



---

**Prise en compte de l'incertitude des caractéristiques  
radiomiques issues d'images médicales pour le calcul de  
prédictions individuelles robustes en cancérologie**

---

Réalisé par  
Mehdy HOUNKONNOU



Supervisé par  
Maitre de stage : Loïc FERRER  
Tuteur universitaire : Adrien RICHOU

UF Mathématiques et Interactions  
Collège des Sciences et Technologiques  
Master de Mathématiques Appliquées, Statistique  
Parcours Modélisation Statistique et Stochastique  
Université de Bordeaux

## Avant-propos

Ce rapport rend compte du stage de six mois que j'ai effectué au sein de l'équipe "Biostatistics Research" à Pessac, au sein du département RADIOMICS de l'entreprise SOPHIA GENETICS. Ce stage s'est effectué dans le cadre de ma formation pour l'obtention du Master de Mathématiques appliquées, Statistiques, parcours Modélisation Statistique et Stochastique de l'Université Bordeaux.

Ce stage fut l'occasion pour moi d'avoir une première réelle expérience dans le monde professionnel de la data science. Il m'a notamment permis de mettre un premier pas dans le monde de la recherche et du développement en Biostatistiques. Les travaux qui y sont effectués ont un impact direct sur la vie des patients atteints de cancers et de maladies rares dans le monde. Ce domaine d'application m'a fasciné, car c'est un mélange harmonieux à la fois des mathématiques et du monde de la santé.

Cette expérience fut très gratifiante pour moi, bien que parfois peu évidente, car j'ai dû m'adapter rapidement à certaines spécificités du domaine médical, assez différentes des connaissances acquises lors de mon parcours universitaire.

L'objectif principal de mon stage était la prise en compte de l'incertitude des caractéristiques radiomiques issues d'imageries médicales pour le calcul de prédictions individuelles robustes en cancer. Les prédictions individuelles robustes se doivent d'être basées sur des caractéristiques radiomiques robustes, qui de par leurs processus d'acquisition sont soumises à divers types de variabilités. Ce rapport relate donc le déroulé de l'ensemble de mon stage ainsi que l'ensemble des réflexions de celui-ci.

Durant mon stage, il m'a parfois été assez frustrant de constater que les jeux de données réelles, dans les applications en santé, sont souvent de taille assez limitée. Ceci a notamment été le cas pour l'application réalisée durant mon stage, ce qui a pu gêner les conclusions de celui-ci.

Effectuer le stage dans ce centre de recherche et de développement fut une expérience très enrichissante, à la fois pour le double enjeu, de mener à bien le stage sachant qu'ensuite les résultats seront intégrés dans le processus d'analyse radiomique de l'entreprise et sera appliqué sur des patients, mais également sur l'apprentissage auprès de mes collaborateurs dont la connaissance du domaine est très pointue.

Les compétences de chacun, l'abondance de projets ainsi que les missions très diversifiées proposées par le département RADIOMICS, m'ont ouvert les yeux sur les possibilités qu'offrent la data science et m'ont convaincu de la voie professionnelle que je souhaite emprunter.

## Remerciements

Avant tout développement sur cette première expérience professionnelle, il apparaît opportun de commencer ce rapport en remerciant toutes les personnes ayant contribué de près ou de loin à la réalisation de ce stage dans les meilleures conditions.

J'adresse tout d'abord mes plus profonds remerciements à mon maître de stage, **Loïc Ferrer**, qui a su me donner tout d'abord la chance d'intégrer son équipe en tant que stagiaire, ensuite pour le temps qu'il m'a consacré, son accompagnement permanent et le partage de son expertise, qui m'ont permis d'acquérir des nouvelles connaissances et de mener à bien mes missions ainsi que la rédaction du rapport.

Je remercie tout particulièrement **Olivier Gallinato** pour avoir apporté son soutien et son expertise en l'imagerie médicale qui m'ont aidé à mieux comprendre les données et à développer une vision essentielle sur l'analyse radiomique.

Mes chaleureux remerciements vont à l'endroit de **Thierry Colin** pour son avis favorable à mon accueil dans l'entreprise. Je remercie également mon tuteur universitaire de stage **Adrien Richou**, pour son suivi et son attention particulière concernant mon environnement de travail.

Je tiens à dire un grand merci à l'équipe Data-Science spécialement celle de Biostatistiques pour son accueil chaleureux et ce, malgré cette crise sanitaire toujours en cours. C'est une équipe avec laquelle une réelle collaboration de travail a pu être créée. Le soutien et la solidarité dont tout le monde a fait preuve m'a permis de vivre mon stage avec le plus grand enthousiasme et avoir confiance en la suite.

