i

Relatives aux méthodes d'arbres

A	Prédicteur ou arbre
A_{rf}	Prédicteur/arbre d'une forêt aléatoire
q	Le nombre de prédicteurs/arbres
Ω	Le noeud d'intérêt
X_{rank}	Vecteur contenant les variable ordonnancées (de la valeur la plus prédictive à la moins prédictive)
n_f	Le nombre de variables à garder testé
n_f^*	Le nombre optimal de variables à garder
$MaxVar$	Le nombre maximal de variables dans le noeud.
$\hat{M}_i$	Mortalité prédite de l'individu i
$\hat{M}_{th}$	Seuil de séparation calculé sur la mortalité lors du test du Log-Rank
$\hat{M}_{th}^*$	Seuil de séparation optimal calculé sur la mortalité lors du test du Log-Rank
err	Erreur de prédiction (1- c-index)
c_j	Valeur de la caractéristique $\mathbf{x}_j$
c_j^*	Valeur optimale de séparation de la caractéristique $\mathbf{x}_j$ dans le noeud Ω
g_{param}	Valeur des paramètres d'arbres testées
g_{param}^*	Valeur des paramètres d'arbres testées

Relatives à l'apprentissage profond

θ	Poids dans le réseau d'apprentissage profond
φ	Fonction d'activation

$W_{ij}^{(k-1)(k)}$	Poids entre le $i^{ème}$ neurone de la couche k-1 et le $j^{ème}$ neurone de la couche k
C	Nombre de cartes de caractéristiques
$\mathbb{V}$	Dernière couche de convolution du modèle
$\mathbb{Z}$	Cartes de caractéristiques
P_p	"Pooling" fixe avec une taille de sortie de $p \times p \times p \times C$
L	Vraisemblance partielle
l	Fonction de coût
$\mathbb{T}$	Vecteur de D Intervalles de temps Δ
$\mathbf{s}_i$	Vecteur des temps de survie (binaire où les intervalles avant le temps de survie sont remplis de uns, et de zéros après)
$\mathbf{e}_i$	Vecteur temps-événement (un Dirac dans l'intervalle où un événement s'est produit, ou rempli de zéros en cas de censure)
α et λ	Constantes permettant de pondérer les fonctions de coûts lors de fonctions combinées
$(\mathbf{x}_k^a, \mathbf{x}_k^p, \mathbf{x}_k^n)$	Triplet contenant une ancre, un positif et un négatif
$(\mathbf{f}_k^a, \mathbf{f}_k^p, \mathbf{f}_k^n)$	Le vecteur de caractéristiques profondes prédites correspondant au triplet $(\mathbf{x}_k^a, \mathbf{x}_k^p, \mathbf{x}_k^n)$
$\mathcal{D}$	Distance euclidienne
(t_k^a, t_k^p, t_k^n)	Temps de survie de l'ancre, le positif et le négatif d'un triplet

Abbreviations

AR	Absolute Resampling
ASCT	Autologous Stem Cell Transplantation
AUC	Area Under Curve
B3D	Baseline 3 Dimensions
BMI	Bone Marrow Involvement
CART	Classification And Regression Trees
CBAM	Convolutional Block Attention Module
CHF	Cumulative Hazard Function
CI	Confidence Interval
CNN	Convolutional Neural Network
EJNMMI	European Journal of Nuclear Medicine and Molecular Imaging
EMD	Extra-Medullar Disease
FDG	Fluorodésoxyglucose
IBSI	Image Biomarker Standardisation Initiative
IJCARS	International Journal of Computer Assisted Radiology and Surgery
IMWG	International Myeloma Working Group
IPCW	Inverse Probability Censoring Weighted
IRM	Imagerie par Résonance Magnétique
LF	Lésions Focales
M2P2	Muliple Myeloma Progression-free survival with PET images
MILCOM	Multi-modal Imaging and Learning for Computational-based Medicine
MM	Myélome Multiple
MSE	Mean Square Error
MTV	Metabolic Tumor Volume
NN	Neural Network
PFS	Progression Free Survival
OAF	Osteoclast Activating Factors
OOB	Out-Of-Bag
ReLU	Rectified Linear Unit
RF	Random Forest
R-ISS	Revised - International Staging System

ROI	Region Of Interest
RR	Relative Ressampling
RSF	Random Survival Forest
RVD	Lenalidomide, bortezomib, et dexamethasone
SDG	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-Sampling technique
SNR	Signal-to-Noise Ratio
SPP	Spatial Pyramidal Pooling
SUV	Standardized Uptake Value
SVM	Support Vector Machine
SV-triplet	Scale-Varying triplet
TDM	Tomodensitométrie
TEP	Tomographie à Emission de Positons
TLG	Total Lesion Glycolysis
T-SNE	T-Distributed Stochastic Neighbor Embedding
TTA	Test-Time Augmentation

Première partie

Introduction et contexte

Chapitre 1

Introduction

La thèse fut réalisée grâce à une collaboration du CRCINA (Centre de recherche en cancérologie et immunologie Nantes-Angers)/Inserm, du Ls2n (Laboratoire des sciences numériques de Nantes)/Ecole Centrale de Nantes, et de l'équipe de médecine nucléaire du CHU (Centre Hospitalier Universitaire) de Nantes. Mes travaux rentrent dans le projet MILCOM (Multi-modal Imaging and Learning for Computational-based Medicine) dont l'objectif est d'aider les médecins à poser un diagnostic, dans le but d'un traitement personnalisé, grâce à des informations plus riches et propres à chaque patient. Mes travaux ont en particulier pour objectif de développer des algorithmes d'apprentissage automatique pour lier de façon quantitative et reproductible les images TEP (Tomographie à Emission de Positon), et la survie de patients atteints de myélome multiple. Le myélome multiple est un cancer de la moelle osseuse caractérisé par un taux de survie à 5 ans d'environ 50% [4] et un haut taux de rechute. Les chances de survie dépendent de la rapidité de prise en charge avec le traitement approprié. Le modèle développé a pour but de déterminer directement quel profil de patient est le plus à risque et donc de potentiellement adapter le traitement en fonction de sa situation clinique. Ainsi le travail réalisé lors de cette thèse est une analyse de survie des patients atteints de myélome multiple à partir de deux bases de données provenant de deux études cliniques. L'analyse de survie peut être définie comme la prédiction du temps écoulé jusqu'à la survenue d'un événement précis. Elle regroupe de nombreuses méthodes qu'elles soient statistiques ou automatiques. La méthode la plus connue est celle de Cox qui permet de prédire un risque et qui est employée lorsque le but est d'évaluer les effets des covariables sur le temps de survie. Si l'on veut simplement comparer la survie de deux populations/groupes (bas et haut risque, hommes/femmes), on peut utiliser la méthode de Kaplan-Meier. Cette méthode a l'avantage d'être simple d'utilisation et d'interprétation, et ne nécessite aucune hypothèse sur les distributions de survie.

Cependant, dans la littérature les Forêts de survie aléatoire (RSF) [5] sont devenues une référence car celles-ci sont moins dépendantes du taux de censure que la méthode de Cox par exemple. C'est une méthode de prédiction de survie à partir d'un ensemble d'arbres de décision qui prend en compte la censure à droite et les données manquantes. Nous avons donc proposé un modèle basé sur ces RSF, où nous donnons en entrée des caractéristiques radiomiques calculées sur des images TEP et des caractéristiques cliniques. Il en ressort une prédiction de la survie du patient mais aussi l'importance des caractéris-

tiques dans le calcul de cette prédiction. Les techniques mises en œuvre sont relativement récentes et originales dans le contexte du myélome multiple. L'élaboration du modèle basé sur les RSF et son analyse sont présentées dans la partie 3.4.2. Deux articles ont été publiés à partir de ces travaux : un article technique dans International Journal of Computer Assisted Radiology and Surgery (IJCARS) [6] et un article applicatif/médical dans l'European Journal of Nuclear Medicine and Molecular Imaging (EJNMMI) [7].

Cependant, il existe maintenant des méthodes d'apprentissage profond, qui ont permis de grande avancées dans de nombreux domaines et notamment en classification et en segmentation [8, 9]. Leur utilisation dans la prédiction de survie reste encore relativement récente et peu présente, bien que celle-ci ait beaucoup augmenté depuis 2016. L'utilisation de l'apprentissage profond pour l'analyse de survie se résume parfois à une classification ou une régression ne prenant pas en compte la censure [10], ou a une extraction de caractéristiques profondes suivie d'une méthode d'apprentissage machine classique (Cox ou RSF) [11]. Les papiers prenant en compte la censure dans la fonction de coût utilisent pour la plupart une simple adaptation de la fonction de Cox [12]. Or de nouvelles fonctions furent proposées [13, 14] et nous souhaitons savoir si celles-ci sont plus prédictives que la fonction de coût Cox. Nous proposons aussi des adaptations de fonctions contrastives par triplets que nous adaptons à la survie.
De plus, la tâche de survie n'est pas la seule limitation à l'utilisation de l'apprentissage profond pour l'analyse de notre base de données. En effet, nous sommes dans un contexte d'études prospectives (et donc soumis à un nombre limité de données), d'images TEP de basse résolution et de petites lésions de tailles variables. Or, l'apprentissage profond nécessite un grand nombre de données et les modèles sont réalisés pour des images de grande taille à haute résolution. Nous avons donc dû composer avec toutes ces limitations afin d'élaborer un modèle d'apprentissage profond permettant l'analyse de survie de nos bases de données. Nous avons également évalué différents modules, présent dans la littérature ou adaptés à notre contexte. Nous avons enfin évalué différentes stratégies permettant de résoudre la question du sur-apprentissage. Ces modèles élaborés sont présentés et analysés dans la partie 7.2. Un premier modèle a été publié dans un article du workshop PRIME de la conférence internationale International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020 [15], et une adaptation de celui-ci est en cours de révision. Ce nouvel article comprend aussi une revue des fonctions de coût de survie, et des stratégies d'adaptation aux petites images et aux petites bases de données (avec notamment les stratégies de pré-entraînement).

Ainsi, nous sommes les premiers à travailler sur l'analyse du survie du myélome multiple grâce aux forêts de survie mais surtout en utilisant de l'apprentissage profond. Nous avons fourni dans chaque cas un modèle et leur analyse approfondie. A notre connais-

sance nous sommes aussi les premiers a avoir utilisé certaines approches dans le contexte de l'analyse de survie (fonction de coût triplet, le pré-entraînement auto-supervisé contrastif, le CBAM (Convolutional Block Attention Module)).
Les contributions et les travaux réalisés sont résumés dans la figure 1.1.

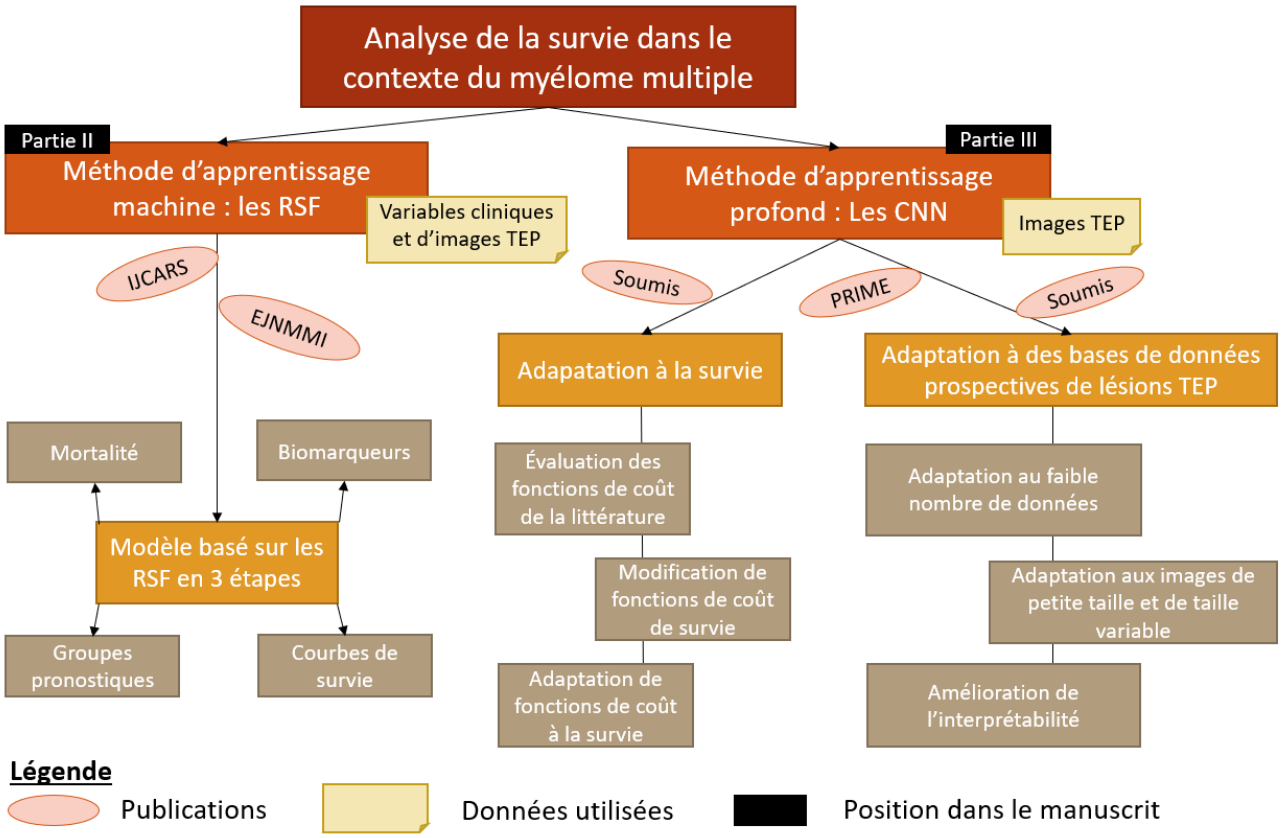


FIGURE 1.1 – Résumé des contributions.

Chapitre 2

Contexte clinique

La thèse se déroule dans le contexte clinique du myélome multiple. Ce cancer hématologique est caractérisé par la multiplication dans la moelle osseuse de plasmocytes anormaux. Les conséquences, présentées dans la figure 2.1, sont :

— L' affaiblissement du système immunitaire en raison de la diminution du nombre de plasmocytes normaux.

— La diminution de la production des cellules sanguines.

— La stimulation de la résorption osseuse ostéoclastique par sécretion de facteurs OAF (Osteoclast Activating Factors). Cela peut ainsi engendrer une hypercalcémie et donc 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 et induire une insuffisance rénale [16].

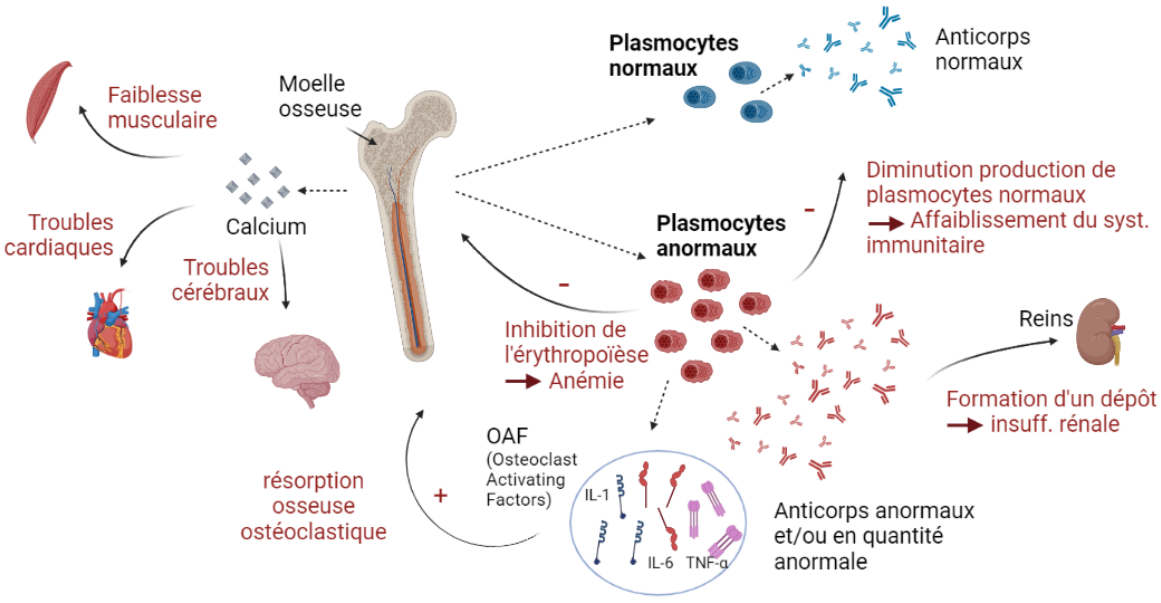


Figure 2.1 – Le myélome multiple est un cancer de la moelle osseuse se caractérisant par la prolifération de plasmocytes anormaux qui induit différents troubles. Figure crée avec BioRender.com.

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% [4].

Pour accroître les chances de survie il faut détecter et soigner au plus tôt la maladie avec le bon traitement. La détection des lésions et le diagnostic peuvent être faits par 18F-FDG (fluorodésoxyglucose) TEP. Cette dernière fut ajoutée aux critères révisés IMWG (International Myeloma Working Group). Une machine TEP/TDM (Tomodensitométrie) est présentée dans la figure 2.2A). Bailly et al. [17] détaillent en profondeur la TEP diagnostique. Les images 3D obtenues avec la 18F-FDG 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 en plus par rapport aux autres techniques d'imagerie [18]. A la différence de l'IRM, la FDG-TEP à l'avantage d'être quantitative. En effet, la fixation du FDG est caractérisée par une indice, le SUV (Standardized Uptake Value), qui prend en compte l'activité injectée et la masse du patient.

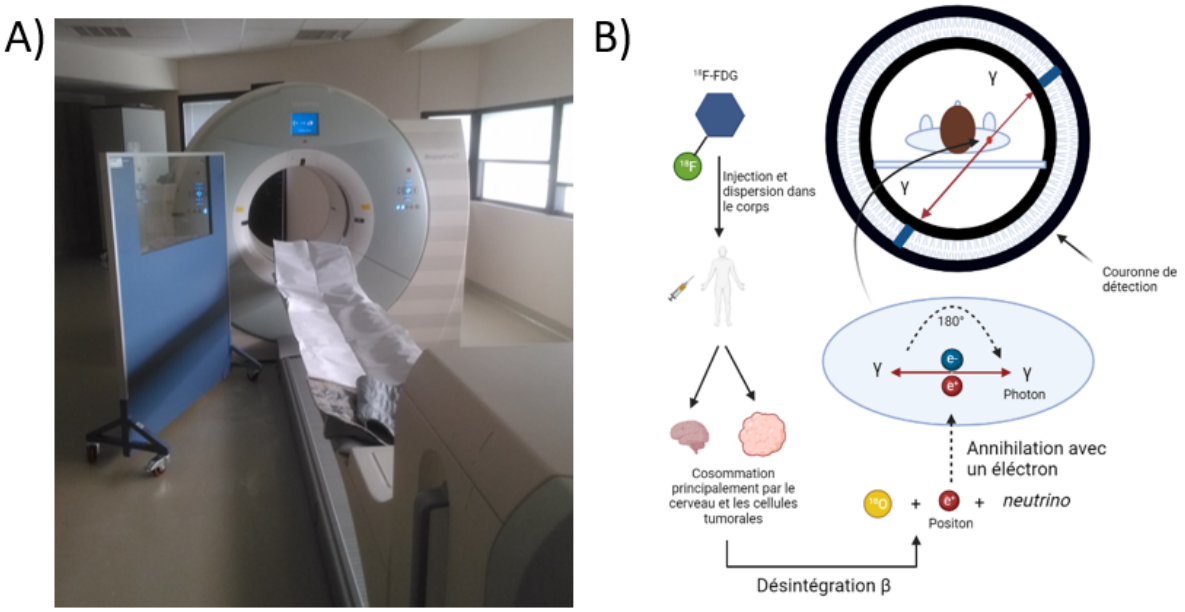


FIGURE 2.2 – A) TEP/TDM(Biograph mCT, SIEMENS) au CHU Nantes. B) Fonctionnement de la 18F-FDG TEP. Les positons vont s'annihiler avec des électrons ce qui produira des photons γ émis à 180°. Ce sont ces photons γ qui seront détectés et donneront une ligne de réponse. Figure crée avec BioRender.com.

La méthode du 18F-FDG TEP résumée dans la figure 2.2B 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 β^+ en formant des positons [voir équation 2.1](et dans 3% des cas

par capture électronique).

$$^{18}_{9}F \rightarrow ^{18}_{8}O + ^{0}_{1}e + neutrino \tag{2.1}$$

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 deux photons γ émis à 180°. Ce sont ces photons γ qui seront détectés par des détecteurs opposés et donneront donc une ligne de réponse. En utilisant ensuite des algorithmes de reconstruction tomographiques adaptés, 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 la distribution du médicament radiopharmaceutique dans le corps du patient dans un espace à trois dimensions.
Le fluor 18 a pour avantage d'avoir une période radioactive raisonnablement longue (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 mm pour l'oxygène-15 par exemple) ce qui permet d'obtenir une carte de fixation radioactive plus proche de la réalité. L'association de la TEP à un TDM, donne lieu à une identification plus aisée des lésions par les médecins dans le cadre du myélome multiple car cela met en évidence les zones à activité se trouvent sur une région osseuse ou non. Un exemple d'images TEP, TDM et TEP/TDM est présenté dans la figure 2.3. Ces images TEP peuvent être utilisées pour le diagnostic et le suivi de la maladie de même que dans le cadre de la recherche.

Ce travail se place dans le contexte de deux études cliniques prospectives, randomisées et multicentriques sur un traitement contre le myélome multiple. Les études, IMAJEM [1] et EMN02/HO95 [2] comportent respectivement 134 et 192 patients. Respectivement 18 et 8 centres ont participé, ce qui induit que nous sommes dans le cas d'études multicentriques. Dans les deux études, pour chaque patient, des images TEP au diagnostic et TDM sont disponibles ainsi que les données cliniques des patients (âge, hémoglobine, traitement, etc.), leur temps de survie depuis la détection de la maladie ou OS (overall survival), et celui sans progression (PFS). Cependant, nous nous sommes concentrés dans ces travaux sur la PFS, étant donné que le taux de censure y est plus faible qu'avec l'OS. La progression de la maladie est définit par l'IMWG (International Myeloma Working Group) comme une augmentation d'au moins 25% d'une valeur de valeurs listées (concentration de la protéine M dans le sérum, dans l'urine, pourcentage de plasmocytes dans la moelle osseuse, etc.) par rapport à la réponse la plus basse. Les données de survie sont disponibles jusqu'à 7 ans. Ces bases de données sont issues d'études cliniques prospectives. Elles ont l'avantage, par rapport aux études rétrospectives, d'être relativement homogène en terme de protocole clinique, de caractéristiques relevées et de réalisation des images TEP.

FIGURE 2.3 – Exemples d'images TDM (gauche), TEP (centre), et TEP/TDM fusionnées (droite). La délimitation verte correspond à la ROI d'une lésion focale. La délimitation rouge correspond à la ROI d'un envahissement diffus de la rate.

Cependant, une étude prospective sera potentiellement plus limitée en nombre de patients. En effet, il ne sera pas possible d'ajouter les patients ayant étaient diagnostiqué/traité avant le début de l'étude ou après. Enfin le taux de censure sera possiblement plus élevé car les évènements ayant eu lieu après la fin de l'étude ne seront pas pris en compte.

Les images FDG-PET/CT ont été acquises dans chaque centre selon leur procédure locale. En résumé, tous les patients étaient à jeun pendant 4 heures avant l'acquisition et le taux de glucose sanguin devait être $\leq 150 mg/dL$. La TEP/TDM-FDG du corps entier a été réalisée entre 54 et 80 minutes après l'injection de 3-7 MB. Les protocoles de reconstruction d'images cliniques de routine ont été utilisés dans chaque centre en utilisant leurs propres paramètres pour la TEP et la TDM. La grille de voxels utilisée pour les images reconstruites, qui est d'importance lorsqu'on traite de la de la délimitation de la tumeur et du calcul des caractéristiques texturales, variait de $(2,7 \times 2,7 \times 3,3)$ à $(5,5 \times 5,5 \times 3,3)$

mm3. Après acquisition, les images TEP ont été récupérées à partir de dcm corps entier. Une ROI (Region Of Interest) a été créée à partir du logiciel Dosisoft pour chaque lésion focale et notamment les plus fixantes (nous nous intéressons uniquement à ces dernières). L'information contenue dans la ROI des lésions les plus fixantes peut avoir un grand intérêt pronostique. Elle peut donc être utilisée, en parallèle des données cliniques, pour déterminer un parcours de soins personnalisé pour chaque patient en fonction de son risque de progression avec chaque traitement.

L'objectif de cette thèse est la création de modèles permettant de prédire la PFS des patients atteints de myélome multiple et la détermination des biomarqueurs de la progression de la maladie, à partir de ces données cliniques et de l'information contenue dans cette ROI des lésions les plus fixantes calculée sur les images TEP. Nous utilisons la TEP de diagnostic car le but est déterminer le meilleur traitement dès le départ et ainsi augmenter les chances de survie. Les bases de données de myélome multiple étant rares et inexistantes publiquement à notre connaissance, nous sommes contraint à l'utilisation de ces deux seules bases de données. Dans la détermination de ces modèles, nous seront donc limités par la censure qui est à un taux de 45% (45% n'ont pas eu d'évènement de progression avant la fin de leur suivi ou de l'étude clinique), par les caractéristiques de ces lésions (taille faible) et des images TEP (faible résolution), et par le nombre de patients.

Chapitre 3

Arrière-plan scientifique

L'objectif de nos travaux étant de réaliser l'analyse de survie dans le contexte du myélome multiple, je présenterai dans ce chapitre l'arrière-plan scientifique, d'abord de l'analyse de survie puis des différentes méthodes de référence par estimation statistique et automatique. Dans les méthodes d'apprentissage automatique nous parlerons essentiellement des forêts de survie aléatoires et des réseaux de neurones convolutionels qui sont à la base de nos méthodes présentées respectivement dans les parties 3.4.2 et 7.2. Nous finirons en présentant les valeurs métriques qui seront utilisées pour évaluer nos modèles.

3.1 L'analyse de survie

L'analyse de la survie peut être utile notamment pour prédire l'efficacité d'un traitement ou pour déterminer la gravité d'une maladie en réalisant des sous-groupes de patients. Kumar et al. [19] utilisent par exemple la méthode de Kaplan-Meier [voir sous-

section 3.2.1] pour évaluer l'impact de nouveaux traitements chez des patients atteints de myélome multiple.

La ***survie*** correspond au temps écoulé jusqu'à la survenue d'un événement précis. Cela peut être le décès mais également un évènement impliquant la progression d'une maladie (PFS). Dans cette section nous prendrons comme exemple d'évènement le décès. La fonction de survie $S_i(t)$ est, pour un individu i et un t fixé, la probabilité de survivre jusqu'à l'instant t, ou équivalent à la probabilité de ne pas avoir eu d'évènement au temps t en fonction de t :

$$S_i(t) = P(t_i > t), \tag{3.1}$$

avec $t > 0$ et t_i la durée de survie de l'individu i.
On peut aussi préciser la fonction de densité de survie qui est définie par :

$$f_i(t) = \frac{dS_i(t)}{dt} \tag{3.2}$$

Cette fonction de survie $S_i(t)$ peut être représentée par des courbes de survie en fonction du temps [voir la figure 3.1].

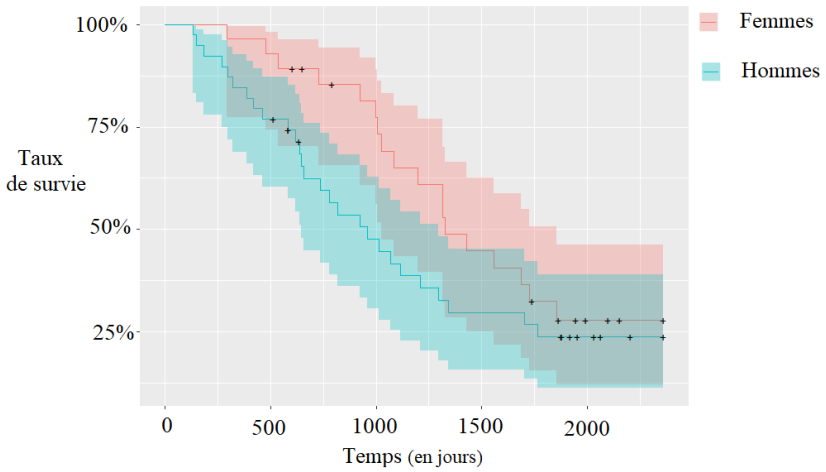


FIGURE 3.1 – Exemple de courbes de Kaplan-Meier présentant la survie de deux groupes. En rose, les femmes et en bleu les hommes. On peut voir ici que les femmes ont, à tout instant, de meilleures chances de survie que les hommes. On voit par exemple qu'à 1000 jours, 50% des hommes sont décédés contre 25% des femmes. Plus les courbes sont séparées et plus la différence de survie entre les deux groupes est grande.

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 a été perdu de vue ou que l'étude s'est terminée avant que l'évènement ne se produise. La figure 3.2 nous présente les différents types de censure. Dans cette thèse, on dénotera la présence d'un

évènement pour le patient i par la variable δ_i. Si δ_i= 1, l'évènement a eu lieu et il n'y a pas de censure. Si $\delta_i = 0$, l'évènement n'a pas eu lieu et il y a censure.

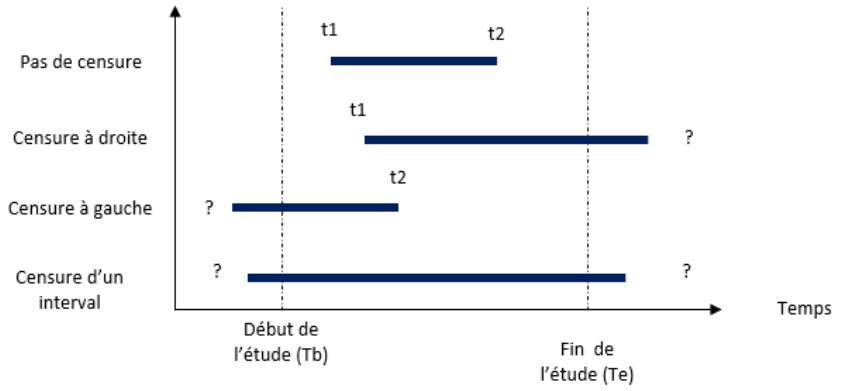


FIGURE 3.2 – Schéma représentant les types de censure. Pas de censure : on connaît la date de début et la date de fin (exemple : la date d'apparition de la maladie est connue et un évènement à été enregistré). Censure à droite : on ne connaît pas la date de l'évènement (exemple : perte de vue du patient avant le décès). Censure à gauche : on ne connaît pas la date de début (exemple : pas de connaissance de la date d'apparition de la maladie). Censure droite et gauche : pas de connaissance ni du début ni de la date de l'évènement.

Un concept important est le ***risque instantané***. Il est définit comme la probabilité d'un évènement dans l'intervalle de temps $[t, t + \Delta t[$ en sachant que l'évènement n'a pas eu lieu avant t. Ceci permet de définir la ***fonction de risque*** (ou fonction de hasard) qui apparaît comme une mesure du risque instantané :

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

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

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

Or, d'après le théorème des probabilités conditionnelles,

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

$\frac{\frac{dS(t)}{d(t)}}{S(t)}$ étant la dérivée de $\log[S(t)]$, on obtient à partir des équations 3.4 à 3.5 :

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

Ainsi, la connaissance de l'une de ces fonctions permet d'estimer les autres. Il existe différentes méthodes pour estimer ces fonctions. Nous allons présenter ces différentes méthodes d'estimation dans la section suivante.

3.2 Les méthodes d'estimation statistique pour l'analyse de survie

L'analyse de survie est possible par l'estimation statistique et l'apprentissage automatique. L'estimation statistique est la méthode la plus courante et la plus utilisée par les médecins. Il existe différentes méthodes pour estimer la fonction de survie :

— Des approches non paramétriques comme le Kaplan-Meier ou l'estimateur de Nelson Aalen, sont utilisées quand aucune hypothèse ne peut être faite sur la distribution des temps de survie.

— Des approches paramétriques comme le model de Weibull, qui nécessite une hypothèse sur la distribution des temps de survie

— Des approches semi-paramétriques comme le model de Cox.

Outre ces modèles d'estimation statistique, il existe aussi des méthodes d'apprentissage automatique qui sont applicables à l'étude de la survie qui seront présentées dans la section 3.3.

3.2.1 Les méthodes non paramétriques

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 [20]. Il permet de décrire la survie d'une population, d'estimer la survie médiane, et le taux de survie à un temps donné et de comparer la survie de différentes populations, souvent par le test du Log-Rank [voir le calcul du Log-Rank dans la sous-section 5.3.4].

3.2.1.1 L'estimateur de Kaplan-Meier

Par soucis de simplification, nous considérerons dans cette section l'évènement comme étant le décès, bien qu'il puisse s'agir par exemple de rechute ou guérison.

Ainsi on considère que survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t. Soit, pour trois temps, $t'' < t' < t$,

$$P(t_i > t) = P(t_i > t | t_i > t') \cdot P(t_i > t' | t_i > t'') \cdot P(t_i > t'') \tag{3.7}$$

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''

De façon générale, on peut considérer un ensemble de N individus, leur temps de survie t_i rangés par ordre croissant avec $i = [0, ..., N]$. On considère les temps t_j avec $j \in [1, ..., J]$ rangés par ordre croissant et J le nombre de temps sur lesquels on veut construire la courbe. Dans le cas de Kaplan-Meier, la probabilité de survie se calcul sur les temps d'évènements (temps auxquels à eu lieu un évènement), donc J sera le nombre de temps distincts où un évènement à eu lieu.

Ainsi, on obtient la probabilité pour le patient i de survivre jusqu'au temps t_j en faisant le produit de la probabilité de survie à chaque temps $t_{k \leq j}$ sachant que le décès n'a pas eu lieu au temps t_{k-1}, avec $t_0 = 0$ et $k \in [1, ..., j]$:

$$P(t_i > t_j) = \prod_{k=1}^{j} P(t_i > t_k | t_i > t_{k-1}), \tag{3.8}$$

Soit Υ_j le nombre d'individus à risque juste avant t_j, et d_j le nombre d'évènements en t_j.

Ainsi, la probabilité $\hat{p_j}$ d'avoir un évènement dans l'intervalle $]t_{j-1}, t_j]$ sachant que l'on était vivant en t_{j-1}, i.e. $\hat{p_j} = P(t_i \leq t_j | t_i > t_{j-1})$, peut être estimée par l'équation 3.9.

$$\hat{p_j} = \frac{d_j}{\Upsilon_j} \tag{3.9}$$

Or, comme les temps t_j sont supposés distincts, $d_j = 0$ quand $\delta_j = 0$ (censure), et $d_j = 1$ quand $\delta_j = 1$. L'équation 3.10 donne ainsi l'estimateur de Kaplan-Meier :

$$\widehat{S(t)} = \prod_{j \in [1,...,n], t_j \leq t} \left(1 - \frac{\delta_j}{\Upsilon_j}\right) \tag{3.10}$$

Les courbes de survie peuvent ainsi être tracées sur un graphique, avec généralement un intervalle de confiance à 95%. 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 un même temps t_j représente la verticale d'une marche (la hauteur de la marche proportionnelle au nombre d'événements survenus à t_j). La figure 3.1 permet de l'illustrer. Ces graphiques permettent de voir de façon claire la différence entre les courbes de survie de deux groupes.
Une méthode équivalente à celle de Kaplan-Meier est la méthode actuarielle. Elle se différencie par le fait que les calculs ne se font pas au temps t_j où des évènements interviennent mais à des temps t_j fixés réguliers, ce qui donne une courbe en segment de droites.

3.2.1.2 L'estimateur de Nelson-Aelen

L'estimateur de Nelson-Aalen est une méthode alternative pour estimer la fonction de survie S(t) en temps continu [21]. Soit H(t) la fonction de hasard cumulée. Dans le cas

continu :

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

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_j où des événements surviennent jusqu'à l'instant t, nous avons :

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{\Upsilon_j} \Rightarrow \hat{S}(t) = e^{-\hat{H}(t)}, \tag{3.13}$$

avec d_j le nombre d'évènements survenant en T_j, et Υ_j le nombre d'individus à risque de subir l'événement juste avant le temps T_j.

3.2.1.3 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. Nous sommes ici dans le cas où le hasard diminue au fil du temps.
D'autres estimateurs existent tels que l'estimateur de Breslow du risque cumulé [22], et l'estimateur de Harrington et Fleming de la survie [23].

3.2.2 Méthode semi-paramétrique : Cox

Le modèle de Cox se retrouve beaucoup dans la littérature. Il est employé lorsque l'objectif est d'évaluer l'effet de covariables sur le temps de survie. Il permet d'expliquer la survenue d'un évènement 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 d'un évènement en fonction de la valeur de la variable, et de son intervalle de confiance. Le hazard ratio est égal au risque relatif instantané de l'évènement pondéré 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 considère l'ensemble de données $\mathcal{X} = \{\mathbb{X}, \mathbb{Y}\}$ avec $\mathbb{Y} = [\mathbf{y}_1, ..., \mathbf{y}_i, ..., \mathbf{y}_N]$, $\mathbf{y}_i =$

$\{t_i, \delta_i\}$, et $\mathbb{X} = [\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_N]$ une matrice contenant les vecteurs $\mathbf{x}_i$ de covariables [1] ou caractéristiques pour chaque patient i. On introduit une fonction de hasard de base h caractérisée par la relation 3.14.
Soit un individu i, et $t > 0$:

$$h(t|\mathbf{x}_i) = h_0(t) \times f(\beta, \mathbf{x}_i), \tag{3.14}$$

avec :

— $\mathbf{x}_i = (x_{i1}, ...x_{iN_c})$ un vecteur de covariables ou caractéristiques de dimensions N_c de l'individu i,

— β le vecteur de paramètres d'intérêt (ne dépend pas du temps t), qui représente l'effet des covariables sur le risque instantané

— f une fonction positive.

Ce modèle est dit à risques proportionnels car, le rapport des fonctions de hasard de deux individus ne varie pas au cours du temps [21]. Soit i_1 et i_2, deux individus $\mathbf{x}_{i_1}$ et $\mathbf{x}_{i_2}$ leur vecteur de covariables respectifs,

$$\frac{h(t|\mathbf{x}_{i_1})}{h(t|\mathbf{x}_{i_2})} = \frac{f(\beta, \mathbf{x}_{i_1})}{f(\beta, \mathbf{x}_{i_2})} \tag{3.15}$$

Dans le cas du modèle de Cox, la fonction f est une fonction exponentielle. Ainsi, l'équation 3.14 devient :

$$h(t|(x_{i1}, ...x_{iN_c}) = h_0(t)e^{\beta_1 x_{i1} + ... + \beta_p x_{iN_c}} \tag{3.16}$$

L'objectif est d'estimer β par la méthode de la vraisemblance partielle qui porte l'information sur les coefficients β_i [21]. 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.

3.3 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 Neighbours), la méthode Bayesienne [24] [25], ou encore les méthodes basées sur les arbres de décision. La majorité de ces méthodes sont des méthodes de classification et de régression qui ne sont pas forcément adaptées pour l'analyse de la survie et notamment

1. Les covariables sont des variables propres aux individus pouvant influer sur la sortie du modèle.

pour les données censurées. En effet, l'utilisation d'une classification impliquerait d'éliminer un bonne partie des patients présentant de la censure. Par exemple, dans le cas d'une classification en 5 classes avec chaque classe correspondante à 1 an, le patient, avec un temps de survie de 2 ans mais sans évènement (censuré), peut être classé dans la classe 2 mais aussi 3, 4 et 5 étant donné que nous ne savons pas quand aura lieu l'évènement. Il est donc inclassable.

Nous présenterons ici les méthodes sur lesquelles se basent les deux méthodes principales de cette thèse. Les forêts de survie aléatoires (RSF) se basent sur les méthodes d'arbres et plus particulièrement de forêts aléatoires, et les méthodes d'apprentissage profond sur les réseaux de neurones (NN) et les réseaux de neurones convolutionnels (CNN).

3.3.1 Les méthodes basées sur les arbres de décision

La méthode des RSF est une méthode d'ensemble d'arbres de décision. Un arbre de décision permet de traiter la régression, la classification bi-classe ou multi-classe ou encore de mélanger des variables explicatives quantitatives et qualitatives. Les méthodes d'arbres de décision sont connues depuis les années 60 mais ont connues 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 de variance. En effet, de petites modifications dans l'échantillon d'apprentissage peuvent avoir des effets importants sur la prédiction. La solution est d'utiliser des forêts, c'est à dire des ensembles d'arbres chacun perturbé de façon aléatoire. Ce sont les forêts aléatoires (Random Forest ou RF) de Breiman [26], basées sur le bagging qui se montrent encore aujourd'hui les plus performantes sur le plan expérimental, et qui sont de plus en plus utilisées pour la survie.

Les méthodes d'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.

3.3.1.1 Les arbres CART

Le principe général de CART est de partitionner récursivement l'espace d'entrée $\mathbb{X}$ de façon binaire (X étant une matrice de dimensions $(N \times Nc)$ avec N le nombre d'individus et Nc le nombre de variables), puis de déterminer une sous-partition optimale afin de regrouper les patients dans des espaces avec une réponse commune. Bâtir un arbre CART se fait en deux étapes. Une première phase est la construction d'un arbre maximal (sans élagage), qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner l'arbre le plus prédictif. L'arbre se construit en commençant par partitionner dans deux noeuds fils, l'entrée X, en fonction d'une variable $\mathbf{x}_j$ et d'une valeur c_j 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. La figure 3.3 présente un exemple d'arbre de décision. 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 [27].

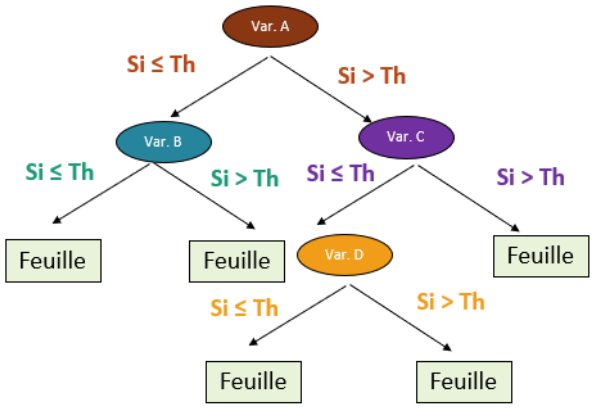


FIGURE 3.3 – Schéma d'un arbre de décision. c_j^* correspond à la valeur optimale de la variable $\mathbf{x}_j$ pour séparer les données en deux noeuds fils.

3.3.1.2 Le bagging

Le bagging est une méthode introduite par Breiman [28] 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. Le principe du Bagging est de tirer un grand nombre d'échantillons, indépendamment les uns des autres, et de construire 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 3.4 [27].
Dans la figure 3.4, $\psi_n = \{\mathbf{x}_i, y_i\}_{i=[1:N]}$ avec N le nombre d'individus, $\mathbf{x}_i$ le vecteur de variables et y_i un vecteur d'entrée (par exemple, la classe à laquelle appartiennent l' individu i). $\psi_N^{S_l} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=[1:L_{se}]}$ avec S_l le sous-ensemble d'individus i, et L_{se} le nombre de sous-échantillons tirés de façon aléatoire de la base d'entraînement $\mathcal{X}_{\text{train}}$. Finalement, $\hat{h}(., S_l)$ est la valeur prédite par les arbres à partir du sous-échantillon S_l, et $\hat{h}_{BAG}(.)$ l'agrégation

de toutes les valeurs prédites (par vote majoritaire ou moyenne par exemple).

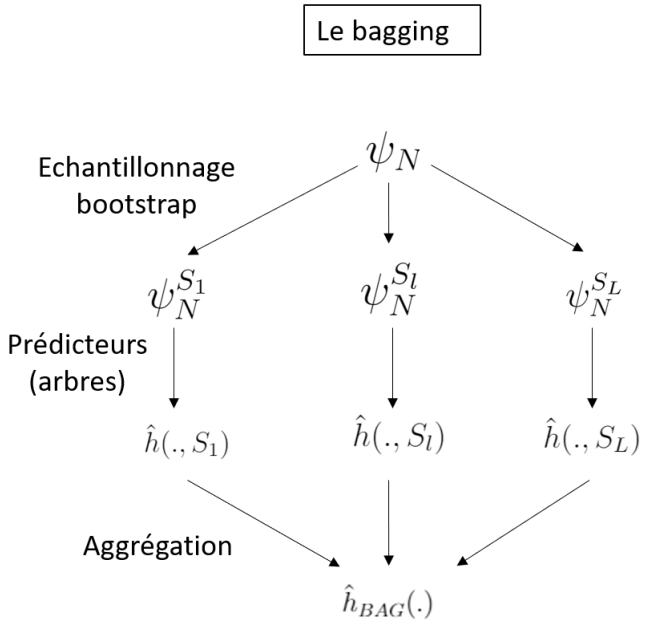


FIGURE 3.4 – Schéma du bagging

3.3.1.3 Les forêts aléatoires

Les forêts aléatoires (RF) sont une méthode d'apprentissage automatique basée sur une collection d'arbres qui sont entraînés sur des sous-ensembles aléatoires et indépendants d'échantillons. La partition à chaque noeud, peut aussi se faire sur un sous-ensemble aléatoire de variables.

La définition très générale de Breiman [26] des forêts aléatoires est la suivante : Soit une matrice $\mathbb{X}$ de dimension $(N \times Nc)$, contenant pour chaque patient les valeurs associées à des variables $\mathbf{x}_j$, $[A(\mathbb{Y}, \mathbb{X}_1), ..., A(\mathbb{Y}, \mathbb{X}_q)]$ une collection de q arbres A, avec $\mathbb{X}_1, ..., \mathbb{X}_q$ des vecteurs de variables aléatoires (indépendantes de l'échantillon d'apprentissage ψ_n), et $\mathbb{Y}$ la valeur cible. Le prédicteur des forêts aléatoires A_{rf} est obtenu en agrégeant cette collection de q arbres aléatoires de la façon suivante :

— $A_{rf}(Y) = \frac{1}{q} \sum_{l=1}^{q} A(Y, \mathbb{X}_l)$ en régression (moyenne des prédictions individuelles des arbres).

— $A_{rf}(Y) = \arg\max_{1 \leq k \leq K} \sum_{l=1}^{q} 1_{A(Y,\mathbb{X}_l)=k}$ en classification (vote majoritaire parmi les prédictions individuelles des arbres).

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

— Pour chaque arbre A :

  — pour chaque noeud Ω :

    1. construire un vecteur $\mathbf{x}_j$ contenant m variables sélectionnées aléatoirement parmi les N_c variables
    2. choisir la meilleure coupure parmi les variables de $\mathbf{x}_j$
    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 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 A (moyenne en régression et vote majoritaire en classification) pour donner le prédicteur RF.

Ainsi, les RF peuvent être vues comme une variante du bagging qui améliore généralement les performances, où la différence intervient dans la construction des arbres individuels. Le tirage, à chaque nœud, des m variables se fait uniformément parmi toutes les variables.

Cette méthode a été reprise pour être appliquée à différents domaines. Pour l'analyse des données de survie ce sont principalement Hothorn et al. [29] et Ishwaran et al. [5].

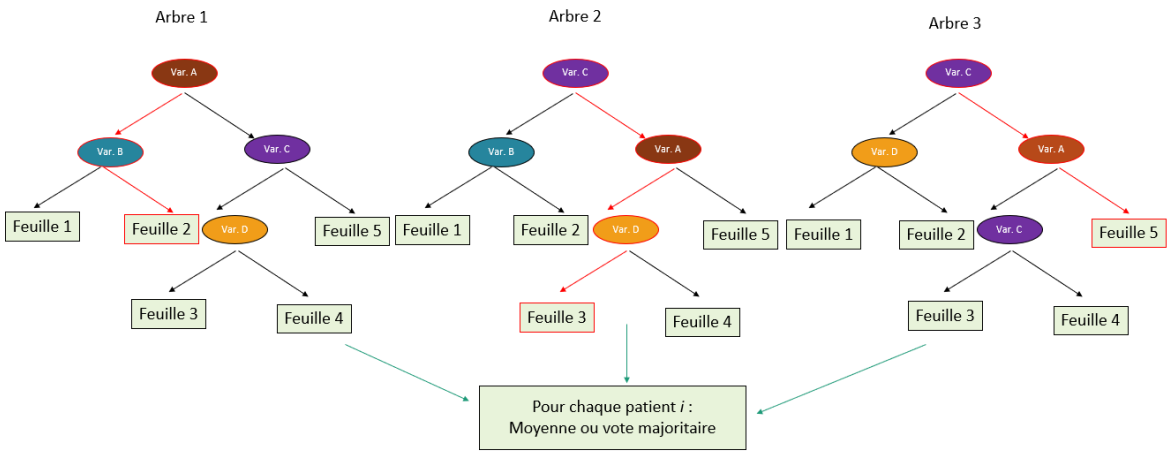


FIGURE 3.5 – Schéma des forêts aléatoires. Construction : A chaque noeud on choisit la variable qui donne la meilleur séparation parmi les m variables $_j$ choisies aléatoirement parmi toutes. Un noeud devient une feuille lorsque les critères ne sont plus respectés. Prédiction : On fait descendre chaque individu dans tous les arbres. Le chemin en rouge correspond à un exemple de prédiction. La prédiction de ce chemin s'écrit dans notre exemple $\hat{h}_i = \text{Moyenne/Vote Maj}(\text{Feuille}_2^{\text{A1}}, \text{Feuille}_3^{\text{A2}}, \text{Feuille}_5^{\text{A3}})$.

3.3.2 L'apprentissage profond

On considère comme des modèles d'apprentissage profond les réseaux de neurones contenant plus de 3 couches. Dans cette section nous commencerons pas définir les réseaux de neurones puis nous parlerons du cas particulier des réseaux de convolutions, largement utilisés dans le cas d'image comme élément d'entrée.

3.3.2.1 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 3.6]. Chaque couche k est composée de neurones $\mathbf{x}_j^k$. A chaque liaison entre deux neurones est associé un poids $W_{ij}^{(k-1)(k)}$, de sorte que les neurones $\mathbf{x}_i^{k-1}$ sont multipliés par ce poids, puis additionnés par les neurones $\mathbf{x}_j^k$. A chaque couche est associée une fonction d'activation φ qui permet d'avoir la sortie. Trois fonctions d'activation courantes sont la fonction linéaire ($\mathbf{x}' = a\mathbf{x} + b$), le ReLU (Rectified Linear Unit, $\mathbf{x}' = \max(0, a\mathbf{x} + b)$) ou la fonction sigmoïde ($\frac{1}{1+e^{-\mathbf{x}}}$).

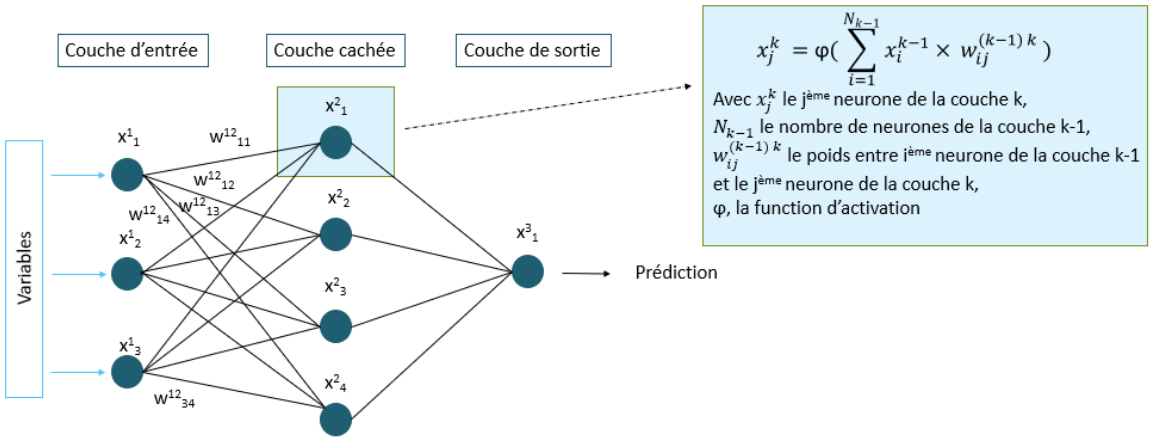


FIGURE 3.6 – Schéma d'un réseau de neurones à 3 couches

3.3.2.2 Les réseaux de neurones convolutionnels

La plupart des réseaux permettant de traiter des images utilisent des couches convolutionnelles et sont appelés CNN. La figure 3.7 rappelle le principe des convolutions. Le fait d'enchaîner les couches de convolution va permettre d'avoir des caractéristiques de plus en plus détaillées. En effet, les premières couches du réseau vont permettre d'obtenir des caractéristiques de bas niveau (exemples : direction, rayures) quand les dernières permettront d'avoir des caractéristiques de haut niveau (exemples : détection des yeux, du bec). D'autres couches que les convolutions sont régulièrement utilisées avec les images. C'est le cas du "pooling". La compréhension de cette couche est importante afin de comprendre

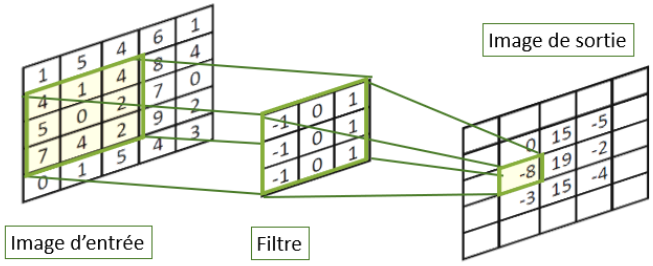


FIGURE 3.7 – Schéma explicatif d'une convolution

certains modules que nous utiliserons par la suite. Ainsi, le "pooling" va permettre de diminuer la taille d'une couche, afin de concentrer l'information et donc de réduire le nombre de paramètres. Deux des "pooling" les plus classiques, le "pooling" max et le "pooling" moyen sont présentés dans la figure 3.8.

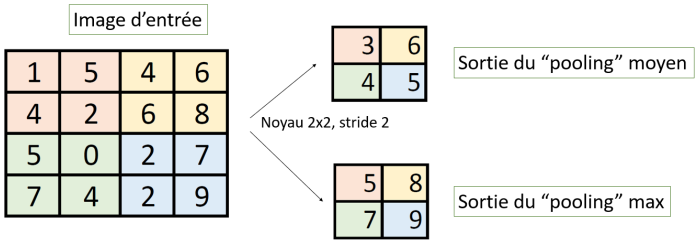


FIGURE 3.8 – Schéma du "pooling". "Pooling" max : on prend le maximum des pixels présents dans le noyau. "Pooling" moyen : on prend la moyenne des pixels dans le noyau.

3.4 Les valeurs métriques d'évaluation

Étant dans un contexte de survie, nous devons évaluer nos modèles grâce à des valeurs métriques prenant en compte la censure. Nous choisissons d'évaluer nos modèles avec le c-index, la valeur métrique de référence de la survie. Nous utilisons dans le même temps la p-value qui permet d'évaluer la séparation entre les groupes de pronostique prédits.

3.4.1 L'indice de Concordance

L'indice de concordance ou 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 [voir figure 3.9].
4. La concordance correspond à la somme sur toutes les paires admissibles.
5. Le c-index est définit par $\mathrm{C} = \frac{Concordance}{N_p}$

On considère pour la partie RSF l'erreur de prédiction au lieu du c-index. L'erreur de prédiction peut être définit comme 1 moins le c-index.

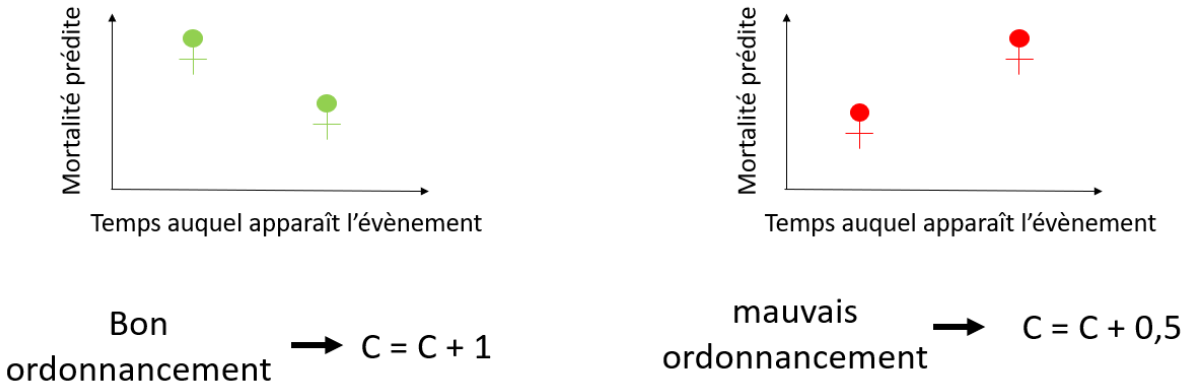


FIGURE 3.9 – Explication du calcul du c-index

3.4.2 La p-value

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 :

— $\mathrm{p} \leq 0.01$: très forte présomption contre l'hypothèse nulle

— $0{,}01 < \mathrm{p} \leq 0{,}05$: forte présomption contre l'hypothèse nulle

— $0{,}05 < \mathrm{p} \leq 0{,}1$: faible présomption contre l'hypothèse nulle

— $\mathrm{p} > 0{,}1$: pas de présomption contre l'hypothèse nulle

La p-value est calculée dans notre cas pour déterminer la meilleure séparation en deux groupes de pronostic (bas et haut risque) réalisée par Log-Rank [voir le calcul du Log-Rank dans la section 5.3.4]. La meilleur séparation sur la valeur prédite (mortalité, temps, risque, etc.) correspondra à la séparation présentant la plus petite valeur de p-value. Elle

évalue si la différence est significative entre les courbes de survie des deux groupes. 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.

Deuxième partie

Analyse de survie par Random Survival Forest

Résumé

Dans cette partie va être abordée la première partie de la thèse. Nous souhaitons prédire la progression des patients atteints de myélome multiple afin de pouvoir réaliser de la médecine personnalisée. Ainsi, l'objectif de cette première partie de thèse était de déterminer un modèle d'apprentissage statistique permettant la prédiction de la progression et des biomarqueurs de la maladie, en utilisant les données cliniques et radiomiques des images TEP. Les méthodes d'estimation statistique et d'apprentissage machine sont des méthodes relativement interprétables et fiable. Elles sont rapides et relativement simples d'utilisation. Parmi les méthodes d'apprentissage machine pour la survie, les forêts de survie aléatoires (RSF) [5], une adaptation des forêts aléatoires, sont devenues la référence ces dernières années en raison de sa robustesse aux données censurées et au bruit dû aux variables. Nous avons choisi de proposer un modèle en trois étapes basé sur les RSF, permettant la prédiction d'un risque et d'un groupe de risque pour chaque patient, ainsi qu'une liste de biomarqueurs de la maladie.
Ainsi, après avoir discuté de l'état de l'art (chapitre 4), nous aborderons les détails de la méthode RSF, des trois étapes de notre modèle, du pré-traitement des données, et du calcul des variables dans le chapitre 5. Nous détaillerons ensuite les expériences et présenterons les résultats (chapitre 6) et nous conclurons dans le chapitre 7.

Chapitre 4

État de l'art

Cette section met en avant l'état de l'art de l'analyse de survie du myélome multiple. Nous parlerons d'abord des méthodes de prédiction de survie présentes dans la littérature, et notamment les RSF et les méthodes de sélection de variables associées. Nous discuterons ensuite de l'utilisation de l'imagerie médicale pour l'étude de la survie et plus particulièrement de l'imagerie TEP. Enfin, nous présenterons l'analyse de survie du myélome multiple, ainsi que les motivation de l'utilisation de l'imagerie médicale qui est associée à cette maladie. Nous conclurons cet état de l'art en expliquant nos choix pour la méthode présentée dans le chapitre 7.

4.1 L'analyse de la survie

Dans la littérature, différentes méthodes ont été présentées pour l'étude de la survie. Lorsqu'il n'y a pas de censure, la tâche de survie peut être remplacée par de la classification ou de la régression. Ainsi, on peut citer l'utilisation des SVM (Machine à Vecteur de Support) [30] ou des réseaux de neurones [31], mais surtout les classifieurs forêts aléatoires (RF) [32]. Parmar et al. [24] comparent 12 méthodes de classification et 14 méthodes de sélection de variables 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 variables, est la méthode des RF. Cependant, ces méthodes présentées dans l'article de Parmar et al. [24] ne permettent pas de prédire un temps de survie, seulement une classe. Lorsque la censure est présente, il est nécessaire d'utiliser des méthodes adaptées. Pour estimer la survie d'un groupe ou la comparer entre deux groupes, la méthode de Kaplan-Meier est souvent utilisée [32]. 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. 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 [33]. Ce modèle reste encore très

utilisé. Raykar et al. [34] proposèrent, de remplacer la maximisation de la vraisemblance partielle du modèle de Cox, par une maximisation de la limite inférieure du C-index (fonction log-sigmoïde). Cependant, ces estimateurs sont de plus en plus remplacés par des méthodes d'arbres et d'apprentissage automatique. L'article de Zhou et al. [35], nous présente les intérêts et inconvénients qu'apportent les différentes méthodes d'arbres dans le contexte de la prédiction du temps avant la ré-incarcération d'ancien prisonniers. Il montre que les résultats des régressions de Cox peuvent être biaisés lorsque la censure est liée aux variables, et lors d'un haut taux de censure. Zhou et al. [35] nous indiquent aussi que les forêts de survie de Breiman sont plus performantes que la régression de Cox, notamment, lorsque la régression de Cox ignore les variables qui sont prédictives sur une période de temps et non tout le temps . Enfin, les forêts de survie sont plus intéressantes que la régression de Cox, dans le cas où il y a grand nombre de prédicteurs et une petite taille d'échantillon [35].

Les arbres de survie peuvent être utiles pour détecter un décalage dans des relations non linéaires entre les variables mais aussi détecter les interactions entre les variables. Parmi les algorithmes d'arbres de survie, les « conditional inference survival trees » développés par Hothorn et al. [36] semblent plus fiables et moins enclins au sur-apprentissage [1] que d'autres méthodes d'arbres comme les "bagging survival trees" [29], et les RSF [5]. Cependant, ce problème est compensé dans ces deux dernières méthodes par l'utilisation d'un grand nombre d'arbres. De plus, il fut montré par Iswharan et al. [5] que, bien que la méthode d'Horthorn et al. [36] soit bonne dans les cas où le taux de censure est faible, elle dépend grandement de celui-ci. Ainsi Iswharan et al. [5] propose les RSF en 2008, un algorithme de forêts aléatoires adaptées à la censure à droite. La méthode est robuste aux données censurées et au bruit dû aux variables.

Les RSF sont donc naturellement utilisées dans de nombreux travaux et sont la méthode de référence pour l'étude de la survie [37–39]. Il est apparu depuis peu des améliorations de la méthode : par exemple, Miao et al. [40] 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. Kumar Dey et al. [41] proposent une méthode basée sur les RSF mais qui utilise les« Extremly randomized trees » et « Adaboost » permettant d'augmenter la rapidité et les performances pour des bases de données de grandes dimensions.

Enfin, de plus en plus d'articles parlent de l'utilisation de l'apprentissage profond pour l'étude de la survie. En 2016, Liao et Ahn [42] proposent un modèle à trois 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 Katzman et al. de 2017 [12], compare

1. On parle de sur-apprentissage lorsqu'un modèle correspond trop étroitement à un ensemble particulier de données et ne peut donc pas généraliser à de nouvelles données.

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. L'état de l'art des méthodes d'apprentissage profond sera présenté plus en détails dans le chapitre 9.

Finalement, lorsque le nombre de variables est grand, ce qui est le cas par exemple lors de l'utilisation de données génomiques ou d'images (radiomiques) comme entrée, il est nécessaire de procéder à une sélection des variables les plus prédictives. Un cas particulier est le cas des problèmes à grande dimension [2]. Pour gérer cette problématique différentes méthodes de sélection de variables existent comme la méthode des moindres carrés ou le modèle de Cox sous pénalisation Lasso [43]. Plus récemment, l'utilisation de l'indice de concordance pour la sélection de variables combinée avec des modèles de Cox "boosting" ou "gradient boosting"(modèle construit de façon progressive avec des prédicteurs Cox non indépendants les uns des autres) ont montré de grandes performances lors d'études comparatives de différentes méthodes d'apprentissage machine et de méthodes de sélection pour la survie [44]. Cependant, au cœur des méthodes RSF, l'optimisation aléatoire effectue une sélection de caractéristiques à chaque nœud. Ce fait a été exploité par Ishwaran et al. [5], qui ont proposé trois méthodes de sélection de variables : VIMP (Variable Importance), MD(Minimal Depth) et VH(Variable Hunting) [voir la section 5.2].
Pour conclure, étant donné notre taux de censure (supérieur à 35%) et le nombre de patients relativement faible nous choisissons de nous baser sur les RSF que ce soit pour la prédiction de la progression ou pour la sélection des variables.

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 des radiomiques comme variable d'entrée des modèles [25, 45]. La radiomique est définie par Bourgier et al. [46] comme 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 ». Les radiomiques peuvent être calculées à partir d'images tomographique provenant de TDM, IRM, TEP, ou n'importe quelle autre modalité. Ces données images sont souvent accompagnées par des données cliniques ou génomiques. Les

2. On est dans un problème de grande dimension lorsqu'il y a plus de variables que de patients.

images sont de plus en plus utilisées comme facteur pronostique, par exemple pour le cancer du poumon [47, 48], le lymphome [49], le cancer tête-cou [32], le cancer de l'œsophage [50] ou encore le carcinome bronchique [51]. Aerts et al. [52] montrent qu'un grand nombre de caractéristiques extraites des images TDM ont un pouvoir pronostique 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.

Concernant l'imagerie TEP (souvent associée à la TDM), les travaux l'utilisant à des fins pronostiques sont nombreux. Lartizien et al. [48], Desseroit et al. [53], Hatt et al. [54] et Bailly et al. [55] l'ont utilisée pour déterminer quelles sont les caractéristiques les plus intéressantes, celles qui dépendent le moins de la segmentation et celles qui sont liées entre elles. Vallières et al. [37] 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 sont généralement plus prédictives que celles extraites de l'IRM, dans le cas des métastases pulmonaires d'un sarcome mais la valeur prédictive est fortement augmentée lors de l'association des deux imageries. L'article de Ben Boullègue et al. [30] montre que la combinaison de facteurs pronostiques habituels avec des paramètres de texture de PET/TDM et de forme appropriés permettent d'améliorer la prédiction d'une réponse métabolique précoce dans plusieurs types de lymphome. Tixier et al. [50] ont démontré que l'analyse texturale d'images FDG-TEP scans peut prédire la réponse à un traitement contre le cancer de l'œsophage. Enfin, des TEP scans ont été utilisées pour montrer la stabilité des caractéristiques radiomiques dans un groupe de patients atteint de NSCLC (Non-Small Cell Lung Cancer) [56].

Les méthode de prédiction de la survie utilisant les images TEP en entrée sont diverses. Certains articles utilisent des classifieurs forêts aléatoires [32], les SVM [30], d'autres de l'apprentissage profond [49]. Cependant aucun d'eux n'associent RSF et imagerie TEP. Seul un article récent, de Steigner et al. [51] 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 d'arbres de survie et au modèle de Cox. En effet, l'application des RSF aux caractéristiques d'images restent encore peu étudiées. Seul, Bhnemann et al. [39] utilisent des RSF avec en entrée des caractéristiques provenant d'images mais celles ci sont des images confocales de puces tissulaires.

L'utilisation des caractéristiques radiomiques, et d'autant plus celles extraites de l'imagerie TEP, pourrait donc améliorer la prédiction de survie par notre modèle, en apportant de l'information nouvelle. Nous choisissons donc de les incorporer en entrée de notre modèle RSF et d'évaluer leur valeur pronostique grâce aux méthodes de sélection de variables.

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 [9]. D'autres comme Decaux et al. [57] et Amin et al. [58] proposent de prédire la survie des patients atteints de myélome multiple à partir de l'expression génique. Comme la majorité des articles médicaux, Decaux et al. [57] utilisent les méthodes de Cox et de Kaplan Meier pour réaliser l'étude. Amin et al. [58] testent plusieurs méthodes d'apprentissage automatique (Prédicteur composé de covariables, Analyse discriminante linéaire, K-NN, méthode des plus proche centroïdes, SVM) pour prédire une réponse complète en fonction du profil d'expression génique. Enfin, Lapa et al. [59] utilisent la méthode de Kaplan-Meier avec des caractéristiques des images TEP/TDM. Outre l'article de Pang et al. [60] qui utilise une base de données de myélome multiple pour montrer la corrélation entre la survie et les polymorphismes du nucléotide simple, aucun papier ne présente pour l'instant l'utilisation de RSF pour l'étude de la survie des patients atteints de myélome multiple.

Concernant les modalités utilisées pour l'étude du myélome multiple, le TDM permet de détecter de petites lésions osseuses qui ne sont pas détectables avec la radiographie conventionnelle [61]. Cependant, les méthodes de référence sont maintenant l'IRM et la TEP en baseline. L'IRM est beaucoup utilisée car plus sensible que le TDM et peut détecter une infiltration de la moelle osseuse diffuse avec une bonne différentiation des tissus mous [62, 63]. L'utilisation de la 18-FDG-TEP combiné au TDM permet aussi une bonne sensibilité [64–67]. L'article de Bodet-Milin et al. [68] 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 l'IRM. De nouveaux traceurs font leur apparition, comme le 68Ga-Pentixafor qui permet une haute sensibilité de détection des lésions du myélome multiple [9, 69] mais ceux ci restent récents. Certains papiers s'attèlent à l'étude des facteurs pronostiques du myélome multiple dans les images TEP. Ainsi, Carlier et al. [70] 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 trois lésions focales, de la SUV Max, des lésions extra-médullaires [2, 71] et du volume métabolique total et de la glycolyse totale [72].

La FDG-TEP reste une des méthode d'imagerie les plus utilisées dans l'exploration clinique du myélome multiple (généralement couplée au TDM). 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, pour la prédiction de survie.

4.4 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 radiomique et son efficacité prouvée, l'utilisation de caractéristiques provenant d'images médicales avec des RSF reste peu commune. Ceci nous amène à nous intéresser à l'utilisation des RSF pour la prédiction de la progression chez les patients atteints de myélome multiple, en y associant des caractéristiques radiomiques. Étant dans un problème à grande dimension (notamment en raison de l'évaluation de différentes implémentations de calcul des radiomiques), nous y associons des méthode de sélection des variables. Dans le but de garder la méthode interprétable et de conserver les caractéristiques qui sont cohérentes avec le modèle prédictif, nous proposerons un cadre où les deux tâches, la sélection des variables et le modèle de prédiction de survie, sont basées sur les RSF.

L'imagerie choisie est 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 TDM. Or, comme l'indique P. Moreau et al. [1], 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.

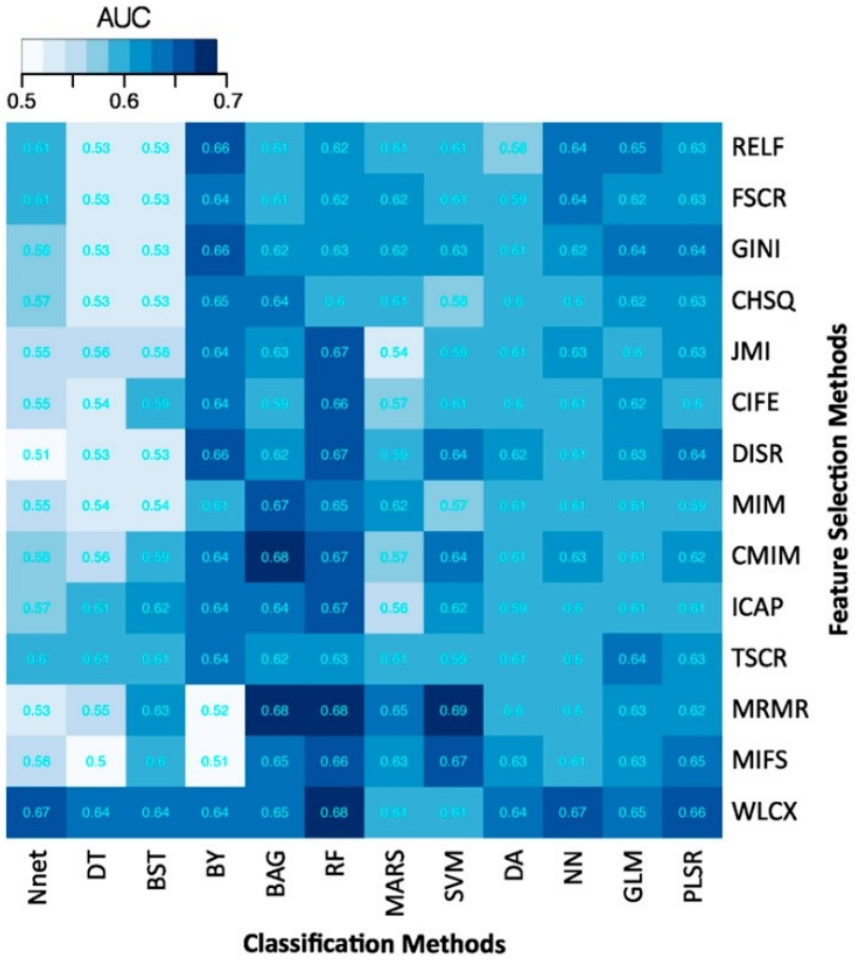


FIGURE 4.1 – Comparaison de 12 méthodes de classification (Nnet : réseau de neurones, DT : arbre de décision, BST : "Boosting", BY : Bayésien, BAG : "Bagging", RF : "Random Forest", MARS : Splines de régression multi-adaptatives, SVM : Machines à vecteurs de support, DA : Analyse discriminante, NN : Plus proche voisin, GLM : Modèles linéaires généralisés, PLSR : Régression partielle des moindres carrés et des composantes principales) et 14 méthodes de sélections de variables. La comparaison se fait sur la valeur de l'AUC (Aire sous la courbe) réalisée dans l'article de Parmar et al. [24]
dans le contexte de la prédiction de la survie ("Overall survival" séparé en deux groupes) du cancer du poumons grâce à des images TDM.

Chapitre 5

Méthodes

A partir d'une base de données prospectives de myélome multiple contenant des images TEP au diagnostic, ainsi que des caractéristiques cliniques, nous proposons de suivre une approche d'apprentissage automatique pour prédire la valeur de la survie sans progression (PFS) pour un nouveau patient. Pour atteindre ce but, nous construisons un modèle unifié basé sur les forêts de survie aléatoires (RSF) pour :

— traiter le grand nombre de caractéristiques cliniques et d'images (de l'ordre d'une centaine),

— identifier les caractéristiques les plus pertinentes pour la prédiction,

— prédire la progression d'un patient en fonction de ses données personnelles (caractéristiques cliniques et d'imagerie).

La prédiction va être réalisée grâce à différentes étapes : la première étape concerne le pré-traitement des données (extraction des images et segmentation). Après cela les caractéristiques radiomiques pourront être calculées et enfin arrivent les méthodes d'apprentissage machine. Ces étapes sont résumées dans la figure 5.1.

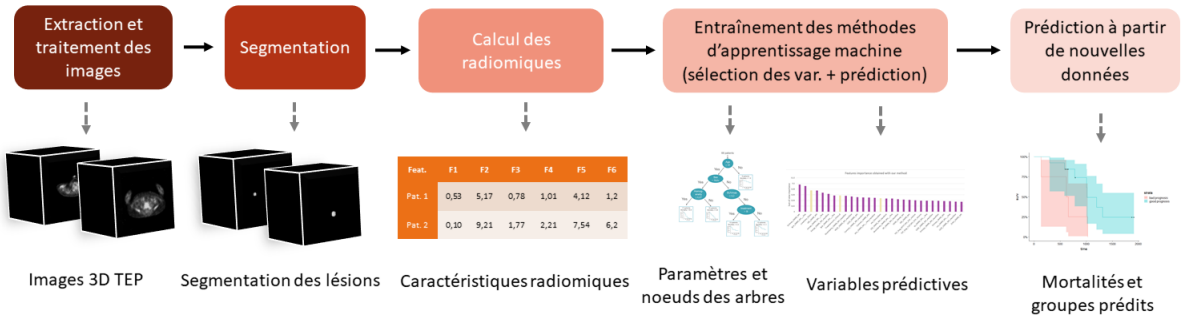


FIGURE 5.1 – Schéma des étapes du modèle de prédiction basé sur les RSF

5.1 La méthode des RSF

La méthode des RSF est une méthode de forêts aléatoires adaptées à l'analyse de survie avec des données censurées à droite, proposée par Ishwaran en 2008 [5]. 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 réside dans le fait que chaque aspect de la construction des RSF prend en compte la survie et la censure. En effet, les RSF prédisent une courbe (risque en fonction du temps) au lieu d'une valeur unique non dépendante du temps, et la dispersion dans les noeuds se fait grâce à la distance entre les courbes de deux groupes, tout en prenant en compte la censure. On peut décomposer la méthode en deux parties : l'entraînement et le test.

5.1.1 L'entraînement

L'entraînement correspond à la création des arbres. Soit une matrice de données $\mathbb{X}$ qui comprend pour chaque individu i un vecteur de caractéristiques $\mathbf{x}_i$ de dimension N_c et égal à $[x_{i1}, ..., x_{ij}, ..., x_{iN_c}]$. A cette base de données est associée pour chaque individu i, un couple de valeurs $\mathbf{y}_i = \{t_i, \delta_i\}_{i=[1:N]}$ avec t_i le temps jusqu'à l'évènement et δ_i la censure (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 données $\mathbb{X}$ et constitue l'in-bag $\mathbb{X}^{in}$. Le reste est placé dans l'Out-Of-Bag (OOB) $\mathbb{X}^{oob}$ et peut être utilisé pour le calcul de l'erreur de prédiction et de l'importance des variables.

Soit un noeud $\Omega = \{\mathbf{x}_j^*, c_j^*\}_{j=[1:m]}$, avec $\mathbf{x}_j^*$ la caractéristique j, c_j^* la valeur de cette caractéristique $\mathbf{x}_j^*$ permettant la meilleur séparation, et m le nombre de variables choisies aléatoirement dans le noeud. A chaque noeud Ω un sous-ensemble des caractéristiques $\mathbf{x}^\Omega$ de $\mathbb{X}^{in}$ est sélectionné aléatoirement parmi toutes les caractéristiques. La meilleure caractéristique $\mathbf{x}_j^*$ 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 $\mathbf{x}_j$ du sous-ensemble X^Ω, et pour différentes valeurs c_j de cette caractéristique, on calcule la valeur du test du Log-Rank qui mesure la séparation du groupe de patients en deux. La caractéristique $\mathbf{x}_j^*$ et le seuil associé c_j^* retenus sont ceux correspondant au score Log-Rank le plus haut. Définissons que lorsque $\mathbf{x}_{ij} \leq c_j$ l'individu i est inscrit dans le noeud fils gauche, et dans le noeud fils droit sinon.

Soit $t_1 < ... < t_z$ les z temps distincts des évènements dans le noeud Ω, $d_{k,l}$ et $\Upsilon_{k,l}$ respectivement le nombre d'évènements et le nombre d'individus à risque au temps $t_k \forall \{1, ..., z\}$ dans le noeud fils gauche l ($d_{k,r}$ et $\Upsilon_{k,r}$ pour le fils droit). On considère à risque au temps t_k un individu qui n'a pas eu d'évènement au temps t_k.

Un noeud Ω distribue les n individus l'atteignant en deux groupes, avec n_l et n_r individus respectivement tel que $n = n_l + n_r$. De plus, pour chaque temps t_k, la somme d'individus à risque et le nombre d'évènements sont préservés. Ainsi, $\Upsilon_k = \Upsilon_{k,l} + \Upsilon_{k,r}$ et $d_k = d_{k,l} + d_{k,r}$. La valeur du test Log-Rank pour la variable $\mathbf{x}_j$ et la valeur c_j est :

$$LR(\mathbf{x}_j, c_j) = \frac{\sum_{k=1}^{z}(d_{k,l} - \Upsilon_{k,l}\frac{d_k}{\Upsilon_k})}{\sqrt{\sum_{k=1}^{z}\frac{\Upsilon_{k,l}}{\Upsilon_k}(1 - \frac{\Upsilon_{k,l}}{\Upsilon_k})\frac{\Upsilon_k - d_k}{\Upsilon_k - 1}d_k}} \tag{5.1}$$

La valeur absolue de $LR(\mathbf{x}_j, c_j)$ mesure la séparation de la survie de populations de deux noeuds fils. Plus elle est élevée et meilleure est la différence entre les deux groupes.

— Le Log-Rank aléatoire : cette méthode est équivalente au Log-Rank mais il n'y a, ici, pas de test sur différentes valeurs c_j de la caractéristique. Seule une valeur aléatoire est prise par caractéristique $\mathbf{x}_j$.

La construction de l'arbre se poursuit jusqu'à ce que le critère d'arrêt ne soit plus respecté [voir la sous-section 3.3.1.3]. Dans les feuilles seront calculées les valeurs qui serviront plus tard à réaliser la prédiction.
Cela peut être comme ici la mortalité, la probabilité de survie, des courbes de survie, un temps ou une classe.

Chaque feuille doit normalement contenir un groupe de patients homogène du point de vue de la survie. Cette dernière est calculée grâce à l'estimateur de la ***fonction de risque cumulative*** (CHF) correspondant à l'estimateur de Nelson-Aalen [Equation 3.13].

$$\hat{H}_\Omega(t) = \sum_{t_{k,\Omega} \leq t} \frac{d_{k,\Omega}}{\Upsilon_{k,\Omega}} \tag{5.2}$$

avec $\hat{H}_\Omega(t)$ la CHF au nœud Ω et temps t, $d_{k,\Omega}$ le nombre de morts au temps $t_{k,\Omega}$ et $\Upsilon_{k,\Omega}$ le nombre d'individus à risque au temps $t_{k,\Omega}$.

On résume l'information de survie de tous les individus dans la feuille Ω, avec un seul CHF. La ***mortalité*** M_i d'un individu i caractérisé par un vecteur $\mathbf{x}_i$ et atteignant la

feuille Ω correspond à la somme des CHF sur chaque temps unique :

$$M_i = \sum_{k=1}^{z} \hat{H}_\Omega(t_k|\mathbf{x}_i) \tag{5.3}$$

Plus la mortalité est élevée par rapport aux autres individus, plus le risque d'évènement est grand. On répète le processus pour chaque arbre de la forêt. Les arbres seront distincts entre eux grâce aux étapes aléatoires lors de la construction.

5.1.2 Le test

Le test correspond à une prédiction de la mortalité sur un ensemble de données choisi. Pour chaque arbre, on prédit la mortalité de chaque individu de l'OOB $\mathbb{X}^{oob}$. 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 prédite finale correspond à la moyenne des mortalités de chaque arbre. Ainsi, en considérant une forêt de q arbres,

$$\hat{M}_i = \frac{\sum_{l=1}^{q} M_i^l}{q}. \tag{5.4}$$

5.2 Les méthodes de calcul de l'importance des variables

L'article de Ishwaran et al. [5] donne une méthode de calcul de l'importance des variables (VIMP). Intuitivement, la méthode mesure l'importance d'une caractéristique $\mathbf{x}_j$ en l'enlevant de tous les arbres et en regardant l'effet sur les prédictions. Un grand changement est un fort indicateur de la valeur prédictive de la variable $\mathbf{x}_j$. Son calcul consiste pour chaque caractéristique $\mathbf{x}_j$, à reconstruire les arbres à l'aide de l'OOB, en remplaçant les séparation utilisant cette variable, par une séparation aléatoire. Ainsi, lorsque un noeud Ω utilise $\mathbf{x}_j$ les individus du noeud Ω sont assignés aléatoirement dans les noeuds fils, et l'erreur de prédiction [1] err_{vimp} associée à la nouvelle forêt est recalculée. La valeur d'importance de la caractéristique $\mathbf{x}_j$ est $\text{VIMP}(\mathbf{x}_j) = \text{err}_{\text{vimp}} - \text{err}_{\text{oob}}$ (avec err_{oob} l'erreur de base, sans assignement aléatoire des individus). Plus la valeur du VIMP est grande, plus la caractéristique a une valeur prédictive. L'algorithme est résumé dans le schéma 5.2.

D'autres méthodes de sélection de variables ont été proposées et notamment "Minimal Depth" (MD) ou méthode de la profondeur minimale et "Variable Hunting" (VH) [73]. La

1. L'erreur de prédiction est calculée en réalisant 1 − c − index.

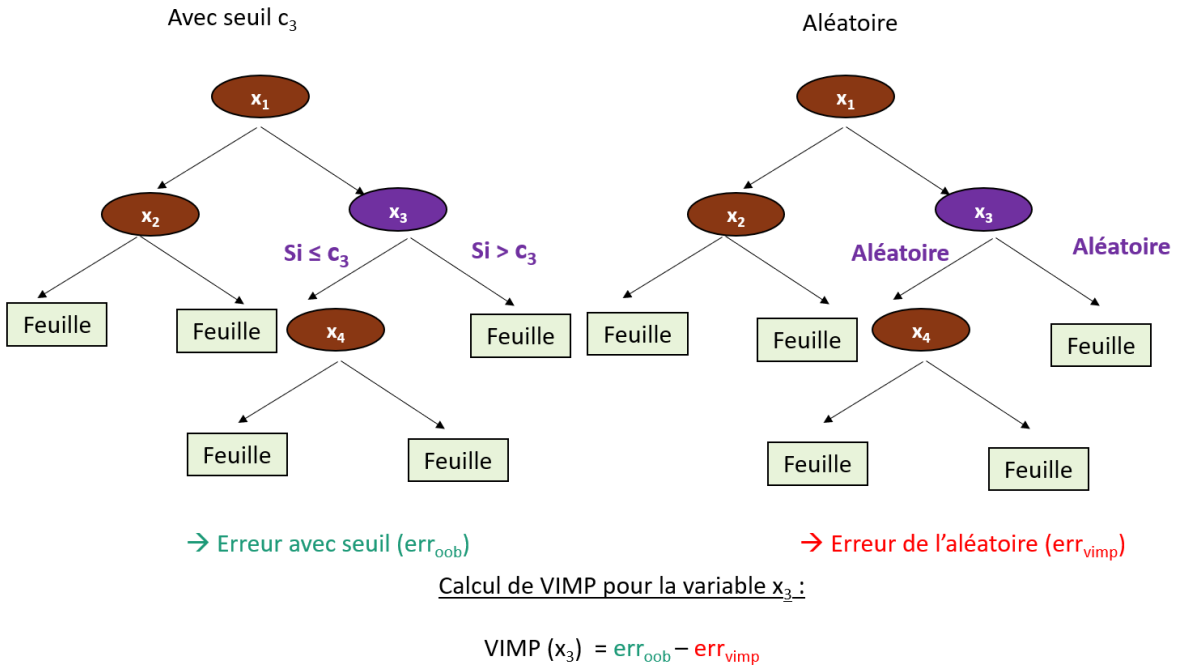


FIGURE 5.2 – Schéma du calcul de l'importance de variable VIMP de la caractéristique $\mathbf{x}_3$, c_3^* correspond au seuil de séparation optimale des variables du noeud.

méthode MD consiste à évaluer la prédictivité d'une variable par la profondeur minimale en supposant que les variables sélectionnées près de la racine sont plus importantes que celles qui sont sélectionnées profondément dans l'arbre. La dernière méthode, VH, a été définie pour les problèmes de très haute dimension. Elle combine les méthodes de MD et VIMP. Une RSF est ajustée à un sous-ensemble aléatoire de données, et un premier groupe de variables est sélectionné en utilisant un seuillage de profondeur minimale (méthode MD). Des variables supplémentaires sont ensuite ajoutées au modèle initial par ordre croissant de profondeur minimale jusqu'à ce que le critère VIMP se stabilise. Le processus est répété plusieurs fois, pour finalement retenir les variables qui apparaissent le plus fréquemment au cours des essais [73]. La méthode est résumée dans le schéma 5.3.

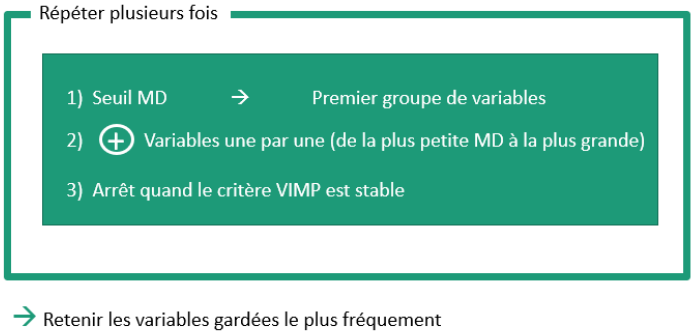


FIGURE 5.3 – Schéma du "Variable Hunting".

5.3 Analyse par RSF : le modèle proposé

Contrairement à l'état de l'art nous souhaitons un modèle considérant et optimisant conjointement la sélection de variables et la prédiction. Nous basons notre modèle sur les RSF en raison de leur robustesse aux données censurées et au bruit dû aux variables. Il peut être décomposé en trois parties :

1. L'optimisation des hyperparamètres des RSF
2. Le calcul de l'importance des variables avec la méthode VIMP
3. L'entraînement du modèle final de RSF avec les paramètres optimaux et les variables les plus prédictives, puis prédiction de la survie à l'aide des RSF précédemment entraînées.

Les détails des expériences varieront en fonction du groupe d'expériences. On peut considérer 2 groupes d'expériences avec les RSF : le groupe d'expériences techniques (publié dans IJCARS [6]) et le groupe d'expériences cliniques (publié dans EJNMMI [7]) [détails dans la section 6.1].

5.3.1 L'optimisation des hyperparamètres des RSF

La première étape correspond à l'optimisation des hyperparamètres des RSF (en jaune dans la figure 5.4). Pour ce faire nous réalisons une recherche par grille sur le nombre d'arbres, le mode de séparation, le nombre minimal d'échantillon dans chaque noeud et le nombre maximal de variables dans chaque noeud. Pour chaque combinaison de paramètres $\mathbf{g}_{\text{param}}$, une validation croisée par k "folds" est réalisée sur l'ensemble des données d'entraînement et l'erreur de prédiction est moyennée sur ces "folds". La combinaison de paramètres donnant l'erreur de prédiction moyenne la plus basse est gardée pour la suite du modèle. L'étape peut être résumé par l'équation suivante :

$$\mathbf{g}^*_{\text{param}} = \min_{\mathbf{g}_{\text{param}}} \left(\frac{\sum_k \text{err}(\mathbb{X}^k_{\text{val}} \mid \mathbf{g}_{\text{param}})}{k} \right) \tag{5.5}$$

avec $\text{err}(\mathbb{X}^k_{\text{val}} \mid \mathbf{g}_{\text{param}})$ l'erreur de prédiction sur les données du jeu de validation du fold k, lorsque la forêt est construit avec le jeu de paramètres $\mathbf{g}_{\text{param}}$ et les données du jeu d'entraînement du fold k $\mathbb{X}^k_{\text{train}}$. Les entrées de cette étape lors de l'entraînement sont donc :

— les données cliniques, radiomiques et volumiques des jeux d'entraînement et de validation. Elles sont contenues dans un vecteur de caractéristiques $\mathbb{X}_t$ avec $\mathbb{X}_t = \{\mathbb{X}^k_{\text{train}}, \mathbb{X}^k_{\text{val}}\}$ avec $k \in [1:5]$ le numéro du "fold",

— un vecteur associé $\mathbb{Y}_t$.

La sortie de cette étape lors de l'entraînement est :

— une liste $\mathbf{g}^*_{\text{param}}$ des valeurs optimales pour chaque hyperparamètre.

Cette étape n'est pas recalculée lors du test. Cette étape n'a donc jamais connaissance du jeu de données de test.

5.3.2 Le calcul de l'importance des variables

La seconde étape (en vert dans la figure 5.4) correspond au classement des variables par VIMP. Ce classement est effectué, compte tenu du vecteur $\mathbf{g}^*_{\text{param}}$ contenant les hyperparamètres optimaux de la première étape (étape 0), sur la base d'une RSF et d'une évaluation de l'importance des variables (VIMP) [voir section 5.2]. VIMP a été exécuté 100 fois (avec 100 séparations aléatoires en jeux d'entraînement et de validation) et chaque caractéristique c_j a été classée en fonction de la somme VS_{c_j} de l'importance de la variable sur les 100 itérations afin de faire face aux instabilités résultant du caractère aléatoire du modèle. Le calcul de VS_{c_j} est présenté dans l'équation suivante :

$$\text{VS}_{c_j} = \sum_{r=1}^{100} \text{VIMP}_{c_j}(\mathbb{X}^r_{\text{val}} | \mathbf{g}^*_{\text{param}}) \tag{5.6}$$

Nous avons donc en entrée de cette étape lors de l'entraînement :

— les données des jeux d'entraînement et de validation $\mathbb{X}_t$,

— un vecteur associé $\mathbb{Y}_t$,

— le vecteur $\mathbf{g}^*_{\text{param}}$ contenant les hyperparamètres optimaux.

Nous avons en sortie de cette étape lors de l'entraînement :

— un vecteur $\mathbb{X}_{\text{rank}}$ contenant les variables ordonnancées de la plus grande à la plus faible valeur d'importance VS_{c_j} avec $\mathbb{X}_{\text{rank}} = \pi_{\text{VS}}(\mathbb{X})$ (π_{VS} étant la permutation des colonnes en fonction de leur valeur VS).

Cette étape n'est pas non plus recalculée lors du test. Cette étape n'a donc jamais connaissance du jeu de données de test.

5.3.3 L'entraînement du modèle final et la prédiction

Dans la dernière étape (bloc rose sur la figure 5.4), un second RSF est entraîné pour la prédiction de la mortalité (reliée à la PFS) grâce aux paramètres sélectionnés lors de la première étape. Il va permettre lors de l'entraînement, de déterminer le nombre optimal de caractéristiques à garder et le seuil de séparation des groupes de pronostique (haut et bas risque).

La recherche du nombre optimal de variables à garder lors de l'entraînement peut être

décrit par l'équation suivant :

$$n_f^* = \min_{n_f} \left(\frac{\sum_k \text{err}(\mathbb{X}_{\text{val}}^k \mid n_f \cap \mathbb{X}_{\text{rank}} \cap \mathbf{g}_{\text{param}})}{k} \right) \tag{5.7}$$

avec n_f^* le nombre de variables à garder optimal, et n_f le nombre de variables à garder testé. Chaque valeur n_f est testée en commençant avec les valeurs les plus prédictives de $\mathbb{X}_{\text{rank}}$ et en ajoutant au fur et à mesure les variables de moins en moins prédictives. Le nombre de variables à garder optimal n_f^* correspond à celui donnant l'erreur de prédiction moyennée sur les "k" folds minimale.

Nous avons donc en entrée de cette étape lors de l'entraînement :

— les données des jeux d'entraînement et de validation $\mathbb{X}_t$,

— un vecteur associé $\mathbb{Y}_t$,

— le vecteur $\mathbf{g}_{\text{param}}^*$ contenant les hyperparamètres optimaux,

— le vecteur $\mathbb{X}_{\text{rank}}$ contenant les variables ordonnancées de la plus grande à la plus faible valeur d'importance VS.

Nous obtenons en sortie de cette étape lors de l'entraînement :

— le nombre optimal de variables à garder n_f^*,

— la mortalité prédite pour les données des jeux d'entraînement et de validation (respectivement $\hat{M}(\mathbb{X}_{\text{train}})$ et $\hat{M}(\mathbb{X}_{\text{val}})$ qui peuvent être associée à la PFS [voir l'équation 5.4]. Cette mortalité est obtenue avec la forêt construite en utilisant les n_f^* meilleures variables de $\mathbb{X}_{\text{rank}}$ et les paramètres d'arbres $\mathbf{g}_{\text{param}}^*$.

Lorsqu'un test est réalisé, la mortalité $\hat{M}(\mathbb{X}_{\text{test}})$ est prédite grâce aux données $\mathbb{X}_{\text{test}}$, $\mathbb{Y}_{\text{test}}$, et à la meilleure forêt F^* construite à partir :

— des données du jeu d'entraînement $\mathbb{X}_{\text{train}}$,

— du vecteur associé $\mathbb{Y}_{\text{train}}$,

— du vecteur $\mathbf{g}_{\text{param}}^*$,

— du vecteur $\mathbb{X}_{\text{rank}}$,

— la valeur n_f^*.

5.3.4 La séparation des patients en groupes pronostiques

Les mortalités prédites sur le jeu de données d'entraînement, obtenues avec la forêt construite en utilisant les n_f^* meilleures variables de $\mathbb{X}_{\text{rank}}$ et les paramètres d'arbres $\mathbf{g}_{\text{param}}^*$, sont utilisées pour déterminer le meilleur seuil de séparation des mortalités en groupe pronostiques. La séparation en deux groupes pronostiques se fait grâce au test du Log-Rank. L'équation du Log-Rank pour la séparation en groupes pronostiques est équivalente à l'équation 5.1 en remplaçant les noeuds fils gauche et droit (respectivement

l et r) par les groupes de haut et bas risque (respectivement h et b). Ainsi, pour différentes valeurs de mortalités $\hat{M}(\mathbb{X}_{\text{train}})$ prédites sur le set d'entraînement [voir l'équation 5.3], nous réalisons le test du Log-Rank sur les sous-groupes déterminés par la valeur seuil M_{th}. Définissons que lorsque la mortalité prédite est inférieure ou égale à M_{th}, le patient ira dans le groupe de bas risque b. Dans le cas contraire, il ira dans le groupe à haut risque h. Soit $t_1 < ... < t_z$ les z temps distincts dans le set d'entraînement, $d_{k,h}$ et $\Upsilon_{k,h}$ respectivement le nombre d'évènements et le nombre d'individus à risque au temps t_k dans le groupe à haut risque ($d_{k,b}$ et $\Upsilon_{k,b}$ pour le groupe à bas risque). On considère à risque au temps t_k un individu qui n'a pas eu d'évènement au temps t_k. Une valeur M_{th} permet de distribuer les n individus en deux groupes, avec n_h et n_b respectivement dans les groupes à haut et bas risque, tel que $n = n_h + n_b$. De plus, pour chaque temps t_k, la somme des individus à risque et le nombre d'évènements sont préservés. Ainsi, $\Upsilon_k = \Upsilon_{k,h} + \Upsilon_{k,b}$ et $d_k = d_{k,h} + d_{k,b}$. La valeur du test Log-Rank pour la valeur seuil de mortalité M_{th} est :

$$LR(M_{th}) = \frac{\sum_{k=1}^{z}(d_{k,h} - \Upsilon_{k,h}\frac{d_k}{\Upsilon_k})}{\sqrt{\sum_{k=1}^{z}\frac{\Upsilon_{k,h}}{\Upsilon_k}(1-\frac{\Upsilon_{k,h}}{\Upsilon_k})\frac{\Upsilon_k - d_k}{\Upsilon_k - 1})d_k}} \tag{5.8}$$

La meilleure séparation M_{th}^* (et donc le meilleur seuil M_{th} sur les mortalités) sera celle qui donnera la valeur absolue de Log-Rank $|LR(M_{th})|$ la plus grande. Ainsi les patients avec une mortalité inférieur au seuil M_{th}^* sera dans le groupe de bas risque et ceux avec une mortalité supérieure à M_{th}^* sera dans le groupe de haut risque. On évalue ensuite la significativité de la séparation grâce à la p-value.