Pour finir, je tiens à remercier toutes les personnes extérieures qui m'ont apporté leur soutien tout au long de cette période de stage et pour la rédaction de celui-ci : Anais, ma famille, et bien évidemment mes camarades de promotion Benjamin, Gauthier et Pol avec qui j'ai passé tant de moments exceptionnels durant ce master.



## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Présentation de l'entreprise</b>	<b>4</b>
2.1	SOPHiA GENETICS . . . . .	4
2.2	Les solutions SOPHiA GENETICS . . . . .	4
2.2.1	SOPHiA for GENOMICS . . . . .	4
2.2.2	SOPHiA for RADIOMICS . . . . .	5
2.2.3	SOPHiA MULTIMODAL . . . . .	6
2.3	La branche RADIOMICS . . . . .	6
2.3.1	Équipe Image Processing . . . . .	8
2.3.2	Équipe IT . . . . .	8
2.3.3	Équipe Biostatistics . . . . .	8
<b>3</b>	<b>Etat de l'art</b>	<b>10</b>
3.1	Coefficients de corrélation et tests statistiques . . . . .	10
3.1.1	Corrélation de Pearson . . . . .	10
3.1.2	Corrélation de Spearman . . . . .	11
3.1.3	Corrélation de Kendall . . . . .	11
3.1.4	Le test de Wilcoxon-Mann-Whithney . . . . .	12
3.1.5	Le test de Kruskal-Wallis . . . . .	13
3.2	Algorithmes de Machine Learning . . . . .	14
3.2.1	Méthodes Linéaires Généralisés avec Régularisation . . . . .	14
3.2.2	Méthodes à bases d'arbres de décision . . . . .	15
3.2.3	Machine à Vecteur Support (SVM) . . . . .	20
3.3	Évaluation des capacités prédictives . . . . .	20
3.3.1	La Validation croisée Leave-One-Out (LOOCV) . . . . .	20
3.3.2	La Validation croisée Emboîtée . . . . .	21
3.4	Critères de performances prédictives . . . . .	22
3.5	Interprétabilité globale . . . . .	24
3.5.1	Importance des variables : . . . . .	24
3.5.2	Graphique de dépendance partielle (PDP) . . . . .	24
3.5.3	Graphique d'Espérance Individuelle Conditionnelle (ICE) . . . . .	25
3.5.4	Effets Locaux Accumulés (ALE) . . . . .	25
3.5.5	Modèle de substitution global (Surrogate models) . . . . .	25
3.6	Interprétabilité Locale . . . . .	25
3.6.1	Les modèles de substitution locale (LIME) . . . . .	25
<b>4</b>	<b>Robustesse des caractéristiques radiomiques</b>	<b>26</b>
4.1	Contexte et motivation . . . . .	26
4.2	Méthodologie d'évaluation de la robustesse des caractéristiques radiomiques . . . . .	27
4.2.1	Représentation 3D d'une image médicale et d'une segmentation . . . . .	27
4.2.2	Description brève du "Modèle déformable" . . . . .	30
4.2.3	Application à la perturbation de segmentation . . . . .	30
4.2.4	Critère d'évaluation de la robustesse des caractéristiques radiomiques . . . . .	31

4.2.5	Modélisation . . . . .	33
4.2.6	Méthodologie d'évaluation de la robustesse . . . . .	37
<b>5</b>	<b>Application : Prédiction de la réponse au traitement chez des patients atteints de cancer du poumon en stade avancé</b>	<b>38</b>
5.1	Contexte . . . . .	38
5.2	Objectifs . . . . .	38
5.3	Données . . . . .	38
5.4	Environnement, importation et gestion des données brutes . . . . .	39
5.5	Analyses descriptive et explicative . . . . .	40
5.5.1	Distribution de la variable réponse . . . . .	40
5.5.2	Associations bi-variées . . . . .	40
5.5.3	Diagramme de corrélation . . . . .	42
5.5.4	Distribution des variables explicatives . . . . .	43
5.6	Prédiction de la progression à la première évaluation . . . . .	44
5.6.1	Modélisation . . . . .	44
5.6.2	Interprétabilité globale . . . . .	47
5.6.3	Interprétabilité locale . . . . .	51
5.7	Impact de la sélection des variables robustes . . . . .	53
5.7.1	Modélisation de l'ICC(1,1) . . . . .	53
5.7.2	Sélection de la configuration des différents DICEs pour les perturbations et seuil de l'ICC la sélection de variables robustes . . . . .	53
5.7.3	Sélection des variables robustes à partir de l'ICC . . . . .	56
5.8	Comparaison des différentes méthodes de sélections . . . . .	61
5.9	Interprétabilité des modèles . . . . .	64
5.9.1	Interprétabilité Globale . . . . .	64
<b>6</b>	<b>Travaux supplémentaires</b>	<b>69</b>
<b>7</b>	<b>Conclusion</b>	<b>70</b>
<b>8</b>	<b>Annexes</b>	<b>76</b>