Lorsque un test est réalisé, on utilise la mortalité $\hat{M}(\mathbb{X}_{\text{test}})$ prédite sur les données de test grâce la meilleure forêt F^* et la meilleure séparation M_{th}^* pour obtenir des groupes de pronostique (haut et bas risque).

5.4 Pré-traitement et récupération des variables

5.4.1 Extraction des images et segmentation

Le calcul des radiomiques, et donc la prédiction de la survie, vont dépendre du bon pré-traitement des images. Après avoir récupéré les images 18F-FDG TEP sous forme de Dicom après reconstruction, ainsi que les ROI des lésions calculées par propagation à partir d'un point dans la lésion sélectionné par les médecins, une segmentation des lésions a été réalisées de façon semi-automatique. Celle-ci est réalisée par vote majoritaire des segmentations 40% , 2.5 du SUV max et k-means. Les détails de ces méthodes sont présentés dans l'annexe A.

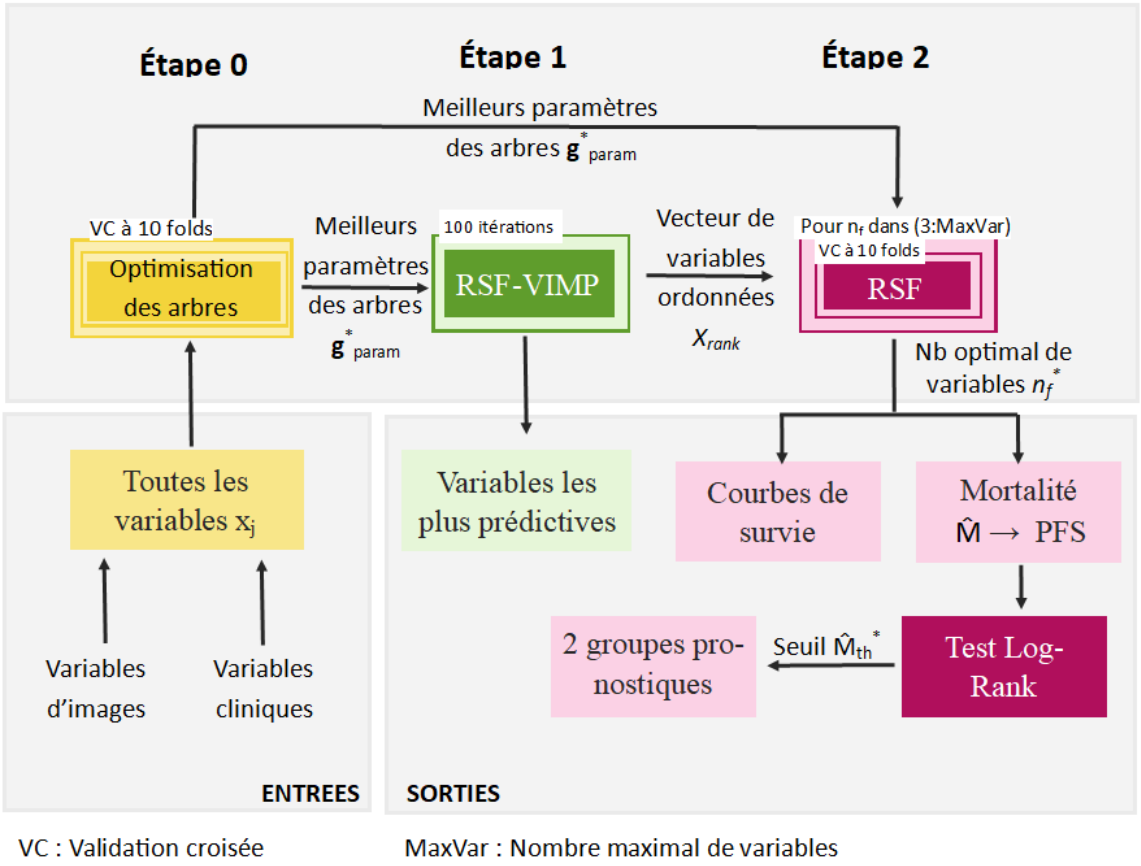


FIGURE 5.4 – Modèle initié pour la prédiction de la survie publié à IJCARS, pour le groupe d'expériences techniques. En jaune : étape d'optimisation, en vert : classement des variables par VIMP, en rose : entraînement et prédiction avec les RSF.

5.4.2 Récupération des données cliniques et d'images (autres que texturales)

Différentes caractéristiques ont été relevées lors des études cliniques IMAJEM [1] et EMN02/HO95 [2]. Outre les données cliniques classiques telles que l'âge, le sexe ou les valeurs sanguines (calcémie, hémoglobine, etc.), des caractéristiques peuvent être calculées grâce aux images TEP, TDM et IRM. Ces caractéristiques sont calculées sur des lésions focales (LF) et/ou des lésions diffuses de la moelle osseuse (BM). Les LF ont été définies comme étant des foyers à l'intérieur des os plus intenses que le fond normal de la moelle osseuse, avec ou sans lésion lytique sous-jacente, et présents sur au moins deux coupes consécutives. L'atteinte diffuse de la moelle osseuse (BM) a été définie comme une captation homogène dans le squelette axial et appendiculaire supérieure à celle du foie ou comme une captation hétérogène quelle que soit l'intensité de la captation. Ainsi nous allons considérer différentes valeurs calculées sur ces lésions :

— Le score **R-ISS** ou Système International de Stadification Révisé permet de définir le stade de la maladie (I, II ou III).

— **Le nombre de LF** a été dichotomisé avec un seuil fixé à 3 [2].

— Une analyse visuelle utilisant l'échelle standard à 5 points de **Deauville**[2] sur les résultats de l'examen a été effectuée. Le score de Deauville peut être calculé sur la moelle osseuse ou sur les lésions focales.

— L'**EMD** (Extra-Medullar Disease) correspond à la valeur binaire de présence de maladie extra-médullaire.

— Les paramètres métaboliques basés sur le volume ont été calculés pour la lésion la plus fixante : Le **MTV** (Metabolic Tumor Volume) désigne le volume métaboliquement actif de la tumeur et quantifie la charge tumorale métabolique totale. La **TLG** (Total Lesion Glycolysis) tient compte du niveau d'accumulation du glucose dans le volume total de toutes les régions d'intérêt et est calculée comme le produit du MTV et du SUV moyen dans la lésion segmentée.

On considèrera par la suite **wbDeauville** score comme le maximum entre le Deauville LF et deauville BM. **wbTLG** sera la somme des TLG de toutes les lésions, et wbSUVmax le maximum entre SUV BM et SUV LF (ou EMD). De plus, à chaque patient est associé le **traitement** qui lui a été administré. Dans IMAJEM [1], les deux bras de traitement sont :

— Bras de traitement A (sans autogreffe de cellules souches) :
  — Induction : 3 cycles de RVD (lenalidomide, bortezomib, et dexamethasone)
  — culture de cellules souches
  — 5 cycles de RVD
  — Maintenance : 1 an de lenalidomide

— Bras de traitement B (avec autogreffe de cellules souches) :
  — Induction : 3 cycles de RVD
  — Culture de cellules souches
  — Transplantation : Melphalan $200mg/m^2$ plus autogreffe de cellules souches (ASCT)
  — 2 cycles de RVD
  — Maintenance : 1 an de lenalidomide

Dans EMN02/HO95 [2] les deux bras de traitement sont :

— Bras de traitement A (sans autogreffe de cellules souches) :
  — Induction : Thalidomide + Dexamethasone

2. Le score de Deauville permet d'évaluer la réponse métabolique au traitement. Il consiste à comparer la fixation à la lésion avec celle au foie.

  - Récupération de cellules souches du sang périphérique
  - Maintenance : Interféron α

- Bras de traitement B (avec autogreffe de cellules souches) :
  - Induction : Thalidomide + Dexamethasone
  - Récupération de cellules souches du sang périphérique
  - Transplantation : Première ASCT (avec Melphalan $200mg/m^2$)
  - Transplantation : Deuxième ASCT (avec Melphalan $200mg/m^2$)
  - Maintenance : Interféron α

5.4.3 Calcul des caractéristiques texturales

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

5.4.3.1 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 (ici segmentation 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 [25]. Les caractéristiques agnostiques peuvent être séparées en caractéristiques de premier, second ordre, voir plus.

- Premier ordre : description de la distribution des valeurs d'intensité des voxels individuels sans prendre en compte les relations dans l'espace (moyenne, max, min, uniformité, entropie, des intensités, asymétrie, kurtosis de l'histogramme des valeurs).
- Second ordre : description de la texture (relation entre les voxels d'intensité similaire) qui sont calculés en utilisant une matrice de co-occurrence.

Le tableau 5.1 donne des exemples de caractéristiques agnostiques et sémantiques. Les définitions peuvent différer. Par conséquent, les définitions données ici sont celles de l'IBSI (Image Biomarker Standardisation Initiative) [3]. Des exemples de calculs de radiomiques (premier et second ordres) sont données dans l'annexe B.

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

5.4.3.2 Les méthodes de calcul

Le calcul de ces matrices et caractéristiques peut se faire de diverses façons. Elles peuvent être calculées en 2D ou 3D (OM : One matrix) et il existe plusieurs méthodes de rééchantillonnage. Deux ont été testées ici :

— AR (Absolute Resampling) : la largeur de bande de l'histogramme d'intensités de voxels reste fixe pour tous les patients (ici 0,3).

— RR (Relative Resampling) : le nombre de bandes de l'histogramme reste fixe (ici 64).

De plus, on peut égaliser l'histogramme ou non pour construire la matrice (on ajoutera le suffixe Heq dans la suite du manuscrit lorsque l'histogramme est égalisé). Enfin, on peut utiliser une taille de voxel qui varie ou non (on ajoutera le suffixe equalsize dans la suite du manuscrit lorsque la taille de voxel ne varie pas). Au total 6 implémentations sont testées :

— OMAR

— OMRR

— OMRR + Heq

— OMAR + equalsize

— OMRR + equalsize

— OMRR + equalsize + Heq

Chapitre 6

Validation expérimentale

La base de données, le nombre de patients, les valeurs d'entrées et la méthode d'évaluation ne sont pas les mêmes en fonction de l'expérience réalisée. On peut considérer deux groupes d'expériences avec les RSF : le groupe d'expériences techniques et le groupe d'expériences cliniques. La section 6.1 comportera les détails de ces expériences et la section 6.2 leurs résultats.

6.1 Détails d'implémentation

Le groupe d'expériences techniques a été réalisé avec IMAJEM [1], lorsque le nombre de patients était relativement faible, mais a permis d'établir le modèle. Il comprend les expériences :

— de comparaison aux méthodes traditionnelles,

— du choix de la méthode de sélection des variables,

— du choix des données d'entrée et d'identification de l'importance des caractéristiques texturales.

Ces expériences ont été publiées dans IJCARS [6].

Le groupe d'expériences cliniques a quant à lui était réalisé par la suite avec les deux bases de données (IMAJEM [1] et EMN02/HO95 [2]), un modèle déjà défini et une connaissance plus approfondie des variables à garder. Il a pour objectifs de prédire le pronostic des patients atteints de myélome multiple, de déterminer les biomarqueurs de la maladie, et de valider la généralisation de la méthode a des données multicentriques. Ces expériences ont été publiées dans EJNMMI [7].

Les différences entre les expériences sont résumées dans le tableau 6.1.

TABLE 6.1 – Différences entres les expériences techniques et cliniques.

Groupe d'expériences	Techniques	Cliniques	Cliniques
Version	Entraînement / Validation	Entraînement / Validation + Test	"Nested"
Nombre de "folds"	10 fois 10 "folds"	10 fois 4 "folds" (70%) + test (30%)	5 "folds" (chaque fold devient le test une fois)
Base de données	IMAJEM	IMAJEM + EMN02/HO95	IMAJEM + EMN02/HO95
Nombre de patients	66	139	139
Nombre de variables	129	22	22
Objectif	Déterminer le modèle	Déterminer les biomarqueurs	Évaluer la généralisation à de nouvelles données

6.1.1 Validation expérimentale des radiomiques par IBSI

Avant le calcul des caractéristiques texturales sur nos données, une étape de validation est nécessaire, afin de garantir une standardisation de leur calcul. En effet, il existe de très nombreuses façons de les calculer, ce qui peut amener des résultats très différents en fonction des choix faits par la personne réalisant le calcul. Cette validation suit les recommandation de l'IBSI. L'IBSI ou Image Biomarker Standardisation Initiative [3], est une collaboration internationale qui travaille à la standardisation des biomarqueurs extraits d'images. Ils utilisent une base de données de référence et fournissent les valeurs de biomarqueurs associées 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 synthétique pour simuler une image médicale [voir la figure C.1 de l'annexe C]. C'est une matrice 3D qui a ici été utilisée afin de calculer les biomarqueurs désirés. La figure C.2 de l'annexe C donne un exemple de présentation des valeurs des biomarqueurs de la matrice GLCM. Notre implémentation de l'extraction des caractéristiques est donc vérifiée dans un premier temps en étant appliquée au fantôme de l'IBSI présent dans la figure C.1 de l'annexe C. Les calculs se font avec les paramètres suivants : une matrice 3D avec échantillonnage absolu, et pas d'égalisation ou de normalisation de la taille des pixels (OMAR).

6.1.2 Détails du groupe d'expériences techniques

6.1.2.1 Détails sur les données d'entrée

Initialement seule la base ***IMAJEM*** [1] était disponible. C'est donc naturellement que les expériences techniques furent réalisées sur cette base. Un total de 134 patients a été inclus dans l'étude, mais seuls ***66 patients*** étaient éligibles pour le calcul des caractéristiques texturales car ils présentaient des LF de taille analysable (64 voxels par lésion).

Pour ces expériences, nous n'avons pas pré-sélectionné les variables (***129 variables gardées***) afin d'évaluer l'importance des données cliniques et d'image. Parmi les 129, 13 variables étaient cliniques et 6 conventionnelles (mesures quantitatives effectuées sur les images TEP, mais non basées sur l'hétérogénéité intra-tumorale). Pour les variables texturales, plusieurs implémentations ont été calculées pour chaque variable, et toutes ont été gardées en entrée du modèle dans le but de déterminer quelle implémentation donne les meilleures prédictions grâce à l'étape de sélection VIMP. La corrélation entre les différentes implémentations de la même variable est automatiquement traitée dans cette étape. Ces implémentations considèrent la normalisation (égalisation absolue, relative ou par histogramme) et l'égalisation des voxels (taille égale ou non). Avec 6 implémentations différentes [voir section 5.4.3.2] de 19 variables nous arrivons à 114 variables texturales. Or, étant donné que les SUVmax sont équivalent dans la majorité des implémentations (car non dépendant du rééchantillonage absolue ou relatif, et le SuvMax reste à 63 lors de l'égalisation des histogrammes avec le rééchantillonage relatif) nous en gardons que 2 (variation de la taille des voxels ou non). Nous arrivons donc à 110 variables texturales. Les variables gardées sont présentées dans le tableau 6.2.

6.1.2.2 Détails d'optimisation et de validation du modèle

La méthode globale proposée (figure 5.4) comportant les trois étapes (d'optimisation d'hyperparamètres, de sélection de variables et de prédiction), est entraînée ***10 fois***, avec une ***validation croisée (VC) de 10 "folds"***, afin d'évaluer la variabilité des RSF (***version Entraînement/validation***).

TABLE 6.2 – Liste des caractéristiques cliniques, conventionnelles (première ligne) et texturales (deuxième et troisième lignes) utilisées dans le groupe d'expériences techniques. Toutes les caractéristiques texturales ont été calculées avec 6 implémentations différentes. ** : variables basées sur les images*

Conventionnelles*	**Variables cliniques**	**Variables cliniques***
SUVmax LF	Age	Nombre de LF (PET baseline)
SUVmax BM	Sexe	Présence de maladie extra-médullaire (EMD)
Total MTV	Hémoglobine	Score Deauville LF
LF MTV	Calcémie	Score Deauville BM
Total TLG (Total lesion glycolysis)	Créatinine	Score Deauville global
TLG LF	R-ISS	Nombre de LF (IRM)
	Bras de traitement	
GLCM* (Gray Level Co-Ocurence Matrix)	**GLRLM*** (Gray Level Run-Length Matrix)	**GLSZM*** (Gray Level Size-Zone Matrix)
Homogeneity	HGRE (High Gray Level Run Emphasis)	HGZE (High Gray Level Zone Emphasis)
Entropy	LGRE (Low Gray Level Run Emphasis)	ZLNU (Size-Zone Non-Uniformity)
Energy	SRE (Short Run Emphasis)	SZHGE (Small Area High Gray Level Emphasis)
Correlation	LRE (Long Run Emphasis)	LZLGE (Low Gray Level Zone Emphasis)
Contrast		SZE (Small Area Emphasis)
Dissimilarity		ZP (Zone percentage)
		RP (Run percentage)
First order*		
Maximum		

Les différentes valeurs de paramètres évaluées dans la partie optimisation du modèle sont :

— Le nombre d'arbres : {20,50,100,500,1000}.

— Le mode de séparation : Log-Rank et Log-Rank aléatoire.

— Nombre minimal d'échantillon dans chaque noeud : [5 : 12].

— Nombre maximal de variables dans chaque noeud : $\{50\%, 70\%, 100\%\}$ du nombre total de variables.

6.1.3 Détails du groupe d'expériences cliniques

6.1.3.1 Détails sur les données d'entrée

Par la suite, lorsque la base de données EMN02/HO95 fut disponible nous avons combiné ces deux bases pour une analyse plus orientée clinique. Nous avions le choix de garder l'une pour l'entraînement et l'autre pour le test ou de les mélanger. Nous avons choisi de les mélanger, car les deux bases était légèrement différentes (le design de l'étude est légèrement différent et la molécule étudiée, bien que similaire n'est pas la même). Une harmonisation des données à été réalisée au niveau de chaque pays (France et Italie), et

non pas au niveau des institutions car elles sont trop nombreuses avec chacune un petit nombre de patients. L'harmonisation a été réalisée grâce à l'approche M-ComBat [74]. La méthode M-ComBat est une modification de la méthode ComBat [75] qui permet d'éliminer les effets de groupes et qui est basée sur un cadre empirique de Bayes. Au contraire de ComBat qui consiste à déplacer les échantillons vers la grande moyenne et la variance groupée, M-Combat les déplace vers la moyenne et la variance du lot de référence "golden-standard".

Parmi tous les patients initialement inclus dans ces études, seuls ceux possédant une FDG-TEP initiale positive et des LF ou de la maladie extra-médullaire (EMD) furent inclus dans ces analyses. Respectivement 134 et 94 patients sont présents dans les bases ***IMAJEM*** [1] et ***EMN02/HO95*** [2] mais seuls 102 and 71 patients avaient des LF ou de l'EMD. De plus, 33 patients avaient des LF et lésions EMD qui ne respectaient pas une taille minimale ou avait des données originales manquantes ou non lisibles. Cela implique donc que seul un sous-groupe de ***139 patients*** fut finalement gardé.

Parmi toutes les caractéristiques texturales calculées en utilisant la lésion LF ou EMD la plus intense, nous avons considéré uniquement les variables texturales les plus robustes selon des études préliminaires [53, 55, 76]. Les images TEP ont été rééchantillonnées à la même taille de voxel ($2 \times 2 \times 2\,mm^3$) en utilisant une interpolation par splines bicubiques, comme suggéré par Hao et al. [77]. Les matrices GLCM (Gray Level Co-Ocurence Matri) et GLRLM (Gray Level Run-Length Matrix) ont été calculées en utilisant une seule matrice, en prenant en compte 13 directions simultanément avec un déplacement d'un voxel. De plus, plusieurs caractéristiques texturales ont été calculées sur la base de la matrice de taille de niveau de gris (GLSZM). Deux méthodes de quantification ont été considérées : une égalisation linéaire utilisant 64 bandes et une quantification absolue utilisant des bandes de largeur fixe de 0,3 SUV. Une taille minimale de 64 voxels (comme dans l'image originale avant rééchantillonnage) était requise pour le calcul de la caractéristique texturale.

La dernière étape de l'extraction des caractéristiques a été consacrée au choix de caractéristiques texturales non corrélées (parmi les 15 initialement calculées). À cette fin, la corrélation de Spearman [1] (ρ) a été utilisée, et lorsque deux caractéristiques étaient fortement corrélées ($> 0{,}8$ [78]), une seule a été retenue au hasard pour être considérée dans le modèle final. Ainsi, seul 5 caractéristiques texturales furent gardées. Sont ajoutées à cela 7 autres variables d'images et 10 variables cliniques et histopathologiques, ce qui donne un total de ***22 variables***. Les variables gardées sont présentées dans le tableau 6.3.

1. Coefficient permettant de mesurer la corrélation entre deux variables. Il est intéressant dans le cas où la relation n'est pas affine, mais il semble y avoir une corrélation.

TABLE 6.3 – Liste des caractéristiques cliniques, conventionnelles (première ligne) et texturales (deuxième et troisième lignes) utilisées dans le groupe d'expériences cliniques. ** : variables basées sur les images*

Conventionnelles*	**Variables cliniques**	**Variables cliniques***
SUVmax LF	Age	Nombre de LF (PET baseline)
SUVmax BM	Sexe	EMD
wb SUVmax	Hémoglobine	wb Deauville
MTV LF	R-ISS	Score Deauville LF
Score Deauville BM		
Total MTV (MTV)	Bras de traitement	
TLG LF		
wb TLG		
GLCM* (Gray Level Co-Ocurence Matrix)	**GLRLM*** (Gray Level Run-Length Matrix)	**GLSZM*** (Gray Level Size-Zone Matrix)
Homogéneité OMAR	LGRE OMAR (Low Gray Level Run Emphasis)	HGZE OMRR (High Gray Level Zone Emphasis)
Entropie OMRR		ZLNU OMRR (Size-Zone Non-Uniformity)

6.1.3.2 Détails d'optimisation et de validation du modèle

Deux approches d'évaluation du modèle ont été réalisées. Une première où la base de données a été séparée en 2 (70% dans le jeu d'entraînement soit 98 patients, et 30% dans le jeu de test soit 41 patients) puis le jeu d'entraînement à été séparé en ***4 sous-ensembles ou "folds"*** pour une validation croisée, est présentée dans la figure 6.1. Nous appellerons cette séparation la version ***Entraînement/Validation + Test*** Cette approche permet d'évaluer la capacité de généralisation du modèle de RSF à des données test observées pour la première fois. Le modèle global (les trois étapes) est finalement répété ***10 fois*** afin d'évaluer la robustesse du modèle. Cependant, cette première approche peut induire un biais qui est souvent négligé dans de nombreuses études, car un ensemble de test donné peut ne pas être représentatif de la diversité des patients/populations.

La deuxième approche est une ***validation croisée imbriquée "nested"*** de ***5 "folds"***, qui sépare l'ensemble de données en 5 sous-ensembles ou "folds". Elle est présentée dans la figure 6.2. Alternativement, chaque fold joue le rôle de l'ensemble de test "extérieur", tandis que les "folds" restants sont utilisés pour la sélection et l'entraînement du modèle. Ce processus est répété 5 fois conduisant à 5 modèles et 5 prédictions (une fois par "fold". Les différentes valeurs de paramètres évalués dans la partie optimisation du modèle sont équivalents à ceux du groupe d'expériences techniques [voir sous-section 6.1.2.2].

6.1.4 L'évaluation du modèle

Les modèles sont évalués par l'erreur de prédiction de la PFS ($1 - c - index$), ainsi que par la séparation en deux groupes pronostiques avec la p-value associée. En effet, après avoir séparé les patients en deux groupes grâce au test du Log-Rank, des courbes de

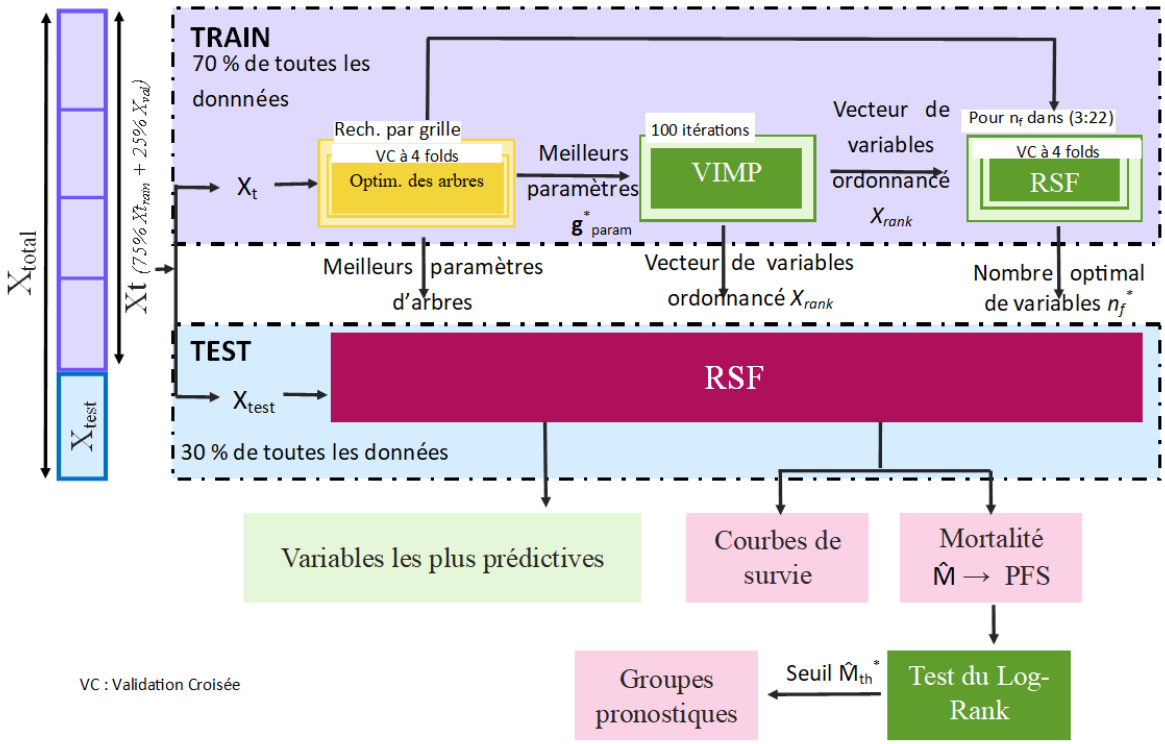


FIGURE 6.1 – Modèle proposé à EJNMMI, pour le groupe d'expériences cliniques, version "Entraînement/Validation + Test".

survie des patients dans chaque groupe sont réalisées grâce à la méthode de Kaplan-Meier et la significativité de la différence entre les deux courbes (avec un intervalle de confiance de 95%) peut être évaluée par la p-value. Ainsi, plus l'erreur de prédiction et la p-value seront faibles et meilleur sera le résultat [voir section 3.4].

6.2 Résultats

6.2.1 Les sorties du modèle

6.2.1.1 Sorties des expériences techniques

Le tableau 6.4 montre les valeurs de mortalité prédite par notre modèle pour 17 patients et leur groupe pronostique prédit. Les résultats de ce tableaux sont associés aux courbes de la figure 6.3. La figure 6.3 montre un exemple de la courbe de Kaplan-Meier obtenue après une séparation avec un test Log-Rank, sur les mortalités prédites, dans le cadre des expériences techniques.

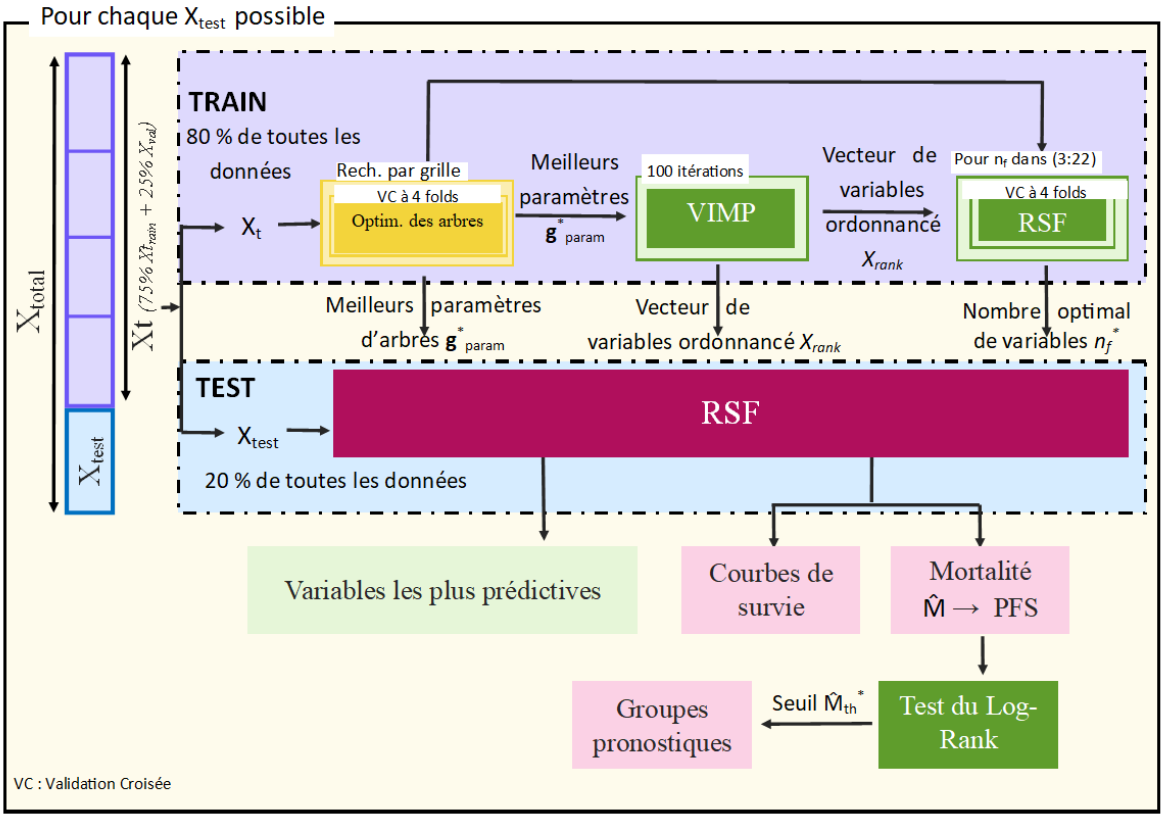


FIGURE 6.2 – Modèle proposé à EJNMMI, pour le groupe d'expériences cliniques, version validation croisée "nested".

TABLE 6.4 – Expériences techniques : Un exemple du résultat de notre méthode. A chaque patient est attribué un groupe pronostique et une mortalité prédite. Le tableau correspond aux courbes de survie de la figure 6.3.

Patient	1	2	3	4	5	6	7	8	9
Groupe pronostique	haut risque	haut risque	bas risque	haut risque	bas risque	bas risque	bas risque	haut risque	bas risque
Mortalité	38.906	39.165	6.489	39.165	31.448	23.185	31.034	39.623	33.499
patient	10	11	12	13	14	15	16	17	
Groupe pronostique	bas risque	bas risque	bas risque	bas risque	bas risque	bas risque	bas risque	bas risque	
Mortalité	10.042	33.486	23.508	6.569	26.998	36.075	22.780	31.530	

6.2.1.2 Sorties des expériences cliniques

Lors des expériences cliniques, un ensemble de données de test (non connu par le modèle lors de l'entraînement) a été utilisé afin d'évaluer sa capacité de prédiction et de généralisation. Le modèle a prouvé ces capacités grâce à une erreur de prédiction moyenne de 0,36 ± 0,03 sur les 10 lancements de l'optimisation randomisée de noeuds. De plus, nous pouvons voir un exemple de courbe de Kaplan-Meier dans la figure 6.4 qui montre la capacité du modèle à séparer les données en groupes de survie (bas risque et haut risque) avec une p-value moyenne calculée sur les 10 lancements de 0,01 ± 0,01. La médiane de la PFS dans ce jeu de données test et pour le faible risque était de 3,7 ans alors qu'elle était de 1,8 an pour le groupe à haut risque. Enfin, la figure 6.5 montre les courbes des

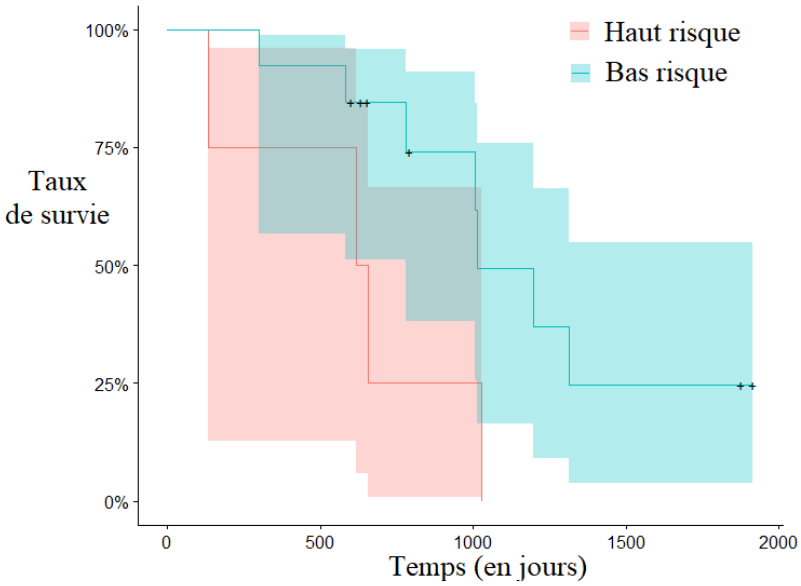


FIGURE 6.3 – Expériences techniques : Un exemple de courbes de survie Kaplan-Meier du meilleur modèle (erreur de prédiction de 0,39), séparées selon la règle du Log-Rank sur la mortalité estimée (p-value de 0,04).

10 lancements de prédiction sur le test et permet de mettre en évidence la stabilité de la prédiction par notre modèle sur un ensemble de données qui n'a pas été observé au préalable par le modèle.

La validation croisée "nested" a montré une erreur de prédiction moyenne sur les 5 "folds" de 0,40 ± 0,05 lorsqu'elle a été appliquée à l'ensemble de données test de chaque "fold". Ainsi, bien que la valeur de l'erreur de prédiction moyenne soit plus faibles qu'avec le modèle Entraînement/Validation + Test, le résultat est plus robuste et il reste tout de même assez bas pour dire que le modèle permet une bonne prédiction, peu importe les données test. Cela montre qu'une généralisation à de nouvelles données est possible. Elle a aussi permis de séparer les patients en deux groupes, car la p-value du Log-Rank moyenne est de 0,03 ± 0,05.

6.2.2 Comparaison des méthodes de survie

6.2.2.1 Comparaison aux méthodes traditionnelles

Lors des expériences techniques, notre modèle fut comparé à des méthodes traditionnelles : la régression de Cox sous pénalisation Lasso (Lasso-cox) [43], et la méthode de gradient-boosting Cox (GB-cox) [79] récemment rapporté comme compétitive [44]. Ces deux méthodes sont des méthodes de Cox avec sélection de variables afin que les modèles soient comparables. Les résultats sont présentés dans la figure 6.6 et la table 6.5.
La figure 6.6 illustre l'erreur de prédiction suggérant que notre modèle surpasse les méthodes de Lasso-Cox et GB-Cox. On peut observer dans la table 6.5 que la moyenne des

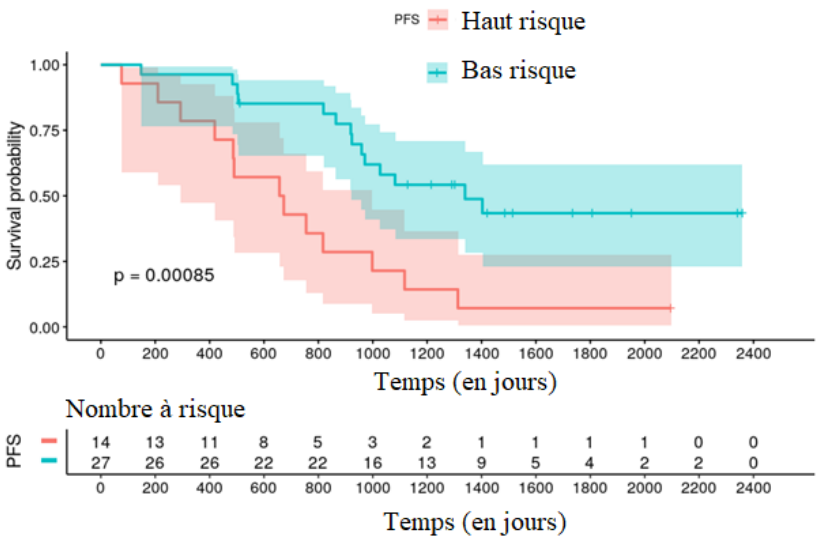


FIGURE 6.4 – Expériences cliniques : Un exemple (parmi les 10 calculés) de courbes de Kaplan- Meier calculées sur l'ensemble de données test pour la stratification de la PFS avec trois caractéristiques conservées dans le modèle (bras de traitement, hémoglobine et SUVmaxBM). La p-value du Log-Rank était de 0,00085 avec un rapport des risques HR = 3,3 (95% CI 1,3-8,2), CI étant l'intervalle de confiance. Le temps est en jours [7].

TABLE 6.5 – Expériences techniques : Erreur moyenne de prédiction de la PFS sur 10 lancements, et p-value moyenne en fonction du modèle.

Méthode	Erreur de prédiction moyenne	p-value moyenne
Notre méthode (RSF+VIMP)	0,36 ± 0,015	0,05
Gradient-Boosting Cox	0,56 ± 0,011	0,27
Lasso-Cox	0,48 ± 0,042	0,4

erreurs de prédiction est très prometteuse (0,36) et quelle est la plus basse des méthodes comparées. En outre, la p-value de la séparation des groupes de pronostic fut calculée. La table 6.5 montre aussi que seule la séparation obtenue après prédiction par notre modèle, donne une p-value moyenne statistiquement significative (0,05).

6.2.2.2 Intérêt de la méthode de sélection

Outre la comparaison avec des méthodes traditionnelles, nous comparons notre modèle (lors des expériences techniques) avec différentes méthodes de sélection de variables proposées par Ishwaran [73]. La comparaison des erreurs de prédiction est présentée dans la figure 6.7a. et est précisée dans le tableau 6.6. On peut y voir que l'utilisation d'une sélection de variable associée aux RSF surpasse largement la méthode RSF seule et montre son intérêt. De plus, on peut voir que l'utilisation de la méthode de sélection VIMP surpasse grandement les autres méthodes et justifie notre choix. Ces résultats sont confirmés par

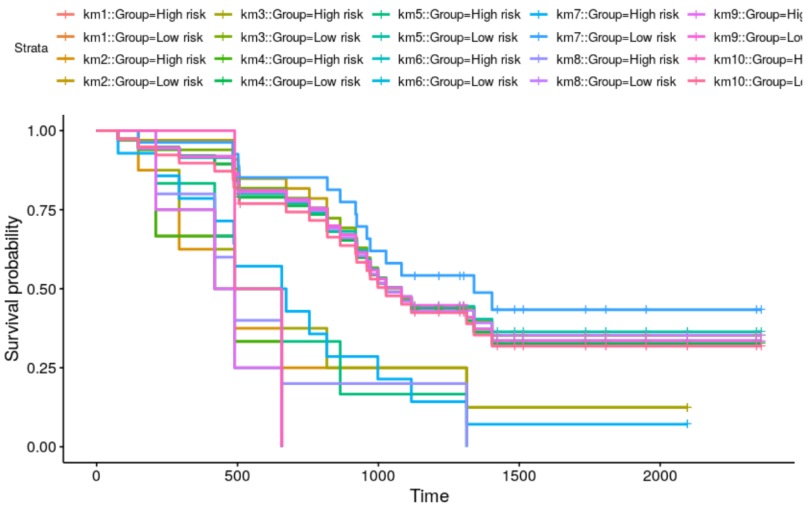


FIGURE 6.5 – Expériences cliniques : Prédiction de la stratification de la PFS calculée sur l'ensemble de données de test pour les 10 lancements [7].

le tableau 6.7 qui indique que les méthodes avec sélection (p-value moyenne entre 0,05 et 0,24) permettent une meilleure séparation que les RSF sans sélection (p-value moyenne de 0,40), et que l'utilisation de VIMP (p-value moyenne de 0,05) améliore encore cette séparation comparé aux autres méthodes, étant donné que c'est la seule méthode avec une séparation significative.

Enfin, un autre résultat intéressant est présenté dans la figure 6.7b. et dans le tableau 6.6. On y voit que la méthode VIMP garde un nombre relativement constant de variables (8,2± 10) quand Variable-Hunting (8,7±34) et Minimal Depth (45±33) ont une bien plus grande variabilité en termes de nombre de variables gardées.

TABLE 6.6 – Expériences techniques : Comparaison des méthodes de sélection de variables avec RSF.

Méthode	Nb. moyen de variables gardées	Er. de préd. moyenne
Notre Méthode (RSF+VIMP)	8,2 ± 10	0,36 ± 0,015
RSF sans sélection	129	0,61 ± 0,027
RSF + Minimal depth	45 ± 33	0,47 ± 0,025
RSF + Variable-Hunting	8,7 ± 34	0,43 ± 0,016

6.2.3 Intêret des radiomiques

La valeur ajoutée prédictive potentielle des caractéristiques basées sur l'image a été analysée lors des expériences techniques à l'aide de trois sous-bases de données : une avec toutes les caractéristiques, une avec les caractéristiques cliniques uniquement et une avec

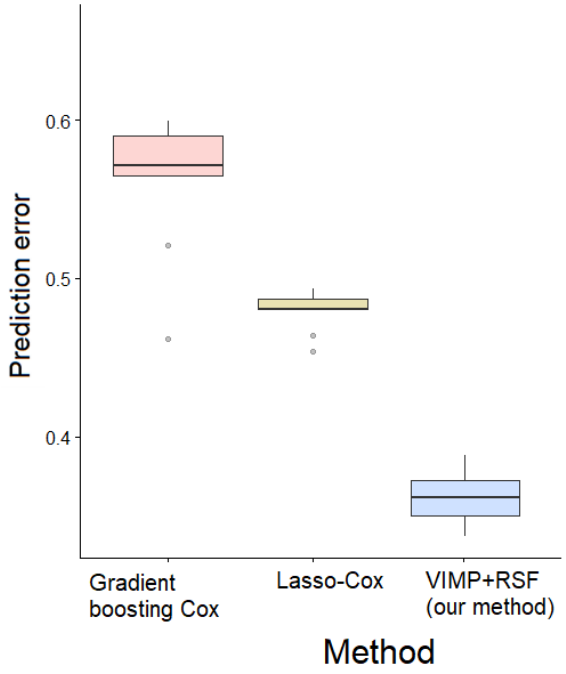


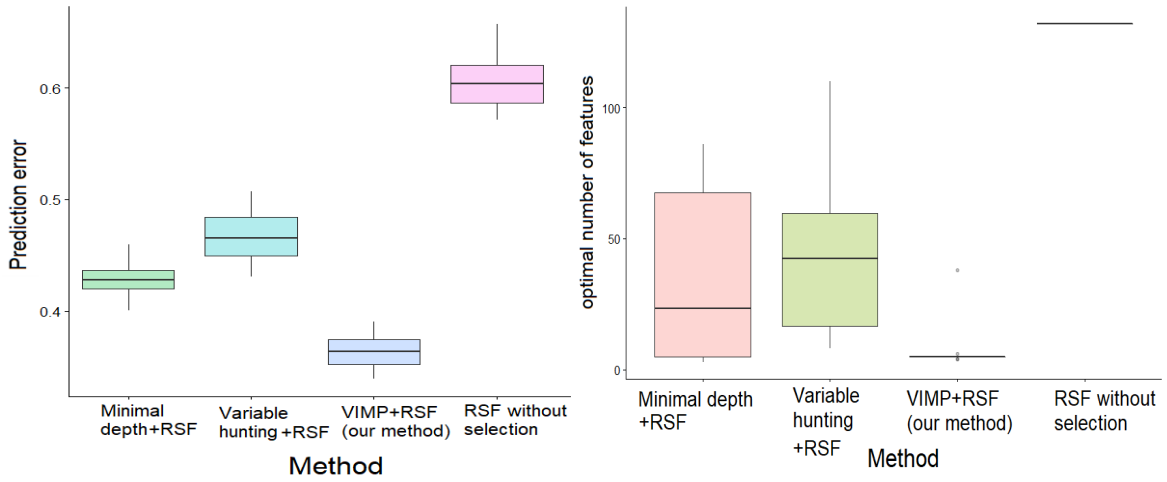
FIGURE 6.6 – Expériences techniques : Erreur de prédiction moyenne pour chaque méthode sur 10 "folds" répétée 10 fois [6]. Les meilleurs résultats sont ceux qui correspondent à une erreur de prédiction faible.

TABLE 6.7 – Expériences techniques : p-value moyenne sur une validation croisée de 5 "folds" [2] en fonction de la méthode, en utilisant toutes les variables.

Méthode	p-value moyenne
Notre méthode	0,05
Sans sélection	0,40
Minimal Depth	0,24
Variable-Hunting	0,11

les caractéristiques d'imagerie uniquement. L'erreur de prédiction a été calculée pour analyser la contribution de chaque sous-base de données à la prédiction du pronostic, comme présenté dans le tableau 6.8. On peut y voir que l'utilisation des caractéristiques d'image fournit de meilleurs résultats que les caractéristiques cliniques seules pour toutes les méthodes avec sélection. En termes de prédiction d'erreur, l'utilisation de caractéristiques cliniques et d'image (y compris les caractéristiques conventionnelles et texturales) donne de meilleurs résultats que l'utilisation des caractéristiques cliniques seules, et des résultats légèrement meilleurs que l'utilisation des caractéristiques d'image seules.

De plus, les variables retenues dans chaque sous-analyse ont également été rapportées dans la figure 6.8. On constate que lorsque l'on considère toutes les caractéristiques disponibles en entrée, la majorité de celles qui sont retenues sont basées sur l'image (en violet).



a)Erreur de prédiction moyenne pour chaque méthode sur 10 "folds" répétée 10 fois.

b) Nombre de variables optimal moyen sur 10 "folds" répété 10 fois.

FIGURE 6.7 – Expériences techniques : Comparaison des méthodes de prédiction de la PFS incluant différentes stratégies de sélection de variables [6].

TABLE 6.8 – Expériences techniques : Erreur de prédiction en fonction du type de caractéristique fournie, et pour différentes méthodes de sélection de variables associées aux RSF.

	Toutes les variables	Variables texturales et conventionnelles	Variables cliniques
Notre méthode (VIMP + RSF)	0.34	0.36	0.45
Minimal depth + RSF	0.45	0.45	0.48
Variable-Hunting + RSF	0.40	0.41	0.45
Sans sélection	0.51	0.67	0.52

6.2.4 Les caractéristiques les plus prédictives

Les variables les plus prédictives ont été étudiées dans les deux groupes d'expériences.

6.2.4.1 Les résultats des expériences techniques

En observant les variables les mieux classées dans la figure 6.8, nous pouvons voir que parmi les 30 meilleures caractéristiques d'image, presque toutes ont été obtenues par réechantillonage relatif (OMRR) au lieu de par un réechantillonage absolue (OMAR).

Cependant, l'ordre exact des meilleures caractéristiques n'est pas toujours pertinent. En effet, lorsqu'il y a peu de variables retenues, les valeurs d'importance sont proches entre les caractéristiques et l'ordre n'est pas toujours stable sur les différentes lancements. Pour un plus grand nombre de caractéristiques retenues, les caractéristiques classées en tête sont souvent les mêmes, mais pas toujours dans le même ordre. Le fait que le nombre de caractéristiques basées sur l'image soit plus important que le nombre de caractéristiques

cliniques (car plusieurs implémentations) peut induire un biais en faveur de la présence d'un plus grand nombre de ces dernières dans les caractéristiques les mieux classées. Nous avons donc aussi évalué cette importance de variable en moyennant les valeurs VIMP, en regardant la fréquence d'apparition de chaque variable dans les 10 meilleures et en regroupant les caractéristiques en moyennant sur l'implémentation. Les résultats sont discutés dans la section 7.

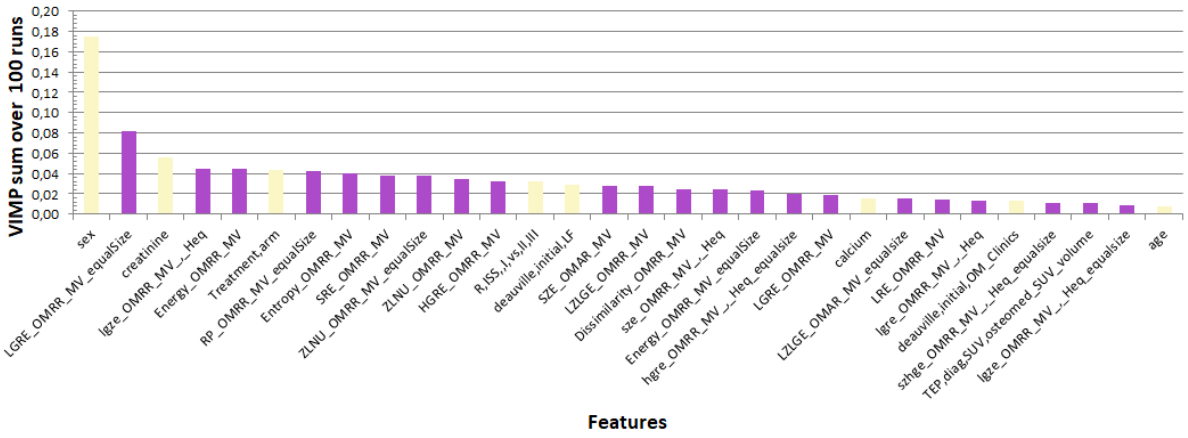


FIGURE 6.8 – Expériences techniques : Histogramme des 30 meilleures caractéristiques selon notre méthode VIMP-RSF. Jaune : variables cliniques, violet : variables basées sur l'image. Les différentes implémentations sont désignées par OMRR (One Matrix relative resampling), OMAR (One Matrix absolute resampling), Heq (histogram equalization), et equalsize (equal size of voxels.)

6.2.4.2 Les résultats des expériences cliniques

Nous avons par la suite évalué les variables dans les expériences cliniques, grâce aux deux bases de données et 22 caractéristiques d'entrée. Sur la base du classement VIMP, la validation croisée RSF a sélectionné 3 caractéristiques parmi les 22 initialement impliquées : le bras de traitement, l'hémoglobine et le SUVmaxBM. Les 7 autres caractéristiques les plus prédictives étaient l'âge, le ZLNU, le ZLNU calculé à l'aide d'une égalisation linéaire, le TMTV, le R-ISS, le score de Deauville pour BM, l'entropie calculée à l'aide d'une égalisation linéaire, et le sexe.

La dépendance partielle des 3 caractéristiques conservées dans le modèle a été étudiée et est présentée dans la figure 6.10. Elle donne un aperçu complet du comportement non linéaire de chaque caractéristique par rapport à la mortalité. La dépendance partielle a été calculée à 2095 jours, ce qui correspond à la durée médiane du suivi. Les patients dont le SUVmaxBM était initialement élevé, qui souffraient d'anémie et qui étaient traités sans greffe de cellules souches présentaient une PFS significativement plus faible et donc un risque plus grand.

La figure 6.11 montre les valeurs moyennes de VIMP sur les 5 "folds" de la validation

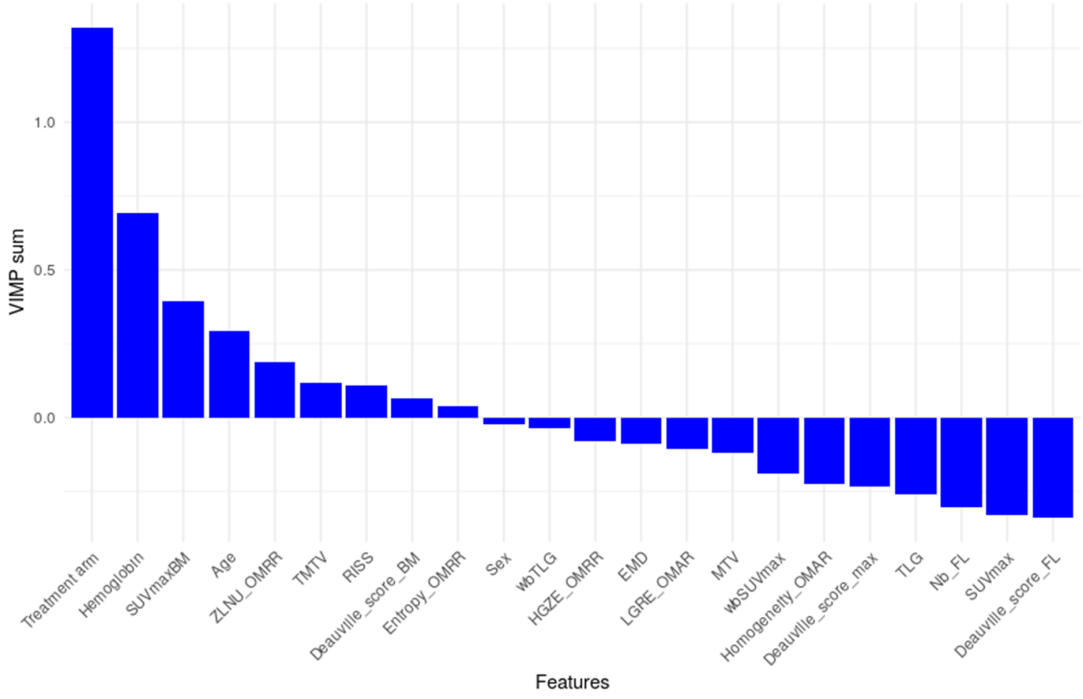


FIGURE 6.9 – Expériences cliniques : Valeurs VIMP sommées sur 100 itérations, calculées sur l'ensemble de données d'entraînement (98 patients) pour chaque variable. OMRR = Egalisation linéaire, OMAR = égalisation absolue, EMD = maladie extra-médullaire, et Nb_FL = nombre de LF >3 ou non. De grandes valeurs indiquent une grande valeur prédictive, alors qu'une valeur de zéro ou négative indique qu'elle n'a pas de valeur prédictive [7].

croisée "nested", ce qui donne une idée des caractéristiques les plus importantes sélectionnées par les RSF. Cette validation croisée "nested" permet de souligner la robustesse du modèle, non seulement par l'erreur de prédiction, mais aussi par les variables choisies qui ne sont pas dépendantes des données de test choisies. Les deux premières variables (bras de traitement et hémoglobine) sélectionnées dans la première approche (figure 6.9), le furent aussi dans la seconde (figure 6.11). La troisième, SUVmax BM est aussi parmi les dix premières de la seconde approche. Finallement, parmi les dix premières variables de la première approche Entraînement/Validation + Test, huit sont aussi dans les dix premières dans l'approche "nested".

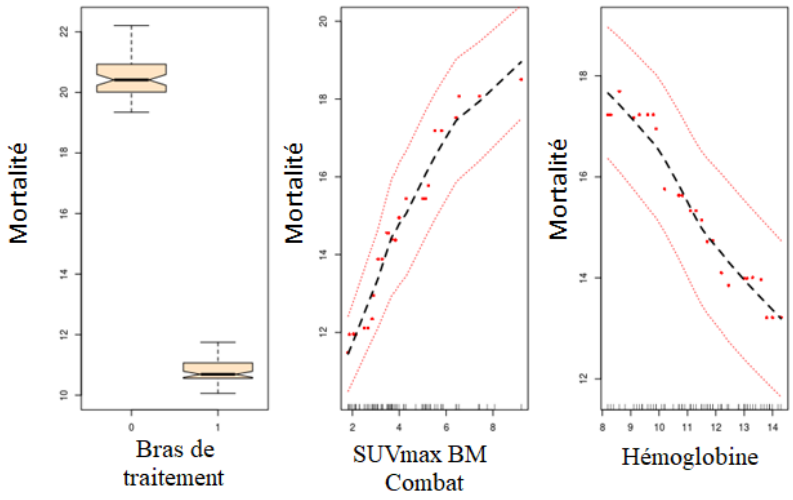


FIGURE 6.10 – Expériences cliniques : Dépendance partielle pour les 3 premières caractéristiques sélectionnées par VIMP. A gauche : bras de traitement (0 = sans ASCT, 1 = avec ASCT), au milieu : SUVmaxBM, à droite : hémoglobine. Les lignes rouges représentent l'intervalle de confiance à 95% [7].

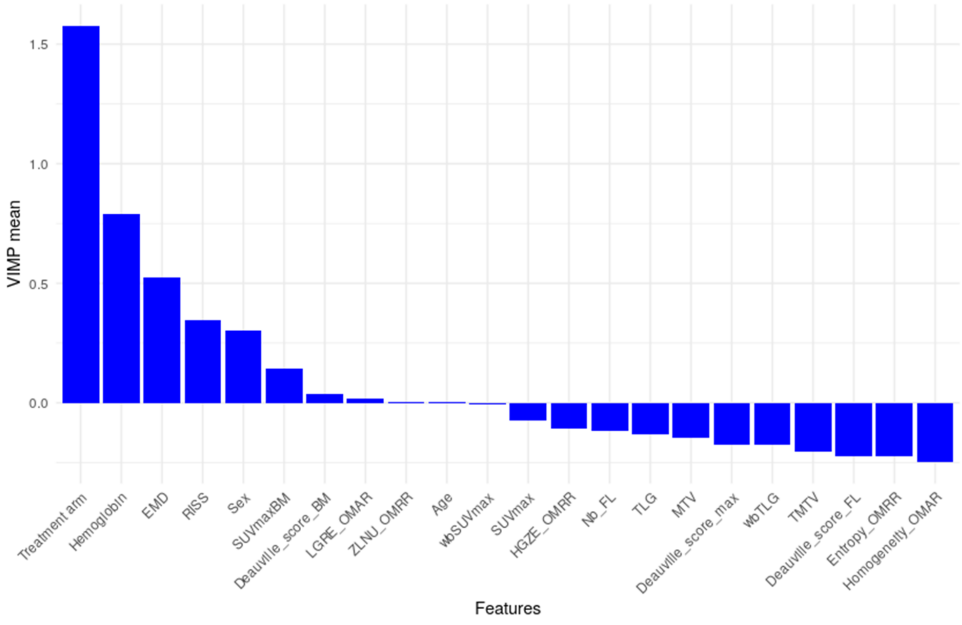


FIGURE 6.11 – Expériences cliniques (nested) : Valeurs VIMP moyennées sur les 5 "folds", calculées sur l'ensemble de données d'entraînement (98 patients) pour chaque variable. OMRR = égalisation linéaire, OMAR = égalisation absolue, EMD = maladie extramédullaire, et Nb_FL = nombre de LF >3 ou non [7].

Chapitre 7

Discussions et conclusion

Outre les résultats principaux dans le chapitre précédent, d'autres observations ont pu être faites lors de ces travaux. Nous en discuterons dans la section 7.1 puis nous conclurons dans la section 7.2.

7.1 Discussions

7.1.1 Influence des paramètres des arbres sur l'erreur de prédiction

Nous avons voulu observer quel était l'impact de chaque paramètre des arbres sur la prédiction afin de déterminer les paramètres qui doivent être choisis attentivement. Pour cela nous présentons l'erreur de prédiction minimale sur le set de validation et d'entraînement (nous réalisons ici une validation croisée Entraînement-Validation) en fonction du mode de séparation (Log-Rank ou "random"), du nombre d'arbres et du nombre minimal d'individus nécessaire pour réaliser une séparation au niveau d'un noeud. Le calcul de cette erreur s'est faîte de la façon suivante :

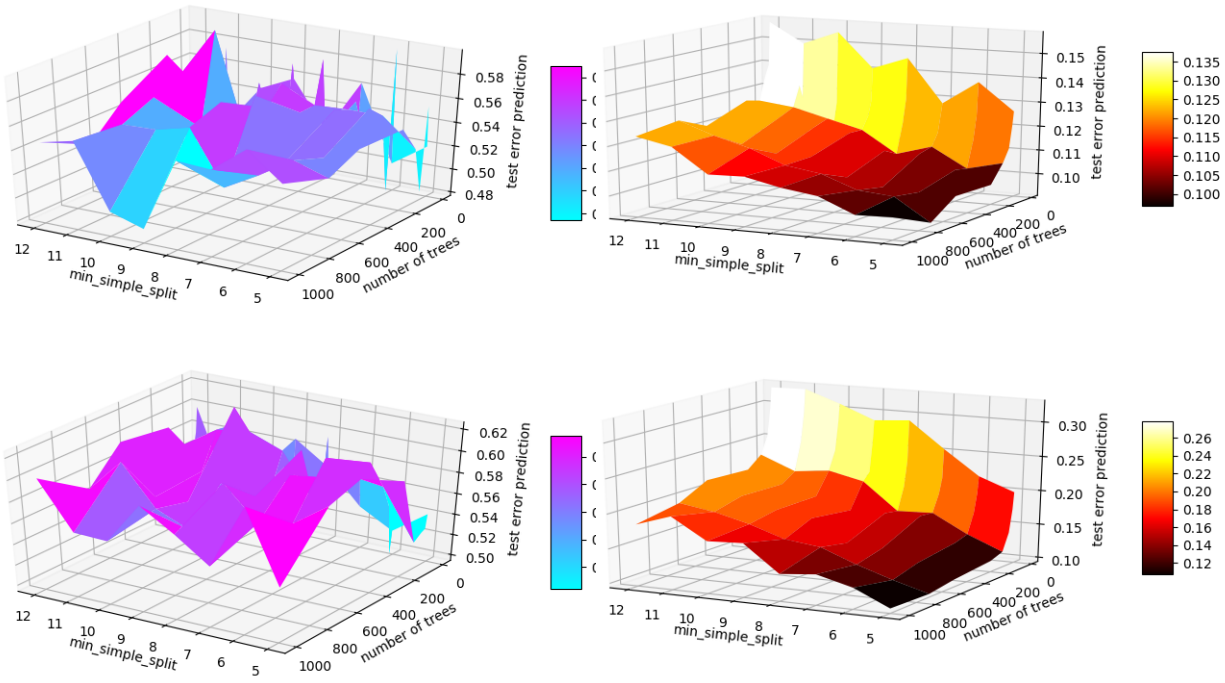
$$err_c(m, a, s) = \frac{\sum_{f \in F} err_o}{3} \tag{7.1}$$

avec :

- err_o : l'erreur observée lors de la construction des arbres avec les différents paramètes,
- m : le mode de séparation $\in \{\text{Log} - \text{Rank}, \text{"random"}\}$,
- s : le nombre d'individus nécessaire pour réaliser une séparation au niveau d'un noeud $\in [5 : 11]$,
- a : le nombre d'arbres $\in \{5, 15, 25, 40, 50, 70, 100, 200, 500, 700, 1000\}$,
- err_c : L'erreur de prédiction moyennée sur le nombre maximal de caractéristiques.

— f : le nombre maximal de caractéristiques tirées aléatoirement au niveau de chaque noeud $\in F$ avec $F = \{0,5; 0,7; 1,0\} \times$ nombre total de caractéristiques

,



(a) set de validation

(b) set d'entraînement

FIGURE 7.1 – Graphiques présentant l'erreur de prédiction minimale sur le set de validation (a) et le set d'entraînement (b) en fonction du nombre d'arbres, du nombre maximal de caractéristiques tirées aléatoirement au niveau de chaque noeud et du mode de séparation. Graphiques du haut : Log-Rank ; Graphiques du bas : "random".

Les graphiques 7.1 représentent l'erreur de prédiction minimale sur les sets de validation (a) et d'entraînement (b) en fonction du nombre d'arbres, du nombre maximal de caractéristiques tirées aléatoirement au niveau de chaque noeud et du mode de séparation. Nous pouvons observer nettement qu'avec nos données, la diminution du nombre d'individus nécessaires à la séparation induit une diminution régulière de l'erreur de prédiction sur le set d'entraînement. L'erreur sur le set d'entraînement semble constante en fonction du nombre d'arbres sauf lorsque le nombre d'arbres est inférieur à 100. Dans ce cas ci, il y a une augmentation relativement importante de l'erreur de prédiction. Cependant, bien que des tendances se distinguent sur le set d'entraînement, l'erreur de prédiction sur le set de validation semble soumis à bien plus de bruit et ne permet pas d'extraire une tendance. Nous pouvons tout de même deviner légèrement que l'erreur de prédiction sur le set de validation est plus basse lorsque le nombre d'arbre est inférieur à 100 et plus élevée lorsque le nombre d'individus nécessaires à la séparation est égal ou supérieur à 11, et donc permettre une meilleure généralisation.

Enfin, nous pouvons voir en observant les échelles de couleurs que les erreurs de prédiction minimale et maximale sur les sets d'entraînement et de validation sont plus élevées dans le cas du mode de séparation "random". Cependant, étant donné qu'on ne peux observer de tendance comme avec les autres paramètres nous ne pouvons savoir si ce résultat est générale ou dépendant du jeu de données.

7.1.2 Cohérence des meilleures variables avec la littérature

Afin de voir la cohérence des biomarqueurs trouvés, nous pouvons les comparer avec la littérature. Ainsi, la caractéristique la plus prédictive dans le contexte de la progression fut le bras de traitement. Les patients sans greffe de cellules souches ont un risque plus grand et ceci est cohérent avec les résultats des études IFM/DFCI 2009 [80] et EMN02/HO95 [81].

Les deuxièmes et troisièmes variables les plus prédictives sont l'hémoglobine et le SUVmaxBM. Cependant, bien que le SUVmaxBM soit ici très corrélé avec la PFS, on ne peut pas le comparer avec la littérature, car celui-ci n'est pas présent dans les précédentes études. Les 7 autres caractéristiques les plus prédictives étaient l'âge, le ZLNU, le ZLNU calculé à l'aide d'une égalisation linéaire, le TMTV, le R-ISS, le score de Deauville pour BM, l'entropie calculée à l'aide d'une égalisation linéaire, et le sexe. Bien que l'âge sorte prédictif, il aurait été attendu qu'un âge plus élevé implique un risque plus grand. Or, l'âge est, ici, inversement proportionnel au risque. Outre la possibilité que cela puisse être dû au hasard, nous pouvons aussi supposer que, le myélome multiple étant une maladie trouvée généralement chez des personnes âgées, le fait de le développer à un âge relativement jeune peut être de mauvais pronostic. Un autre point intéressant, est que le sexe sorte comme prédictif avec une forte corrélation avec la PFS mais aucune étude n'en parle. Ici les femmes ont moins de risques que les hommes [voir la figure 7.2].

Le SUVmax des lésions focales lui n'est pas sortie dans cette étude ce qui n'est pas cohérent avec les deux premières grandes études prospectives [2, 82, 83] mais l'est avec les résultats de l'étude IMAJEM [1]. La maladie extra-médullaire (EMD) aurait du sortir mais ce manque est probablement dû au fait qu'il y a très peu de patients qui en sont atteints dans cette étude (12 patients) [voir la figure 7.2]. Les paramètres dérivés du volume métabolique (MTV, TLG, TMTV, wbTLG) ne sont pas sortis non plus, ce qui n'est pas cohérent avec les précédentes études [72, 84–86]. Ceci est peut être dû à la différence de segmentation utilisée pour le calcul, au fait que les études précédentes étaient rétrospectives, et au fait de ne pas inclure l'atteinte diffuse de la moelle osseuse dans le calcul.

Enfin, de nombreuses variables basées sur les images TEP (SUVmaxBM, caractéristiques volumiques et texturales) n'étaient pas présentes dans les précédentes études ce qui peut expliquer les différences.

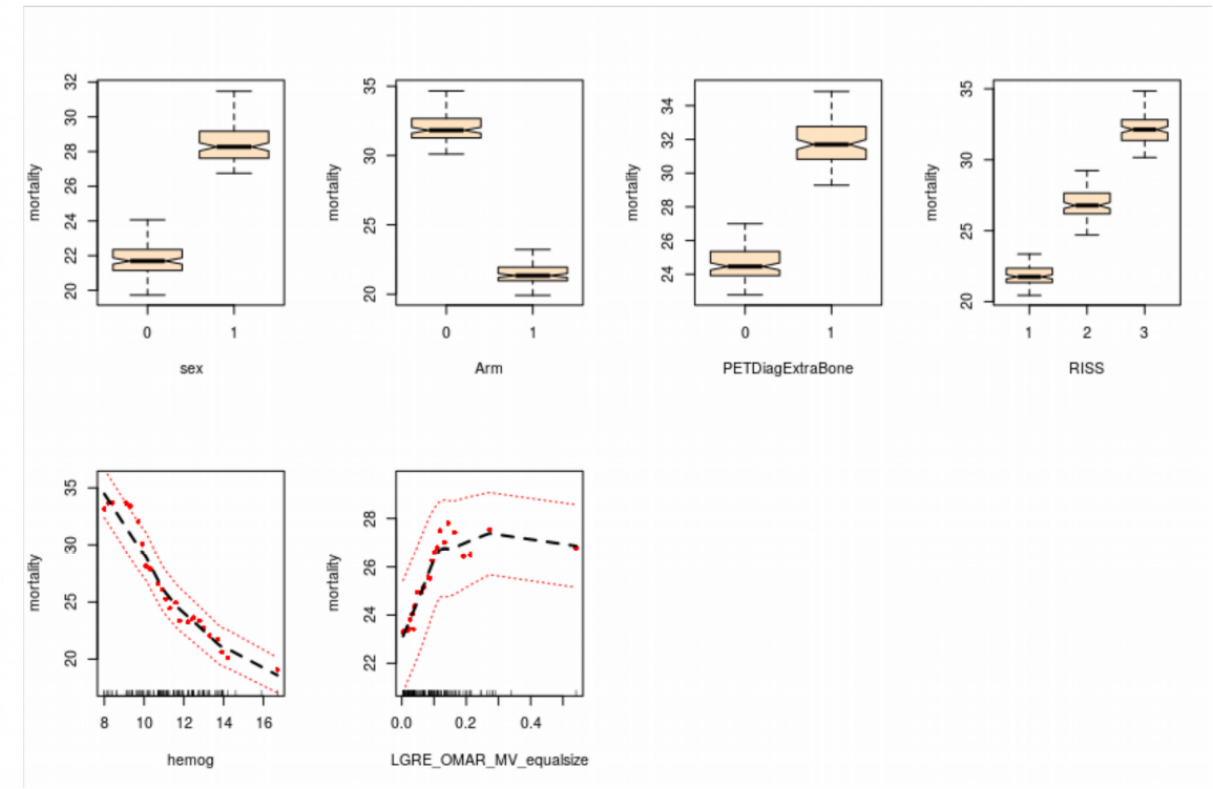


FIGURE 7.2 – Corrélation des variables les plus prédictives, mais non gardées par le modèle, avec la mortalité [7].

7.2 Conclusions

Ainsi, le but de cette partie était la production et l'utilisation d'un modèle d'apprentissage automatique, basé sur les RSF, permettant la prédiction de la PFS et l'identification de biomarqueurs dans le contexte du myélome multiple.

Des premières expériences ont été réalisées afin de déterminer le meilleur modèle de prédiction de la progression des patients atteints de MM. Celui-ci se compose d'une partie d'optimisation automatique des paramètres d'arbre, d'une partie d'ordonnancement des variables par VIMP (Variable Importance) permettant de déterminer les biomarqueurs de la maladie, et une partie de prédiction par RSF permettant d'avoir un groupe pronostique (haut ou bas risque) et un risque pour chaque patient, ainsi qu'un courbe de survie pour chaque groupe. Nous avons mis en entrée de ce modèle la base de données IMAJEM [1] avec les données cliniques, conventionnelles, volumiques et texturales des images TEP. La méthode surpasse les méthodes classiques (Lasso-Cox et GB-Cox) et les autres méthodes de sélection de variable (Minimal Depth et Variable-Hunting) avec une erreur de prédiction de 0,36 et une p-value de 0,05 pour la séparation des groupes. Dans ces travaux, nous avons aussi montré l'importance de la texture des images et la stabilité du nombre de variables gardées grâce à VIMP. Ces travaux ont été publiés dans IJCARS [6].

Par la suite nous avons réalisé des expériences cliniques grâce aux deux bases de données (IMAJEM [1] et EMN02/HO95 [2]) afin de déterminer précisément les biomarqueurs de la maladie, à l'aide de la méthode précédemment décrite. La caractéristique la plus prédictive est le bras de traitement ce qui est cohérent avec la littérature. En plus du bras de traitement, le modèle a gardé l'hémoglobine et le SUVmax de la moelle osseuse. Enfin, outre le fait que le SUVmax des lésions focales ne soit pas sortie, et que le sexe soit sorti, les autres caractéristiques prédictives sont en cohérence avec la littérature. Ces expériences ont été publiées dans EJNMMI [7].

Plusieurs pistes ont commencé à être étudiées et d'autres restent à voir entièrement, pour améliorer ce modèle. C'est notamment le cas de l'équilibre des données et la corrélation des variables. En effet, pour de meilleurs résultats, il faudrait un nombre équivalent de patients dans chaque temps d'évènement. Une augmentation des patients non-censurés, pourrait par exemple, équilibrer ces données, mais aussi augmenter le nombre de données d'entrées et diminuer le taux de censure. De plus, comme vu précédemment il reste un problème de corrélation entre les variables, et notamment celles de textures lorsqu'elles proviennent de la même caractéristique avec des méthodes d'implémentation différentes. Cela implique qu'une pré-sélection reste encore nécessaire dans certains cas.
Les différentes pistes étudiées sont les suivantes :

— Afin d'éviter la corrélation des variables, une solution serait de réaliser la sélection VIMP de façon récursive [87].

— Une autre solution serait le calcul du VIMP groupé [88].

— Pour équilibrer les données en terme de censure ou de temps, la méthode SMOTE (Synthetic Minority Over-Sampling technique) [88] peut être utilisée au préalable sur les données.

— Pour diminuer le problème des variables présentes chez peu de patients (comme la maladie extra-médullaire) l'IPCW (Inverse Probability Censoring Weighted) [89,90] peut être une solution. Il peut aussi être utilisé pour la censure.

— Enfin, toujours pour diminuer le problème des variables présentes chez peu de patients nous pouvons tester les RSF améliorées de Miao et al. [40] (iRSF) qui consisent en une modification du critère de séparation (Log-Rank pondéré [91]) et d'arrêt (fonction de séparation décroissante)

Troisième partie

Analyse de survie par apprentissage profond

Résumé

Dans cette partie va être abordée la seconde moitié de la thèse. Après avoir créé un modèle par apprentissage automatique basé sur les RSF, nous souhaitons prédire la progression des patients atteints de myélome multiple grâce à l'apprentissage profond. En effet, l'apprentissage profond à permis de grandes avancées pour de nombreuses tâches, mais reste relativement peu étudié dans le contexte de l'analyse de survie. De plus, son utilisation dans un contexte de bases de données prospectives d'images TEP induit plusieurs défis. Ainsi, l'objectif de cette seconde partie de thèse était de déterminer un modèle d'apprentissage profond, basé sur les réseaux de neurones convolutionnels (CNN), permettant la prédiction de la progression du myélome multiple à l'aide de l'imagerie TEP. Les défis à prendre en compte sont, le faible nombre de patients, la taille des lésions (petite et variable), la faible résolution des images TEP, l'interprétabilité, et l'adaptation à une tâche d'analyse de survie. Pour ce faire nous proposons différentes stratégies tels que l'attention sur les filtres, le pré-entraînement auto-supervisé ou la création de fonctions de coût basées sur les triplets et adaptées à la survie. Ainsi, après avoir discuté du contexte (chapitre 8) et l'état de l'art (chapitre 9), nous aborderons les stratégies d'adaptation de l'apprentissage profond à notre contexte dans le chapitre 10. Nous détaillerons les expériences et les résultats dans le chapitre 11, puis nous conclurons dans le chapitre 12.

Chapitre 8

Contexte

Les travaux utilisant l'apprentissage profond pour l'analyse de survie dans une étude clinique ne sont pas encore courants. Cette analyse est pratiquement tout le temps réalisée à l'aide de méthodes d'estimation statistique ou l'apprentissage machine telles que Cox, Kaplan-Meier ou RSF. Différentes raisons sont à l'origine de ce manque. Dans notre cas les méthodes par apprentissage profond, en particulier l'architecture mais aussi les fonctions de coût pour les optimiser, devront être adaptées à nos données et à la tâche de survie.

8.1 L'analyse de survie par apprentissage profond

Dans les domaines de la vision par ordinateur et de l'analyse d'images médicales, l'apprentissage profond a permis d'améliorer considérablement certaines tâches comme la classification et la segmentation [8, 9]. Cependant, son utilisation pour l'analyse de survie à partir d'images reste encore très limitée. Une grande partie des analyses de survie par apprentissage profond est remplacée par de la classification [92] [1] ou de la régression sans prendre en compte la censure. Or, cela implique d'évincer la majorité des patients censurés. En effet, lors de la classification, les patients censurés qui n'ont pas un temps de survie dans la dernière classe ne peuvent être classés [2].

Une autre technique est d'utiliser des réseaux convolutionnels pré-entraînés (par exemple avec une tâche de classification) pour extraire un vecteur de caractéristiques. En effet, les CNN sont connus pour extraire des représentations hiérarchiques des images. De plus, le pré-entraînement du modèle permet de réduire le besoin en nombre de données. Puis, ces caractéristiques sont données en entrée de RSF entraînées sur la tâche de survie. Bien que cette technique permette d'obtenir des caractéristiques automatiquement (sans sélection préalable), son entraînement est plus long à calculer que le calcul à la main des

1. Séparation des patients en classes en fonction du temps de survie sans prendre en compte la censure ou en ne prenant que les patients non censurés

2. Exemple : si on a 6 classes, une classe par année, et qu'un patient présente une censure à droite à 3 ans, on ne peut savoir si l'évènement aura lieu à 3 ans, 4 ans ou plus. Il pourra donc être dans toutes les classes supérieures ou égales à 3 ans

radiomiques. Cela peut être pallié par le pré-entraînement sur des bases de données plus larges et disponibles mais celles-ci doivent idéalement avoir des textures et des tailles équivalentes afin d'être discriminantes pour la survie.

Enfin, la méthode qui est en train de se développer est celle basée sur l'adaptation du modèle de Cox à l'apprentissage profond. Elle est devenue la référence en matière d'analyse de survie par apprentissage profond grâce à DeepSurv [12] ou DeepConvSurv [93]. Néanmoins, il reste encore beaucoup à faire, les c-index ne dépassant que très rarement 0,7. La figure 8.1 résume les trois méthodes principales d'adaptation de l'apprentissage profond à l'analyse de survie.

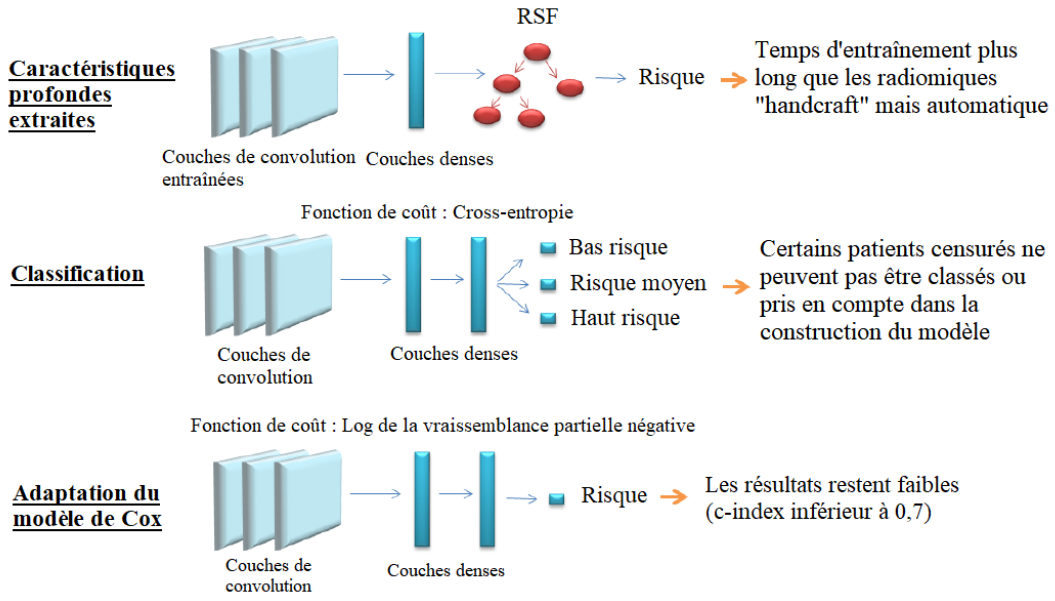


FIGURE 8.1 – Les trois méthodes principales d'adaptation de l'apprentissage profond à l'analyse de survie.

8.2 Les défis des bases de données TEP prospectives

Les modèles d'apprentissage profond ont généralement entre cent mille et plusieurs millions de paramètres. Pour éviter le sur-apprentissage (et donc permettre la généralisation du modèle), il faudrait un grand échantillon d'entrée (en incluant l'augmentation de données). Le nombre de données devrait être d'autant plus grand si le problème est très complexe. Fort heureusement, le nombre de données peu être diminué dans certains cas (facilité de séparation des données/de la tâche) ou grâce à certaines méthodes (pré-initialisation des poids). Néanmoins, le nombre de données reste un facteur déterminant pour le bon fonctionnement de la prédiction, ce qui n'est pas toujours compatible avec des bases de données provenant d'études cliniques comme c'est le cas dans cette thèse. Des stratégies que nous avons mis en place pour améliorer le taux de succès des méthodes par apprentissage profond avec des données images en faible nombre sont donc :

— de s'appuyer sur les techniques d'augmentation de données [voir la sous-section 11.1.1.3]

— d'adapter la quantité de paramètres [voir la sous-section 10.1.1.1]

— de s'intéresser au méthodes d'initialisation des poids (pré-entraînement) [voir la sous-section 10.1.3]

— de favoriser la régularisation de la fonction de coût et les techniques permettant la généralisation (dropout) [voir la sous-section 10.1.1.1]

En plus d'être adaptés aux grandes bases de données, les modèles d'apprentissage profond basés sur des images comme les CNN, sont construits principalement pour de grandes images d'entrée de haute résolution. Ceci pose deux problématiques : la quantité d'information disponible et le pré-entraînement. Concernant le premier problème, des images de basse résolution et de petite taille pourraient ne pas contenir assez d'informations discriminantes à extraire afin de réaliser la prédiction. En effet, nous savons que plus un CNN a de couches convolutionnelles, plus il pourra extraire de détails. Or, la taille de l'image d'entrée limite la profondeur des réseaux de neurones classiques s'appuyant sur les couches de "pooling" visant à l'amélioration de l'invariance spatiale. Ainsi, pour pallier à ce problème d'images de petites taille nous proposons l'utilisation de la SPP (Spatial Pyramidal Pooling) [voir la section 10.1.2.1] et nous diminuons la taille de nos noyaux de convolution [voir la section 10.1.1.1]. La deuxième problématique des images de petite taille est le pré-entraînement. Or, nous venons de mentionner que l'initialisation des poids par pré-entraînement est important dans le cas par exemple d'une petite base de données. En effet, la plupart des bases de données d'images disponibles publiquement sont des bases de données avec des images de plus haute résolution et de plus grande taille, ou non adaptées à nos images du point de vue textural. Ceci est important car les images de pré-entraînement doivent être de même taille que les images que l'ont veut utiliser dans notre modèle CNN. Un exemple de base de données très utilisée est ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [94], qui a des images de grande taille et très différentes du point de vue textural (images non médicales). Un autre, plus proche car contenant des images médicales (IRM) est BRATS [95]. Là encore les images sont grandes (de l'ordre de 512×512 pixels). La base de données MNIST [96] contient bien des images de plus petite taille (28×28 pixels) mais celles-ci sont très différentes de nos lésions TEP d'un point de vue textural (chiffres). De plus, quand la plupart des modèles sont construits pour des images de taille supérieure à $(200 \times 200$ pixels), les coupes de nos lésions TEP ont, comme indiquées dans le tableau 8.1, une taille 3D variant de $(3 \times 3 \times 3)$ à $(32 \times 40 \times 53)$.

Un autre point important dans le pré-entraînement est que les images soient proches (d'un point de vue textural) de nos images. En effet, si les images utilisées pour le pré-entraînement sont trop différentes de nos images, les poids ne seront pas initialisés cor-

TABLE 8.1 – Les tailles minimales, maximales et médianes de nos images TEP (en nombre de voxels).

	Minimum	Maximum	Médiane
Images TEP corps entier	128 × 128 × 250	200 × 200 × 915	140 × 140 × 357
Lésions	3 × 3 × 3	32 × 40 × 53	12 × 12 × 11

rectement. Les meilleures images dans notre cas seraient des images TEP de lésions de myélome multiple mais il n'en existe aucune en libre accès à notre connaissance. A défaut, une base de données idéale contiendrait des images TEP liées à une autre condition médicale, ce qui est le cas de la base Head and Neck [32] mais les images y sont trop grandes (128 × 128 × 128 voxels). Enfin, en dernier recours, des images médicales (IRM, TDM) pourrait être suffisantes mais celles si ont des tailles plus grandes que les images TEP corps entier. Ainsi, pour pallier à ce problème d'images de pré-entraînement trop différentes et grandes nous proposons deux pré-entraînements basés sur nos propres images [voir la sous-section 10.1.3].

Une limitation de l'apprentissage profond est le coût que demande ces modèles contrairement aux méthodes d'estimation statistique ou aux autres méthodes d'apprentissage machine. Ces coûts sont multiples : temps de calcul, coût matériel. En effet, en raison du nombre conséquent de paramètres, avec un même matériel, les temps de calculs de l'entraînement vont être beaucoup plus grands (peut-être de l'ordre de plusieurs jours même avec du matériel adapté). Ce temps d'entraînement peut être réduit par du matériel (GPU, CPU, etc.) mais cela a un coup très élevé. La figure 8.2 présente une expérience de classification réalisée avec la base de données DTD [97], qui consiste à comparer le temps de calcul d'un entraînement [3] avec un CNN 2D classique de 4 588 103 paramètres et un SVM de 4 521 paramètres, en fonction du nombre d'images. Bien que la comparaison ne permette pas de déterminer la vitesse pour un nombre de paramètres équivalents, on peut voir que le temps d'entraînement du SVM est moins dépendant de la quantité de données d'entrée (or un grand nombre d'images est nécessaire pour le CNN).

Pour finir, les méthodes d'apprentissage profond sont plus difficilement interprétables que les méthodes comme Cox ou RSF. Or, une bonne interprétabilité permet de contribuer à l'acceptabilité du modèle si on souhaite son utilisation dans un contexte d'étude clinique ou de routine. Nous proposons donc l'inclusion de méthodes d'attention (spatiale et sur les filtres) [voir la sous-section 10.1.2.2].

Pour conclure, comme les RSF, les méthodes d'apprentissage profond permettrons de prédire un risque pour les patients atteints de myélome multiples grâce aux images TEP.

3. Un entraînement pour le CNN correspond à l'entraînement sur une combinaison de paramètres contenant 30 "epochs". Pour le SVM, cela correspond à une recherche par grille sur 7 paramètres (648 combinaisons).

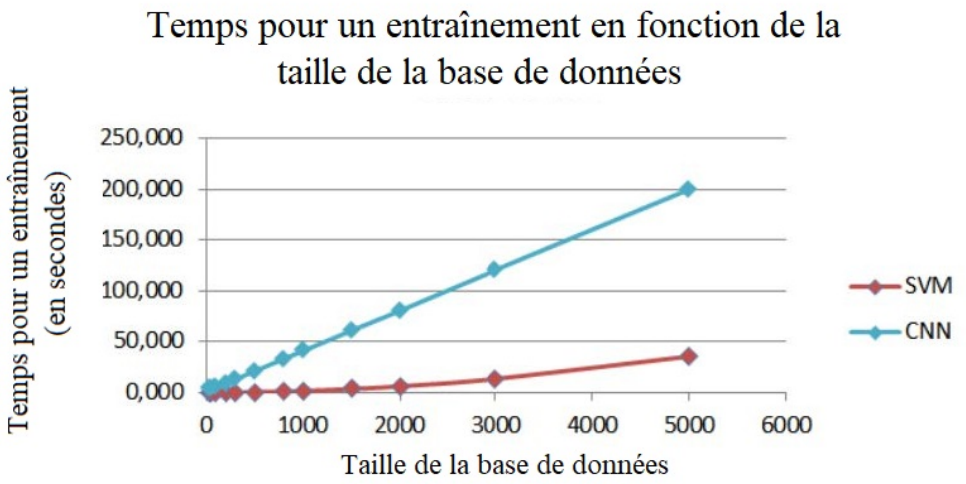


FIGURE 8.2 – Temps pour un entraînement en fonction du nombre d'images de la base de données DTD [97] avec un CNN 2D simple (4 588 103 paramètres) and SVM (4 521 paramètres)

Toutefois, l'entraînement risque d'être plus long, cela demandera beaucoup d'adaptation du modèle et l'interprétation des filtres de sortie (biomarqueurs) un travail plus poussé. Cependant, de nombreuses techniques existent pour pallier à ces problèmes, telles que celles proposées plus haut. De plus, les CNN vont permettre une potentielle analyse spatiale de l'image (importance de chaque partie de la lésion), qui reste difficile avec des radiomiques calculés à la main. Les méthodes d'apprentissage profond permettent une bien plus grande personnalisation du modèle afin d'améliorer constamment les résultats de prédiction. Enfin, l'analyse de survie par apprentissage profond reste encore relativement peu étudié et laisse une grande place à l'innovation.

Chapitre 9

État de l'art

9.1 Apprentissage profond et analyse de survie

Les premières adaptations de l'apprentissage profond aux tâches de survie permettent d'extraire des caractéristiques issues d'un CNN à l'aide d'un réseau de neurones pré-entraîné et d'alimenter des modèles de prédiction tels que Lasso Cox [98] ou RSF [99]. Le problème de la survie a également été simplifié par de la classification en différents groupes de risque [92] ou par de la régression du temps sans l'événement [99]. Cependant, ces formulations ne traitent pas nativement les données censurées.

9.1.0.1 Les méthodes de prédiction de risque

Les méthodes de prédiction du risque adaptent la fonction de coût pour tenir compte de la censure. Faraggi et Simon [100] ont adapté le modèle linéaire des risques proportionnels de Cox (CPH)[voir la sous-section 3.2.2] avec un réseau ascendant plus flexible. Katzman et al. revisitent (dans l'approche dite DeepSurv) la fonction de coût de Faraggi-Simon dans le contexte de réseaux plus profonds [12]. D'autres variantes s'en tiennent au modèle linéaire CPH, mais utilisent le réseau soit pour la réduction non linéaire de la dimensionnalité de l'entrée [101], soit pour la prédiction des poids [102].

Un problème avec le modèle de Cox est que la vraisemblance partielle pour chaque individu dépend non seulement de la sortie du modèle pour cet individu, mais aussi de la sortie pour tous les individus ayant une survie plus longue. Cela rend l'utilisation de la descente de gradient stochastique (SGD) pour l'optimisation de poids du réseau problématique, car avec la SGD, seul un petit nombre d'individus est visible pour le modèle à un moment donné. Il faudrait donc utiliser l'ensemble de la base de données pour chaque étape de descente de gradient. Ceci n'est pas souhaitable car cela ralentit la convergence. D'autres travaux [13, 103–107], dont ceux de Gensheimer et al. [13], proposent un modèle de survie en temps discret qui traite les risques non proportionnels, contrairement au

modèle de Cox. Ce modèle est également mieux adapté à l'entraînement par SGD par batch et donc plus rapide que celui de DeepSurv [12, 101] lorsque moins de données sont utilisées. Le modèle permettant la prédiction d'un temps discret à la place d'un risque (unité arbitraire), les sorties sont plus interprétables.

9.1.0.2 Les méthodes d'ordonnancement

Une deuxième façon de traiter les problèmes de la vraisemblance partielle de Cox sont les méthodes d'ordonnancement. Elles consistent à formuler le problème en termes de relation ordinale entre les temps de survie individuels, ce qui permet de gérer implicitement la censure lorsque l'on ne considère que les paires admissibles.

Dans cette direction, Jing et al. [14] proposent la méthode RankDeepSurv avec une fonction de coût de régression MSE (erreur quadratique moyenne) sur le temps continu et une régularisation d'ordonnancement par paire ("ranking") prenant en compte la censure. D'autres méthodes d'ordonnancement que ce soit par pairs, par triplets [108, 109] ou plus (N-pair-mc [110], proxy NCA [111], Lifted Struct [112], Ranked List loss [113]) existent et n'ont pas encore été utilisées pour la survie. Nous avons décidé de tester la combinaison d'une fonction de coût de survie avec des fonctions d'ordonnancement par paire en s'inspirant de Jing et al. [14], mais aussi par triplets. Pour l'ordonnancement par triplets nous avons d'abord étudié la fonction originale [108], où des triplets sont tirés de l'ensemble de données afin de comparer un échantillon "ancre" avec un second élément de la même classe et un troisième de classe différente. Nous avons par la suite étudié la fonction SV-triplet (Scale-Varying triplet) [109] qui permet de ne pas prendre seulement la classe en compte mais aussi l'ordre entre les classes pour des variables ordinales tel que l'âge. Ces deux fonctions de coût n'ont jamais était utilisées pour de la prédiction de survie et n'avaient donc jamais été adaptées à la censure avant notre travail.

9.2 Apprentissage profond et données en faible nombre et de petite taille

En ce qui concerne le type de données, Xinliang et al. [114] ont été les premiers à adapter DeepSurv aux couches convolutives (DeepConvSurv) et à traiter des images en entrée. Zhu et al. [93] étendent par exemple DeepConvSurv pour traiter de très grandes images histopathologiques de lames entières (512×512 pixels).

Nous nous intéressons plutôt aux problèmes soulevés par la faible résolution des images TEP. Amyaret et al. [92] et Li et al. [115] proposent tous deux des modèles CNN 3D pour l'analyse de survie à partir d'images TEP. Amyaret et al. [92] ciblent la réponse à la radio-chimiothérapie dans le cancer de l'œsophage en simplifiant le problème de survie à

une classification sans censure et en utilisant une couche d'entrée de taille relativement importante ($100 \times 100 \times 100$).

Pour prédire la réponse au traitement dans le cancer colorectal, Li et al [115] modifient le modèle DeepConvSurv avec une couche supplémentaire de Spatial Pyramidal Pooling (SPP) [116] afin de traiter les petites lésions à échelle variable. Les données TEP et TDM sont toutes deux considérées comme entrée du modèle, montrant que la multi-modalité améliore les performances. Bien que notre étude ne porte que sur les images TEP, nous empruntons l'idée d'utiliser les SPP pour traiter les tailles variables de nos lésions pour l'analyse de survie du myélome multiple. En outre, nous proposons une nouvelle stratégie pour gérer la variabilité des petites lésions, au moyen d'un modèle d'attention spatiale. Les modèles d'attention furent utilisés avec succès dans une grande variété d'applications médicales, notamment la reconstruction [117], la segmentation, la détection [118] ou la classification [119], mais presque jamais avec des tâches de survie. Les seuls articles trouvés se concentrent sur l'attention spatiale seule [120, 121]. Ainsi, nous proposons dans ces travaux l'ajout d'une partie d'attention (spatiale et sur les filtres) en s'inspirant de la méthode CBAM (Convolutional Block Attention Module) [122] pour améliorer l'interprétabilité du modèle. La partie attention sur les filtres va permettre de déterminer ultérieurement les filtres "radiomiques appris" les plus prédictifs. Une version préliminaire du modèle a été présentée à PRIME 2020 [15].

En plus de ces modules, nous avons étudié des méthodes de pré-entraînement n'utilisant aucunes données extérieures à notre base de données, y compris une méthode de pré-entraînement auto-supervisé dont l'efficacité a été prouvée pour des tâches visuelles [123]. Un grand nombre de tâches prétextes pour le pré-entraînement auto-supervisé avec des images médicales ont été proposées, comme la tâche de puzzle [124], ou la reconstruction d'image [125]. Cependant, à notre connaissance, nous sommes les premiers à utiliser le pré-entraînement auto-supervisé avec des images TEP.

Le pré-entraînement contrastif est un type de pré-entraînement auto-supervisé. Des pré-entraînement contrastif avec fonction de coût triplet ont déjà été proposés [126, 127] mais n'était pas adapté à la survie ou aux images médicales. Nous avons étendu le modèle proposé à PRIME 2020 [15] dans un article de journal envoyé en révision, en évaluant l'intérêt individuel de chaque partie de l'attention, en ajoutant différentes méthodes de pré-entraînement et en comparant plusieurs fonctions de coût. A notre connaissance, nous sommes les premiers à utiliser un pré-entraînement auto-supervisé et contrastif avec des images TEP. Une autre nouveauté est l'utilisation de l'attention sur les filtres dans l'analyse de survie avec des images. Un dernier aspect original est d'adapter les CNN pour la survie du myélome multiple à partir d'images de lésions TEP.