## Table des figures

1	Prises de vue axiale, sagittale et coronale d'un cerveau . . . . .	5
2	Test de Mann-Whitney - Exemples de mélanges d'échantillons . . . . .	12
3	Approche du Gradient Boosting . . . . .	17
4	Représentation de l'algorithme Adaboost [Trevor Hastie, 2009] . . . . .	18
5	Etape de l'algorithme Adaboost [Freund and Schapire, 1996] . . . . .	19
6	Classification et séparation linéaire . . . . .	20
7	La Validation croisée Leave-one-out . . . . .	20
8	La Validation croisée Emboîtée . . . . .	21
9	Image 3D d'un parenchyme . . . . .	28
10	Segmentation ou contour 2D d'une tumeur sur un parenchyme . . . . .	28
11	Segmentation 3D d'une tumeur sur un parenchyme . . . . .	29
12	Maillage d'une segmentation en 3D transparente et non transparente . . . . .	29
13	Contour original (rouge) et ses perturbations en 2D avec un Dice = 0.90 . . . . .	31
14	De gauche à droite un maillage originale et une perturbation en 3D de Dice = 0.90 . . . . .	31
15	Distribution de la variable réponse : Progression à la première évaluation . . . . .	40
16	Distribution de la variable Compactness 1 suivant la variable d'intérêt . . . . .	42
17	Matrice de corrélation (bleu = 1, rouge = -1, blanc = 0) . . . . .	43
18	Top 10 des variables les plus importantes pour la prédiction . . . . .	48
19	De gauche à droite : PDP, ICE et ALE pour la variable Asphérité . . . . .	49
20	De gauche à droite : PDP, ICE et ALE pour la variable Dependence count entropy . . . . .	50
21	De gauche à droite : PDP, ICE et ALE pour la variable Minor axis length . . . . .	50
22	Arbre de décision obtenu comme modèle de substitution global . . . . .	51
23	Interprétabilité locale LIME pour une prédiction individuelle . . . . .	52
24	Le nombre de variables robustes en fonction du seuil de l'ICC suivant les Dices dans l'ordre : 0.87, 0.90, 0.93, 0.96, 0.99 . . . . .	54
25	Comparaison du PRAuc suivant les différents seuils de l'ICC pour le Dice 0.90 . . . . .	55
26	Nombre de variables robustes par classes . . . . .	56
27	Shapes features selection . . . . .	57
28	Le nombre de variables sélectionnées en fonction du seuil de l'ICC (borne inférieure de l'IC à 95%) . . . . .	58
29	Nombre de variables robustes par classes . . . . .	59
30	Histogram features selection . . . . .	59
31	Bilan des sélections de variables avec l'ICC . . . . .	61
32	Comparaison du PRAuc selon les trois méthodes de sélection . . . . .	62
33	Comparaison du F1score selon les trois méthodes de sélection . . . . .	63
34	Comparaison de la Sensibilité selon les trois méthodes de sélection . . . . .	63
35	De gauche à droite le top 10 des variables importantes du modèle avec l'ICC 0.90 et avec l'IC de l'ICC 0.90 . . . . .	64
36	De gauche à droite : PDP, ICE et ALE pour le top 3 des variables importantes : ICC 0.90 . . . . .	65
37	De gauche à droite : PDP, ICE et ALE pour le top 3 des variables importantes : IC de l'ICC 0.90 . . . . .	66
38	Arbre de décision obtenu comme modèle de substitution global . . . . .	67
39	Arbre de décision obtenu comme modèle de substitution global . . . . .	68

40	First orders features selection . . . . .	76
41	Textures features selection . . . . .	76
42	Histogram features selection . . . . .	77
43	Original features selection . . . . .	77
44	Emphasis features selection . . . . .	78
45	GL matrix features selection . . . . .	78
46	Original features selection . . . . .	79
47	Textures features selection . . . . .	79
48	Shapes features selection . . . . .	80
49	First orders features selection . . . . .	80
50	Emphasis features selection . . . . .	81
51	GL matrix features selection . . . . .	81