Chapitre 10

Méthodes

Les méthodes présentées dans ce chapitre ont pour but d'adapter un modèle CNN à la prédiction de la PFS [1] (Progression Free Survival) en présence de censure, à partir de nos images TEP de lésions provenant de bases de données prospectives (et donc d'un faible nombre de patients avec de petites images à taille variable). Les étapes du modèle de prédiction de la survie par apprentissage profond que nous avons proposé sont exposées dans la figure 10.1. Après segmentation de la ROI (polygone délimitant de façon large

1. Temps avant/risque de la prochaines progression de la maladie

la lésion) produit par un expert, nous concevons un réseau et deux techniques de pré-entraînement pour apprendre a extraire des caractéristiques d'images discriminantes pour la tâche de survie. Le réseau intègre un modèle d'attention. Finalement, les poids du réseau sont raffinés pour la tâche de la survie à l'aide de fonctions de coût spécifiques.

Dans cette section nous présenterons les différentes étapes plus en détails. Nous commencerons par parler de l'adaptation des modèles d'apprentissage profond à nos données, soit une base de données avec un nombre limité de patients, de petites images avec une faible résolution et des lésions de tailles variables. Dans le même temps, nous présenterons les méthodes utilisées pour améliorer l'interprétabilité du modèle [voir la section 10.1]. Nous présenterons ensuite les différentes fonctions de coût de la littérature, ainsi que celles que nous avons adaptées [voir section 10.2].

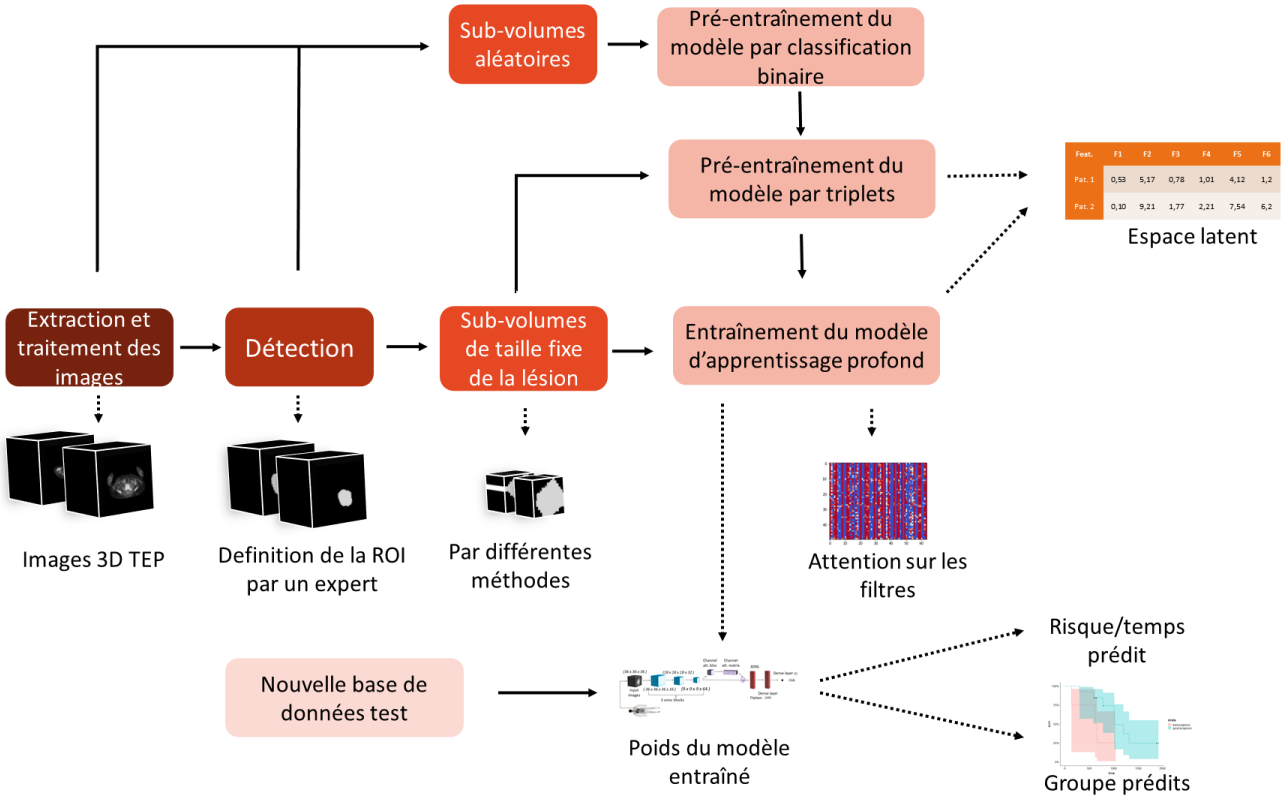


FIGURE 10.1 – Schéma des étapes de prédiction par apprentissage profond.

10.1 Adapter un modèle d'apprentissage automatique aux données TEP de bases prospectives

10.1.1 Définition d'un réseau CNN de base

L'entrée de nos méthodes est un ensemble de données $\mathcal{X} = \{\mathbb{X}_i, t_i, \delta_i\}_{i=1}^{N}$, appartenant à N patients, chacun consistant en une matrice $\mathbb{X}_i$ caractérisant l'image du patient i associé à une valeur cible de temps (time-to-event) t_i et une valeur binaire indiquant la censure δ_i ($\delta_i = 1$ si la valeur cible est censurée et $\delta_i = 0$ sinon). Nous avons étudié plusieurs dimensions de $\mathbb{X}_i$ (2D, 2.5D [2], 3D) mais seul le modèle en 3D sera présenté dans cette sous-section par soucis de simplicité.

10.1.1.1 Bloc d'apprentissage des caractéristiques radiomiques

Ce bloc est un CNN standard, dont l'architecture a été inspiré par Li et al. [115]. Il permet d'obtenir un espace latent $\mathbb{Z}$ en prenant en entrée du bloc, des images $\mathbb{X}$ de lésion 3D. Dans un premier temps il transforme ces images $\mathbb{X}$ en $\mathbb{V}$ (contenant C carte des caractéristiques) par une suite de blocs de convolution (trois dans notre modèle). Subséquemment, les cartes de caractéristiques sont vectorisées et concaténées afin d'obtenir un vecteur contenant l'espace latent $\mathbb{Z}$. Contrairement à Li et al. [115], nous avons choisi d'ajouter à chacun de ces blocs une régularisation L2 dans le but de diminuer le sur-apprentissage. L'utilisation de la normalisation classique par batchs ne convient pas dans notre contexte car ayant sélectionné un petit batch (10 images) le résultat pourrait être trop bruité. C'est pourquoi nous ajoutons une normalisation par instance [128] qui a été proposé initialement pour être indépendante du contraste. En effet, bien que la TEP soit une modalité quantitative, les normalisations liées au sujet (poids, temps d'injection, machine utilisée, etc.) ne garantissent pas une uniformité entre les sujets. Ainsi, la normalisation par instance sera sensible au contraste relatif (entre les patients) mais pas absolu.

La fonction d'activation ReLU (Rectified Linear Unit) a aussi été remplacée par la fonction d'activation "Leaky ReLU". Cette fonction d'activation permet d'éviter le problème appelé "dying ReLU" [3]. Cela donne lieu à moins de dépendance à l'initialisation des poids et à la normalisation des données. La fonction "Leaky ReLU" est définit par Maas et al. [129] comme :

$$\varphi_{\text{LeakyReLU}}(x) = \begin{cases} 0.01x & \text{, si } x < 0 \\ x & \text{, sinon} \end{cases} \tag{10.1}$$

2. Le 2.5D correspond ici à utiliser 3 images centrales du volume dans chaque direction qui représenterons chacune un canal.

3. Le ReLU met à zéro les valeurs inférieures à zéro. Cela est intéressant en terme de parcimonie (seuls les neurones les plus importants auront des valeurs supérieures à zéro). Cependant, le gradiant de zéro étant zéro, les neurones arrivant à des valeurs négatives larges ne peuvent plus être mis à jour. On dit que le neurone meurt. Pour éviter cela, il faut faire très attention à l'initialisation et à la normalisation des données.

avec x, la valeur d'entrée de la fonction d'activation.

Détails de l'implémentation

La couche d'entrée du modèle est de taille $(36 \times 36 \times 36)$. Le choix de cette taille a été fait en prenant en compte la distribution de la taille des lésions de myélome multiple dans notre base de données où le polygone autour des lésions a une taille entre $(3 \times 3 \times 3)$ et $(32 \times 40 \times 53)$ pixels [voir le tableau 8.1]. La taille des noyaux de convolution est de $(3 \times 3 \times 3)$, $(5 \times 5 \times 5)$ et $(3 \times 3 \times 3)$ avec un padding et un stride de 1. La dernière couche a donc vu sa taille de noyau diminuer par rapport à Li et al [115] en raison de la plus petite taille de nos lésions. Le nombre de filtres est de 16, 32 et 64 pour chaque convolution.

10.1.1.2 Bloc de prédiction du risque ou du temps

Le bloc de prédiction va permettre, à partir de l'espace latent $\mathbb{Z}$, de prédire un risque $h_\theta(\mathbb{Z})$ ou un temps. Il contient une couche de sortie et peut avoir entre l'espace latent et la couche de sortie une ou plusieurs couches denses. La couche de sortie aura un seul neurone qui prédit le risque $h_\theta(\mathbb{Z})$ dans le cas des fonctions de coût Cox et Rank&cox, ou le temps de l'évènement dans le cas de RankDeepSurv. La couche de sortie contient D neurones qui prédisent le temps discret de l'évènement dans le cas des fonctions de coût de survie discrète et Rank&discret [voir la section 10.2].

A partir de la prédiction $h_\theta(\mathbb{Z})$, nous déterminons deux groupes (haut et bas risque) à l'aide de la séparation par Log-Rank décrite dans la partie 5.3.4 (en remplaçant la mortalité par le risque ou le temps en fonction de la fonction de coût). Le Log-Rank part de l'idée que la meilleure séparation entre deux groupes sera celle ayant obtenue le score de Log-Rank le plus haut.

Détails d'implémentation

Le bloc de prédiction du modèle de Li et al. [115] contient deux couches denses (de tailles 1000 et 200) avec dropout et une couche de sortie. Cependant, au vue de notre relativement faible nombre de données et après quelques tests, nous avons décidé de diminuer le nombre de neurones en ne gardant qu'une couche dense de taille 100 et une couche de sortie.

10.1.2 Variabilité de taille et sélection de filtres

Une fois le modèle de base défini nous avons ajouté des modules (entre le bloc de convolutions et le bloc de prédiction) visant une meilleure question de la variabilité de taille de lésions mais aussi en conditionnant le choix de filtres à chaque images. Deux méthodes principales furent proposées : la SPP et l'attention.

10.1.2.1 Spatial pyramidal pooling (SPP)

La méthode SPP fut créée pour gérer les images multi-échelles de façon efficace. Elle est utilisée dans Li et al. [115] pour traiter des images TEP/TDM de lésions de tailles variables de cancer rectal. Cette méthode à ainsi pour but de régler le problème de la taille variable de nos lésions.

La méthode SPP consiste en plusieurs couches de "pooling" à différentes échelles qui sont appliquées aux cartes de caractéristiques de la dernière couche de convolution appelée ici $\mathbb{V}$, pour être ensuite aplaties et concaténées [voir la figure 10.2]. Contrairement au "pooling" classique, la taille des cartes de caractéristiques résultantes de chaque "pooling" est fixe, et donc la dimension du vecteur de caractéristiques de sortie reste constante quelle que soit la taille de l'image d'entrée. On considère ici que la base d'entraînement contient en plus du sub-volume $\mathbb{X}$ autour de la lésion, une région d'intérêt $\mathbb{M}$ qui segmente la lésion plus précisément, avec une boîte englobante ou une segmentation par pixel. L'entrée du modèle dans cette sous-section est $\mathcal{X} = \{\mathbb{X}_i, t_i, \delta_i, \mathbb{M}_i\}_{i=1}^{N}$. Les étapes de la SPP sont les suivantes :

Pour chaque image $\mathbb{X}_i \in \mathbb{X}$ de taille $(36 \times 36 \times 36)$ en entrée du modèle :

- Récupérer la sortie de la dernière convolution $\mathbb{V}_i$ de taille $(H \times W \times L \times C)$ (dans notre cas $36 \times 36 \times 36 \times 64$), avec C le nombre de canaux de la dernière convolution. A cette étape, toutes les cartes de caractéristiques ont la même taille.
- Re-cadrer $\mathbb{V}_i$ grâce au masque $\mathbb{M}_i$, tel que $\bar{\mathbf{v}}_i$ retienne seulement la partie des cartes de caractéristiques appartenant à la ROI. Ainsi, la sortie de cette étape $\bar{\mathbf{v}}_i$ a une taille $(\bar{H}_i \times \bar{W}_i \times \bar{L}_i \times C)$ variable, en fonction de la taille de la ROI/lésion.
- Afin d'explorer l'information texturale à différentes échelles tout en construisant un descripteur du sub-volume $\mathbb{X}_i$ de taille fixe, plusieurs opérations de pooling sont utilisés pour ré-échantillonner les cartes $\bar{\mathbf{v}}_i$ de taille $(\bar{H}_i \times \bar{W}_i \times \bar{L}_i \times C)$ vers une pyramide de cartes $\bar{\bar{\mathbf{v}}}_i^{(p)}$ de taille croissante ($|\bar{\bar{\mathbf{v}}}_i^{(3)}| > |\bar{\bar{\mathbf{v}}}_i^{(2)}| > |\bar{\bar{\mathbf{v}}}_i^{(1)}|$, avec $|.|$ le cardinal) mais fixe pour tous les patients ($|\bar{\bar{\mathbf{v}}}_i^{(p)}| = |\bar{\bar{\mathbf{v}}}_j^{(p)}|, \forall i, j \in [1 : N]$). L'étape suivante est donc l'estimation de la grille d'échantillonage pour chaque échelle de la pyramide. Soit P_p un "pooling" permettant d'avoir une taille de sortie $(p \times p \times p)$. La taille du noyau de "pooling" sera alors $(\lceil \frac{\bar{H}_i}{p} \rceil \times \lceil \frac{\bar{W}_i}{p} \rceil \times \lceil \frac{\bar{L}_i}{p} \rceil)$ et le stride $(\lfloor \frac{\bar{H}_i}{p} \rfloor \times \lfloor \frac{\bar{W}_i}{p} \rfloor \times \lfloor \frac{\bar{L}_i}{p} \rfloor)$ (avec $\lfloor . \rfloor$ la partie entière inférieure et $\lceil . \rceil$ la partie entière supérieure).
- Une fois les tailles recalculées, on appliquer les "poolings" P_p à $\bar{\mathbf{v}}_i$ en fonction de la taille de noyau et du stride calculés précédemment. Si on considère deux "poolings" P_4 et P_2 et $C = 64$, les tailles $(\bar{\bar{H}}_i^{(4)} \times \bar{\bar{W}}_i^{(4)} \times \bar{\bar{L}}_i^{(4)} \times C)$ et $(\bar{\bar{H}}_i^{(2)} \times \bar{\bar{W}}_i^{(2)} \times \bar{\bar{L}}_i^{(2)} \times C)$ des sorties $\bar{\bar{\mathbf{v}}}_i^{(4)}$ et $\bar{\bar{\mathbf{v}}}_i^{(2)}$ seraient respectivement $(4 \times 4 \times 4 \times 64)$ et $(2 \times 2 \times 2 \times 64)$.
- Vectoriser et concaténer le résultat de tous les "poolings" P_p. Avec notre exemple de P_4 et P_2 nous aurions donc un vecteur final $\mathbf{z}_i$ pour chaque patient i de taille $4 \times 4 \times 4 \times 64 + 2 \times 2 \times 2 \times 64 = 4608$ pour décrire l'image $\mathbb{X}_i$.

Lorsque le modèle ne contient pas de SPP, une couche de "poolling" moyenneur est ajoutée à chaque bloc de convolution afin de diminuer le nombre de paramètres. Une alternative est d'utiliser un modèle d'attention spatiale.

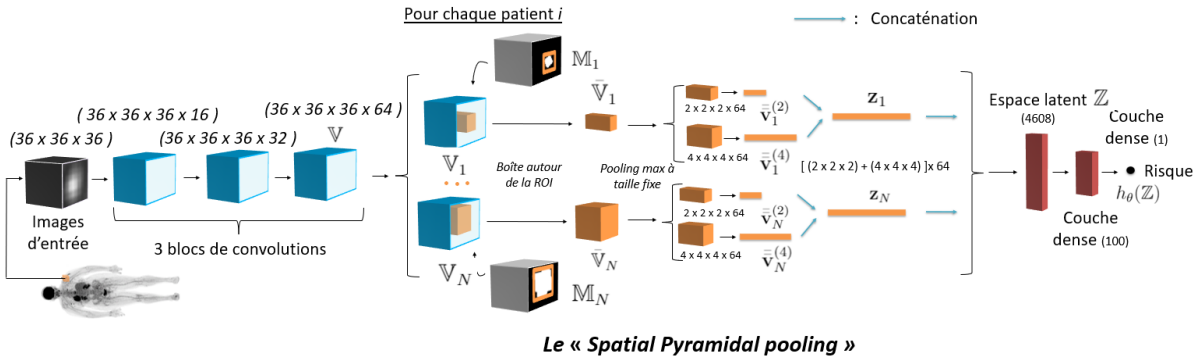


FIGURE 10.2 – Schéma de la méthode SPP (Spatial Pyramidal Pooling). On peut voir en bleu les blocs de convolution, en orange la SPP et en rouge le bloc de prédiction.

10.1.2.2 Module d'attention CBAM

Un mécanisme d'attention apprend à attribuer des poids aux caractéristiques extraites en fonction de leur pertinence pour une tâche donnée, et permet ainsi d'améliorer la performance de la tâche. Le mécanisme est flexible dans le sens où les poids s'adaptent à l'image d'entrée. Nous avons, lors de ces travaux, évalué le modèle CBAM [122] (Convolutional Block Attention Module) qui comprend une partie d'attention spatiale et une partie d'attention sur les filtres. Ce bloc d'attention fut placé après les blocs de convolution (ou après la SPP en fonction de la conformation du modèle). La figure 10.3 schématise la méthode.

L'attention sur les filtres est calculée en compressant l'information spatiale des couches convolutionnelles avec des opérations de "pooling" max et moyenneur (figure 10.4A.). Les valeurs compressées sont ensuite transmises à un réseau partagé de perceptron multicouche (MLP) pour prédire les poids d'attention. Les poids d'attention spatiale sont calculés à chaque emplacement de l'image. Ils sont calculés en appliquant les "pooling" moyen et max pour compresser les canaux (figure 10.4B.). Les deux sont concaténés et une couche convolutionnelle à un filtre est appliquée.

Les deux matrices d'attention sont ensuite appliquées aux cartes de caractéristiques par une multiplication par éléments. De cette façon, le CBAM concentre l'attention spatiale à l'intérieur de la lésion et non autour. Pour cette raison, elle constitue une alternative intéressante à la SPP. Les cartes d'attention spatiale peuvent également, dans certains cas, montrer la partie la plus importante de la lésion. L'attention sur les filtres fournit des informations sur les filtres les plus informatifs et donne ainsi une certaine interprétabilité au modèle. A la fin, la sortie du bloc CBAM est vectorisée pour constituer le vecteur de

caractéristiques $\mathbf{z}_i$ représentant chaque patient dans un espace latent. Les deux approches, SPP et CBAM, furent intégrées à notre modèle afin de répondre à la grande variabilité et la petite taille de nos lésion.

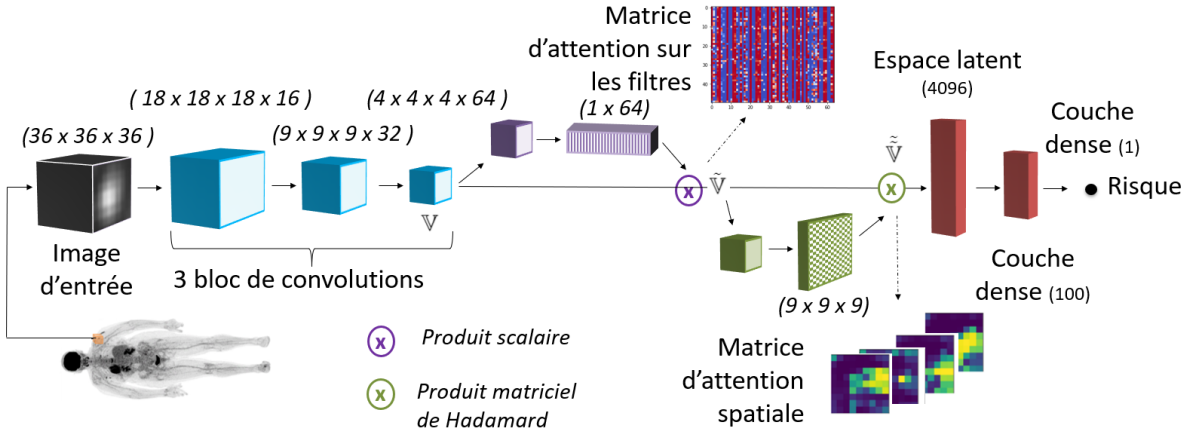


FIGURE 10.3 – Modèle 3D contenant les blocs d'attention spatiale et sur les filtres. Bleu : bloc de convolution ; Violet : bloc d'attention sur les filtres [122] ; Vert :Bloc d'attention spatiale ; Rouge : bloc de prédiction (cas de l'utilisation de la fonction de coût Cox). Les détails des blocs d'attentions sont présentés dans la figure 10.4. La matrice d'attention sur les filtres est présentée en détails dans l'annexe F.

10.1.3 Pré-entraînement du modèle

Avant de s'intéresser au bloc de prédiction nous allons décrire la première stratégie pour contrer la faible taille de notre base de données. Une méthode très utilisée pour contrer le problème du manque de données (et donc éviter le sur-apprentissage) mais aussi pour accélérer l'apprentissage du réseau est l'apprentissage par transfert, qui utilise une étape de pré-entraînement.

10.1.3.1 Les étapes d'un pré-entraînement

Un pré-entraînement consiste à prendre un jeu de données, généralement différent du jeu de données d'intérêt, et de réaliser une tâche prétexte afin d'initialiser les poids. Pour ce faire, le modèle est modifié en remplaçant la dernière couche (couche de prédiction) par une couche permettant de réaliser la tâche prétexte. Par exemple, pour une tâche prétexte de classification binaire, la couche de prédiction de risque de notre modèle va être remplacée par une couche dense à 2 neurones et une activation Softmax. Le réseau est ensuite entraîné afin de prédire les deux classes grâce au jeu de données de pré-entraînement. Une fois le pré-entraînement terminé, les poids vont être utilisés pour initialiser le modèle d'intérêt. La dernière couche de prédiction de la valeur prétexte est remplacée par notre

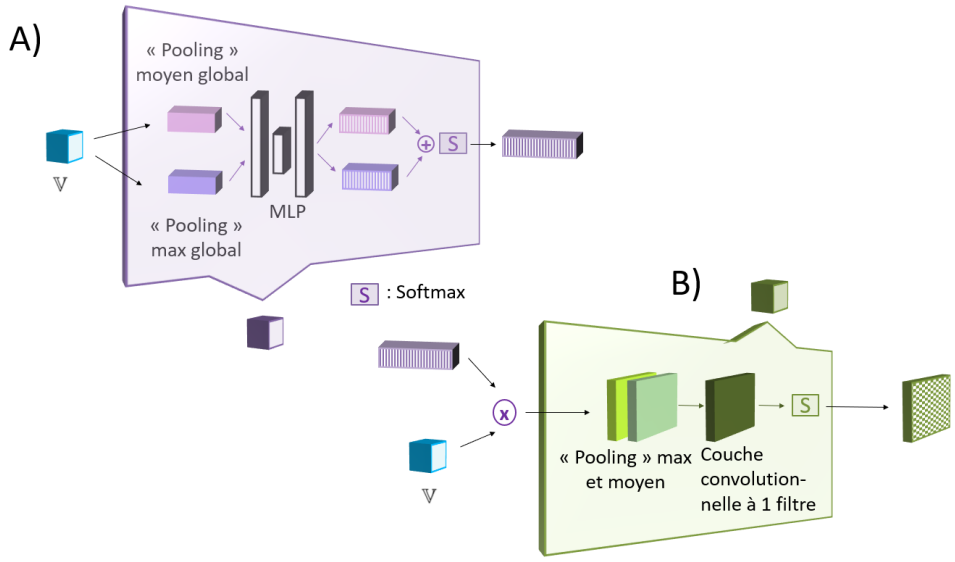


FIGURE 10.4 – Bloc d'attention CBAM [122]. A) Bloc d'attention sur les filtres. B) Bloc d'attention spatiale.

couche de prédiction de la valeur d'intérêt (initialisée aléatoirement). Par la suite, plusieurs possibilités s'offrent à nous pour entraîner notre modèle pour la tâche principale.

— **L'extraction de caractéristiques** : Dans cette méthode, toutes les couches du réseau pré-entraîné autres que la couche de prédiction sont gelées. Ainsi, lors de l'apprentissage de notre tâche d'intérêt, seule la dernière couche sera mise à jour. Cette technique est conseillée lorsque l'ensemble de données est très petit et que les images de pré-entraînement sont très proches des images de notre jeu de données.

— **Le "finetuning" total** : Dans cette méthode, aucune couche n'est gelée. Le pré-entraînement aura simplement servi à initialiser les couches du réseau. Lors de l'apprentissage avec nos données, les poids de toutes les couches vont être mis à jour. Cette méthode est principalement utilisée lorsque le jeu d'images est grand, cela permet alors d'accélérer l'apprentissage du réseau par rapport à un réseau initialisé aléatoirement ou par des zéros.

— **Le "finetuning" partiel** : Cette dernière méthode est un mélange des deux premières. Seules certaines couches du réseau vont être gelées. Cette technique est utilisée lorsque la nouvelle base de données est petite mais très différente des images du pré-entraînement. Dans ce cas là, ce sont les couches basses (début du réseau) qui vont être fixées, et les couches hautes (fin du réseau) qui vont être entraînées. En effet, les caractéristiques des couches basses sont simples et génériques et peuvent se retrouver dans des données très différentes tandis que les caractéristiques des couches hautes sont plus spécifiques et discriminantes pour chaque tâche.

Le pré-entraînement se fait généralement à l'aide d'un autre jeu de données et permet d'initialiser les poids du réseau. Cependant, afin que cela soit efficace il faut que les données

soit relativement équivalentes à nos données en taille et en texture. En effet, il est courant d'utiliser des bases de données comme ImageNet [94], CIFAR [130] or MNIST [96] qui ont un grand nombre d'images mais qui sont relativement éloignées des caractéristiques des images TEP et sont souvent de trop grande taille pour notre réseau (ImageNet, CIFAR). Il serait intéressant d'utiliser des bases de données médicales qui se rapprochent plus de nos données mais elles sont souvent bien plus petites en terme de quantité d'images. Certaines bases de données comme BRATS [131] contiennent des images médicales en grand nombre mais sont des images IRM et sont trop grandes pour notre réseau. Une dernière solution est l'utilisation de nos propres données.

10.1.3.2 Pré-entraînement avec une classification binaire de nos données

En effet, bien que nous ayons un nombre restreint de patients, en prenant des patchs dans chaque image 3D et en se concentrant sur une tâche plus simple telle que la classification binaire, il est possible d'en tirer avantage. Nous avons donc choisi de pré-entraîner notre réseau à partir de patchs de nos images (de la lésion la plus fixante pour chaque patient) et de les classifier en fonction du ratio de pixels appartenant à la lésion qu'ils contiennent. L'algorithme de récupération de notre jeu de données de classification binaire est le suivant :

1. Pour chaque patient i,
   - Tous les patchs positifs possibles sont extraits (un patch est considéré comme positif lorsque le ratio de pixels appartenant à la lésion est supérieur à un ratio r_+) [voir la figure 10.5].
   - $N_{\text{patch}}/2$ patchs positifs sont ajoutés à la liste des patchs $\mathbb{X}_{\text{patch}_i}$ et nous complétons avec $N_{\text{patch}}/2$ patchs négatifs pris aléatoirement dans l'image.
2. Les patchs positifs et négatifs de tous les patients sont ensuite réunis dans $\mathbb{X}_{\text{patch}}$
3. Le vecteur $\mathbb{Y}_{\text{bin}} = [\mathbf{y}_{\text{bin}_1}, ..., \mathbf{y}_{\text{bin}_k}, ..., \mathbf{y}_{\text{bin}_{N_{\text{patch}}}}]^T$ avec les étiquettes pour l'ensemble des patchs est créé avec $\mathbf{y}_{\text{bin}_k} = \{0, 1\}^N$ où chaque étiquette est égale à 1 lorsque le pourcentage de pixels du patchs appartenant à la lésion est supérieur ou égal au ratio r_+ et 0 sinon.

Le set de données obtenues peut s'écrire $\mathcal{X}_{\text{bin}} = \{\mathbb{X}_{\text{patch}_k}, \mathbf{y}_{\text{bin}_k}\}_{k=1}^{N_{\text{patch}}}$. Les données sont séparées en deux sets : un set d'entraînement $\mathcal{X}_{\text{patch}_{\text{train}}}$ et un set de validation $\mathcal{X}_{\text{patch}_{\text{val}}}$ (en faisant attention à ce que les patchs d'un même patient restent dans un seul set). Grâce à ce set de classification binaire, nous allons pré-entraîner notre réseau. Pour ce faire, nous modifions le modèle en remplaçant la dernière couche (couche de prédiction du risque) par une couche dense à 2 neurones et une activation Softmax. On entraîne ensuite le réseau grâce à notre set d'entraînement de données binaires $\mathcal{X}_{\text{patch}_{\text{train}}}$. Les poids sont initialisés avec les poids du modèle donnant les meilleurs résultats sur le set de validation. Les trois

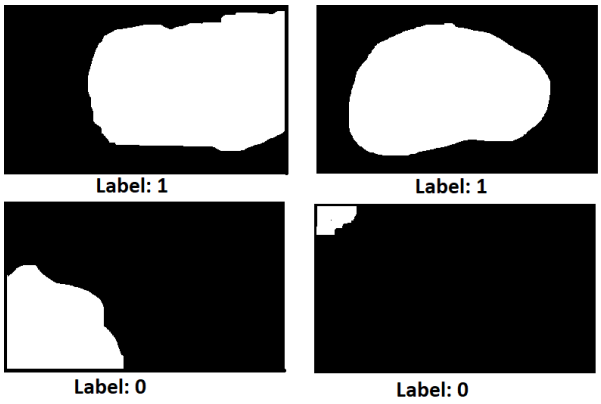


FIGURE 10.5 – Schéma des patchs utilisés pour le pré-entraînement binaire. Les patchs sont considérés comme positifs lorsque le rapport entre les pixels de la lésion et du fond est supérieur à un rapport r_+ (ici 50%).

méthodes de congélations furent testées (extraction de caractéristiques, "finetuning" total et "finetuning" partiel).

10.1.3.3 Pré-entraînement auto-supervisé contrastif grâce à des fonctions de coût triplet

Un groupe de méthodes de plus en plus utilisé est l'apprentissage auto-supervisé. L'auto-apprentissage peut lui même être divisé en trois groupes : les tâches prétextes conçues par les experts (colorisation, prédiction de la rotation, puzzle, etc.), l'apprentissage basé sur le regroupement, et l'apprentissage contrastif. C'est ce dernier que nous avons choisi. L'apprentissage contrastif consiste à contraster les distances dans l'espace latent entre deux échantillons d'une même classe, et l'un d'entre eux avec un échantillon d'une classe différente. Il est très utilisé car les caractéristiques qui en résultent sont plus invariantes (à la rotation, à la position etc) qu'avec les autres groupes de méthodes. Nous choisissons de réaliser le pré-entraînement contrastif avec les fonctions de coût tripletSurv et SV-tripletSurv basées sur des triplets extraits de données d'entraînement. Comme expliqué dans la section suivante, nous modifions les fonctions de coût triplet [108] et SV-triplet [109] pour les adapter à la censure. Celles-ci deviendront respectivement tripletSurv et SV-tripletSurv [voir les sous-sections 10.2.4 et 10.2.4.2]

Lors de ce second pré-entraînement, nous supprimons provisoirement les couches denses de taille 100 et finale. La couche de sortie sera donc la couche de vectorisation de taille 4096 contenant l'espace latent $\mathbb{Z}$. Le pré-entraînement est adapté pour prédire l'espace latent en fonction du temps de survie et du statut de censure des patients, grâce à une fonction de coût tripletSurv (ou SV-tripletSurv). Les trois méthodes de congélation (extraction de caractéristiques, "finetuning" total et "finetuning" partiel) sont testées.

10.2 Adapter l'apprentissage automatique à la survie

Comme décrit dans la section 8, il existe plusieurs défis liées à l'adaptation de réseaux de neurones à la tâche de survie, tels que la censure et la dépendance de sujets à l'ensemble de la population. Dans cette section, nous nous intéressons à la définition d'une fonction de coût adaptée à la survie. Une première contribution est une revue des fonctions de coût existantes pour l'apprentissage de la survie. Une deuxième contribution est la proposition de deux adaptations de fonctions par triplets à la survie.

10.2.1 Le modèle de Cox adapté à une fonction de coût

L'une des principales méthodes d'analyse de survie est le modèle de Cox qui est employé pour évaluer les effets des covariables sur le temps de survie. Une approche commune de l'analyse de survie avec les CNN est de dériver une fonction de coût à partir du modèle de Cox, et plus précisément, à partir de la fonction de vraisemblance qui permet d'estimer ses paramètres. La vraisemblance partielle pour chaque coefficient est le produit des probabilités de risque dans le temps. Pour chaque événement-temps t_i la probabilité est calculée comme le rapport entre le risque du patient i et le risque cumulé de tous les individus encore à risque au temps t_i. Comme vu dans la section 3.2.2, le risque du patient i est définit par $h(t|X_i) = h_0(t)e^{\beta^T X_i}$. Au vue de cette définition, la vraisemblance partielle sur tous les coefficients β est :

$$L_{\text{cox\_lin}}(\mathcal{X}, \beta) = \prod_{\{i|\delta_i=1\}} \frac{e^{\beta^\top X_i}}{\sum_{\{j|t_j \geq t_i\}} e^{\beta^\top X_j}}, \tag{10.2}$$

avec un ensemble d'entraînement $\mathcal{X} = \{\mathbb{X}_i, \mathbf{y}_i\}_{i=1}^N$ où est associé à chaque patient i, le temps d'évènement t_i et la censure δ_i. Ici, le produit est effectué sur les événements temporels définis (c'est-à-dire non censurés). On peut voir que l'élément $h_0(t)$ a disparu lors de la division. Ainsi, le risque n'est plus dépendant du temps auquel il est calculé et nous considérons maintenant $h_\beta(\mathbb{X}_i) = e^{\beta^\top \mathbb{X}_i}$. Finalement, suivant [12, 100], le modèle linéaire $\beta^\top X_i$ est remplacé par la sortie d'un réseau de neurones $h_\theta(\mathbb{X}_i)$ paramétré par des poids θ tout en conservant la même vraisemblance partielle, avec $\mathbb{X}_i$ l'image d'entrée du patient i. Le calcul du logarithme négatif de la vraisemblance à partir de l'équation 10.2 conduit à la fonction de coût suivante pour optimiser les paramètres θ (équation de DeepSurv et Li et al. [115]) :

$$l_{\text{cox\_nn}}(\mathcal{X}, \theta) = -\sum_{\{i|\delta_i=1\}} [h_\theta(\mathbb{X}_i) - \log \sum_{\{j|t_j \geq t_i\}} e^{h_\theta(\mathbf{x}_j)}] \tag{10.3}$$

avec θ les poids du réseau, h_θ la fonction de risque prédit, et $\mathbf{x}_i$ le vecteur de données décrivant le patient i. Cette fonction de coût pousse le réseau à prédire, dans une couche de sortie linéaire de un neurone, les risques continus qui expliquent l'ordre des événements

dans l'ensemble de données. Notez que la sortie du patient actuel dépend de la sortie de tous les patients à risque au moment de l'événement.

10.2.2 Fonction de coût de discrète

A la place du hasard continu, Gensheimer et al. [13] proposent un modèle de survie statistique permettant d'obtenir une sortie plus interprétable qui prédit un temps de survie au lieu d'un risque. Le temps d'évènement est défini sur D intervalles de temps discrets $\mathbb{T} = [\Delta_1, \Delta_2, ..., \Delta_D]$. La couche de sortie du réseau est configurée pour avoir D neurones, chacun prédisant la probabilité $p_{d,\theta}(\mathbf{x}_i)$ de survivre à l'intervalle $\Delta_d \in \mathbb{T}$. Cette probabilité est complémentaire du risque car $h_{d,\theta}(\mathbf{x}_i) = 1 - p_{d,\theta}(\mathbf{x}_i)$. La vraisemblance est définie différemment pour un patient censuré ou non censuré. Pour un individu i non censuré, la vraisemblance d'avoir un évènement dans l'intervalle Δ_k est le risque $h_{k,\theta}(\mathbf{x}_i)$ multiplié par la probabilité de survivre aux intervalles Δ_1 à Δ_{k-1}, qui est $\prod_{d=1}^{k-1}(1-h_{d,\theta}(\mathbf{x}_i))$. Ainsi, la vraisemblance pour les individus non censurés peut être exprimée comme :

$$L(\mathbf{x}_i, \theta, k) = h_{k,\theta}(\mathbf{x}_i) \prod_{d=1}^{k-1} (1 - h_{d,\theta}(\mathbf{x}_i)) \tag{10.4}$$

En considérant la vraisemblance $L(\mathbf{x}_i, \theta, k)$ de l'équation 10.4, le logarithme de la vraisemblance des individus non censurés $\log L(\mathbf{x}_i, k, \theta)$ est :

$$\log L(\mathbf{x}_i, \theta, k) = \log h_{k,\theta}(\mathbf{x}_i) + \sum_{d=1}^{k-1} \log(1 - h_{d,\theta}(\mathbf{x}_i)) \tag{10.5}$$

Pour les individus censurés, seul le produit sur les temps de survie est conservé (deuxième terme). Ainsi le logarithme de la vraisemblance des individus censurés $\log \tilde{L}(\mathbf{x}_i, k, \theta)$ est :

$$\log \tilde{L}(\mathbf{x}_i, \theta, k) = \sum_{d=1}^{k-1} \log(1 - h_{d,\theta}(\mathbf{x}_i)) \tag{10.6}$$

Si nous considérons que la sortie du réseau est un vecteur de D probabilités de survie prédites $p_{d,\theta}(\mathbf{x}_i)$, où la probabilité de survie pour l'intervalle d est $p_{d,\theta}(\mathbf{x}_i) = 1 - h_{d,\theta}(\mathbf{x}_i)$, alors pour chaque individu i, cette vraisemblance peut être exprimée en termes de vecteurs binaires de vérité-terrain $\mathbf{s}_i, \mathbf{e}_i \in \{0, 1\}^D$, indiquant respectivement le vecteur des temps de survie $\mathbf{s}_i$ (codé comme un vecteur binaire où les intervalles avant le temps de survie sont remplis de uns, et de zéros après), et le vecteur temps-événement $\mathbf{e}_i$ (codé comme un vecteur avec un Dirac dans l'intervalle où un événement s'est produit, ou rempli de zéros en cas de censure). La figure 10.6 illustre un exemple de vecteur d'entrée (vérité terrain) composé des deux vecteurs binaires $\mathbf{s}_i$ et $\mathbf{e}_i$ pour deux patients. Une autre façon de présenter ces vecteurs d'entrée est exposé dans la figure 10.7.

En utilisant $\mathbf{s}_i$ et $\mathbf{e}_i$ la fonction de coût (log de la vraisemblance partielle négative)

	Time										Censorship									
Patient 1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
Patient 2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 10.6 – Exemple de vecteur d'entrée (vérité terrain) composé des deux vecteurs binaires $\mathbf{s}_i$ et $\mathbf{e}_i$. Si on considère 10 intervalles de 100 jours, Le patient 1 a un évènement au temps de 700 jours et n'est pas censuré. Le patient 2 a été suivi 300 jours et est censuré

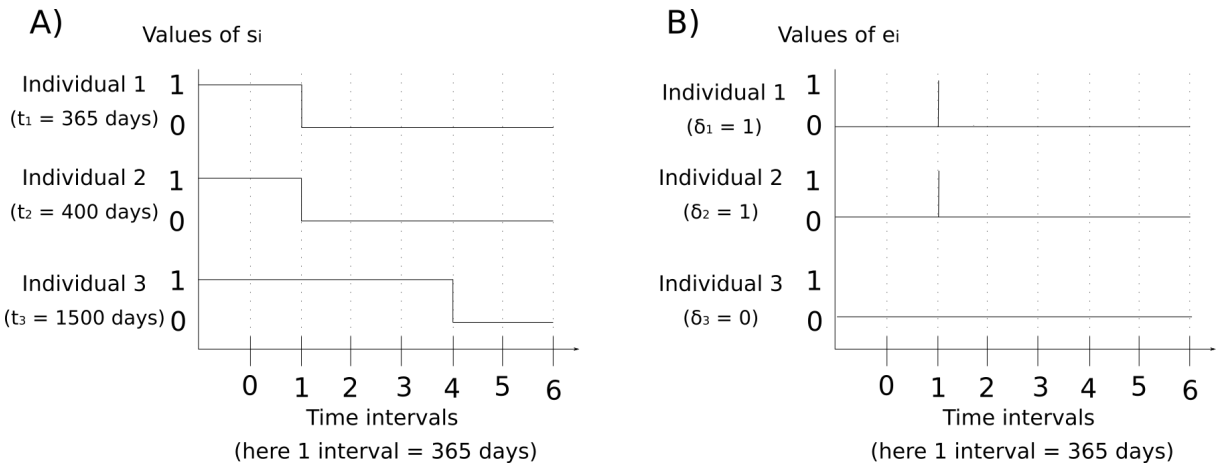


FIGURE 10.7 – Exemple de vecteur d'entrée de la fonction de coût de survie discrète. Dans notre cas, chaque intervalle représente 365 jours. A) Exemple de vecteur s_i (temps de survie). B) Exemple de vecteur e_i (temps d'évènement).

pour tout individu peut être définie par l'équation suivante :

$$l_{\text{ind}}(\mathbf{x}_i, \theta) = -\sum_{d=1}^{D} \left[\log\left[1 - \mathbf{e}_i(d)(1 - h_{d,\theta}(\mathbf{x}_i))\right] + \log\left[1 - \mathbf{s}_i(d) h_{d,\theta}(\mathbf{x}_i)\right] \right] \tag{10.7}$$

En sommant sur les N individus et en exprimant la sortie $p_{d,\theta}$ du réseau, on obtient la fonction de coût de survie discrète suivante :

$$L_{\text{dsurv}}(\mathcal{X}, \theta) = -\sum_{i=1}^{N} \sum_{d=1}^{D} \left\{ \log\left[1 - \mathbf{e}_i(d) p_{d,\theta}(\mathbf{x}_i)\right] + \log\left[1 - \mathbf{s}_i(d)(1 - p_{d,\theta}(\mathbf{x}_i))\right] \right\} \tag{10.8}$$

Outre l'interprétabilité de ses résultats, ce modèle n'est plus proportionnel, ni dépendant d'un risque de base h_0. De plus, ce choix rend la vraisemblance de chaque individu indépendante des autres patients, favorisant l'entraînement avec de petites tailles de batchs et par opposition au jeu de données complet pour la fonction de coût Cox [13]. Finalement, la méthode fournit des temps plutôt que des risques. La plage de valeurs de sortie ne dépendra donc pas des valeurs de la plage d'apprentissage. Cela facilite l'interprétation de la valeur prédite pour un nouvel individu.

10.2.3 Combinaison avec une fonction de coût d'ordonnancement par paires

Une manière directe de formuler l'analyse de survie est la régression du temps d'événement t_i pour chaque patient i non censuré, avec les paramètres du modèle optimisés par une fonction de coût d'erreur quadratique moyenne (MSE). Afin de tenir compte des données censurées et de restreindre davantage le problème, Jing et al. [14] ont proposé un réseau qui prédit le temps de survie $\hat{t}_i$, entraîné avec la fonction de coût RankDeepSurv (que nous appelleront aussi Rank&MSE) composée d'une fonction de coût MSE étendue l_{reg} et d'une fonction de coût d'ordonnancement par paire l_{rank}, équilibré par des paramètres α et λ :

Ainsi, la fonction de coût proposée est $l_{\text{RankSurv}} = \alpha l_{\text{reg}} + \lambda l_{\text{rank}}$, avec :

$$l_{\text{reg}} = \frac{1}{N} \sum_{i,j \in \mathcal{C}_{\text{reg}}} (t_i - \hat{t}_i)^2 \tag{10.9}$$

où $\mathcal{C}_{\text{reg}}$ est l'ensemble des individus admissibles, qui comprend les individus non censurés et les individus dont le temps prédit $\hat{t}_i$ est supérieur ou égal au temps censuré t_i. Pour la fonction de coût d'ordonnancement, les distances sont comparées par paires. La fonction de coût d'ordonnancement l_{rank} est donc :

$$l_{\text{rank}} = \frac{1}{N} \sum_{i,j \in \mathcal{C}_{\text{rank}}} \Big[(t_j - t_i) - (\hat{t}_j - \hat{t}_i) \Big]^2 \tag{10.10}$$

L'ensemble $\mathcal{C}_{\text{rank}}$ contient les paires considérées comme valides. Une paire est considérée valide si la distance entre les vérités-terrains des deux individus est plus grande que la distance entre les deux prédictions. De plus, il faut que les deux individus soient non censurées ou que l'individu qui a le temps de survie le plus court soit non censuré.

Nous avons testé des adaptations de la fonction de coût RankDeepSurv, en remplaçant la fonction de régression MSE par les fonctions de coût Cox et de survie discrète. Ainsi, la fonction Rank&cox correspond à $l_{\text{Rank\&cox}} = \alpha l_{\text{cox-nn}} + \lambda l_{\text{rank}}$ et la fonction Rank&discret correspond à $l_{\text{Rank\&discret}} = \alpha l_{\text{dsurv}} + \lambda l_{\text{rank}}$.

Finalement, les fonction de coût Rank&MSE et Rank&discret tirent avantage de la fonction de Cox en permettant la prédiction d'un temps de survie au lieu d'un risque. De plus, elles prennent davantage en compte les patients censurés que les fonctions de Cox et de survie discrète. Cependant, contrairement à la fonctions de survie discrète, elles sont continues et prennent en compte l'ordonnancement des individus, ce qui les rapproches du c-index. Néanmoins, elles présentent l'inconvénient d'avoir des paramètres qui doivent être déterminés à la main et qui peuvent avoir une grande influence sur le résultat.

10.2.4 Combinaison avec une fonction de coût d'ordonnancement par triplets

Nous utiliserons dans ce travail deux fonctions de coût de "triplet" différentes. Celle de base [108] et celle appelée "Scale-Varying triplet" (ou SV-triplet) [109]. Nous les modifions pour les adapter à la survie et elles deviennent respectivement les fonctions tripletSurv et SV-tripletSurv.

10.2.4.1 La fonction de coût triplet

La fonction de coût triplet est une fonction qui se base sur l'ordonnancement de triplets, par opposition à la fonction Rank&MSE qui se base sur l'odonnancement par paires. Un triplet est composé de trois images : une ancre $\mathbf{x}^a$, un échantillon positif $\mathbf{x}^p$ et un échantillon négatif $\mathbf{x}^n$. Une image est considérée comme positive si elle appartient à la même classe que l'ancre, et négative sinon. Le but est de construire un espace de caractéristiques discriminantes en rapprochant l'ancre de l'échantillon positif et en l'éloignant de l'échantillon négatif.

Dans la définition de la fonction de coût "triplet" utilisée par Schroff *et al.* [108], un triplet $(\mathbf{x}_k^a, \mathbf{x}_k^p, \mathbf{x}_k^n) \in \tau$ est considéré quand $\| f(\mathbf{x}_k^a) - f(\mathbf{x}_k^p) \|_2^2 + \mu \leq \| f(\mathbf{x}_k^a) - f(\mathbf{x}_k^n) \|_2^2$ avec $f(x)$ le vecteur de caractéristiques issues d'un CNN ou espace latent (dans notre cas la couche aplatie avant les couches denses), μ la marge imposant une séparation minimale entre le négatif et le positif, et τ l'ensemble de tous les triplets possibles. Remplaçons $(f(\mathbf{x}_k^a), f(\mathbf{x}_k^p), f(\mathbf{x}_k^n))$ par $(\mathbf{f}_k^a, \mathbf{f}_k^p, \mathbf{f}_k^n)$ par la suite. Par conséquent, la fonction de coût à minimiser devient :

$$l_{\text{trip}} = \sum_{k \in \tau} \| \mathbf{f}_k^a - \mathbf{f}_k^p \|_2^2 - \| \mathbf{f}_k^a - \mathbf{f}_k^n \|_2^2 + \mu \tag{10.11}$$

Cependant, pour éviter un trop grand nombre de triplets et pour s'assurer d'une convergence suffisamment rapide, deux méthodes de génération de triplets ("online" or "offline") et deux méthodes de sélection de triplets ont été définies ("hard" et "all") [108].

Dans la méthode hors ligne ("offline"), tous les triplets possibles sont générés et ensuite mis par batch. A l'inverse, l'approche en ligne ("online") génère des triplets pour chaque nouveau batch, en vérifiant et en éliminant les triplets invalides par la suite. Pour la sélection, l'option "batch hard" ne conserve pour chaque ancre que les échantillons positifs et négatifs les plus durs du batch (respectivement la plus grande distance et la plus petite distance). Pour la méthode "batch all", la fonction de coût est moyennée sur tous les triplets valides. Nos expériences sont réalisées avec la méthode "batch hard" en ligne, qui est plus efficace et favorise la convergence.

Comme cette fonction de coût triplet n'est pas adaptée à la censure mais conçue pour la classification, nous adaptons cette fonction à l'analyse de survie. Nous commençons par

discrétiser le temps de survie t_i en un nombre de classes y_i fixe dans le but de déterminer les échantillons positifs et négatifs. Dans un second temps nous définissons des règles de vérification de la validité des triplets en fonction de leur censure. Pour une image ancre donnée, nous ne prenons donc le triplet seulement lorsque :

— Pour les images positives : l'ancre et le positif sont distincts et ont tous deux un événement, ou, le positif et l'ancre appartiennent à la dernière classe. Ceci peut être définit comme suit : $\{a \neq p\} \cap (\{\delta_k^a = \delta_k^p = 1\}$ OU $\{y_k^a = y_k^p = Y_{\max}\})$ avec $Y_{\max}$ la classe maximale.

— Pour l'image négative : l'ancre et le négatif ont une classe différente et les deux ont eu un événement, ou, celle avec le temps de survie minimal a un événement. Ceci peut être définit comme suit : $\{y_k^a \neq y_k^n\} \cap (\{\delta_k^a = \delta_k^n = 1\}$ OU $\{\delta_k^q = 1\})$
avec $q = \begin{cases} a, & \text{si } y_k^a < y_k^n. \\ n, & \text{sinon.} \end{cases}$

Ceci n'est applicable que dans le cas de la censure de droite. Nous appelons notre fonction triplet adaptée à la survie tripletSurv.

10.2.4.2 La fonction de coût Scale-Varying triplet (SV-triplet)

Plus récemment, Im et al. [109] ont proposé la fonction de coût Scale-Varying triplet (SV-triplet), qui considère des différences continues (âge) au lieu des différences de classe pour générer des triplets. Cette fonction de coût SV-triplet peut être considérée comme une adaptation des fonctions de coût triplet [108] pour la classification des valeurs continues. Les différences résident dans :

— la détermination des éléments positifs et négatifs du triplet

— l'ajout d'un poids qui s'adapte à l'échelle (différence d'âge)

— la suppression du besoin de la marge μ.

Ainsi, Im et al. [109] définissent la fonction de coût comme :

$$l_{\text{sv-triplet}} = \frac{1}{|\tau|} \sum_{k \in \tau} w_k \cdot l_\tau(d_k^p, d_k^n) \tag{10.12}$$

Avec, τ l'ensemble des triplets choisis, $\tau = \{(\mathbf{f}_k^a, \mathbf{f}_k^p, \mathbf{f}_k^n) | a \neq p \neq n \cap \parallel t_k^a - t_k^p \parallel < \parallel t_k^a - t_k^n \parallel\}$; $\mathbf{f}_k^a$, $\mathbf{f}_k^p$, et $\mathbf{f}_k^n$ les espaces latents de respectivement l'ancre, le positif et le négatif, et, t_k^a, t_k^p, et t_k^n les vrais temps de survie. Les poids w_k donnent la priorité aux triplets où l'ancre et le positif ont des évènements proches :

$$w_k = \frac{1 + \epsilon}{\parallel \overline{t_k^a} - \overline{t_k^p} \parallel + \epsilon} - 1 \tag{10.13}$$

avec ϵ une petite constante pour éviter la division par zéro, $\overline{t_k} = (t_k - t_{\min})/(t_{\max} - t_{\min})$ les valeurs vraies normalisées avec $t_{\min}$ et $t_{\max}$ calculés sur les N patients. Finalement la

fonction de coût individuelle pour un triplet est

$$l_\tau(d_k^p, d_k^n) = -\log(d_k^n) \tag{10.14}$$

avec avec d_k^n la sortie de la fonction softmax calculées sur les distances et $\mathcal{D}$ la distance euclidienne [voir l'équation 10.15].

$$d_k^n = \frac{\exp(\mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^n))}{\exp(\mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^n)) + \exp(\mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^p))} \tag{10.15}$$

Pour adapter cette fonction de coût 10.12 à la survie nous modifions le critère de sélection $\| t_k^a - t_k^p \| < \| t_k^a - t_k^n \|$, et donc τ, pour considérer la censure. Ainsi, les triplets acceptés sont ceux vérifiants :

— $t_k^a \leq t_k^p < t_k^n$ et $\delta_k^a = \delta_k^p = 1$ ($\delta_k^n = 1$ ou 0)

— $t_k^p < t_k^a < t_k^n$ et $\delta_k^a = \delta_k^p = 1$ et $\mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^p) < \mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^n)$ ($\delta_k^n = 1$ ou 0)

— $t_k^n < t_k^a < t_k^p$ et $\delta_k^n = \delta_k^p = 1$ et $\mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^p) < \mathcal{D}(\mathbf{f}_k^a, \mathbf{f}_k^n)$ ($\delta_k^a = 1$ ou 0)

— $t_k^n < t_k^p \leq t_k^a$ et $\delta_k^n = \delta_k^p = 1$ ($\delta_k^a = 1$ ou 0)

Les cas des triplets valides sont résumés dans la figure 10.8.

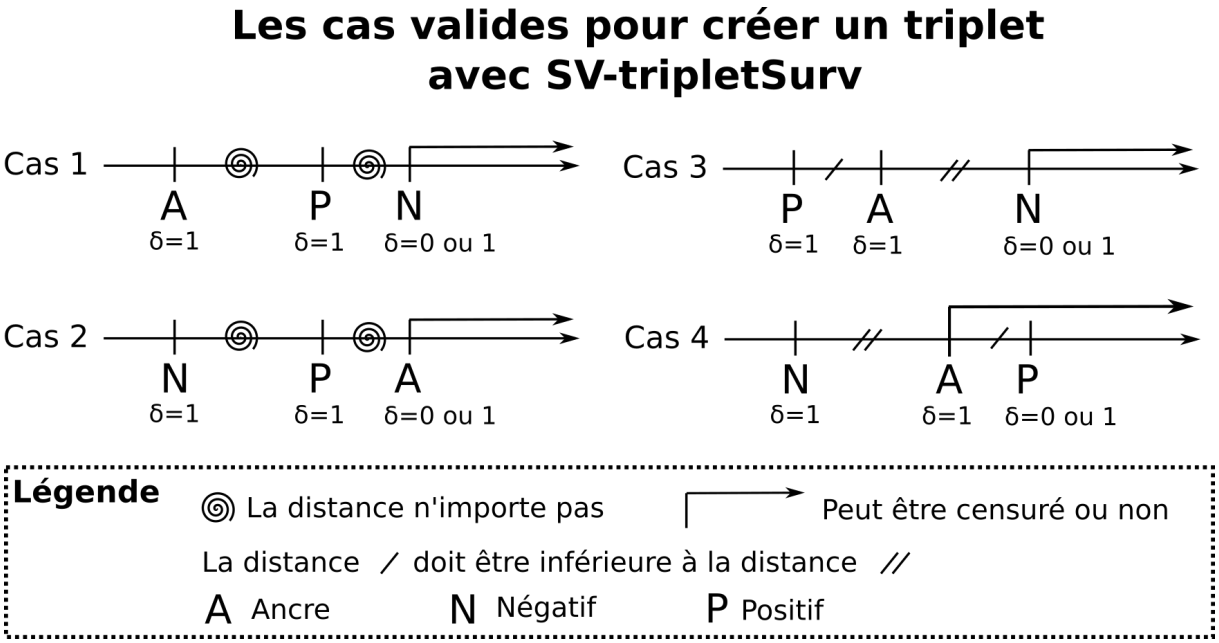


FIGURE 10.8 – Résumé des triplets valides pour la fonction de coût SV-tripletSurv.

De plus, nous avons ajouté une étape de normalisation des sorties du Softmax qui est présentée dans l'équation 10.16 pour éviter les petites valeurs de sortie.

$$\tilde{d_k^n} = \frac{d_k^n - \min_{q\in\psi}(d_q^n)}{\max_{q\in\psi}(d_q^n) - \min_{q\in\psi}(d_q^n)} \tag{10.16}$$

en considérant le min et le max de d^n sur le batch, et ψ le batch ($\psi \subset \tau$).

La fonction SV-triplet adaptée à la survie est appelée SV-tripletSurv.

10.2.4.3 Combinaison des fonctions de coût tripletSurv et SV-tripletSurv avec la fonction de Cox

Ces fonctions de coût tripletSurv et SV-tripletSurv seront utilisées en combinaison avec la fonction de Cox 10.2.1 afin de donner les fonctions de coût $l_{\text{tripletSurv\&Cox}}$ et $l_{\text{SV-tripletSurv\&Cox}}$ telles que $l_{\text{tripletSurv\&Cox}} = \alpha l_{\text{cox\_nn}} + \lambda l_{\text{tripletSurv}}$ et $l_{\text{SV-tripletSurv\&Cox}} = \alpha l_{\text{cox\_nn}} + \lambda l_{\text{SV-tripletSurv}}$, avec α et λ des valeurs choisies afin d'avoir des valeurs dans un même ordre de grandeur pour chaque fonction. Les fonctions de coût $l_{\text{tripletSurv}}$ et $l_{\text{SV-tripletSurv}}$ seront de plus utilisées pour les pré-entraînements constrastifs [voir la sous-section 10.1.3.3].

Chapitre 11

Validation expérimentale

Afin d'évaluer les méthodes discutées dans le chapitre 10 nous réalisons différentes expériences. L'évaluation se fait sur nos données provenant des bases IMAJEM [1] et EMN02/HO95 [2] décrites dans le chapitre 2. Ainsi, nous présenterons dans ce chapitre

les détails de ces expériences (section 11.1) puis nous discuterons de leurs résultats (section 11.2).

11.1 Cadre expérimental

Lors de l'entraînement et l'évaluation de nos modèles d'apprentissage profond, nous utilisons les deux bases de données IMAJEM [1] et EMN02/HO95 [2] avec respectivement 87 et 65 patients, et un taux de censure global de 45%. Nous utilisons en entrée de nos modèles les images TEP des lésions les plus fixantes (une par patient).

11.1.1 Le traitement des données

11.1.1.1 Définition de la dimension d'entrée

Tout d'abord, nous évaluons trois types d'entrées (une entrée 2D, une 2.5D et une 3D) dans le modèle B3D* sans attention (les modèles sont présentés dans la sous-section 11.1.4.1), auquel on change la couche d'entrée et la taille des filtres convolutifs pour s'adapter à la dimension d'entrée.

Modèle 2D Le modèle 2D reçoit en entrée une ou plusieurs coupes de la ROI (polygone délimitant grossièrement la lésion) traitées comme des images 2D différentes. Différentes configurations furent testées [voir la figure 11.1] :

- 9 coupes : on récupère 3 coupes 2D par axe (axiale, sagittale et coronale). Pour chaque axe, on prend une coupe au centre et deux coupes autour de la façon suivante : Soit x, y et z les positions respectivement dans les axes axiale, sagittale et coronale. Pour chaque axe k, avec $k \in \{1, 2, 3\}$, la taille maximale L_k de la ROI est $L_k = \max(\mathbb{X}) - \min(\mathbb{X})$. Les emplacements C_{ki} des coupes i, avec $i \in \{1, 2, 3\}$ pour l'axe k sont $C_{ki} = \min(\mathbb{X}) + \lfloor \frac{i \times L_k}{4} \rfloor$.
- 3 coupes : Pour chaque axe, on récupère, des coupes de la même façon que pour 9 coupes, mais en gardant seulement le coupe du centre ($i = 2$).
- 1 coupe : On ne garde simplement que le coupe centrale de l'axe sagittale ($k = 2$ et $i = 2$).

Chaque coupe va être associée aux valeurs de temps et de censure de la lésion d'origine mais traitée de façon indépendante dans la base de données. La figure 11.1 présente l'emplacement des coupes récupérés dans l'image.

Modèle 2.5D Pour le modèle 2.5D, de la même façon que la méthode 2D où l'on garde 3 coupes, on récupère la coupe du centre dans chaque axe. La différence réside dans

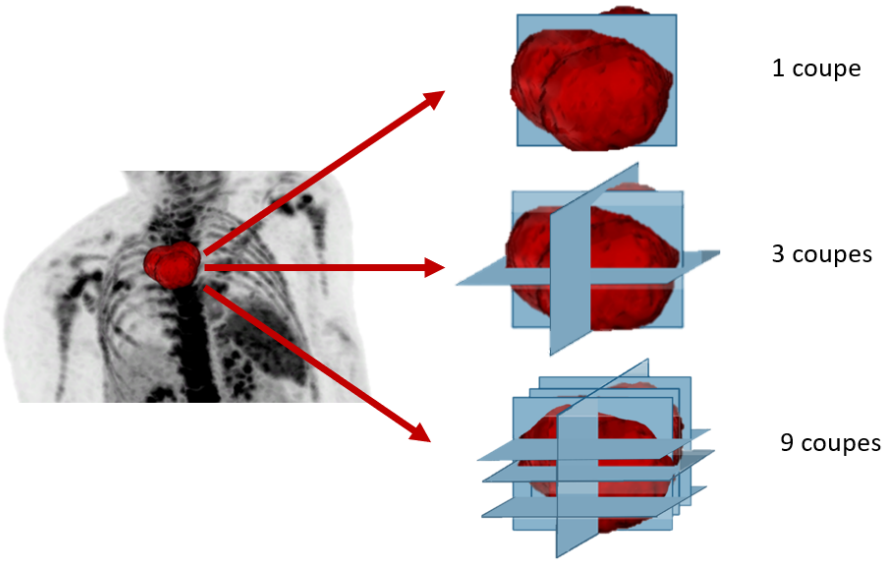


FIGURE 11.1 – Récupération des coupes dans la lésion 3D afin de produire des images 2D.

l'utilisation de ces coupes. Ainsi, dans le cas du 2.5D, chacune des trois coupes correspondra à un canal. La dimension 2.5D correspond donc à une image d'entrée à trois canaux d'un seul patient dans un modèle 2D.

Modèle 3D En entrée des modèles 3D nous avons un volume 3D par patient. Bien que cette dimension implique un nombre moins important d'entrées qu'avec 3 ou 9 coupes, et un plus grand nombre de paramètres, elle a pour avantage de garder l'information spatiale dans chaque direction pour chaque image.

11.1.1.2 L'extraction et le pré-traitement des images d'entrée

A partir de ces images 2D, 2.5D ou 3D, deux méthodes ont été utilisées pour obtenir des images de lésions de taille équivalentes et contrer le soucis de variabilité de taille des lésions. La taille de $(36 \times 36 \times 36)$ ou (36×36) fut choisie en prenant en compte les valeurs maximales et minimales de tailles des lésions [voir table 8.1]. Dans un soucis de simplification, nous aborderons le sujet en parlant d'un modèle 3D. De plus, il est intéressant de préciser que ce n'est pas la segmentation qui est ici utilisée pour extraire des images de lésions mais la ROI (polygone délimitant plus largement la lésion qu'une segmentation).

— Lorsque le module SPP est appliqué, on utilise la méthode que l'on appellera "Box", afin de garder l'information de contexte de la lésion avant de couper autour de la lésion grâce au module SPP. La méthode consiste à prendre une boîte de taille fixe $(36 \times 36 \times 36)$ autour des lésions (peu importe la taille de la lésion), avec pour centre le centre de la boîte englobante de la ROI.

— Lorsque le SPP n'est pas appliqué, on utilise la méthode "Interpolate", qui consiste à prendre une boîte englobante de la ROI de la taille de la lésion (et donc de taille variable) et d'appliquer une interpolation 3D cubique afin d'arriver à la taille fixe $(36 \times 36 \times 36)$.

Après leur extraction, les intensités des images sont normalisées par $\bar{\mathbb{X}} = \frac{\mathbb{X} - min(\mathbb{X})}{max(\mathbb{X}) - min(\mathbb{X})}$, avec $\mathbb{X}$ l'image à normaliser et $\bar{\mathbb{X}}$ l'image normalisée.

11.1.1.3 L'augmentation des données

L'augmentation des données à pour but de réduire le sur-apprentissage. Quatre opérations (rotation, translation, zoom, retournement) ont été appliquées de la même façon dans chaque expérience pour obtenir 15 ou 30 images différentes par patient. Par exemple, dans le cas du 3D, l'augmentation entraîne respectivement un total de 2280 et 4560 images 3D. Le détail des paramètres des opérations réalisées se trouve dans l'annexe D.

11.1.2 Détails d'implémentation des modèles

Outre le format de données d'entrée, nous précisons maintenant les détails d'implémentation de la version SPP de l'architecture [voir la figure 10.2], ainsi que les couches de sortie selon le modèle (discret ou continu) de la survie.

11.1.2.1 SPP et attention

Avec l'attention spatiale et/ou l'attention sur les cartes de caractéristiques intermédiaires, notre modèle B3D* reste fidèle à celui de la figure 11.2.

Avec la méthode SPP nous avons restreints nos "poolings" P_4 et P_2 à $(4 \times 4 \times 4)$ et $(2 \times 2 \times 2)$ voxels. En effet, contrairement à Li et al. [115] nous n'avons pas gardé le "pooling" de taille $(8 \times 8 \times 8)$, nos lésions étant trop petites [voir le tableau 8.1]. Le "pooling" P_8 de taille $(8 \times 8 \times 8)$ serait dominé par les informations de fond [1]. Cependant se limiter aux seuls "poolings" P_4 et P_2 implique de conserver moins d'information.

Finalement, lorsque nous avons avons combiné la SPP et l'attention, nous avons évalué deux configurations : la SPP avant et après l'attention. Hors, le module d'attention nécessite en entrée des images 2D ou 3D (et non un vecteur). La SPP, nécessitant une vectorisation afin de concaténer les deux "poolings" P_4 et P_2, nous avons décidé de remplacer la SPP par un "pooling" fixe P_4 unique (de taille $4 \times 4 \times 4$) lorsque celui-ci était placé avant l'attention. En effet, cet emplacement du "pooling" restait intéressant car permettait de diminuer le temps de calcul par rapport au placement de la SPP après l'attention.

1. L'image d'entrée devrait être plus grande, ce qui impliquerait de prendre de l'information du fond. L'intérêt de la SPP est perdu.

11.1.2.2 La discrétisation des temps

Lors de l'utilisation de la fonction de coût discrète de Cox, nous avons choisit de séparer les temps en 7 intervalles contenant chacun 365 jours (ainsi chaque intervalle correspond à un an). Cette même discrétisation des intervalles de temps a été utilisée lorsque la discrétisation des temps de survie était nécessaire (fonction de coût tripletSurv, pré-entraînement contrastif).

11.1.2.3 Dimension de l'espace latent pour les fonctions de coût tripletSurv et SV-tripletSurv

Lors de la prédiction de l'espace latent par la fonction tripletSurv (ou SV-tripletSurv) il est possible de choisir le nombre de caractéristiques à prédire. En effet, la dernière couche correspondra au nombre de caractéristiques prédites. Nous avons choisi de tester avec 4096 et 100 neurones (4096 correspondant au nombre de neurones une fois la sortie du bloc convolutif vectorisé, et 100 un nombre plus faible permettant d'utiliser les caractéristiques choisies dans un modèle de prédiction simple par la suite).

11.1.3 Détails d'optimisation et d'évaluation des modèles

Nous utilisons pour l'optimisation et l'évaluation une validation croisée en 4 "folds" (la séparation des patients reste la même pour tous les modèles et toutes les évaluations).. La séparation des patients se fait de façon à ce que la distribution des temps de survie et de la censure soit équivalente dans chaque "fold". Les modèles sont tous évalués de façon homogène avec une recherche par grille contenant les paramètres suivants :

— Taux d'apprentissage $\in \{1e^{-3}, 1e^{-4}, 1e^{-5}, 1e^{-6}\}$

— Décroissance du taux d'apprentissage $\in \{1e^{-8}, 1e^{-10}, 1e^{-12}\}$

— Taille de l'augmentation des données $\in \{15, 30\}$ par patient

— Taux de dropout $\in \{0.17, 0.5, 0.83\}$

— Taille du batch $\in \{10, 32, 64\}$

— Nombre de couches de convolution à geler $\in \{0, 1, 2, 3\}$ (seulement dans le cas du "finetining" après le pré-entraînement).

Après de nombreuses expériences préliminaires, nous avons décidé de réaliser les entraînements avec 70 epochs. Toutefois, le modèle gardé n'est pas celui obtenu à la dernière epoch mais celui à l'epoch qui permet le meilleur c-index de validation.

L'optimisation utilise l'optimiseur ADAM car selon Kingma et al [132], la méthode est "efficace sur le plan informatique, a peu de besoins en mémoire, est invariante à la remise à l'échelle diagonale des gradients et est bien adaptée aux problèmes qui sont grands en termes de données/paramètres". L'évaluation se fait sur le c-index de la validation

par 'Test-Time Augmention'(TTA) [133]. Lors d'une TTA, les prédictions de toutes les augmentations d'un même patient sont moyennées afin de donner une seule prédiction moyenne par patient. Le c-index est alors calculé sur les prédictions moyennées.

Pour toutes les méthodes nous reportons le c-index avec l'augmentation TTA. Finalement, nous évaluons aussi la séparation de la population en deux groupes avec les courbes de Kaplan Meier [134] et calculons la p-value pour évaluer la séparation.

11.1.4 Détails sur les expériences de comparaison

11.1.4.1 Dénomination des modèles CNN

Lors de nos expériences nous allons comparer les différents modules, stratégies et fonctions de coût à l'aide de différentes configurations de notre meilleur modèle M2P2 (Multiple Myeloma Prognosis Prediction). Nous partons d'un modèle CNN simple 3D présenté dans la figure 11.2 que nous nommons B3D (Baseline 3 Dimensions). B3D est composé de trois blocs de convolutions et du bloc de prédiction (couche dense de taille 100 et couche de sortie dépendante de la fonction de coût). A partir de ce modèle nous définissons le modèle B3D*, une adaptation de B3D à laquelle nous avons ajouté une couche de "pooling" à chaque bloc de convolution et un module d'attention sur les filtres. Enfin, notre meilleur modèle est appelé M2P2. Il est basé sur B3D* avec un pré-entraînement binaire puis par tripletSurv, et la fonction de coût Cox. Les trois modèles sont résumés dans les figures 11.3 et 11.2.

11.1.4.2 La base DTD pour comparer les fonctions de coût

Afin de comparer les fonctions de coût sans l'influence de notre base de données nous avons réalisé une étude à l'aide de la base de données texturale DTD [97]. Cette base comprend 5640 images divisées en 47 catégories. La taille des images est entre (300×300) et (640×640) pixels. Nous allons utiliser cette base pour évaluer nos performances sur une base de données contrôlée. Nous pourrons ainsi vérifier si la difficulté à obtenir des valeurs de c-index supérieures à 0,7 est due principalement à nos images, à l'analyse de survie ou au contexte de petites données à faible résolution, et au nombre limité de patients. Dans ce but, plusieurs adaptations ont été faites sur cette base de données.

Pour commencer, nous avons adapté la base de données à la survie en créant artificiellement des temps de survie et de la censure. En effet, 7 des 47 classes furent gardées. Nous avons rangé ces classes en allant des textures les plus uniformes aux plus aléatoires et nous leur avons attribué un temps de survie artificiel comme indiqué dans la figure 11.4. Concernant la censure, nous avons décidé d'appliquer de la censure aléatoire tout en respectant un taux de censure variable. En effet, il nous paraissait intéressant d'évaluer l'impact du taux de censure sur les différentes fonctions de coût.

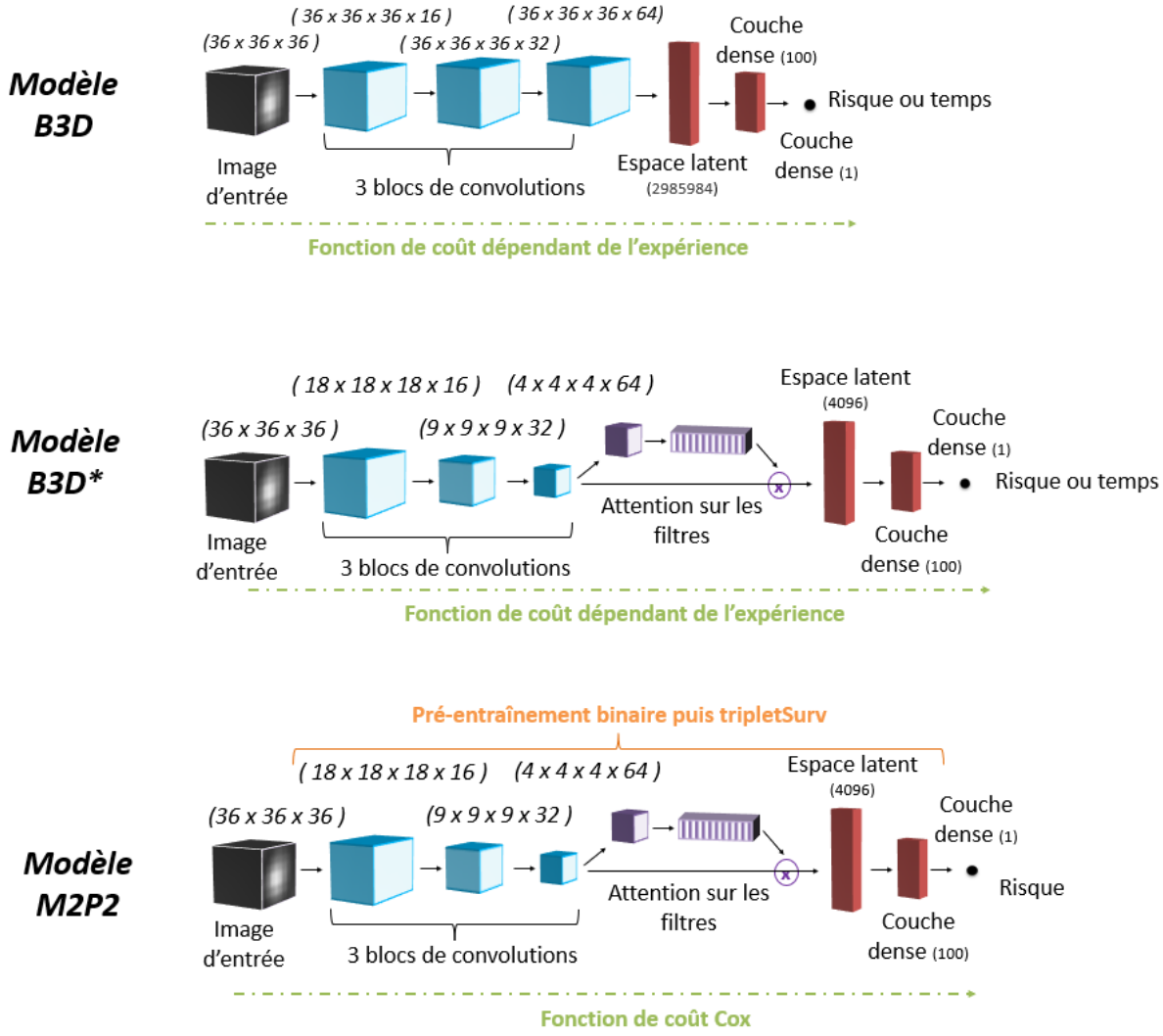


FIGURE 11.2 – Schémas des trois modèles utilisés lors des expériences.

Modèle	CNN de base	"Pooling" au niveau des convolutions	Attention sur les filtres	Pré-entraînement binaire	Pré-entraînement tripletSurv
B3D	X				
B3D*	X	X	X		
M2P2	X	X	X	X	X

FIGURE 11.3 – Tableau résumant les modules de chaque modèle. La fonction de coût de B3D (Baseline 3 Dimensions) et B3D* dépend de l'expérience. La fonction de coût de M2P2 (Multiple Myeloma Prognosis Prediction) est celle de Cox. Une croix dans le tableau signifie que le module est présent. Le CNN de base correspond aux blocs de convolutions + le bloc de prédiction.

FIGURE 11.4 – Exemple d'images présentes dans les 7 classes gardées dans la base de données DTD et leur temps de survie artificiel associé.

De plus, pour avoir une taille régulière et adaptée à notre modèle, nous diminuons la taille de toutes les images à une taille fixe de (288×288). Enfin, pour avoir un nombre de patients équivalent, nous ne prenons que 22 images aléatoires par classes. En gardant 7 classes nous obtenons donc 154 images, soit un nombre équivalent au nombre d'images de notre base de données (152). Par la suite, ces images sont soumises à la même augmentation de données que notre base.

11.1.4.3 Caractéristiques issues d'un CNN en entrée des RSF

Afin de comparer avec une méthode de RSF et ainsi de couplet l'extraction de caractéristiques et le modèle de prédiction, nous extrayons les caractéristiques profondes de nos modèles et nous les utilisons comme entrée d'une analyse en composantes principales (PCA). Suite à la PCA, nous décidons de garder le nombre de variables qui permet d'avoir 80% de l'information. Ces variables gardées prédites par la PCA sont utilisées comme entrée des RSF.

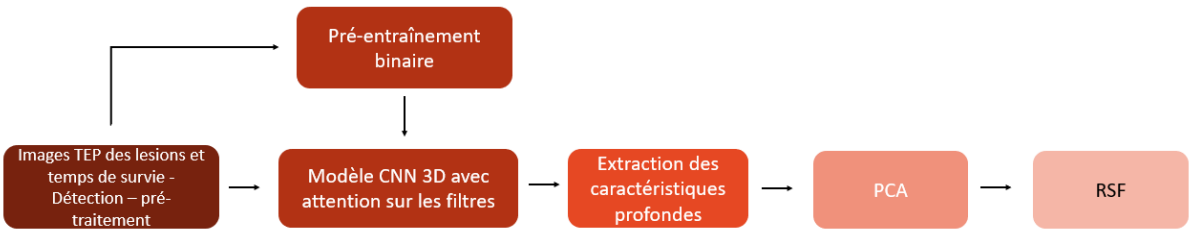


FIGURE 11.5 – Schéma de la méthode de prédiction par RSF avec en entrée le résultat de la PCA appliquée aux caractéristiques profondes extraites de nos modèles.

11.2 Résultats

Après souhaitons évaluer au travers de ces expériences, différents modules et configurations. Après avoir montré l'intérêt de l'apprentissage profond, nous présenterons les expériences ayant permis de définir notre meilleur modèle (dimension des images d'entrée, ajout de la SPP, de l'attention, des méthodes de pré-entraînement). Nous exposerons par la suite l'évaluation des fonctions de coût de survie, puis nous finirons avec les résultats obtenus sur notre meilleur modèle (M2P2).

Nous avons choisi de présenter, dans ce chapitre, le c-index d'entraînement et le c-index de validation pour chaque expérience. En effet, même si nous comparons les méthodes en utilisant le c-index de validation, le c-index d'entraînement peut aider à la détection du sur-apprentissage. Afin d'avoir un ordre d'idées des valeurs de c-index qui existent dans la littérature nous présentons le tableau 11.1 contenant la valeur de c-index trouvée dans les articles les plus pertinents (analyse de survie par CNN avec des images).

TABLE 11.1 – C-index de validation de modèles de la littérature réalisant de la prédiction de survie à l'aide d'images et d'un CNN.

Article	Modèle	Images (taille)	nombre de patients	C-index
Li et al. [115]	CNN 3D + SPP	TEP	84	0,6
Liu et al. [121]	CNN 3D avec attention	IRM ($128 \times 128 \times 128$)	309	0,68
Li et al. [135]	Graph CNN	Images pathologiques (lames entières)	entre 263 et 463	0,62-0,7

11.2.1 Évaluation de l'intérêt du modèle d'apprentissage profond

Afin d'évaluer l'intérêt du modèle d'apprentissage profond par rapport aux modèles traditionnels, nous avons testé les méthodes de Lasso-Cox, RSF (avec les 11 meilleurs caractéristiques en entrée), un modèle 2D simple (modèle B3D* sans attention au format 2D) et un modèle 3D simple (modèle B3D* sans attention). Les résultats sont présentés dans la table 11.2.

On peut voir dans la table 11.2 qu'un CNN 2D simple avec une fonction de coût de Cox permet d'améliorer significativement le c-index par rapport aux modèles de base comme RSF ou Lasso-Cox. Cette amélioration est d'autant plus grande lors de l'utilisation d'un modèle 3D. Outre l'amélioration de la prédiction, les modèles d'apprentissage profond permettent de ne pas être dépendant de l'ensemble des variables mises en entrée du modèle comme ceci est le cas avec les méthodes classiques de Cox et RSF. En effet, même en utilisant des méthodes de sélection, les variables gardées vont dépendre des variables mises en entrée de cette méthode de sélection, notamment en raison de la corrélation qui peut y avoir entre les variables. Il est a noter que, dans la plupart des méthodes par apprentissage profond, il y aura un écart entre résultats d'entraînement et de validation.

TABLE 11.2 – Comparaison avec les méthodes de base. C-index TTA moyen sur une validation croisée à 4 "folds" (± l'écart-type standard).

Modèle	C-index d'entraînement	C-index de validation
Lasso-Cox	0,525 (±0, 053)	0,511(±0, 050)
RSF	0,790(±0, 0133)	0,480(±0, 116)
CNN 2D	0,581 (±0, 0569)	0,573 (±0, 031)
CNN 3D	0,809 (±0, 105)	0,599 (±0, 012)

TABLE 11.3 – Comparaison des dimensions du CNN simple. C-index TTA moyen sur une validation croisée à 4 "folds" (± l'écart-type standard).

Modèle	C-index d'entraînement	C-index de validation	Nombre de paramètres
CNN 2D (1 patch)	0,581 (±0, 0569)	0,573 (±0, 031)	134 091
CNN 2D (3 patchs)	0,545 (±0, 023)	0,504 (±0, 068)	134 091
CNN 2D (9 patchs)	0,566 (±0, 035)	0,542 (±0, 048)	134 091
CNN 2,5D	0,579 (±0, 035)	0,570 (±0, 039)	134 379
CNN 3D	0,809 (±0, 105)	0,599 (±0, 020)	529 647

Malgré cette chute de performances, il y a apprentissage et les réseaux généralisent mieux les données non vues lors de l'entraînement.

11.2.2 Définition de la dimension d'entrée des images

Une fois la capacité prédictive de l'apprentissage profond démontrée, il fut nécessaire de déterminer la dimension des entrées du modèle. Différentes dimensions furent testées afin de déterminer laquelle permet de garder suffisamment d'information pour permettre la prédiction de survie sans pour autant avoir un modèle trop grand (et donc trop de paramètres) qui induirait potentiellement d'avantage de sur-apprentissage. Chaque méthode d'extraction des images, et notamment la définition de dimensions d'entrée (2D, 2.5D et 3D), est présentée dans la sous-section 11.1.1.1. Ici, les expériences ont été réalisées grâce à des modèles simple 2D (modèle B3D* sans attention, au format 2D) et 3D (modèle B3D* sans attention), avec une fonction de coût Cox.

On peut voir dans le tableau 11.3 que le modèle avec des images 3D permet de meilleures prédictions. Bien que plus volumineux en terme de paramètres, une plus grande quantité d'informations est gardée, dont l'information relative à l'organisation spatiale des voxels caractérisant une lésion. Ensuite, viennent les modèles avec des images 2.5D et 2D avec 1 patch, qui sont proches d'un point de vue de la structure et du résultat. Pour rappel, ces deux modèles reçoivent en entrée une image 2D par patient avec un ou trois canaux respectivement. Les résultats les plus faibles sont ceux des modèles 2D avec 3 et 9 patchs traités comme provenant de plusieurs individus. Différentes explications sont possibles. Par exemple, peut-être que seul l'axe axial possède de l'information pertinante. Au vue de ces résultats, nous gardons pour la suite des expériences le modèle 3D.

11.2.3 Évaluation de la SPP et de l'attention

11.2.3.1 Évaluation de la SPP

Nous souhaitons évaluer la méthode SPP utilisée par Li et al. [115] et permettant de régler le problème des lésions de taille variable. Lors de l'ajout d'un module de Spatial Pyramidal Pooling (SPP) nous utilisons la méthode "Box" (lésions dans leur taille d'origine sans ré-échantillonage), car l'information de fond sera supprimée par la suite lors de la SPP. Les résultats de cette évaluation sont présentés dans la table 11.4. L'expérience avec SPP correspond à l'architecture décrite en section 10.1.2.1 utilisant 2 "poolings" de taille différente P_4 et P_2 après les couches convolutionnelles, donnant respectivement des cartes de tailles $(4 \times 4 \times 4)$ et $(2 \times 2 \times 2)$, qui sont ensuite concaténées après avoir été mises sous forme de vecteur. L'expérience avec "pooling" fixe correspond à un "pooling" P_4 unique donnant une carte de taille $(4 \times 4 \times 4)$, sans le mettre sous forme de vecteur par la suite. Le modèle utilisé est le modèle B3D avec une fonction de coût de Cox et un pré-entraînement binaire. Lorsque ni la SPP ni le "pooling" fixe ne sont appliqués, nous ajoutons des couches de "pooling" à chaque bloc de convolution afin de diminuer le nombre de paramètres de la même façon qu'on le ferai avec la SPP ou le "pooling" fixe, mais progressivement (B3D* sans attention).

TABLE 11.4 – Évaluation de l'ajout d'un module de SPP. Modèle B3D avec pré-entraînement binaire et fonction de coût de Cox. Moyenne des c-index TTA sur une validation croisée de 4 "folds" (± l'écart-type standard). Avec SPP : SPP avec 2 "poolings" fixes P_4 et P_2. Avec "pooling" fixe : un "pooling" fixe P_4 unique.

Modèle	C-index d'entraînement	C-index de validation	Nombre de paramètres
B3D* sans attention + pré-entr. binaire	0,834 (±0,176)	0,607 (±0,010)	529 647
B3D + pré-entr. binaire + SPP	0,862 (±0,173)	0,582 (±0,073)	580 847
B3D + pré-entr. binaire + "pooling" fixe	0,807 (±0,189)	0,594 (±0,043)	529 647

On peut voir dans le tableau 11.4 que la meilleur prédiction est obtenue sans la SPP, puis avec le "pooling" fixe. Ce résultat peut être expliqué potentiellement par une perte d'information de l'organisation spatiale des pixels lors de la vectorisation avec la SPP. Ainsi, la SPP pourrait être plus intéressante lorsque les images d'entrée sont de plus grande taille. De plus, les moins bonnes performances de la SPP et du "pooling" fixe peuvent aussi potentiellement s'expliquer par une trop grande perte d'information. En effet, on passe d'une image de taille $(36 \times 36 \times 36)$ à une image $(4 \times 4 \times 4)$ et $(2 \times 2 \times 2)$ par un "pooling" max. Un "pooling" moyen aurait peut être permis de rendre cette perte d'information moins brutale.

11.2.3.2 Évaluation de l'importance de l'attention

Nous évaluons ensuite la pertinence du module d'attention. Les résultats sont présentés dans le tableau 11.5. Tous les tests de cette section sont basés le modèle B3D* avec pré-

TABLE 11.5 – Évaluation de l'attention. Modèle B3D* avec pré-entraînement binaire et fonction de coût de Cox. Moyenne des c-index sur une validation croisée de 4 "folds" (± l'écart-type standard).

Modèle	C-index d'entraînement	C-index de validation	Nb de paramètres
B3D* sans attention + pré-entr bin.	0,834 (±0, 176)	0,607 (±0, 010)	529 647
B3D* (attention sur les filtres) + pré-entr bin.	0,837 (±0, 148)	0,625 (±0, 015)	530 743
B3D* + att. spatiale (spatiale en 2nd)+ pré-entr bin.	0,942 (±0, 017)	0,585 (±0, 020)	531 429
B3D* + att. spatiale (spatiale en 1er)+ pré-entr bin.	0,795 (±0, 213)	0,606 (±0, 012)	531 429

TABLE 11.6 – Évaluation du positionnement du "pooling" fixe par rapport à l'attention sur les filtres. Modèle B3D + pré-entraînement binaire + fonction de coût Cox. Attention puis SPP : SPP après l'attention avec 2 "pooling"s fixes P_4 et P_2 ; "Pooling" fixe puis attention spatiale : SPP avant l'attention avec un "pooling" fixe P_4 unique. C-index TTA moyen sur une validation croisée à 4 "folds" (± l'écart-type standard)

Modèle	c-index d'entraînement	c-index de validation	Nb de paramètres
B3D + pré-entr. bin. + "Pooling" fixe puis att.	0,714 (±0, 180)	0,590 (±0, 040)	530 743
B3D + pré-entr. bin. + att. puis SPP	0,831 (±0, 140)	0,573 (±0, 040)	581 943

entraînement binaire et fonction de coût Cox, auquel nous enlevons ou ajoutons un module d'attention. Les valeurs présentées sont le c-index TTA moyen sur une validation croisée de 4 "folds" avec l'écart-type standard sur ces "folds".

Les résultats montrent qu'il est plus intéressant d'utiliser l'attention sur les filtres seul plutôt que de ne pas utiliser d'attention ou d'utiliser l'attention spatiale plus l'attention sur les filtres. Il est même plus intéressant de n'avoir aucun module d'attention plutôt qu'un module avec de l'attention spatiale bien que la différence soit faible. Il semble donc que l'apprentissage et le choix des filtres (processus semblable a la sélection de variables) ai une plus grande importance que l'attention spatiale. Une autre raison pourrait être que les régions d'intérêt sont suffisantes et qu'une localisation plus fine n'est pas compatible avec la résolution de nos images.

11.2.3.3 Évaluation de la combinaison de l'attention et de la SPP

Enfin, nous avons évalué la combinaison des modules de SPP et d'attention. En effet, cette combinaison pourrait permettre au modèle de forcer l'attention sur les filtres à se concentrer sur l'intérieur des lésions. Nous n'évaluons que l'utilisation de l'attention sur les filtres, étant donné les résultats précédents [voir le tableau 11.5]. Nous testons le modèle où un "pooling" P_4 se situe avant l'attention ("pooling" fixe puis attention) et un modèle où la SPP (P_4 + P_2) se situe après l'attention (attention puis SPP). Une expérience avec la SPP (P_4 + P_2) avant l'attention n'est pas pertinente car elle nécessite la vectorisation pour la concaténation des deux "poolings" mais cela fausse l'entrée de l'attention qui demande une entrée en 3D avec des canaux. Les expériences sont réalisées avec le modèle B3D + pré-entraînement binaire + fonction de coût Cox.

On remarque dans les résultats du tableau 11.6 qu'il est préférable de ne pas utiliser

TABLE 11.7 – Évaluation de l'utilisation du pré-entraînement. Modèle B3D* (avec la fonction de coût Cox). c-index TTA moyen sur une validation croisée à 4 "folds" (± l'écart-type standard). Lorsque cela n'est pas précisé, les résultats présentés sont ceux où il n'y a pas eu de couches gelées.

Modèle	C-index d'entraînement	C-index de validation
B3D*	0,760 (±0, 133)	0,6138 (±0, 0179)
B3D* + pré-entr. binaire	0,837 (±0, 148)	0,625 (±0, 015)
B3D* + pré-entr. contrastif tripletSurv	0,858 (±0, 096)	0,639 (±0, 038)
B3D* + pré-entr. contrastif SV-tripletSurv ($1^{ère}$ couche gelée)	0,685 (±0, 179)	0,617 (±0, 019)
B3D* + pré-entr. binaire puis tripletSurv	0,775 (±0, 179)	0,662 (±0, 025)
B3D* + pré-entr. binaire puis SV-tripletSurv	0,791 (±0, 168)	0,624 (±0, 026)

de "pooling" fixe ou de SPP dans la cas de l'utilisation de l'attention. Ainsi le meilleur résultat est obtenu avec attention sur les filtres et sans "pooling" fixe. Il faut noter que l'expérience de l'attention suivie de la SPP (581 943 paramètres) sera plus demandeuse en coût de calcul que l'expérience du "pooling" fixe puis attention (530 743 paramètres) car l'attention se fera sur des matrices plus grandes (non réduites par du "pooling"). De plus, l'expérience sans SPP est la méthode la plus rapide en terme de temps de calcul. En effet, la SPP est une méthode plus lente, et l'expérience sans SPP a des "poolings" dans les convolutions ce qui permet tout de même de diminuer le nombre de paramètres.

Finalement, le modèle sélectionné (M2P2) ne sera composé que de l'attention sur les filtres, l'attention spatiale et la SPP ne permettant pas d'améliorer les résultats de prédiction.

11.2.4 Évaluation des méthodes de pré-entraînement

Pour améliorer notre c-index de validation et réduire le sur-apprentissage, nous avons conçu différentes techniques de pré-entraînement [voir la section 10.1.3]. Les expériences pour valider ces techniques sont réalisées avec le modèle B3D* (avec la fonction de coût Cox). Les valeurs présentées sont le c-index TTA moyen sur une validation croisée de 4 "folds" avec l'écart-type standard sur ces "folds". Pour chaque méthode de pré-entraînement seul le meilleur résultat parmi les différentes méthodes de "finetuning" est présenté.

Les résultats présentés dans le tableau 11.7 montrent l'intérêt du pré-entraînement. Le pré-entraînement binaire et le pré-entraînement par tripletSurv seuls permettent d'améliorer les résultats. Cependant, on constate que la succession des deux améliore le c-index de validation. Une hypothèse, est que le pré-entraînement binaire aide la prédiction des vecteurs de caractéristiques profondes qui est effectuée avec le pré-entraînement par tripletSurv.

Toutefois, le pré-entraînement par SV-tripletSurv ne semble pas améliorer la prédiction. En effet, lors du calcul de la fonction de coût SV-tripletSurv de nombreux triplets sont considérés comme non valides en raison de la censure [voir la section 10.2.4.2]. Ce

TABLE 11.8 – Évaluation de différentes fonctions de coût de survie. Modèle B3D* avec pré-entraînement binaire. C-index TTA moyen sur une validation croisée à 4 "folds" (± l'écart-type standard)

Fonction de coût	C-index d'entraînement	C-index de validation	Temps/risque
Cox	0,837 (±0, 148)	0,625 (±0, 015)	Risque
Discrète	0,953 (±0, 082)	0,621 (±0, 021)	Temps
Rank&cox	0,785(±0, 111)	0,627 (±0, 011)	Risque
Rank&discret	0,986 (±0, 014)	0,619 (±0, 010)	Temps
Rank&MSE	0,591 (±0, 169)	0,610 (±0, 060)	Temps
TripletSurv&cox (4096)	0,864 (±0, 052)	0,606 (±0, 031)	Risque
SV-tripletSurv&cox (4096)	0,786 (±0, 141)	0,611 (±0, 048)	Risque
SV-tripletSurv&cox (100)	0,784 (±0, 165)	0,601 (±0, 027)	Risque

pré-entraînement pourrait être intéressant si le taux de censure était plus bas ou le nombre de patients plus grand.

11.2.5 Évaluation des différentes fonctions de coût

Nous avons évalué l'état de l'art des fonctions de coût (fonction de coût Cox [12], fonction de coût de survie discrète [13] et fonction de coût Rank&MSE adaptée à la survie (RankDeepSurv) [14]), des modifications de ces fonctions (Rank&cox, Rank&discret) et des adaptations de fonctions de coût qui sont ne sont normalement pas dédiées à la survie (tripletSurv&cox et SV-tripletSurv&cox). Ainsi, nous souhaitons déterminer quelles sont les meilleures fonctions de coût de survie, évaluer l'impact de l'ajout d'une fonction de coût d'ordonnancement et comparer la fonction de coût d'ordonnancement par paire avec la fonction d'ordonnancement par triplet. Les résultats sont présentés dans le tableau 11.8. Les expériences sont basées sur le modèle B3D* avec pré-entraînement binaire. Les valeurs présentées sont le c-index TTA moyen sur une validation croisée de 4 "folds" et l'écart-type standard sur ces "folds". Dans le cas des fonctions tripletSurv et SV-tripletSurv, il est indiqué à côté du nom de la fonction un chiffre entre parenthèses. Ce chiffre correspond au nombre de neurones de la dernière couche lors de l'entraînement. Cela correspond donc au nombre de caractéristiques profondes prédites.

Dans le tableau 11.8, Rank&MSE a un c-index de validation légèrement inférieur aux autres fonctions de la littérature et aux autres fonctions d'ordonnancement par paire. De plus, Rank&cox donne une prédiction légèrement meilleure que les autres. Cependant, aucune des fonctions de coût ne semble surpasser les autres de façon nette, au vu du c-index de validation. Bien que le pré-entraînement par tripletSurv ait amélioré les valeurs de c-index, la combinaison de tripletSurv ou SV-tripletSurv avec la fonction de Cox n'a pas amélioré la prédiction. Nous pouvons aussi constater que les fonctions de survie discrète et Rank&discret semblent présenter un sur-apprentissage plus important. Leur valeur de

c-index de validation pourraient donc potentiellement être encore améliorée.

De plus, comme la fonction de coût Rank&cox semble être légèrement meilleure que les autres, nous l'avons testée avec le meilleur modèle M2P2 mais le c-index de validation (0,6315 $\pm 0,0269$) n'a pas surpassé celui avec la fonction de coût Cox (0,6621 $\pm 0,0246$).