## Liste des tableaux

1	Matrice de confusion . . . . .	23
2	Caractéristiques de l'ordinateur . . . . .	39
3	Extrait des associations entre les variables et la variable d'intérêt . . . . .	41
4	Les variables ayant les p-valeurs les plus significatives i.e p-valeurs < 0.05 . . . . .	41
5	Tableau des performances prédictives de la sélection de variable Kendall à 0.90 . . . . .	46
6	Tableau des performances prédictives de la sélection de variable Kendall à 0.80 . . . . .	46
7	Matrice de confusion pour le modèle GBM . . . . .	47
8	Prédictions obtenues par le modèle complexe et le modèle simple . . . . .	52
9	Perturbations de la variable "Grey level variance" . . . . .	53
10	Variables non robustes . . . . .	57
11	Résultats de la sélection de variable suivant ICC à 0.90 . . . . .	62
12	Résultats de la sélection de variable suivant la borne inférieure de l'intervalle de confiance de l'ICC à 0.90 . . . . .	62
13	Morphological Features . . . . .	82
14	First order features . . . . .	83
15	Histogram features . . . . .	84
16	Original data Features . . . . .	85
17	Emphasis features . . . . .	86
18	GLMatrix features . . . . .	87
19	Textures features 1 . . . . .	88
20	Textures features 2 . . . . .	89

## Acronymes et terminologie

Voici une liste des abréviations utilisées dans ce rapport :

- LOOCV : Leave-One-Out Cross Validation (validation croisée leave-one-out)
- CART : Classification and Regression Trees (arbres de régression et de classification)
- NPV : Negative Predictive Value (valeur prédictive négative)
- LIME : Local Interpretable Model-agnostic Explanations (explications agnostiques du modèle interprétable local)
- ALE : Accumulated Local Effects (effet locaux accumulés)
- PDP : Partial Dependence Plot (graphique de dépendance partielle)
- ICC : Intraclass correlation coefficient (coefficient de corrélation intra-classe)
- IRMs : Imagerie par résonance magnétique
- ADN : Acide désoxyribonucléique
- IBSI : Initiative de Standardisation de Biomarqueurs d'Images

Ci-dessous quelques termes spécifiques qui seront employés dans ce rapport :

1. **Cancer** : C'est une prolifération cellulaire anarchique et non contrôlée par l'organisme.
2. **Segmentation d'image** : C'est une stratégie de traitement d'image qui consiste à regrouper des pixels d'une image selon certaines caractéristiques (distance, couleur, angle, etc.).
3. **Maillage ou mesh** : Il s'agit de la discrétisation spatiale d'un milieu continu, ou aussi, une modélisation géométrique d'un domaine par des éléments proportionnés finis et bien définis. L'objet d'un maillage est de procéder à une simplification d'un système par un modèle représentant ce système et, éventuellement, son environnement (le milieu), dans l'optique de simulations de calculs ou de représentations graphiques.
4. **Voxel** : C'est un pixel en 3D.
5. **Immunothérapie** : Traitement relatif à l'injection d'un produit qui va principalement réactiver le système immunitaire contre les cellules cancéreuses.
6. **Coupe** : Une coupe ou "slice" est une section tomographique d'un objet ici les scanners de radiologiques. Une coupe est définie par sa position et son épaisseur ; chaque coupe est divisée en une matrice de voxels. En tomographie assistée par ordinateur, une table motorisée fait glisser le patient à travers le portique et les coupes sont réalisées lorsque le tube à rayons X tourne en cercle autour du patient.
7. **Features radiomiques ou Caractéristiques radiomiques** : C'est un ensemble de variables extraites de la segmentation d'une zone d'intérêt, des caractéristiques quantitatives d'images médicales qui décrivent une maladie. Dans le rapport j'utiliserai le mot "features" ou "caractéristiques" comme "variables".
8. **Coefficient de similarité du Dice** : Egalement connu sous le nom d'indice "Sørensen-Dice" ou simplement de "coefficient de Dice", c'est un outil statistique qui mesure la similarité entre deux ensembles de données. Cet indice est devenu sans doute l'outil le plus largement utilisé dans la validation des algorithmes de segmentation d'images créés avec l'Intelligence Artificielle, mais il s'agit d'un concept beaucoup plus général qui peut être appliqué à des

ensembles de données pour une variété d'applications, y compris le traitement automatique des langues. Étant donné deux ensembles,  $X$  et  $Y$ , il est défini comme suit :

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|},$$

où  $|X|$  et  $|Y|$  sont les cardinalités des deux ensembles (c'est-à-dire le nombre d'éléments dans chaque ensemble).

9. **Fiabilité test-retest** : Elle reflète la variation des mesures prises par un instrument sur le même sujet dans les mêmes conditions. Elle est généralement indicative de la fiabilité dans les situations où les évaluateurs ne sont pas impliqués ou lorsque l'effet des évaluateurs est négligeable, comme dans le cas d'un instrument d'enquête d'auto-évaluation.
10. **Fiabilité Inter-évaluateur ou inter-opérateur** : Elle reflète la variation entre 2 ou plusieurs évaluateurs qui mesurent le même groupe de sujets.
11. **Fiabilité intra-évaluateur ou intra-opérateur** : Il reflète la variation des données mesurées par un évaluateur sur 2 essais ou plus.
12. **Bootstrap** : Supposons que nous ayons un modèle ajusté à un ensemble de données d'apprentissage. L'idée de base derrière cette technique est de tirer au hasard  $B$  fois (nombre d'échantillons bootstrap) des ensembles de données avec remise à partir des données d'apprentissage, chaque échantillon étant de la même taille que les données d'apprentissage originales. Ainsi, certains individus pourraient être présent plusieurs fois dans un échantillon bootstrap tandis que d'autres ne seront pas présent du tout. De ces nouveaux ensembles le modèle est réajusté ce qui permet d'estimer correctement une quantité d'intérêt (la moyenne ou la variance).
13. **Observations et erreurs OOB** : Les observations non utilisées à la suite du processus bootstrap sont appelées les observations out-of-bag (OOB). Il est possible d'évaluer la performance de modèle pour prédire la variable d'intérêt à partir de ces observations et l'erreur associée à cette prédiction s'appelle l'erreur Out-of-Bag.
14. **Bagging** : Le processus d'agrégation de plusieurs processus bootstrap s'appelle le bagging (Bootstrap Aggregating). Le Bagging est une méthode d'apprentissage d'ensemble utilisée pour réduire la variance d'un estimateur. Dans cette méthode les modèles "faibles"<sup>1</sup> sont entraînés indépendamment sur des échantillons bootstrap et, selon le type de tâche - régression ou classification, par exemple - la moyenne ou la majorité de ces prédictions donnent une estimation plus précise.
15. **Approche Enveloppante ou Wrappers** : Les méthodes de sélection de variables enveloppantes ont été introduites par [John et al., 1994]. Leur principe est de générer des sous-ensembles de variables candidats et de les évaluer grâce à un algorithme de classification. Cette évaluation est faite par le calcul d'un score, par exemple un score d'un ensemble sera un compromis entre le nombre de variables éliminées et le taux de réussite de la classification sur un fichier de test. L'appel de l'algorithme de classification est fait plusieurs fois à chaque évaluation (c'est-à-dire qu'à chaque sélection d'une variable, nous calculons le taux de classification pour juger la pertinence d'une caractéristique) car un mécanisme de validation croisée est fréquemment utilisé. Le principe des "wrappers" est de générer un sous-ensemble bien adapté à l'algorithme de classification. Les taux de reconnaissance sont élevés car la sélection prend en

---

1. c'est-à-dire à peine plus efficaces qu'une classification aléatoire.

compte le biais intrinsèque de l'algorithme de classification. Un autre avantage est sa simplicité conceptuelle ; nous n'avons pas besoin de comprendre comment l'induction est affectée par la sélection des variables, il suffit de générer et de tester. Cependant, trois raisons font que les "wrappers" ne constituent pas une solution parfaite. D'abord, ils n'apportent pas vraiment de justification théorique à la sélection et ils ne nous permettent pas de comprendre les relations de dépendances conditionnelles qu'il peut y avoir entre les variables. D'autre part la procédure de sélection est spécifique à un algorithme de classification particulier et les sous-ensembles trouvés ne sont pas forcément valides si nous changeons de méthode d'induction. Finalement, c'est l'inconvénient principal de la méthode, les calculs deviennent de plus en plus longs, voire irréalisables lorsque le nombre de variables est très élevé.