11.2.5.1 Comparaison des matrices de confusion

Les fonctions de coût de survie discrète, Rank&discret et Rank&MSE prédisent des temps. Il est donc possible de produire une matrice de confusion afin de comparer les temps prédits aux temps réelles. Pour cela, nous avons discrétisé les temps prédits en classes (1 an par classe) sans prendre en compte la censure. Ces matrices sont présentés dans la figure 11.6. Les fonction de survie discrète et Rank&discret sont équivalentes en terme

Discret

Temps de survie discrétisé \ validation TTA	prédiction 0	1	2	3	4	5	6
0	0	3	2	9	3	1	0
1	0	5	5	5	9	5	3
2	0	1	4	8	6	7	2
3	2	3	5	5	8	5	5
4	1	2	2	2	3	8	2
5	0	1	2	4	6	2	4
6	0	0	0	0	3	2	0

validation \ prédiction	0	1	2	3	4	5	6
0	18	104	40	74	26	10	133
1	54	108	65	167	45	18	263
2	14	85	41	132	67	4	287
3	81	85	15	123	69	18	359
4	27	30	65	57	19	4	248
5	23	38	32	78	55	0	194
6	1	4	6	14	1	24	70

Entraînement \ prédiction	0	1	2	3	4	5	6
0	1026	4	29	63	38	1	54
1	2	1534	94	105	75	23	327
2	0	0	1239	103	89	9	450
3	4	11	1	1348	126	13	717
4	0	4	2	9	551	9	775
5	5	1	0	0	2	96	1186
6	0	0	0	1	0	0	329

Rank&Discret

Temps de survie discrétisé \ validation TTA	prédiction 0	1	2	3	4	5	6
0	0	2	4	3	7	1	1
1	0	2	8	3	10	7	2
2	0	2	6	5	7	6	2
3	1	1	4	5	12	7	3
4	2	1	4	1	5	4	3
5	0	0	0	7	3	5	4
6	0	0	1	1	1	2	0

validation \ prédiction	0	1	2	3	4	5	6
0	27	102	74	103	31	0	203
1	41	102	94	199	92	6	426
2	23	138	97	130	35	1	416
3	57	83	95	143	65	1	546
4	55	47	82	99	24	3	290
5	11	40	18	94	83	0	324
6	7	13	18	16	11	9	76

Entraînement \ prédiction	0	1	2	3	4	5	6
0	1374	21	52	58	7	1	107
1	1	2127	155	122	66	5	404
2	0	0	1827	138	78	0	477
3	3	9	7	1949	138	3	861
4	0	0	16	28	727	1	1028
5	0	5	1	11	0	108	1585
6	0	0	0	2	0	0	448

Rank&MSE

Temps de survie discrétisé \ validation TTA	prédiction 0	1	2	3	4	5	6
0	0	5	9	4	0	0	0
1	0	8	18	3	3	0	0
2	0	7	17	2	2	0	0
3	0	10	16	4	2	1	0
4	0	7	10	1	1	1	0
5	0	5	10	2	0	2	0
6	1	1	2	1	0	0	0

validation \ prédiction	0	1	2	3	4	5	6
0	0	82	310	65	20	3	0
1	0	120	542	89	72	17	0
2	0	132	490	57	44	12	0
3	0	160	505	71	84	35	0
4	0	121	314	20	36	34	0
5	0	75	317	27	22	54	0
6	30	15	63	15	9	3	0

Entraînement \ prédiction	0	1	2	3	4	5	6
0	30	306	1035	39	0	0	0
1	1	535	1844	84	40	16	0
2	0	396	1500	187	90	32	0
3	0	407	1743	225	191	44	0
4	0	225	912	83	218	137	0
5	0	210	840	20	124	305	1
6	0	60	240	0	2	76	12

FIGURE 11.6 – Matrice de confusion (sommée sur les 4 folds) des fonctions de coût de survie discrète, Rank&discret et Rank&MSE après discrétisation des temps prédits. Les lignes de la figure correspondent aux fonctions de coût et les colonnes aux différents c-index. Les lignes de chaque matrice de confusion correspondent aux classes du temps réel discrétisé et les colonnes au temps prédit discrétisé. Plus la valeur est verte et plus le nombre est grand.

de distribution des prédictions. Les prédictions ont tendances à être plus grandes que la valeur réelle (voir les matrices de confusion de l'entraînement). Cela peut s'expliquer par

le fait que les classes de temps discrétisées ne prennent pas en compte la censure. Ainsi, si on prend l'exemple d'un individu ayant un temps à 2 ans et de la censure, il est possible que le temps auquel aura réellement lieu l'évènement de ce patient soit plus grand que 2 ans. Dans ce contexte, une prédiction d'un temps de 3, 4, 5 ou 6 ans peut être possible.

A contrario, la fonction Rank&MSE semble prédire pratiquement que des temps entre 1 et 3 ans. Ceci est d'autant plus flagrant lorsque l'on observe le détails pour chaque "fold" [voir l'annexe G]. Cela pourrait s'expliquer par le fait que nous avons choisit le modèle ayant le meilleur c-index afin de pouvoir comparer les fonctions entre elles. Or celui ci ne prend pas en compte la valeur absolue mais seulement relative. Ainsi, un résultat donnant un bon ordonnancement mais une mauvaise distribution/valeur des temps prédits peut donner le même c-index qu'avec le même ordonnancement et des valeurs ayant une bonne distribution (même distribution de valeurs entre temps prédit et temps réel). Une solution serait d'utiliser une valeur de mesure prenant en compte à la fois l'ordonnancement et la distance entre temps prédit et temps réel. Pour garder le c-index afin de comparer les fonctions de coût, une autre solution serait de prendre en compte la distribution des prédictions dans la fonction de coût, comme par exemple en adaptant la fonction de Wasserstein [136] à la survie.

11.2.5.2 Comparaison des fonctions de coût de survie grâce à la base de données texturale DTD

Étant donné que nos données ne permettent que difficilement d'évaluer les différentes fonctions de coût, nous avons utilisé la base de données texturale DTD [97] dans la configuration décrite dans la sous-section 11.1.4.2. Avec cette base de données nous pouvons simuler le taux de censure pour évaluer la dépendance des fonctions de coût à la censure. Nous comparons dans le tableau 11.9 les fonctions grâce à un modèle 2D contenant l'attention sur les filtres et un "pooling" fixe unique P_4 avant. Nous évaluons toujours le c-index TTA moyen sur une validation croisée à 4 "folds" ($\pm$ l'écart-type standard).

Afin de simplifier la comparaison entre cette base et notre base, nous avons réhaussé sur le tableau les chiffres correspondant à un taux de censure de 40%, similaire au taux de 45% sur nos données. Nous pouvons observer que le c-index de validation à ce taux de censure varie entre 0.661 et 0.737 (avec notre base de données il varie entre 0.610 et 0.627). De manière générale, les valeur de c-index sont plus grandes avec la base DTD. Il semble donc que pour une taille de base et un taux de censure équivalent notre base soit toujours plus difficile à utiliser pour la prédiction de la survie. Ceci est cohérent car l'inspection visuelle de textures par un humain est possible pour la base de textures DTD mais pas pour les lésions. De plus, d'autres différences peuvent résider dans la taille des images ou la distribution des temps de survie. Un autre point à prendre en compte est la distribution de la censure. En effet, si la distribution de la censure est aléatoire dans la base DTD, elle est répartie différemment dans notre base (le taux de censure est plus

TABLE 11.9 – Comparaison des fonctions de coût grâce à la base de données DTD, en fonction du taux de censure. En rouge : le c-index de validation à 40% de censure.

Fonction de coût	Taux de censure	C-index d'entraînement	C-index de validation
Cox	20	0,843 ($\pm$0,129)	0,700 ($\pm$0,002)
	30	0,779 ($\pm$0,142)	0,712 ($\pm$0,015)
	40	0,823 ($\pm$0,101)	0,691 ($\pm$0,029)
	50	0,847 ($\pm$0,122)	0,698 ($\pm$0,024)
	70	0,789 ($\pm$0,146)	0,702 ($\pm$0,026)
	80	0,816 ($\pm$0,105)	0,692 ($\pm$0,037)
Discrète	20	0,900 ($\pm$0,068)	0,753 ($\pm$0,037)
	30	0,904 ($\pm$0,071)	0,729 ($\pm$0,047)
	40	0,847 ($\pm$0,106)	0,732 ($\pm$0,052)
	50	0,824 ($\pm$0,157)	0,726 ($\pm$0,061)
	70	0,917 ($\pm$0,035)	0,732 ($\pm$0,046)
	80	0,938 ($\pm$0,039)	0,731 ($\pm$0,041)
Rank&cox	20	0,842 ($\pm$0,141)	0,703 ($\pm$0,013)
	30	0,853 ($\pm$0,116)	0,684 ($\pm$0,030)
	40	0,748 ($\pm$0,136)	0,702 ($\pm$0,026)
	50	0,851 ($\pm$0,119)	0,697 ($\pm$0,022)
	70	0,859 ($\pm$0,115)	0,699 ($\pm$0,013)
	80	0,802 ($\pm$0,159)	0,688 ($\pm$0,031)
Rank&discret	20	0,892 ($\pm$0,022)	0,729 ($\pm$0,060)
	30	0,939 ($\pm$0,042)	0,740 ($\pm$0,044)
	40	0,842 ($\pm$0,103)	0,737 ($\pm$0,066)
	50	0,903 ($\pm$0,061)	0,732 ($\pm$0,054)
	70	0,888 ($\pm$0,060)	0,742 ($\pm$0,064)
	80	0,934 ($\pm$0,020)	0,736 ($\pm$0,061)
Rank&MSE	20	0,638 ($\pm$0,030)	0,651 ($\pm$0,059)
	30	0,632 ($\pm$0,024)	0,654 ($\pm$0,056)
	40	0,640 ($\pm$0,014)	0,661 ($\pm$0,066)
	50	0,641 ($\pm$0,012)	0,656 ($\pm$0,059)
	70	0,638 ($\pm$0,022)	0,653 ($\pm$0,057)
	80	0,640 ($\pm$0,033)	0,650 ($\pm$0,058)

haut lorsque les temps de survie sont grands). Pour une meilleur comparaison, il faudrait appliquer une distribution équivalente de la censure dans les deux bases.

Ensuite, en comparant les valeurs à 40% de censure entre les différentes fonctions de coût, nous remarquons que le c-index le plus bas est obtenu avec Rank&MSE ce qui est cohérent avec les résultats du tableau 11.8. Cependant, contrairement aux résultats obtenus avec notre base de données, ici les fonctions basées sur la fonction de Cox donnent des c-index de validation plus bas que ceux basés sur la fonction de survie discrète. Cela pourrait rester cohérent si on prend en compte le fait que ces fonctions sont soumises à du sur-apprentissage avec notre base de données et peuvent dont être améliorées. De plus, cette légère supériorité de la fonction de survie discrète et Rank&discret avec DTD peut aussi s'expliquer par le fait que, bien que les valeurs de temps soient des valeurs numériques, elles restent discrètes (7 classes) et non continues comme avec notre base de données. Il semble donc que la fonction de Cox classique soit plus efficace avec des données continues et la fonction discrète de Cox avec des données discrètes. Nous pouvons aussi noter que la combinaison de la fonction d'ordonnancement avec les fonctions de survie (Rank&cox et Rank&discret) semblent améliorer très légèrement le c-index de validation
.

Contrairement à ce qui serait attendu, nous constatons que toutes les fonctions sont très peu dépendantes du taux de censure. Cependant, au vu des différences de c-index, ce résultat n'est pas extrapolable.

11.2.5.3 Conclusion de l'étude sur les fonctions de coût

Pour conclure, les résultats sur notre base de données montrent une supériorité légère de Cox et Rank&cox. Il faudrait étudier plus en profondeur les fonctions de survie discrète et Rank&discret afin d'évaluer la possibilité de diminuer le sur-apprentissage. De même, la supériorité de la fonction de survie discrète et Rank&discret avec la base DTD, montre qu'il serait intéressant d'approfondir les expériences afin de déterminer si cela est du à la base de donnée contenant des classes ou si ces résultats légèrement plus bas avec notre base sont seulement dus au sur-apprentissage avec notre base.

Nous avons vu également que la fonction Rank&MSE donnait des c-index de validation plus bas que ce soit pour notre base de données ou avec DTD. Le problème de cette fonction de coût semble venir, au vu des matrices de confusion, d'une distribution des prédiction trop éloignée des valeurs de temps de survie réels. Un changement de valeur de α pourrait améliorer cet aspect en donnant plus d'importance à MSE qu'à la fonction d'ordonnancement.

Enfin, les combinaisons avec l'ordonnancement par pair semblent améliorer très légèrement les prédictions mais pas celles avec les triplets. Une technique pour améliorer ces prédictions seraient de déterminer automatiquement le meilleur coefficient entre les deux fonctions combinées. De plus, les fonctions de coût tripletSurv et SV-tripletSurv pour-

raient être intéressantes dans le cas de bases de données plus grandes ou avec moins de censure. Néanmoins, la fonctions de coût tripletSurv reste intéressante dans le contexte de pré-entraînement.

11.2.6 Les résultats obtenus avec notre meilleur modèle M2P2

En réunissant les résultats des expériences précédentes nous avons composé le meilleur modèle M2P2 (Multiple Myeloma Prognosis Prediction) qui contient un CNN 3D, de l'attention sur les filtres, du pré-entraînement binaire et tripletSurv, et une fonction de coût Cox. M2P2 permet de prédire non seulement un risque mais aussi des groupes pronostiques. La séparation en groupes peut être évaluée grâce aux courbes de Kaplan-Meier et à la p-value. Nous présentons ici les résultat du modèle M2P2 qui donna les meilleurs résultats (c-index de 0,66).

La p-value moyenne sur les 4 "folds" est de 3,40E-03($\pm 6,80E-03$) pour le set de validation et de 1,08E-06 ($\pm 2,17E-06$) pour le set d'entraînement. Cette p-value de validation est bien inférieure à 0,05, pour tous les "folds", ce qui implique que la séparation est significative. Les quatre courbes de Kaplan-Meier sont présentées dans la figure 11.7.

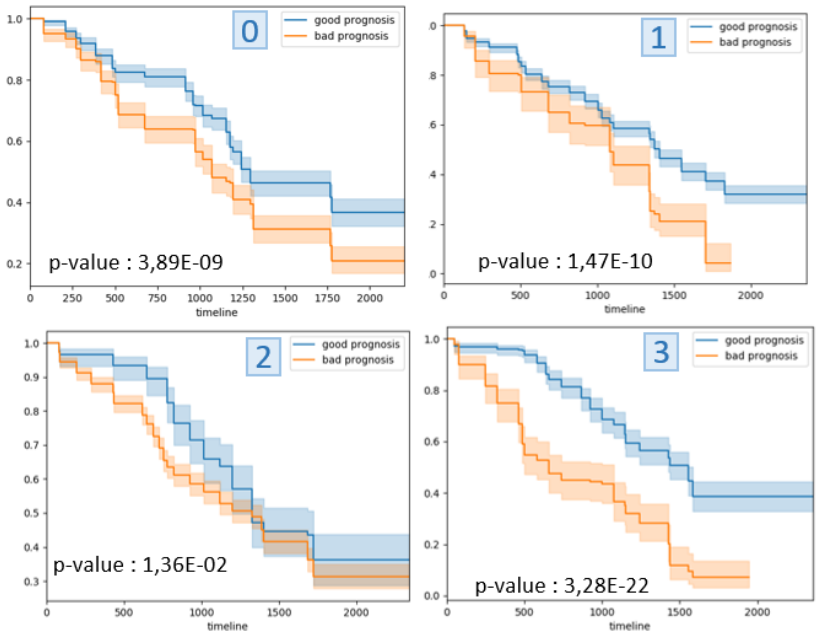


FIGURE 11.7 – Courbes de Kaplan-Meier de validation du modèle M2P2. Une courbe et une p-value par "fold".

Chapitre 12

Discussions et conclusions

12.1 Discussions

Nous discuterons dans cette section d'alternatives à nos évaluations et de points qui auraient le mérite d'être plus approfondis.

12.1.1 Séparation en groupe pronostique avec ou sans augmentation TTA

Lorsque nous présentons les c-index nous les présentons avec une augmentation TTA. Or, la p-value et les courbes de Kaplan-Meier sont elles présentées sans TTA. En effet, la TTA implique le regroupement des prédictions par patient. Ainsi, le nombre d'individus dans les courbes de validation passe de 585 ou 1170 (pour respectivement 15 et 30 images par patient après augmentation) à 39. Le nombre d'individus étant plus faible, l'intervalle de confiance et la p-value deviennent plus grands. Nous présentons dans cette section les résultats de sortie du modèle M2P2 avec la fonction de coût Cox. La p-value de validation moyenne passe de 3,40E-03 (sans TTA) à 1,73E-01 (avec TTA). La figure 12.1 présente chaque courbe et p-value individuellement. En comparant avec la figure 11.7, on devine des courbes similaires mais avec un intervalle de confiance à 95% bien plus grand. Les courbes présentant une grande différence sont celles avec très peu de patients (comme la courbe du groupe "bas risque" du "fold" 2). Nous remarquons avant tout que la courbe du "fold" 1 n'est pas présente. En effet, aucun patient n'étant prédit dans le groupe de mauvais pronostique, il n'est pas possible de réaliser une courbe de Kaplan-Meier. Ainsi, pour obtenir une répartition plus équilibrée des patients dans les groupes de validation, nous améliorons le seuil de séparation. Le seuil de séparation étant basé sur la meilleure séparation des données d'entraînement, nous avons testé d'exclure les patients ayant les prédictions les plus extrêmes (prédiction inférieure ou supérieure à 2 écart-types lorsque

la prédiction est un risque, 400 jours lorsque la prédiction est un temps [1]). Cette action pourrait dans le même temps améliorer la séparation des "folds" non significatifs (0 et 2). Après calcul d'une nouvelle valeur de séparation nous avons les courbes et p-values de la figure 12.2. Nous pouvons observer que la valeur de séparation fut modifiée pour le "fold" 1 qui présente maintenant une courbe de Kaplan-Meier dans chaque groupe. Cependant, cette méthode de mise à l'écart de prédictions "aberrantes" n'a pas amélioré la séparation des "folds" 0 et 2 qui reste non significative (supérieure à 0,05).

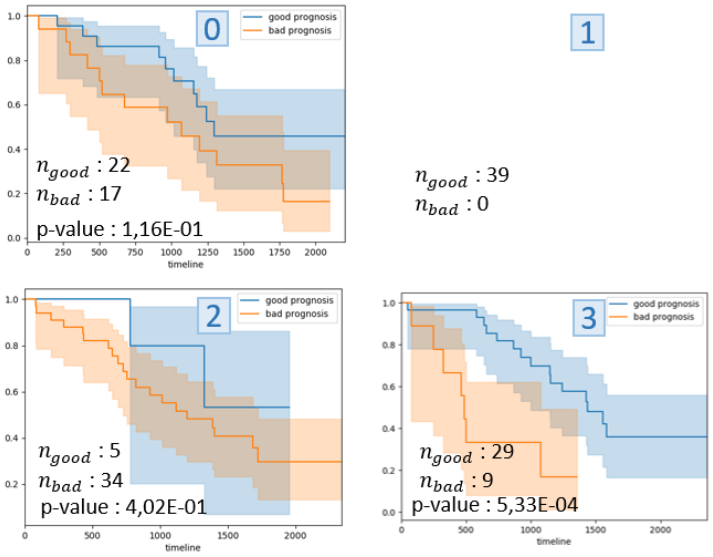


FIGURE 12.1 – Courbes de Kaplan-Meier de validation du modèle M2P2 avec la fonction de coût Cox lorsque la TTA est appliquée (groupement des prédiction par patient). Une courbe et une p-value par "fold". n_{good} et n_{bad} correspondent respectivement au nombre d'individus dans les groupe "bas risque" et "haut risque".

Finalement, nous avons choisi de présenter les courbes de Kaplan-Meier de validation sans TTA afin de pouvoir comparer les différents modèles. En effet, plus de données seraient nécessaires pour resserrer la valeur des intervalles de confiance entre les deux types d'évaluation.

12.1.2 Impact de chaque classe sur la prédiction globale

Lors de la définition de notre modèle, nous avons pu observer que les classes de temps de survie (discrétisation des temps de survie) n'avaient pas toutes le même impact sur le c-index. Pour évaluer quel était l'impact de chaque classe, nous avons réalisé la prédiction en enlevant successivement chaque classe de la base de données avant l'entraînement (sur le modèle B3D avec attention sur les filtres plus SPP, la fonction de coût de Cox et le

1. Le choix de 2 écart-types et 400 jours est fait au regard du nombre de prédictions qui sont supérieurs ou inférieurs à ces valeurs.

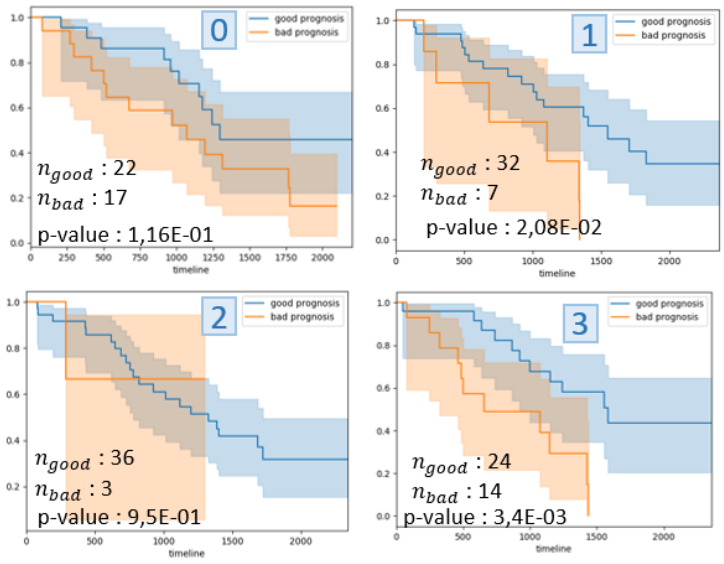


FIGURE 12.2 – Courbes de Kaplan-Meier de validation du modèle M2P2 avec la fonction de coût Cox lorsque la TTA est appliquée (groupement des prédiction par patient) et que les prédictions extrêmes sont enlevées des prédiction de l'entraînement. Une courbe et une p-value par "fold". n_{good} et n_{bad} correspondent respectivement au nombre d'individus dans les groupe "bas risque" et "haut risque".

pré-entraînement binaire). Nous avons remarqué que lorsque l'on enlève la classe 0 (temps inférieurs ou égaux à 365 jours) on passe d'un c-index de 0,57 (avec toutes les données) à un c-index de 0,64. Cette grande augmentation du c-index implique que le fait de garder les patients ayant le temps de survie le plus court, diminue la qualité de la prédiction. Cela peut éventuellement s'expliquer par le fait que la classe 0 puisse être un groupe moins homogène que les autres classes. Nous avons aussi évalué le c-index en supprimant les autres classes mais aucune suppression de classes n'augmente autant le c-index (va de 0,56 pour la classe 6 à 0,60 pour la classe 5). Seule la suppression de la classe 6 (temps supérieurs à 2190 jours) diminue la valeur de c-index.

De plus, ayant remarqué que notre modèle donnait de meilleurs résultats sans la classe 0 (temps inférieurs ou égaux à 365 jours), nous avons séparé en deux classes les prédictions [2] des patients ayant un temps de survie inférieur à 365 jours (classe A) et les autres (classe B). Le c-index moyen de la classe A est de 0,27 ($\pm 0,18$) et celui de la classe B est de 0,59 ($\pm 0,05$). Cela implique que l'ordonnancement dans la classe A est très compliqué comparé à celui dans la classe B. Cela pourrait expliquer un meilleur c-index en écartant les patients avec un temps de survie inférieur à 365 jours. Cependant, ce résultat peut être simplement dû au fait que le c-index d'une classe sera plus bas (car l'ordonnancement plus difficile quand les données sont proches) que celui d'une classe avec des prédictions

2. Prédiction réalisée avec le modèle M2P2 avec fonction de coût Cox.

TABLE 12.1 – Valeur de c-index à l'intérieur de chaque classe (discrétisation des temps). La classe 6 ne contient pas assez de patients pour permettre un c-index significatif.

Classe	0	1	2	3	4	5	6
C-index	0,36 (±0, 26)	0,66 (±0, 12)	0,60 (±0, 06)	0,59 (±0, 05)	0,61 (±0, 02)	0,63 (±0, 04)	0,0

TABLE 12.2 – Prédiction par RSF, précédé par une PCA, avec en entrée les caractéristiques profondes extraites de nos modèles. C-index TTA moyen sur 4 "folds". Le nombre de variables gardées correspond au nombre de variables à garder pour avoir 80% de l'information avec la PCA. Le modèle CNN utilisé est le 3D avec attention sur les filtres et pré-entraînement binaire.

Fonction de coût	C-index d'entraînement	C-index de validation	nb de var. gardées
Cox	0,808 (±0, 049)	0,530 (±0, 061)	91 (±21)
Discrète	0,740 (±0, 060)	0,533 (±0, 044)	130 (±59)
Rank&cox	0,813 (±0, 042)	0,463 (±0, 029)	95(±49)
Rank&discret	0,838 (±0, 062)	0,490 (±0, 038)	119(±44)
Rank&MSE	0,847 (±0, 025)	0,514 (±0, 070)	218(±50)
TripletSurv	0,877 (±0, 036)	0,485 (±0, 072)	89(±18)
TripletSurv&cox	0,827 (±0, 018)	0,466 (±0, 060)	79(±15)

plus éloignées. De ce fait, nous avons évalué le c-index à l'intérieur de chaque classe séparément, afin d'évaluer si l'ordonnancement à l'intérieur de la classe 0 est plus difficile que dans les autres classes mais aussi pour évaluer si les erreurs dans le c-index étaient dues à l'ordonnancement entre classes ou à l'ordonnancement à l'intérieur d'une classe (le deuxième étant moins important que la première). Les résultats sont présentés dans le tableau 12.1. On peut voir une différence significative entre le c-index à l'intérieur de la classe 0 et celui à l'intérieur des autres classes. Cette dernière expérience confirme que les prédictions des patients ayant un temps de survie inférieur à 365 jours sont plus difficiles et ont un impact négatif sur le c-index global.

12.1.3 Qualité de l'espace latent appris

Afin d'évaluer l'effet de fonctions de coût de survie sur l'espace latent, nous avons entraîné le modèle B3D* avec pré-entraînement binaire et plusieurs fonctions de coût. Nous avons ensuite extrait les caractéristiques profondes de chaque modèle et nous les avons utilisées comme entrée d'une analyse en composantes principales (PCA). Finalement, la sortie de cette PCA (variables contenant 80% de l'information) est utilisée en entrée du modèle RSF. Les c-index moyens sur 4 "folds" sont présentés dans le tableau 12.2.

Nous remarquons dans le tableau 12.2 que les résultats ne sont pas en accord avec ce que nous avons vu avec l'expérience décrite dans la sous-section 11.2.5. Ce sont Cox et de survie discrète qui ici donnent les meilleures prédictions. Cependant au vue des valeurs de c-index, ce résultat n'est pas très pertinent ($< 0, 54$). La méthode PCA + RSF ne donne pas une bonne prédiction avec nos caractéristiques profondes. Une approche

sans PCA permettrait d'avantage d'interprétabilité, cependant l'utilisation de la PCA reste nécessaire car la quantité de variables (4096) est trop grand pour permettre de l'utiliser dans l'étape de sélection des variables par VIMP. Enfin, une dernière solution pour remplacer la PCA serait de prédire directement 100 caractéristiques profondes au lieu de 4096 (comme réalisé avec SV-tripletSurv dans la sous-section 11.2.5). Cette évaluation des caractéristiques profondes demande donc à être approfondie.

12.2 Conclusions

Après avoir démontré l'intérêt de l'utilisation de l'apprentissage automatique, nous avons déterminé plusieurs méthodes permettant de résoudre les problématiques liées à nos données :

— Afin de résoudre la difficulté liée au nombre limité de patients, nous avons proposé une architecture adaptée en taille avec une normalisation par instance. Nous avons aussi intégré plusieurs types de régularisation (dropout, L1 et L2). Finalement, notre approche originale à été la conception de deux techniques de pré-entraînements successifs (par classification binaire de patchs et par apprentissage contrastif d'un espace latent reflétant la survie avec une fonction de coût tripletSurv). Cette dernière étape permet d'augmenter les valeurs de c-index de 7.7%. Une méthode de pré-entraînement par apprentissage contrastif avec une fonction de coût SV-tripletSurv a été testée mais n'a pas été retenue.

— Afin de résoudre la difficulté liée à la taille faible des images, nous avons diminué la taille des convolutions par rapport au modèle de Li et al. [115]. Nous avons également testé l'attention spatiale et la SPP mais ceux ci n'ont pas permis d'améliorer le c-index.

— Afin de résoudre la difficulté liée à la variabilité de taille des lésions, nous avons retenue une méthode de pré-traîtement dite "interpolate" qui consiste à couper l'image autour de la ROI, puis d'interpoler à une taille fixe ($36 \times 36 \times 36$) par une interpolation spline bicubique. Les méthodes de SPP et d'attention spatiale furent appliquées en première instance (en parallèle de la méthode de pré-traitement "Box" qui consiste à couper l'image d'une taille fixe de ($36 \times 36 \times 36$) autour de la lésion) mais donnèrent de moins bons résultats.

— Enfin, l'interprétabilité fut améliorée par l'attention CBAM (sur les filtres et spatiale) mais seule l'attention sur les filtres améliora la prédiction.

A l'aide de nombreuses expériences, nous avons déterminé le modèle le plus prédictif, qui s'adapte le mieux à nos données. Ce modèle est composé d'un bloc de convolutions 3D qui permet de prédire les caractéristiques profondes, de l'attention sur les filtres qui permet d'améliorer l'interprétabilité, et de deux pré-entraînements successifs (par classification

binaire de patchs et l'apprentissage contrastif d'un espace latent reflétant la survie avec une fonction de coût tripletSurv). Cette dernière partie est une contribution originale de nos travaux.

Afin d'adapter les méthodes d'apprentissage automatique à la survie, nous avons proposé une revue et une comparaison des fonctions de coût de survie trouvées dans la littérature (Cox, de survie discrète et RankDeepSurv). Nous avons de plus, combiné la partie ordonnancement de RankDeepSurv avec les fonctions de Cox et de survie discrète. Enfin nous avons testé la combinaison de Cox avec un ordonnancement par tripletSurv et par SV-tripletSurv que nous avons adaptés à la censure. Bien que la différence entre les c-index de chaque fonction ne soit pas évidente, les fonction de Cox et Rank&cox se détachent légèrement. Parmi les deux fonctions, celle donnant les meilleurs résultats de c-index avec notre meilleur modèle reste Cox. Les fonctions par triplet se sont avérées plus pertinentes dans l'étape de pré-entraînement qu'en tant que fonction de coût pour notre modèle. Toutes ces méthodes sont résumées dans le tableau 12.3 en fonction de la problématique qu'elles traitent.

Perspectives

Bien que nous ayons un modèle complet, il reste de nombreuses pistes à explorer. En effet, la SPP pourrait par exemple devenir plus efficace en remplaçant le "pooling" max avec un "pooling" moyen afin de perdre moins d'information.
Une exploration plus en profondeur des filtres prédis par notre modèle M2P2 pourrait être associée à la matrice d'attention sur les filtres afin de déterminer quels sont les filtres les plus importants. Nous pourrions aussi comparer ces filtres avec les radiomiques calculées manuellement afin d'en extraire de potentielles similitudes.
Un autre point intéressant serait d'utiliser les radiomiques calculées à la main, soit afin de réaliser un pré-entraînement (prédiction des radiomiques), soit pour une fusion avec les caractéristiques profondes. Ces deux techniques furent investiguées lors de la thèse mais ne donnèrent pas de résultats satisfaisants.

De plus, les fonctions de coût demanderaient une investigation plus approfondie afin de déterminer leurs avantages et inconvénients en fonction du contexte des données. Les constantes α et λ des fonctions de coût RankDeepSurv, Rank&cox, Rank&discret, tripletSurv&cox et SV-tripletSurv&cox sont pour le moment déterminées à la main afin d'équilibrer l'ordre de grandeur des fonctions combinées mais la recherche automatique des meilleurs α et λ afin de savoir à quelle fonction donner le plus d'importance, pourrait améliorer grandement les prédictions. Enfin, d'autres fonctions pourraient être adaptées et testées telle que la fonction de Wasserstein [136] qui calcule le coût minimal de transport de la masse d'une probabilité de distribution à une autre.

Finalement, l'évaluation avec un set de test, idéalement provenant d'une base de données extérieure, serait plus robuste qu'une validation croisée Entraînement-Validation

TABLE 12.3 – Liste des méthodes gardées dans le modèle (meilleures méthodes) ou testées (autres méthodes) en fonction de la problématique. * : Contribue au traitement de la censure mais de façon implicite, en favorisant l'apprentissage de caractéristiques discriminantes à partir de données censurées.

Problématique	Censure	Petite base de données	Petites images
Meilleures méthodes	Cox	Nombre et taille faibles des couches denses	Taille faibles des couches convolutives
		Dropout, norm. par instance	
		Régularisation L1 et L2	
		Augmentation des données	
		Pré-entraînement. classif. binaire de patches	
		Pré-entraînement par tripletSurv*	
Autres méthodes	Rank&cox	Pré-entraînement par SV-tripletSurv*	Attention spatiale
	Discret		SPP
	Rank&MSE		
	Rank&discret		
	TripletSurv&Cox		
	SV-TripletSurv&Cox		
Problématique	**Images de tailles variables**	**Interprétabilité**	
Meilleures méthodes	Méthode de pré-traitement "Interpolate"	Attention sur les filtres	
Autres méthodes	méthode de pré-traitement "Box"	Attention spatiale	
	SPP		
	Attention spatiale		

seule. En effet, pour le moment le nombre de patients est trop faible pour séparer notre base en Entraînement-Validation-Test car l'apprentissage profond demande un grand nombre de patients. D'autant plus dans le cas de la survie où la valeur de c-index dépendra du nombre de patients inclus, et différentes fonctions de coût (et notamment SV-tripletSurv) demanderaient un nombre encore plus important de patients.

Malgré les limites et les pistes encore à explorer, nous avons démontré pour la première fois la faisabilité de l'adaptation de méthodes d'apprentissage profond à l'analyse de survie de patients atteints de myélome multiple à partir de données de deux bases multicentriques prospectives. La qualité de prédiction surpasse largement le c-index atteint par les approches classiques et de référence.

Quatrième partie

Conclusions et perspectives

Le but principal de la thèse était de prédire la survie des patients atteints de myélome multiple grâce à des images TEP et des données cliniques. Cette objectif devait être atteint en prenant en compte certaines contraintes : des bases de données relativement petites, un taux de censure d'environ 45%, modèle qui doit être à relativement faible coût de calcul et faible complexité, l'interprétabilité, et la faible résolution des images TEP de lésions de petite taille.

Nous avons souhaité atteindre cet objectif en proposant deux modèles différents : l'un basé sur la nouvelle méthode de référence, les RSF, qui est adapté à nos données et à notre contexte, l'autre, basé sur l'apprentissage profond, un type de méthodes très prometteuses dans bien des domaines mais peu présent en analyse de survie et peu adapté pour le moment à notre contexte. Le premier modèle serait utilisable directement lorsque le deuxième serait un plus grand défi mais permettrait une plus grande avancée dans l'état de l'art.

Première partie de thèse : Les RSF

La première partie de la thèse concerna donc l'analyse de survie grâce à des méthodes d'apprentissage automatique par RSF. Afin de prédire la valeur de la PFS pour un nouveau patient, nous construisons un modèle unifié pour :

— traiter le grand nombre de caractéristiques cliniques et d'images (de l'ordre d'une centaine),

— déduire les caractéristiques les plus pertinentes pour la prédiction,

— prédire la progression d'un patient en fonction de ses données personnelles (caractéristiques cliniques et d'imagerie).

Les contributions apportées dans ces travaux sont les suivantes :

— Nous proposons un modèle unifié pour la sélection de variables et l'analyse de survie quand ils sont normalement utilisés séparément.

— Nos travaux montrent l'intérêt d'utiliser les radiomiques pour la prédiction de la PFS.

— Nous déterminons des biomarqueurs qui sont relativement cohérents avec la littérature.

— Nous sommes les premiers à investiguer l'utilisation des RSF dans le contexte du myélome multiple et des images TEP.

Finalement, les RSF semblent être la méthode de prédiction de survie la plus performante comparée aux méthodes classiques. Elles sont, de plus, un bon compromis entre performance et coût (de calcul, de données etc.) comparées aux méthodes d'apprentissage profond qui, elles, demandent beaucoup de données, de temps, et de mémoire, ce qui n'est

pas toujours compatible avec la réalisation d'études cliniques. Néanmoins, ces dernières ont permis de grandes avancées dans de nombreux domaines tels que la segmentation et la classification. Il nous paraissait donc important de réaliser des travaux afin de déterminer un modèle qui s'adapterait à nos contraintes (petite base de données, petites images, censure etc.) tout en améliorant les performances.

Seconde partie de thèse : l'apprentissage profond

Nous avons donc travaillé sur la conception de différents aspects des modèle d'apprentissage profond afin d'obtenir le modèle qui s'adaptera le mieux à notre contexte. Nous avons commencé par créer un modèle de base, sur lequel nous avons testé différentes configurations de données d'entrée, modules d'attention et de SPP, pré-entraînements, fonctions de coût et pré-traitements des données. Le tableau 12.3 résume les méthodes que nous avons testées pour répondre à chaque problématique. Les méthodes principales, qui ont été gardées, sont l'attention sur les filtres, le pré-entraînement par classification binaire de patches et celui par apprentissage contrastif avec tripletSurv.
Les contributions apportées dans ces travaux sont les suivantes :

— En ce qui concerne l'application médicale, ce travail étudie pour la première fois les approches d'apprentissage profond dans le contexte de l'analyse de la survie des patients atteints de MM à partir d'images TEP. Pour étayer la signification de nos résultats, nous nous appuyons sur les données de deux études cliniques prospectives recueillies sur sept ans.

— Nous présentons une étude des fonctions de perte adaptées à l'analyse de survie avec censure. L'étude comprend deux approches d'apprentissage contrastif, jamais utilisées auparavant pour l'analyse de survie. Une étude expérimentale montre leur faisabilité et met en évidence les avantages potentiels de leur combinaison.

— Pour faire face à la taille limitée des ensembles de données prospectifs, nous proposons deux stratégies de pré-entraînement efficaces qui évitent le besoin d'annotations supplémentaires tout en améliorant la convergence et les performances.

— Enfin, nous montrons l'intérêt d'inclure un module d'attention sur les canaux pour identifier des filtres CNN prédictifs pour la survie. Ceci peut être vu comme une première étape pour la détermination de biomarqueurs.

Les contributions principales

Pour conclure, nous sommes partis d'une base de données et de l'idée d'utiliser la survie. Puis nous avons :

1. Déterminé une méthode d'extraction et de labellisation des données (images et cliniques) afin de pouvoir les incorporer à des méthodes d'apprentissage automatique.

2. Inclus une deuxième étude sur le myélome multiple, que nous avons homogénéisé avec la première.
3. Produit un algorithme d'apprentissage machine basé sur les RSF et permettant la prédiction d'un risque de survie, d'un groupe pronostique et de biomarqueurs.
4. Appliqué ce modèle pour l'analyse des études cliniques du myélome multiple et ces biomarqeurs.
5. Produit un modèle d'apprentissage profond M2P2 adapté à la survie et à de petites lésions, avec un nombre limité de patients. Ce modèle permet la prédiction d'un risque, d'un groupe de survie et d'une matrice d'importance des filtres.
6. Étudié la plupart des fonctions de coût de survie de la littérature et apporté de nouvelles (tripletSurv, SV-tripletSurv, Rank&cox, Rank&discret).
7. Présenté des méthodes permettant de gérer les petites images de taille variable (SPP, attention spatiale, méthodes de pré-traitement).
8. Proposé des méthodes de pré-entraînement avec nos propres données afin de pallier du nombre de patients limité et du manque de bases de données publiques adaptées.
9. Publié ces travaux dans deux journaux et deux conférences, et un troisième article de journal est en cours [voir la liste des publications se trouve dans le tableau 12.4].
10. Proposé de nombreuses pistes d'amélioration pour le modèle RSF mais surtout pour le modèle d'apprentissage profond M2P2.

Perspectives

Bien que les deux modèles améliorent l'erreur de prédiction par rapport aux méthodes de la littérature, ils peuvent encore être améliorés. De nombreuses pistes ont été exposées dans les chapitres 7.2 et 12.2. Ainsi, le modèle RSF pourrait être perfectionné, notamment en se concentrant sur la corrélation des variables, et la distribution de celles-ci ce qui renforcerait la prédiction de biomarqueurs. Cependant, le travail le plus grand et qui peut avoir le plus d'impact sur l'état de l'art reste l'amélioration du modèle d'apprentissage profond. En effet, la marge de progression en terme de c-index reste grande et beaucoup de possibilités s'offrent encore à nous. Outre, l'utilisation des radiomiques pour le pré-entraînement par exemple, ou l'interprétation des filtres plus en profondeur comme vu dans la section 12.2, il nous reste encore des données qui n'ont pas été utilisées dans ce contexte. En effet, les données cliniques (autres que les images) possèdent une grande quantité d'informations qui ont été montrées prédictives, et les images TDM peuvent donner de nouvelles informations. L'intégration des images TDM et des données cliniques au modèle d'apprentissage profond pourrait faire l'objet d'une thèse entière tant les possibilités sont grandes. Un autre point important qu'il reste à voir, est l'application de ces modèles à de nouvelles bases de données telle que celle du Lymphome. Cette base

Table 12.4 – Liste des publications.

Type	Date	Citation	Partie
Journal	En cours	**L. Morvan**, C. Nanni, A.-V. Michaud, et al. Multiple myeloma prognosis from PET images : deep survival losses and contrastive pretraining. *Soumis*	III
Journal	2020	B. Jamet, **L. Morvan**, C. Nanni, et al. Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including FDG-PET radiomics : a combined analysis of two independent prospective European trials. *European Journal of Nuclear Medicine and Molecular Imaging*	II
Conférence	2020	**L. Morvan**, C. Nanni, A.-V. Michaud, et al. Learned Deep Radiomics for Survival Analysis with Attention. *Predictive Intelligence in Medicine. PRIME 2020. Lecture Notes in Computer Science*, vol 12329. Springer	III
Journal	2020	**L. Morvan**, T. Carlier, B. Jamet, et al. Leveraging RSF and PET images for prognosis of multiple myeloma at diagnosis. *International Journal of Computer Assisted Radiology* 15, 129-139	II
Conférence	2019	**L. Morvan**, D. Mateus, C. Bailly, et al. Prédiction de la progression chez des patients atteints de myélome multiple avec des Random Survival Forest, *Médecine Nucléaire*, V. 43, Issue 2	II

pourrait être non seulement utilisée comme base externe de validation mais aussi en tant que base de pré-entraînement.

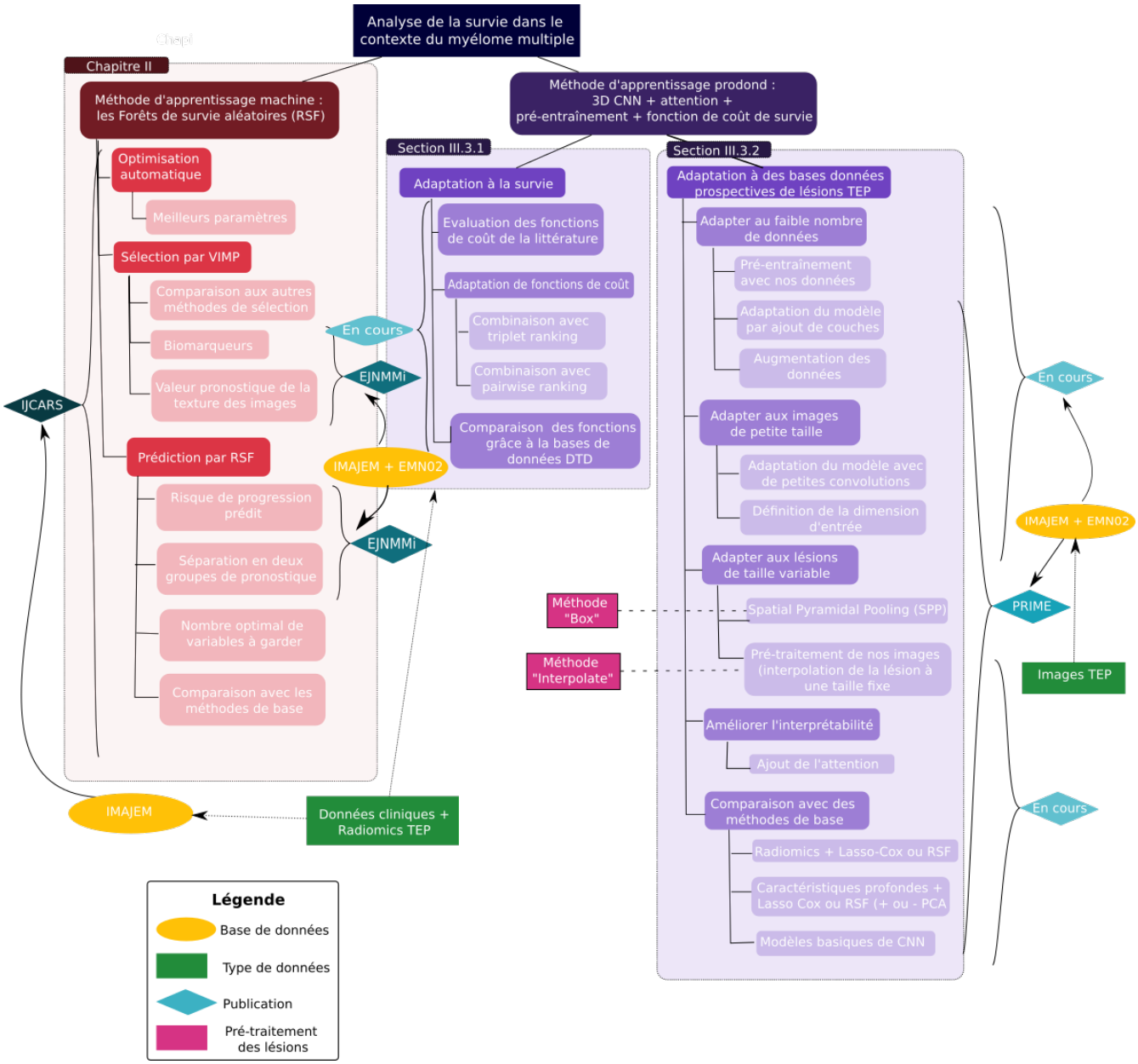


FIGURE 12.3 – Résumé technique des travaux.

Annexes

Annexe A

Les méthodes manuelles de segmentation utilisées

Les segmentation 40% et 2.5

La segmentation à 40% revient à considérer dans la ROI autour de la lésion tous les voxels qui ont une valeur de SUV supérieur à 40% du SUV max. Soit $\mathbf{p}$ un voxel dans la ROI, $I(\mathbf{p})$ son intensité en niveau de gris, et $I_{\max}$ la valeur d'intensité la plus large dans la ROI : $I_{\max} = \max_{\mathbf{p}} I(\mathbf{p})$. Le masque de segmentation à 40%, $M_{40}(\mathbf{p})$, est une image binaire telle que

$$\mathbb{M}_{40}(\mathbf{p}) = \begin{cases} 1 & \text{si } I(\mathbf{p}) \geq 0,4\dot{I}_{\max} \\ 0 & \text{sinon.} \end{cases} \tag{A.1}$$

La segmentation à 2.5 retient dans la lésion tous les voxels qui ont une valeur de SUV supérieur à 2,5 SUV.

$$\mathbb{M}_{2.5}(\mathbf{p}) = \begin{cases} 1 & \text{si } I(\mathbf{p}) > 2,5 \\ 0 & \text{sinon.} \end{cases} \tag{A.2}$$

La segmentation par k-means

La méthode des k-means est un outil de classification non supervisée qui permet de répartir un ensemble de données en k classes homogènes, dans notre cas K classes d'intensité similaire, tel que ROI $= \{\text{Classe}_1, ..., \text{Classe}_k\}$ L'algorithme des k-means vise à minimiser la variance intra-classe, qui se traduit par la minimisation de l'énergie suivante :

$$E = \frac{1}{2} \sum_{k \in K} \sum_{\mathbf{p} \in \text{Classe}_k} ||I(\mathbf{p}) - \bar{I_k}||_2 \tag{A.3}$$

avec $\bar{I_k}$ l'intensité moyenne du cluster Classe_k

La minimisation de cette énergie peut se réaliser par une optimisation itérative est traduite par les étapes suivantes :

1. Initialisation des noyaux.
2. Mise à jour de l'appartenance de chaque voxel à un cluster.

3. Réévaluation des noyaux.
4. Itération sur les étapes 2 et 3 jusqu'à stabilisation des noyaux.

On applique cet algorithme afin d'obtenir une segmentation en deux classes (lésion/ fond). A la sortie nous avons :

$$\mathbb{M}_{\text{kmeans}}(\mathbf{p}) = \begin{cases} 1 & \text{si } \mathbf{p} \in \text{Classe}_{\text{lésion}} \\ 0 & \text{sinon.} \end{cases} \tag{A.4}$$

Le vote majoritaire

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 trois masques. Si un pixel a une valeur supérieure à 0,67, alors il appartiendra à la lésion. Formellement,

$$\mathbb{M}(\mathbf{p}) = \begin{cases} 1 & \text{si } \frac{\mathbb{M}_{40}(\mathbf{p})+\mathbb{M}_{2.5}(\mathbf{p})+\mathbb{M}_{\text{kmeans}}(\mathbf{p})}{3} \geq 0,67 \\ 0 & \text{sinon.} \end{cases} \tag{A.5}$$

Exemple de calculs de radiomiques

Les caractéristiques de premier ordre

Voici la définition de quelques exemples de caractéristiques du premier ordre que l'on retrouve régulièrement.
Soit :

- X_v un ensemble de N_v voxels inclus dans la région d'intérêt (ROI),
- X_g l'ensemble des N_g intensités discrétisées (niveaux d'intensité) des N_v voxels de la ROI,
- $\mathcal{H} = \{n_1, ..., n_{N_g}\}$ l'histogramme des fréquences d'apparition n_i de chaque intensité discrétisée i dans X_g,
- $\mathcal{P}_i = \frac{n_i}{N_v}$ la probabilité d'occurrence de chaque intensité discrétisée i.