16. **Approche Filtre ou Filter** : L'approche filtre sélectionne un sous-ensemble de variables en pré-traitement des données d'un modèle (lors de l'étape de l'analyse des données), Le processus de sélection est indépendant du processus de classification. Un de ces avantages est d'être complètement indépendant du modèle de données que nous cherchons à construire. Elle propose un sous-ensemble de variables satisfaisant pour expliquer la structure des données qui se cachent et que le sous-ensemble est indépendant de l'algorithme d'apprentissage choisi. Ce contexte est aussi adaptatif dans la sélection de variables non supervisées (Géurif [.S08], Mitra et al.[MP02], Bennani et Géurif [Hal98]). De plus les procédures filtres sont généralement moins coûteuses en temps de calcul puisqu'elles évitent les exécutions répétitives des algorithmes d'apprentissage sur différents sous-ensembles de variables. En revanche, leur inconvénient majeur est qu'elles ignorent l'impact des sous-ensembles choisis sur les performances de l'algorithme d'apprentissage.
17. **Approche Embarquée ou Embedded** : Les méthodes "Embedded" intègrent directement la sélection dans le processus de l'apprentissage, les arbres de décisions sont l'illustration la plus emblématique. On peut également citer d'autres méthodes telles que la classification naïve bayésienne ou encore les méthodes sparses... Mais, en réalité, nous classons dans ce groupe toutes techniques qui évaluent l'importance d'une variable en cohérence avec le critère utilisé pour évaluer la pertinence globale du modèle. Les méthodes embarquées, comme les méthodes enveloppantes, permettent de bien prendre en compte les interactions entre variables mais elles ont également l'avantage d'être généralement plus rapides en évitant le processus d'aller-retour entre sélection et évaluation par apprentissage. Toutefois, contrairement aux méthodes enveloppantes, le choix de ces méthodes est limité et l'influence du choix de la méthode d'apprentissage est d'autant plus grande.



## 1 Introduction

Le cancer est une maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive. En se multipliant de façon anarchique, les cellules cancéreuses donnent naissance à des tumeurs de plus en plus grosses qui se développent en envahissant puis détruisant les zones qui les entourent (organes). En France, les quatre principaux cancers sont ceux de la prostate, du sein, du poumon et du côlon-rectum. De nos jours divers moyens sont mis en œuvre dans la lutte contre le cancer conduisant ainsi à des améliorations et des perspectives optimistes dans ce domaine.

Afin d'établir un diagnostic précis, de définir le traitement le plus approprié, d'évaluer un effet thérapeutique ou de prévoir l'évolution de la maladie, les cliniciens peuvent effectuer différents examens à savoir des bilans cliniques et biologiques, des examens d'imagerie médicale (par exemple tomodensitogrammes ou imagerie par résonance magnétique (IRMs) ) et/ou des séquençages d'ADN (acide désoxyribonucléique). Concernant l'imagerie médicale , celle-ci passe progressivement de l'analyse visuelle conventionnelle des images à une médecine personnalisée quantitative, grâce au développement récent de méthodes d'analyse axées sur les données, comme l'analyse radiomique.

L'analyse radiomique consiste à extraire des caractéristiques quantitatives des images médicales, qui, couplées à d'autres informations individuelles (données cliniques, biologiques etc.), peuvent aider le clinicien dans le diagnostic, le pronostic ou la prédiction de l'évolution de la maladie du patient. En pratique, l'analyse radiomique commence avec l'acquisition d'images médicales (IRM ou CT-scan ou PET-scan). Par la suite, une segmentation des images a lieu et des caractéristiques de l'image peuvent être extraites, telles que la taille de la tumeur (surface, volume, etc.), la distribution de l'intensité des voxels (moyenne, variance, nombre de compartiments, etc.) ou des indicateurs de texture tumorale (contraste, homogénéité, etc.). L'ensemble des caractéristiques extraites sont appelées caractéristiques radiomiques. Ces dernières années, des études ont montré que les caractéristiques radiomiques ont le potentiel d'améliorer de manière significative notre capacité à stratifier les patients en fonction de la réponse probable au traitement, au-delà des facteurs pronostiques conventionnels, conduisant ainsi à des soins du cancer véritablement personnalisés [Limkin et al., 2017].

**SOPHiA GENETICS** a développé une solution logicielle nommée **SOPHiA Radiomics** qui permet la visualisation et la segmentation 3D de lésions cibles dans de nombreuses applications cliniques, à partir d'images médicales. En particulier, la procédure de segmentation semi-automatique aide le clinicien à délimiter la zone d'intérêt de l'image (ROI : Region Of Interest). Dans ce type de segmentation, le clinicien a une participation active. La démarche commence par la sélection d'un point de départ dans la zone d'intérêt. Ensuite, des algorithmes itératifs permettent de délimiter la ROI, avant vérification et possible modification manuelle de celle-ci. Des caractéristiques quantitatives de la ROI (indicateurs de taille, de forme, de texture, etc.) sont ensuite extraites à partir des définitions de l'IBSI (Initiative de Standardisation de Biomarqueurs d'Images) [Zwanenburg et al., 2020]. L'IBSI fournit des définitions consensuelles pour le traitement des images et le calcul des indicateurs d'image, afin de résoudre le manque de reproductibilité et de validation des études radiomiques.

Les informations extraites sur la tumeur sont couplées à d'autres informations individuelles telles que des données cliniques (sex, age, etc.) et biologiques (marqueurs d'expressions géniques, autres

marqueurs tumoraux, etc.) dans le but de construire des modèles diagnostiques, pronostiques et/ou prédictifs de l'évolution de la maladie pour chaque patient. L'analyse radiomique permet ainsi de coupler différentes informations issues de technologies innovantes et performantes. Elle devient un outil d'aide à la décision très important pour les cliniciens, notamment pour les pathologies dont les outils décisionnels traditionnels ne sont pas suffisamment performants.

Néanmoins la radiomique aujourd'hui présente certains défis techniques dans son processus multi-système comprenant l'acquisition d'images, le traitement d'images, la segmentation d'images et le développement de modèles [Granzier et al., 2020]. Les caractéristiques radiomiques ont montré qu'elles sont très vulnérables et sensibles aux différents modes d'acquisition, de traitement et de reconstruction des images : elles sont sujettes à diverses variations durant ces processus [Owens et al., 2018]. En effet, les caractéristiques radiomiques peuvent varier suivant le logiciel utilisé pour la segmentation de la zone d'intérêt (variabilité due aux logiciels), aussi elles peuvent varier suivant la personne qui effectue la segmentation (variabilité inter-opérateur) et encore plus loin elles peuvent varier suivant les différents points de temps de segmentation d'une même personne (variabilité intra-opérateur) [Zwanenburg et al., 2019]. Afin d'envisager une utilisation des outils prédictifs d'analyse radiomique en pratique clinique, il est primordial d'avoir des modèles qui sont basés sur des caractéristiques radiomiques robustes : invariantes à tous ces paramètres de fluctuation.

En rejoignant l'équipe SOPHiA Radiomics, j'ai principalement participé au développement de l'axe de recherche sur la sélection des caractéristiques radiomiques robustes, par rapport aux variabilités inter-opérateurs, dans le but de créer une signature robuste face à cette incertitude [Duron et al., 2019]. Cela permettra de proposer aux cliniciens partenaires une analyse radiomique plus robuste, avec des prédictions précises, reproductibles et plus fiables.

Chez les patients traités pour un cancer, le clinicien peut évaluer l'efficacité du traitement (chimiothérapie, radiothérapie, immunothérapie, ...) sur le patient à travers plusieurs critères (RECIST<sup>1</sup> ou PERCIST<sup>2</sup>). Un exemple incontournable reste l'étude de l'évolution de la tumeur lors d'une consultation quelques temps après le début du processus de guérison : ce qu'on appelle la première évaluation. En fonction de l'évolution ou de la régression de la taille de la tumeur il est dit que le patient a progressé ou non. Mes premiers travaux ont porté sur la prédiction du statut à la première évaluation chez des patients atteints de cancer du poumon en stade avancé traité par immunothérapie (Pembrolizumab), uniquement à partir des données d'imagerie collectées en pré-traitement. L'objectif principal est ici d'identifier quels patients sont susceptibles de ne pas répondre au traitement.

Dans un premier temps, l'analyse prédictive a été faite sans identification des caractéristiques radiomiques robustes. Par la suite, je me suis intéressé à l'incertitude des caractéristiques radiomiques extraites selon la qualité de l'image et selon les paramètres d'image processing qui ont été considérés. Sachant qu'il y a différentes sources de variabilité possibles, mon étude s'est focalisée sur la variabilité inter-opérateur. Celle-ci se définit par la fluctuation des caractéristiques radiomiques

---

1. la maladie est catégorisée en « progression » si la somme des grands axes des lésions cibles augmente de 20% et 5 mm [Eisenhauer et al., 2009] par rapport à l'examen de baseline, ou si une nouvelle lésion apparaît.