L'énergie

$$\text{É}nergie = \sum_{i=1}^{N_v} (X_{v,i})^2 \tag{B.1}$$

L'énergie traduit la magnitude des valeurs de voxels de l'image.

L'entropie

$$Entropie = -\sum_{i=1}^{N_g} \mathcal{P}_i \times \log_2(\mathcal{P}_i) \tag{B.2}$$

L'entropie (ici celle de Shannon) traduit le "désordre" dans dans l'image.

Asymétrie

$$Asymétrie = \frac{\frac{1}{N_v}\sum_{i=1}^{N_v}(X_{v,i} - \bar{X}_v)^3}{(\sqrt{(\frac{1}{N_v}\sum_{i=1}^{N_v}(X_{v,i} - \bar{X}_v)^2})^{3/2}} \tag{B.3}$$

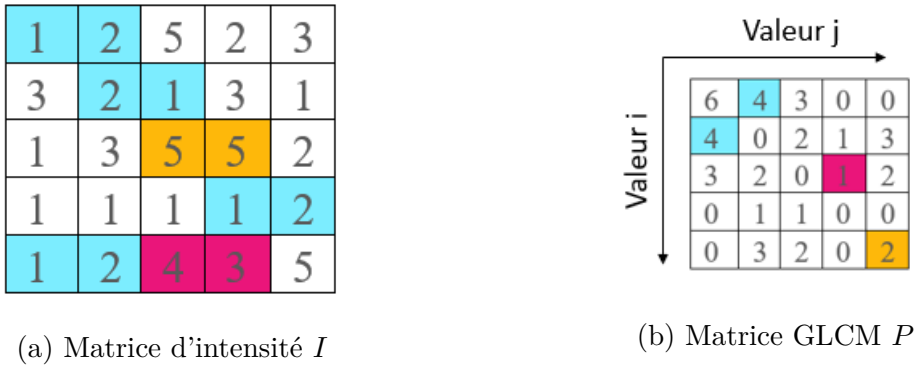
L'asymétrie mesure l'asymétrie de la distribution des valeurs autour de la valeur moyenne. $\bar{X}_v$.

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 GLCM de taille $N_g \times N_g$ GLCM traduit la probabilité jointe $P(i, j|\sigma, \Phi)$ 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 σ pixels et d'un angle Φ. Un exemple est présenté dans la figure B.1.



(a) Matrice d'intensité I

(b) Matrice GLCM P

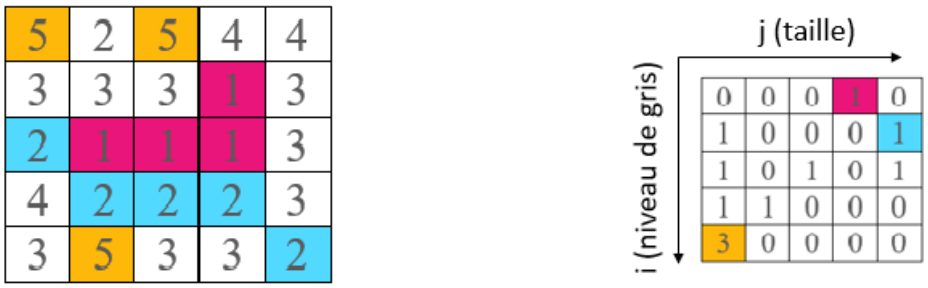
FIGURE B.1 – Exemple illustrant la construction de la matrice GLCM. Soit une matrice d'intensité I. Pour une distance de 1 et un angle de 0° (horizontal plan, de gauche à droite) la matrice de co-occurrence P sera celle présentée dans la figure (b). En effet, si on prend l'exemple du couple $(i, j) = (1, 2)$, la valeur de $P(1, 2|1, 0)$ sera 4, car de façon horizontale le couple (1,2) apparaît quatre fois dans la matrice I (en bleu dans la matrice I). Le couple (4,1) n'apparaît pas, donc la valeur de $P(4, 1|1, 0)$ sera de 0.

Sur cette matrice P seront ensuite faits différents calculs. Les caractéristiques calculées sur cette matrice sont l'entropie [voir l'équation B.2], la corrélation, le contraste, l'énergie [voir équation B.1] 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.Un exemple est présenté dans la figure B.2.

Un exemple de caractéristique calculée sur cette matrice est la Small Zone High Grey



(a) Matrice intensité I

(b) Matrice GLSZM P

FIGURE B.2 – Exemple illustrant la construction de la matrice GLSZM. Soit une matrice I à 5 niveaux de gris. Il y a 1 zone de taille 4 à valeur 1 (en rose) donc $P(4,1|1,0) = 1$.

level Emphasis (SZHGE ou SAHGLE) :

$$SZHGE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \frac{i^2 P(i,j)}{j_2}}{N_z} \tag{B.4}$$

Avec N_s le nombre de tailles de zones dans l'image et N_z 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.

Annexe C

La validation expérimentale des radiomiques par IBSI

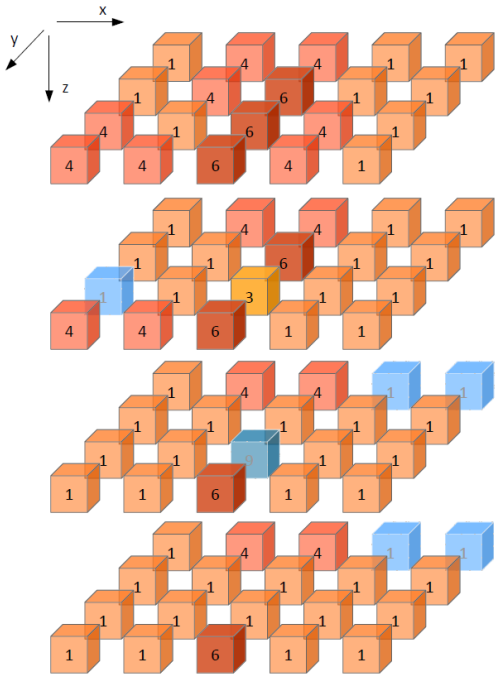


Figure C.1 – Matrice fantôme utilisée pour la vérification des caractéristiques [3].

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

FIGURE C.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 [3].

Annexe D

Paramètres de l'augmentation de données

Quatre opérations ont été prises en compte pour l'augmentation des données

- Translation dans les 3 directions ρ_{trans}
- Zoom avec le facteur ρ_{scale}
- Rotation avec un angle de ρ_{rotation} degrés dans l'axe frontale
- Retournement horizontal (ρ_{flipH}) et/ou vertical (ρ_{flipV})

Nous avons commencé par créer une matrice M_{da} de 30 combinaisons de paramètres de façon aléatoire (30 étant le nombre maximum d'images que nous souhaitons par patient). Une fois cette matrice créée, nous l'utilisons pour tous les modèles et tous les patients afin d'avoir des résultats reproductibles et de pouvoir réaliser des comparaisons entre les modèles. Lorsque 15 images sont gardées par patients, nous ne réalisons que les 15 premières augmentations de la matrice. La matrice créée et utilisées est présente dans le tableau D.1.

TABLE D.1 – Matrice M_{da} contenant les 30 combinaisons de paramètres d'augmentation des données obtenues de façon aléatoire et utilisées pour tous les modèles et tous les patients. ρ_{trans} : translation dans les 3 directions, ρ_{scale} : zoom avec le facteur , ρ_{rotation} : rotation avec un angle de degrés dans l'axe frontale, (ρ_{flipH}, ρ_{flipV}) : respectivement retournement horizontal et vertical (1 = vrai et 0 = faux).

No. de l'image	1	2	3	4	5	6	7	8	9	10
ρ_{trans}	5	4	4	0	3	4	4	5	2	1
ρ_{scale}	1,4	1,2	1,3	1,2	1	1	1,1	1	1	1,1
ρ_{rotation}	-10	-20	30	30	-20	10	0	10	30	10
(ρ_{flipH}, ρ_{flipV})	(0,0)	(1,1)	(1,0)	(0,0)	(1,0)	(1,1)	(1,0)	(1,1)	(1,1)	(1,1)
No. de l'image	11	12	13	14	15	16	17	18	19	20
ρ_{trans}	1	3	5	4	2	3	0	1	4	4
ρ_{scale}	1,3	0,9	0,9	1	1	1,2	1,3	1,2	0,8	1,4
ρ_{rotation}	20	20	30	0	-20	-10	0	30	30	0
(ρ_{flipH},ρ_{flipV})	(0,1)	(0,1)	(1,1)	(1,0)	(0,1)	(1,0)	(1,0)	(1,1)	(1,1)	(0,0)
No. de l'image	21	22	23	24	25	26	27	28	29	30
ρ_{trans}	4	2	5	2	1	2	0	5	5	0
ρ_{scale}	1,4	1,1	1	1	1,4	1,2	1	1,2	1	0,8
ρ_{rotation}	30	10	0	10	-20	0	20	30	-10	20
(ρ_{flipH},ρ_{flipV})	(0,1)	(0,1)	(1,0)	(1,0)	(0,1)	(1,0)	(0,0)	(1,0)	(0,0)	(1,0)

Annexe E

Poids attribués aux fonctions de coût combinées

Lorsque nous combinons des fonctions de coût, nous attribuons à chaque fonction un poids. Si nous considérons les fonctions combinées telles que $l_{\text{combinée}} = \alpha l_{\text{ordonnancement}} + \lambda l_{\text{survie}}$, avec $l_{\text{ordonnancement}} \in \{l_{\text{rank}}, l_{\text{tripletSurv}}, l_{\text{SV-tripletSurv}}\}$ et $l_{\text{survie}} \in \{l_{\text{cox-nn}}, l_{\text{dsurv}}, l_{\text{reg}}\}$ alors nous obtenons le tableau E.1.

Table E.1 – Valeurs de α et λ lors de la combinaison de deux fonctions de coût.

Fonction	α	λ
Rank&MSE	3	1E-05
Rank&cox	1	2
Rank&discret	1	2E-05
TripletSurv&cox	1	1E03
SV-tripletSurv&cox	1	1

Annexe F

Les sorties du module d'attention

Les modules d'attention ont pour but de fournir des matrices de poids qui seront appliquées au modèle afin de concentrer le modèle sur certains pixels ou certains filtres. L'attention sur les filtres fournira une matrice de poids pour chaque filtre/canal, à chaque image d'entrée. Plus le poids est élevé, plus le pouvoir prédictif du filtre est important. Nous avons récupéré la matrice d'attention d'un "fold" obtenue avec le modèle M2P2. Sur la figure F.1, sont présentés en blanc les poids les plus importants. Les colonnes représentent les 64 filtres et les lignes une partie des individus du set d'entraînement. On peut observer que les valeurs de poids sont relativement constantes sur les images. Pour évaluer l'importance d'un filtre indépendamment des images d'entrée, nous allons faire la somme (ou la moyenne) des valeurs de poids sur les images et ainsi obtenir un poids par filtre. Ce vecteur obtenu après la somme des poids de la matrice d'attention sur les patients est présenté dans la figure F.2. Ainsi, les filtres avec un poids blanc sont les filtres les plus importants et les filtres avec un poids noir les moins importants. L'importance est très tranchée pour la majorité des filtres.

Figure F.1 – Sortie de l'attention sur les filtres du "fold" 0 du modèle M2P2. En ligne : images du set de validation, en colonnes : le filtre. De noir à blanc : du plus petit au plus grand poids.

FIGURE F.2 – Somme sur les patients de la matrice d'attention sur les filtres du "fold" 0 du modèle M2P2. En ligne : images du set de validation, en colonnes : le filtre. De noir à blanc : du plus petit au plus grand poids.

Annexe G

Détails de la matrice de confusion de Rank&MSE

Lorsque l'on observe les détails des matrices de confusion pour chaque fold des prédictions faites par Rank&MSE, nous pouvons voir que le modèle prédit principalement une seule classe pour chaque fold (exepté le fold où k = 2).

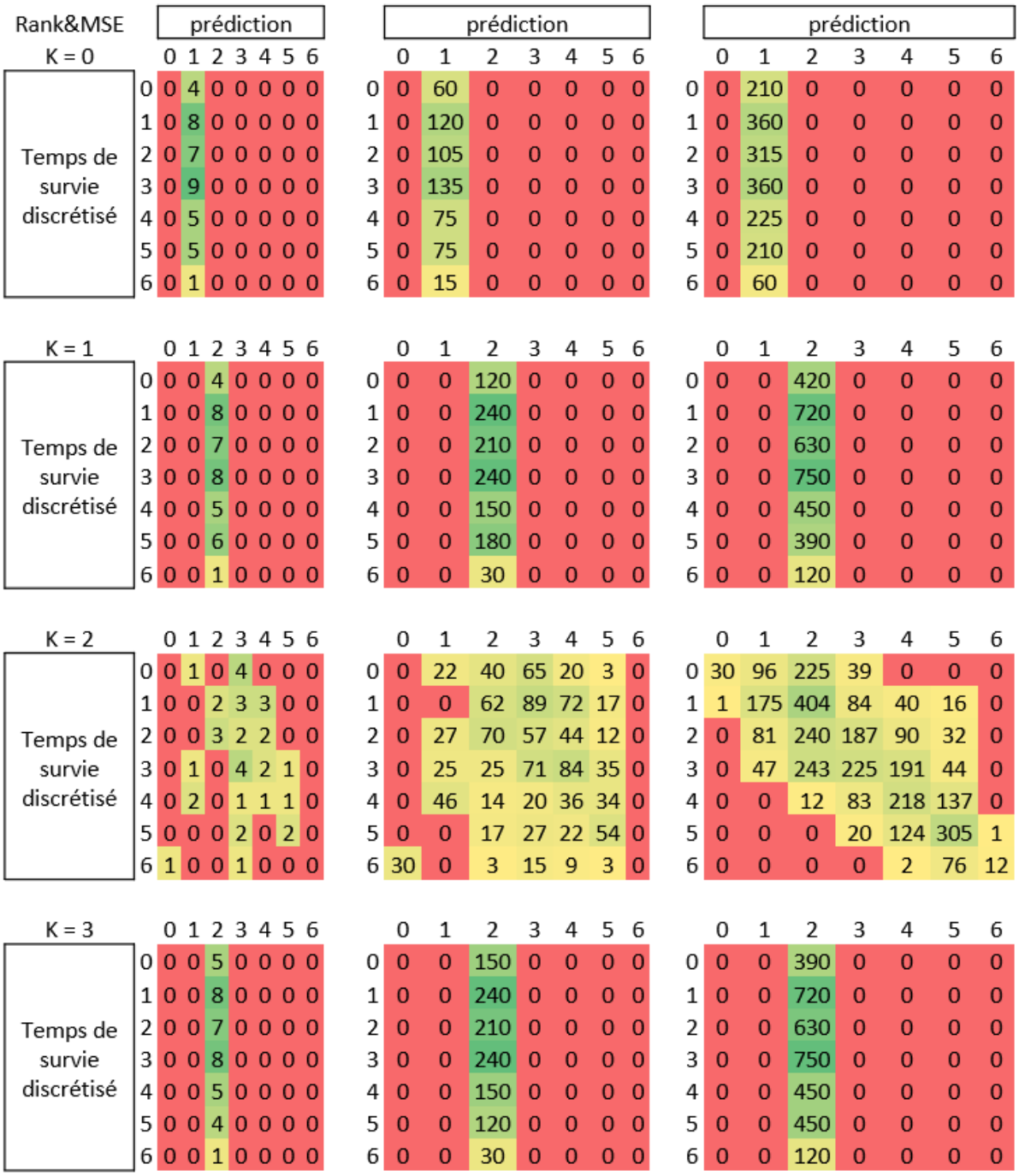


FIGURE G.1 – Détails de la matrice de confusion de Rank&MSE pour chaque fold.

Bibliographie

[1] Moreau, P., Attal, M., Caillot, D., Macro, M., Karlin, L., Garderet, L., Facon, T., Benboubker, L., Escoffre-Barbe, M., Stoppa, A., Laribi, K., Hulin, C., Perrot, A., Marit, G., Eveillard, J., Caillon, F., Bodet-Milin, C., Pégourié, B., Dorvaux, V., Chaleteix, C., Anderson, K., Richardson, P., Munshi, N., Avet-Loiseau, H., Gaultier, A., Nguyen, J., Dupas, B., Frampas, É., and Kraeber-Bodéré, F. (2017) 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.

[2] Zamagni, E., Patriarca, F., Nanni, C., Zannetti, B., Englaro, E., Pezzi, A., Tacchetti, P., Buttignol, S., Perrone, G., Brioli, A., Pantani, L., Terragna, C., Carobolante, F., Baccarani, M., Fanin, R., Fanti, S., and Cavo, M. (2011) 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.

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

[4] Oostvogels, R., Uniken Venema, S. M., de Witte, M., Raymakers, R., Kuball, J., Kröger, N., and Minnema, M. C. (2017) In search of the optimal platform for Post-Allogeneic SCT immunotherapy in relapsed multiple myeloma : a systematic review. *Bone Marrow Transplantation,* **52**(9), 1233–1240.

[5] Ishwaran, H., Kogalur, U., Blackstone, E., and Lauer, M. (2008) Random Surival Forest. *The annals of applied statistics,* **2**(3), 841–860.

[6] Morvan, L., Carlier, T., Bailly C.and Jamet, B., Bodet-Milin, C., Moreau, P., Touzeau, C., Kraeber-Bodere, F., and Mateus, D. (2020) Leveraging RSF and PET images for prognosis of multiple myeloma at diagnosis. *International Journal of Computer Assisted Radiology and Surgery (IJCARS),* p. 129–139.

[7] Jamet, B., Morvan, L., Nanni, C., Michaud, A.-V., Bailly, C., Chauvie, S., Moreau, P., Touzeau, C., Zamagni, E., Bodet-Milin, C., Kraeber-Bodéré, F., Mateus, D., and Carlier, T. (2020) Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including FDG-PET radiomics : a combined analysis of two independent prospective European trials. *European Journal of Nuclear Medicine and Molecular Imaging,* **48**, 1005 – 1015.

[8] Vyshnav, M., Sowmya, V., Gopalakrishnan, E., Variyar VV., S., Menon, V. K., and Soman, K. (2020) Deep Learning Based Approach for Multiple Myeloma Detection.

In *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* pp. 1–7.

[9] Zhou, Y., Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., Piraud, M., Buck, A., Shi, K., and Menze, B. H. (2018) Automated Whole-Body Bone Lesion Detection for Multiple Myeloma on 68Ga-Pentixafor PET/CT Imaging Using Deep Learning Methods. In *Contrast Media Molecular Imaging* pp. 1555–4309.

[10] Lu, C., Wang, X., Prasanna, P., Corredor, G., Sedor, G., Bera, K., Velcheti, V., and Madabhushi, A. (2018) Feature Driven Local Cell Graph (FeDeG) : Predicting Overall Survival in Early Stage Lung Cancer. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., (eds.), *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 407–416.

[11] Baek, S., He, Y., Allen, B. G., Buatti, J. M., Smith, B. J., Tong, L., Sun, Z., Wu, J., Diehn, M., Loo, B. W., Plichta, K. A., Seyedin, S. N., Gannon, M., Cabel, K. R., Kim, Y., and Wu, X. (2019) Deep segmentation networks predict survival of non-small cell lung cancer. *Scientific Reports,* **9**(1).

[12] Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018) DeepSurv : personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology,* **18**.

[13] Gensheimer, M. F. and Narasimhan, B. A scalable discrete-time survival model for neural networks. *PeerJ,* **2019**(1), 1–17.

[14] Jing, B., Zhang, T., Wang, Z., Jin, Y., Liu, K., Qiu, W., and Li, C. (2019) A deep survival analysis method based on ranking. *Artificial Intelligence In Medicine,* **98**(June), 1–9.

[15] Morvan, L., Nanni, C., Michaud, A.-V., Jamet, B., Bailly, C., Bodet-Milin, C., Chauvie, S., Touzeau, C., Moreau, P., Zamagni, E., Kraeber-Bodéré, F., Carlier, T., and Mateus, D. (2020) Learned Deep Radiomics for Survival Analysis with Attention. pp. 35–45.

[16] Albagoush, S. and Azevedo, A. (2021) Multiple Myeloma. *StatPearls Publishing,*.

[17] Bailly, C., Leforestier, R., Jamet, B., Carlier, T., Bourgeois, M., Guérard, F., Touzeau, C., Moreau, P., Chérel, M., Kraeber-Bodéré, F., and Bodet-Milin, C. (2017) PET Imaging for Initial Staging and Therapy Assessment in Multiple Myeloma Patients. *International Journal of Molecular Sciences,* **18**(2).

[18] Granier, D. P. Le myélome multiple. *MN-net,*.

[19] Kumar, S. K., Rajkumar, S. V., Dispenzieri, A., Lacy, M. Q., Hayman, S. R., Buadi, F. K., Zeldenrust, S. R., Dingli, D., Russell, S. J., Lust, J. A., Greipp, P. R., Kyle, R. A., and Gertz, M. A. (2008) Improved survival in multiple myeloma and the impact of novel therapies. *Blood,* **111**(5), 2516–2520.

[20] Colletaz, G. MODÈLES DE SURVIE. *Notes de Cours MASTER 2 ESA voies professionnelle et recherche, université d'Orleans,*.

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

[22] Breslow, N. and McCann, B. (1971) Statistical Estimation of Prognosis for Children with Neuroblastoma. *Cancer Research,* **31**(12), 2098–2103.

[23] Harrington, D. P. and Fleming, T. R. (1982) A Class of Rank Test Procedures for Censored Survival Data. *Biometrika,* **69**(3), 553–566.

[24] Parmar, C. A., Grossmann, P., Bussink, J., Lambin, P., and Aerts, H. J. W. L. (2015) Machine Learning methods for Quantitative Radiomic Biomarkers. In *Scientific reports*.

[25] Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016) Radiomics : Images Are More than Pictures, They Are Data. In *Radiology* Vol. 278, pp. 563–577.

[26] Breiman, L. (2001) Random Forests. *Machine Learning,* **45 1**, 5–32.

[27] Robin Genuer, J.-M. P. Arbres CART et Forêts aléatoires, Importance et sélection de variables. *HAL,*.

[28] Breiman, L. (1996) Bagging predictors. *Machine Learning,* **42 2**, 123–140.

[29] Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004) Bagging survival trees.. *Statistics in medicine,* **23 1**, 77–91.

[30] Bouallègue, F. B., Tabaa, Y. A., Kafrouni, M., Cartron, G., Vauchot, F., and Mariano-Goulart, D. (2017) 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**, 4608–4619.

[31] Choi, E., Schuetz, A., Stewart, W., and Sun, J. (2017) Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association : JAMIA,* **24**, 361 – 370.

[32] Vallières, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J. W. L., Khaouam, N., Nguyen-Tan, P. F., Wang, C.-S., Sultanem, K., Seuntjens, J., and El Naqa, I. (2017) Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports,* **7**(10117), 2911–2918.

[33] Cox, D. R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B,* **34**(2), 187–220.

[34] Raykar, V. C., Steck, H., Krishnapuram, B., Dehing-Oberije, C., and Lambin, P. (2007) On Ranking in Survival Analysis : Bounds on the Concordance Index. p. 1209–1216.

[35] Zhou, Y. and Mcardle, J. (2015) Rationale and Applications of Survival Tree and Survival Ensemble Methods.. *Psychometrika,* **80**, 811–33.

[36] Hothorn, T., Hornik, K., and Zeileis, A. (2006) Unbiased Recursive Partitioning : A Conditional Inference Framework. *Journal of Computational and Graphical Statistics,* **15**(3), 651–674.

[37] Vallières, M., Freeman, C. R., Skamene, S. R., and El Naqa, I. (2015) 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.

[38] Nasejje, J. B. and Mwambi, H. (2017) 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.

[39] Bühnemann, C., Li, S., Yu, H., White, H. B., Schäfer, K., Llombart-Bosch, A., Machado, I., Picci, P., Hogendoorn, P., Athanasou, N., Noble, J., and Hassa, A. (2014) Quantification of the Heterogeneity of Prognostic Cellular Biomarkers in Ewing Sarcoma Using Automated Image and Random Survival Forest Analysis. *Plos one,*.

[40] Miao, F., Cai, Y.-P., Zhang, Y.-X., Fan, X.-M., and Li, Y. Predictive Modeling of Hospital Mortality for Patients With Heart Failure by Using an Improved Random Survival Forest. *IEEE Access,*.

[41] Dey, A. K., Suhas, N., Sai Teja, T., and Juneja, A. (2018) Some variations on Ensembled Random Survival Forest with application to Cancer Research. *arXiV :1709.05515,*.

[42] Liao, L. and Ahn, H.-I. (2016) Combining deep learning and survival analysis for asset health management. *International Journal of Prognostics and Health Management,* **7**(Special Is).

[43] Tibshirani, R. (1997) The lasso method for variable selection in the cox model. *Statistics in Medicine,*.

[44] Wenzheng, S., Jiang, M., Dang, J., Chang, P., and Yin, F.-F. (2018) Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. *Radiation Oncology,* **13**(1), 197.

[45] Hatt, M., Tixier, F., Visvikis, D., and le Rest, C. C. (2017) Radiomics in PET/CT : More Than Meets the Eye ?. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine,* **58 3**, 365–366.

[46] Bourgier, C., Colinge, J., Aillères, N., Fenoglietto, P., Brengues, M., Pèlegrin, A., and Azria, D. (2015) Définition et applications cliniques des radiomics. *Cancer/Radiothérapie,* **19**(6), 532–537 26e Congrès national de la Société française de radiothérapie oncologique SFRO).

[47] Kumar, D., Chung, A., Shafiee, M., Khalvati, F., Haider, M., and Wong, A. (2017) Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction.

[48] Lartizien, C., Rogez, M., Niaf, E., and Ricard, F. Computer-Aided Staging of Lymphoma Patients With FDG PET/CT Imaging Based on Textural Information. *IEEE Journal of Biomedical and Health Informatics,* **18**(3), 946–955.

[49] Bi, L., Kim, J., Kumar, A., Wen, L., Feng, D. D., and Fulham, M. J. (2017) 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.

[50] Tixier, F., Le Rest, C. L., Hatt, M., Albarghach, N., Pradier, O., Metges, J., Corcos, L., and Visvikis, D. (2011) Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. *The Journal of Nuclear Medicine,* **52**, 369–378.

[51] Steiger, S., Arvanitakis, M., Sick, B., Weder, W., Hillinger, S., and A. Burger, I. (2017) 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.

[52] Aerts, H., Velazquez, E. R., Leijenaar, R., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R., and Lambin, P. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Vol. 5, .

[53] Desseroit, M., Tixier, F., Weber, W., Siegel, B., Rest, C., Visvikis, D., and Hatt, M. (2017) Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non–Small Cell Lung Cancer Tumors : A Repeatability Analysis in a Prospective Multicenter Cohort. *The Journal of Nuclear Medicine,* **58**, 406 – 411.

[54] Hatt, M., Majdoub, M., Vallières, M., Tixier, F., Le Rest, C. L., Groheux, D., Hindié, E., Martineau, A., Pradier, O., Hustinx, R., Perdrisot, R., Guillevin, R., El Naqa, I., and Visvikis, D. (2015) 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. *The Journal of Nuclear Medicine,* **56**, 38 – 44.

[55] Bailly, C., Bodet-Milin, C., Couespel, S., Necib, H., Kraeber-Bodéré, F., Ansquer, C., and Carlier, T. (2016) Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. *PLoS ONE,* **11**.

[56] Larue, R., Defraene, G., De Ruysscher, D. K. M., Lambin, P., and van Elmpt, W. J. C. (2017) Quantitative radiomics studies for tissue characterization : a review of technology and methodological procedures. *The British journal of radiology,*.

[57] Decaux, O., Lodé, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jézéquel, P., Attal, M., Harousseau, J., Moreau, P., Bataille, R., Campion, L., Avet-Loiseau, H., and Minvielle, S. (2008) 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**, 4798–805.

[58] Amin, S. B., Minvielle, S., Hanlon, B., Shah, P. K., Li, C., Li, Y., Swanson, D., Moreau, P., Magrangeas, F., Anderson, K. C., Avet-Loiseau, H., and Munshi, N. C. (Nov., 2014) Gene Expression Profile Alone Is Inadequate In Predicting Complete Response In Multiple Myeloma. In *Leukemia* Vol. 28, pp. 2229–2234.

[59] Lapa, C., Lückerath, K., Malzahn, U., Samnick, S., Einsele, H., Buck, A., Herrmann, K., and Knop, S. (2014) 18FDG-PET/CT for prognostic stratification of patients with multiple myeloma relapse after stem cell transplantation. *Oncotarget,* **5**, 7381 – 7391.

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

[61] Horger, M., Claussen, C., Bross-Bach, U., Vonthein, R., Trabold, T., Heuschmid, M., and Pfannenberg, C. 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.

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

[63] Dimopoulos, M. A., Hillengass, J., Usmani, S., Zamagni, E., Lentzsch, S., Davies, F. E., Raje, N., Sezer, O., Zweegman, S., Shah, J., Badros, A., Shimizu, K., Moreau, P., Chim, C.-S., Lahuerta, J. J., Hou, J., Jurczyszyn, A., Goldschmidt, H., Sonneveld, P., Palumbo, A., Ludwig, H., Cavo, M., Barlogie, B., Anderson, K., Roodman, G. D., Rajkumar, S. V., Durie, B. G., and Terpos, E. (2015) Role of Magnetic Resonance Imaging in the Management of Patients With Multiple Myeloma : A Consensus Statement. *Journal of Clinical Oncology,* **33**(6), 657–664.

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

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

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

[67] Cavo, M., Terpos, E., Nanni, C., Moreau, P., Lentzsch, S., Zweegman, S., Hillengass, J., Engelhardt, M., Usmani, S., Vesole, D., San-Miguel, J., Kumar, S. K., Richardson, P., Mikhael, J., da Costa, F. D., Dimopoulos, M., Zingaretti, C., Abildgaard, N., Goldschmidt, H., Orlowski, R., Chng, W., Einsele, H., Lonial, S., Barlogie, B., Anderson, K., Rajkumar, S. V., Durie, B., and Zamagni, E. (2017) Role of 18F-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.

[68] Bodet-Milin, C., Eugène, T., Bailly, C., Lacombe, M., Frampas, E., Dupas, B., Moreau, P., and Kraeber-Bodéré, F. (2013) FDG-PET in the evaluation of myeloma in 2012.. *Diagnostic and interventional imaging,* **94 2**, 184–9.

[69] Lapa, C., Schreder, M., Schirbel, A., Samnick, S., Kortüm, K. M., Herrmann, K., Kropf, S., Einsele, H., Buck, A., Wester, H., Knop, S., and Lückerath, K. (2017) [68Ga]Pentixafor-PET/CT for imaging of chemokine receptor CXCR4 expression in multiple myeloma - Comparison to [18F]FDG and laboratory values. *Theranostics,* **7**, 205 – 212.

[70] Carlier, T., Bailly, C., Leforestier, R., Touzeau, C., Moreau, P., Kraeber Bodéré, F., and Bodet Milin, C. (2017) Valeur pronostique des paramètres de texture TEP au diagnostic dans le myélome multiple symptomatique (MM). *Médecine Nucléaire,* **41**(3), 143.

[71] Bartel, T., Haessler, J., Brown, T., Shaughnessy, J., van Rhee, F., Anaissie, E., Alpe, T., Angtuaco, E., Walker, R., Epstein, J., Crowley, J., and Barlogie, B. (2009) F18-fluorodeoxyglucose positron emission tomography in the context of other imaging techniques and prognostic factors in multiple myeloma.. *Blood,* **114 10**, 2068–76.

[72] McDonald, J., Kessler, M., Gardner, M., Buros, A., Ntambi, J., Waheed, S., van Rhee, F., Zangari, M., Heuck, C., Petty, N., Schinke, C., Thanendrarajan, S., Mitchell, A., Hoering, A., Barlogie, B., Morgan, G., and Davies, F. (2016) Assessment of Total Lesion Glycolysis by 18F FDG PET/CT Significantly Improves Prognostic Value of GEP and ISS in Myeloma.. *Clinical Cancer Research,*.

[73] Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010) High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association,* **105**(489), 205–217.

[74] Stein, C., Qu, P., Epstein, J., Buros, A., Rosenthal, A., Crowley, J., Morgan, G., and Barlogie, B. (2015) Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics,* **16**.

[75] Fonseca, R., Leif Bergsagel, P., Drach, J., Shaughnessy, J. D., Gutiérrez, N. C., Stewart, A. K., Morgan, G. J., Van Ness, B., Chesi, M., Minvielle, S., Neri, A., Barlogie, B., Kuehl, W. M., Liebisch, P., Davies, F. E., Chen-Kiang, S., Durie, B. G. M., Carrasco, R., Sezer, O., Reiman, T., Pilarski, L. M., and Avet-Loiseau, H. (2009) International Myeloma Working Group molecular classification of multiple myeloma : spotlight review. *Leukemia,* **23**, 2210–2221.

[76] van Velden, F. H. P., Kramer, G. M., Frings, Virginie Nissen, I. A., Mulder, E. R., de Langen, A. J., Hoekstra, O. S., Smit, E. F., and Boellaard, R. (2016) Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies : Impact of Reconstruction and Delineation. *Molecular Imaging and Biology,* pp. 788–795.

[77] Hao, H., Zhou, Z., Li, S., Maquilan, G., Folkert, M. R., Iyengar, P., Westover, K. D., Albuquerque, K., Liu, F., Choy, H., Timmerman, R., Yang, L., and Wang, J. (2018) Shell feature : a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer. *Physics in Medicine & Biology,* **63**(9).

[78] Mukaka, M. M. (2012) Statistics corner : A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal,* **24**(3), 69–71.

[79] Ridgeway, G. Generalized Boosted Models : A guide to the gbm package. (2007).

[80] Attal, M., Lauwers-Cances, V., Hulin, C., Leleu, X., Caillot, D., Escoffre, M., Arnulf, B., Macro, M., Belhadj, K., Garderet, L., Roussel, M., Payen, C., Mathiot, C., Fermand, J. P., Meuleman, N., Rollet, S., Maglio, M. E., Zeytoonjian, A. A., Weller, E. A., Munshi, N., Anderson, K. C., Richardson, P. G., Facon, T., Avet-Loiseau, H., Harousseau, J.-L., and Moreau, P. (2017) Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. *New England Journal of Medicine,* **376**(14), 1311–1320.

[81] Cavo, M., Gay, F., Beksac, M., Pantani, L., Petrucci, M., Dimopoulos, M., Dozza, L., van der Holt, B., Zweegman, S., Oliva, S., van der Velden, V. V. D., Zamagni, E., Palumbo, G., Patriarca, F., Montefusco, V., Galli, M., Maisnar, V., Gamberi, B., Hansson, M., Belotti, A., Pour, L., Ypma, P., Grasso, M., Croockewit, A., Ballanti, S., Offidani, M., Vincelli, I., Zambello, R., Liberati, A., Andersen, N. F., Broijl, A., Troia, R., Pascarella, A., Benevolo, G., Levin, M., Bos, G., Ludwig, H., Aquino, S.,

Morelli, A., Wu, K., Boersma, R., Hájek, R., Durian, M., von dem Borne, P. A., di Toritto, T. C., Driessen, C., Specchia, G., Waage, A., Gimsing, P., Mellqvist, U., van Marwijk Kooy, M., Minnema, M., Mandigers, C., Cafro, A., Palmas, A., Carvalho, S., Spencer, A., Boccadoro, M., and Sonneveld, P. (2020) Autologous haematopoietic stem-cell transplantation versus bortezomib-melphalan-prednisone, with or without bortezomib-lenalidomide-dexamethasone consolidation therapy, and lenalidomide maintenance for newly diagnosed multiple myeloma (EMN02/HO95) : a multicentre, randomised, open-label, phase 3 stud. *The Lancet. Haematology,*.

[82] Bartel, T. B., Haessler, J., Brown, T. L. Y., Shaughnessy, John D., J., van Rhee, F., Anaissie, E., Alpe, T., Angtuaco, E., Walker, R., Epstein, J., Crowley, J., and Barlogie, B. (2009) F18-fluorodeoxyglucose positron emission tomography in the context of other imaging techniques and prognostic factors in multiple myeloma. *Blood,* **114**(10), 2068–2076.

[83] Usmani, S. Z., Mitchell, A., Waheed, S., Crowley, J., Hoering, A., Petty, N., Brown, T., Bartel, T., Anaissie, E., van Rhee, F., and Barlogie, B. (2013) Prognostic implications of serial 18-fluoro-deoxyglucose emission tomography in multiple myeloma treated with total therapy 3. *Blood,* **121**(10), 1819–1823.

[84] Fonti, R., Larobina, M., Vecchio, S. D., Luca, S. D., Fabbricini, R., Catalano, L., Pane, F., Salvatore, M., and Pace, L. (2012) Metabolic Tumor Volume Assessed by 18F-FDG PET/CT for the Prediction of Outcome in Patients with Multiple Myeloma. *The Journal of Nuclear Medicine,* **53**, 1829 – 1835.

[85] Fonti, R., Pellegrino, S., Catalano, L., Pane, F., Vecchio, S. D., and Pace, L. (2019) Visual and volumetric parameters by 18F-FDG-PET/CT : a head to head comparison for the prediction of outcome in patients with multiple myeloma. *Annals of Hematology,* **99**, 127–135.

[86] Terao, T., Machida, Y., Tsushima, T., Miura, D., Narita, K., Kitadate, A., Takeuchi, M., and Matsue, K. (2020) Pre-treatment metabolic tumour volume and total lesion glycolysis are superior to conventional positron-emission tomography/computed tomography variables for outcome prediction in patients with newly diagnosed multiple myeloma in clinical practice. *British Journal of Haematology,* **191**(2), 223–230.

[87] Nicodemus, K., Malley, J., Strobl, C., and Ziegler, A. (2009) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics,* **11**, 110 – 110.

[88] Gregorutti, B., Michel, B., and Saint-Pierre, P. (Oct, 2015) Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis,* **90**, 15–35.

[89] Gerds, T. A. and Schumacher, M. (2006) Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal,* **48**(6), 1029–1040.

[90] Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine,* **18**(17-18), 2529–2545.

[91] Yang, S. and Prentice, R. L. (2005) Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika,* **92**, 1–17.

[92] Amyar, A., Ruan, S., Gardin, I., Chatelain, C., Decazes, P., and Modzelewski, R. (March, 2019) 3-D RPET-NET : Development of a 3-D PET Imaging Convolutional Neural Network for Radiomics Analysis and Outcome Prediction. *IEEE Transactions on Radiation and Plasma Medical Sciences,* **3**(2), 225–221.

[93] Zhu, X., Yao, J., Zhu, F., and Huang, J. (2017) WSISA : Making Survival Prediction from Whole Slide Histopathological Images. In *IEEE Conference on computer vision and pattern recognition (CVPR)* pp. 970–975.

[94] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009) Imagenet : A large-scale hierarchical image database. In *IEEE Conference on computer vision and pattern recognition (CVPR)* pp. 248–255.

[95] Feng, X., Tustison, N., and Meyer, C. (2019) Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. In *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* pp. 279–288.

[96] Deng, L. (2012) The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine,* **29**(6), 141–142.

[97] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. (2014) Describing Textures in the Wild. In *IEEE Conference on computer vision and pattern recognition (CVPR).*

[98] Lao, J., Chen, Y., Li, Z. C., Li, Q., Zhang, J., Liu, J., and Zhai, G. (2017) A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports,* **7**(1).

[99] Shboul, Z. A., Alam, M., Vidyaratne, L., Pei, L., Elbakary, M. I., and Iftekharuddin, K. M. (2019) Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction. *Frontiers in Neuroscience,* **13**, 966.

[100] Faraggi, D. and Simon, R. (1995) A neural network model for survival data. *Statistics in Medicine,* **14**(1), 73–82.

[101] Ching, T., Zhu, X., and Garmire, L. X. (2018) Cox-nnet : An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology,* **14**(4), 1–18.

[102] Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J., and Cooper, L. A. D. (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences,* **115**(13), 2970–2979.

[103] Lee, C., Zame, W. R., Yoon, J., and Van Der Schaar, M. (2018) DeepHit : A deep learning approach to survival analysis with competing risks. *Association for the Advancement of Artificial Intelligence (AAAI) conference,* pp. 2314–2321.

[104] Fotso, S. (2018) Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv :1801.05512,*.

[105] Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. (2011) Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* p. 1845–1853.

[106] Kvamme, H. and Borgan, (2019) Continuous and Discrete-Time Survival Prediction with Neural Networks. *arXiv :1910.06724,*.

[107] Kvamme, H., Borgan, , and Scheel, I. (2019) Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research,* **20**(129), 1–30.

[108] Schroff, F., Kalenichenko, D., and Philbin, J. (2015) FaceNet : A unified embedding for face recognition and clustering. In *IEEE Conference on computer vision and pattern recognition (CVPR)* pp. 815–823.

[109] Im, W., Hong, S., Yoon, S.-E., and Yang, H. S. (2019) Scale-Varying Triplet Ranking with Classification Loss for Facial Age Estimation. In *Computer Vision - ACCV* pp. 247–259.

[110] Sohn, K. (2016) Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., (eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc. Vol. 29, .

[111] Movshovitz-Attias, Y., Toshev, A., Leung, T., Ioffe, S., and Singh, S. (2017) No Fuss Distance Metric Learning Using Proxies. *2017 IEEE International Conference on Computer Vision - ICCV,* pp. 360–368.

[112] Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. (2016) Deep Metric Learning via Lifted Structured Feature Embedding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4004–4012.

[113] Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., and Robertson, N. M. (2019) Ranked List Loss for Deep Metric Learning.

[114] Zhu, X., Yao, J., and Huang, J. (2016) Deep convolutional neural network for survival analysis with pathological images. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* Number 1 pp. 544–547.

[115] Li, H., Boimel, P., Janopaul-Naylor, J., Zhong, H., Xiao, Y., Ben-Josef, E., and Fan, Y. (2019) Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. In *International Symposium on Biomedical Imaging (ISBI)* pp. 846–849.

[116] He, K., Zhang, X., Ren, S., and Sun, J. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Computer Vision - ECCV* pp. 346–361.

[117] Huang, Q., Yang, D., Wu, P., Qu, H., Yi, J., and Metaxas, D. (2019) MRI Reconstruction Via Cascaded Channel-Wise Attention Network. In *International Symposium on Biomedical Imaging (ISBI)* pp. 1622–1626.

[118] Qianqian, T., Caizi, L., Weixin, S., Xiangyun, L., Yaliang, T., Zhiyong, Y., and Pheng, A. H. (2019) RIANet : Recurrent interleaved attention network for cardiac MRI segmentation. *Computers in Biology and Medicine,* **109**, 290 – 302.

[119] Herent, P., Schmauch, B., Jehanno, P., Dehaene, O., Saillard, C., Balleyguier, C., Arfi-Rouche, J., and Jégou, S. (2019) Detection and characterization of MRI breast lesions using deep learning. *Diagnostic and Interventional Imaging,* **100**(4), 219 – 225.

[120] Kaji, D. A., Zech, J., Kim, J. S., Cho, S., Dangayach, N., Costa, A. B., and Oermann, E. (2019) An attention based deep learning model of clinical events in the intensive care unit. *PLoS ONE,* **14**.

[121] Liu, Z., Sun, Q., Bai, H., Liang, C., Chen, Y., and Li, Z. (2019) 3D Deep Attention Network for Survival Prediction from Magnetic Resonance Images in Glioblastoma. In *IEEE International Conference on Image Processing (ICIP)* pp. 1381–1384.

[122] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018) CBAM : Convolutional Block Attention Module. In *Computer Vision - ECCV* pp. 3–19.

[123] Newell, A. and Deng, J. (June, 2020) How Useful Is Self-Supervised Pretraining for Visual Tasks ?. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[124] Taleb, A., Lippert, C., Klein, T., and Nabi, M. (2021) Multimodal Self-supervised Learning for Medical Image Analysis. In *Information Processing in Medical Imaging* pp. 661–673.

[125] Zhou, Z., Sodha, V., Rahman Siddiquee, M. M., Feng, R., Tajbakhsh, N., Gotway, M. B., and Liang, J. (2019) Models Genesis : Generic Autodidactic Models for 3D Medical Image Analysis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)* pp. 384–393.

[126] Leenstra, M., Marcos, D., Bovolo, F., and Tuia, D. (2021) Self-supervised Pre-training Enhances Change Detection in Sentinel-2 Imagery. In *Pattern Recognition. ICPR Int. Workshops and Challenges* pp. 578–590.

[127] Sirinam, P., Mathews, N., Rahman, M. S., and Wright, M. (2019) Triplet Fingerprinting : More Practical and Portable Website Fingerprinting with N-Shot Learning. Association for Computing Machinery CCS '19 p. 1131–1148.

[128] Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2016) Instance Normalization : The Missing Ingredient for Fast Stylization. *arXiv :1607.08022,* **abs/1607.08022**.

[129] Maas, A. L. (2013) Rectifier Nonlinearities Improve Neural Network Acoustic Models.

[130] Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 (Canadian Institute for Advanced Research). ,.

[131] Menze, B. H., Jakab, A., Bauer, S., and al. (2015) The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). **34**(10), 1993–2024.

[132] Kingma, D. P. and Ba, J. (2015) Adam : A Method for Stochastic Optimization. *CoRR,* **abs/1412.6980**.

[133] Shanmugam, D., Blalock, D., Balakrishnan, G., and Guttag, J. When and Why Test-Time Augmentation Works. (2020).

[134] Kleinbaum, D. G. and Klein, M. (2012) Survival Analysis : A Self-Learning Text, Springer Science and Business Media, LLC, .

[135] Li, R., Yao, J., Zhu, X., Li, Y., and Huang, J. (2018) Graph CNN for survival analysis on whole slide pathological images.

[136] Luck, M., Sylvain, T., Cohen, J. P., Cardinal, H., Lodi, A., and Bengio, Y. (2018) Learning to rank for censored survival data. *arXiv :1806.01984,*.

CENTRALE NANTES

Titre : Prédiction de la progression du myélome multiple par imagerie TEP : adaptation des forêts de survie aléatoires et de réseaux de neurones convolutionnels

Mot clés : Myélome multiple, Réseaux de neurones convolutionnels, Analyse de survie, Forêts de survie aléatoires, Tomographie à Emission de Positons

Résumé : L'objectif de ces travaux est de fournir un modèle permettant la prédiction de la survie et l'identification de biomarqueurs dans le contexte du myélome multiple (MM) à l'aide de l'imagerie TEP (Tomographie à émission de positons) et de données cliniques. Cette thèse fut divisée en deux parties : La première permet d'obtenir un modèle basé sur les forêts de survie aléatoires (RSF). La seconde est basée sur l'adaptation de l'apprentissage profond à la survie et à nos données. Les contributions principales sont les suivantes : 1) Production d'un modèle basé sur les RSF et les images TEP permettant la prédiction d'un groupe de risque pour les patients atteints de MM. 2) Détermination de biomarqueurs grâce à ce modèle 3) Démonstration de l'intérêt des radiomiques TEP 4) Extension de l'état de l'art des méthodes d'adaptation de l'apprentissage profond à une petite base de données et à de petites images 5) Étude des fonctions de coût utilisées en survie. De plus, nous sommes, à notre connaissance, les premiers à investiguer l'utilisation des RSF dans le contexte du MM et des images TEP, à utiliser du pré-entraînement auto-supervisé avec des images TEP et, avec une tâche de survie, à adapter la fonction de coût triplet à la survie et à adapter un réseau de neurones convolutionnels à la survie du MM à partir de lésions TEP.

Title: Prediction of multiple myeloma progression by PET imaging: adaptation of random survival forests and convolutional neural networks

Keywords: Multiple myeloma, Convolutional Neural Network, Survival analysis, Random Survival Forest, Positron Emission Tomography

Abstract: The aim of this work is to provide a model for survival prediction and biomarker identification in the context of multiple myeloma (MM) using PET (Positron Emission Tomography) imaging and clinical data. This PhD is divided into two parts: The first part provides a model based on Random Survival Forests (RSF). The second part is based on the adaptation of deep learning to survival and to our data. The main contributions are the following: 1) Production of a model based on RSF and PET images allowing the prediction of a risk group for multiple myeloma patients. 2) Determination of biomarkers using this model. 3) Demonstration of the interest of PET radiomics. 4) Extension of the state of the art of methods for the adaptation of deep learning to a small database and small images. 5) Study of the cost functions used in survival. In addition, we are, to our knowledge, the first to investigate the use of RSFs in the context of MM and PET images, to use self-supervised pre-training with PET images, and, with a survival task, to fit the triplet cost function to survival and to fit a convolutional neural network to MM survival from PET lesions.