2. les critères percists sont basés sur des paramètres métaboliques et utilisent la variation du SUV (standardized uptake value) et l'évolution du nombre de lésions hypermétaboliques.

extraites, générées par la délimitation d'une tumeur par plusieurs médecins/observateurs utilisant le même logiciel. En effet, la récupération des données issues d'une expérience mettant en situation cette variabilité implique un coût très élevé. Ainsi, nous avons réalisé des perturbations des contours d'un clinicien à l'aide un modèle déformable créant ainsi des contours perturbés ayant un coefficient de similarité de Dice [Dice, 1945] avec le contour initial fixé à des seuils arbitraires (0.87, 0.90, 0.99 etc.), tout en vérifiant que le contour généré n'est pas totalement dénué de sens. Nous avons ainsi pu détecter les caractéristiques radiomiques robustes (invariant suivant le mode d'acquisition et de segmentation) et un nouveau modèle de prédiction a été développé. Les modèles de prédiction avec et sans sélection des caractéristiques radiomiques robustes ont ensuite été comparés.

J'ai également réalisé un état de l'art des différentes méthodes dans la littérature dans l'optique de pouvoir quantifier l'incertitude des prédictions estimées sur les modèles de machine learning utilisant les caractéristiques radiomiques robustes.

La suite du rapport est structurée comme suit : nous présentons dans un premier temps l'entreprise SOPHiA GENETICS, son domaine de compétences, ses départements et leurs missions en Section 2. Un bref état de l'art est présenté en Section 3 afin d'introduire les notions statistiques requises au sein de ce rapport. En section 4 et 5 sont respectivement présentés la méthodologie d'analyse de la robustesse des caractéristiques radiomiques (déttection des caractéristiques sujettes à de fortes variabilités) ainsi que les résultats d'une application du processus d'analyse radiomique sur les patients atteint d'un cancer du poumon couplé avec la quantification de l'impact des incertitudes des caractéristiques radiomiques sur les préditions estimées. Nous présentons en Section 6 des travaux supplémentaires effectués durant cette période de stage. Le rapport se conclut en Section 7.

## 2 Présentation de l'entreprise

### 2.1 SOPHiA GENETICS

Fondée par Jurgi Camblong, Pierre Hutter et Lars Steinmetz en 2011, **SOPHiA GENETICS** est un leader mondial de la médecine basée sur les données. Avec un effectif d'environ 400 employés (30 % ayant un doctorat) et présent sur plusieurs continents (près de 40 pays), sa mission est de démocratiser la médecine basée sur les données et de transformer les données en informations précieuses, permettant finalement une meilleure gestion des maladies. Afin de fournir une compréhension multidimensionnelle des maladies, l'entreprise a réuni la puissance de la science et de la technologie. Elle combine une expertise approfondie des sciences de la vie et des disciplines médicales avec des compétences mathématiques et informatiques. Elle a développé une intelligence collective dans le domaine de la santé, offrant un accès mondial à sa technologie et facilitant le partage des connaissances entre pairs. Ces technologies ont été approuvées et implémentées dans plus de 1000 hôpitaux à travers une centaine de pays dans le monde. **SOPHiA GENETICS** développe et commercialise une plateforme "Software-as-a-Service" basée sur le "cloud" : la plateforme SOPHIA DDM™, qui permet aux établissements de santé d'obtenir des informations rapides et robustes à partir de leurs données. Il y a des applications de cette technologie à des maladies telles que le cancer et les troubles héréditaires, où la combinaison des informations génomiques et phénotypiques est essentielle pour soutenir les découvertes, les décisions de traitement et les efforts de développement de médicaments.

### 2.2 Les solutions SOPHiA GENETICS

Le but de l'entreprise est de soutenir l'analyse d'un large éventail de modalités de données numériques sur la santé afin de générer de nouvelles perspectives. Elle s'est développée grâce à son expertise sur l'analyse de données génomiques. Aujourd'hui **SOPHiA GENETICS** a diversifié son offre notamment au niveau de l'analyse radiomique et multimodale.

#### 2.2.1 SOPHiA for GENOMICS

La plateforme SOPHiA DDM™ rationalise l'analyse, l'interprétation et l'écriture des données génomiques. Ses performances analytiques avancées et la visualisation des données favorisent la prise de décisions éclairées basées sur les altérations génomiques. Les échantillons biologiques (y compris les tissus frais congelés, l'ADN tumoral circulant liquide et fixé au formol) peuvent être traités par la plateforme SOPHiA DDM™, ce qui permet d'identifier en toute confiance les altérations génomiques. Les résultats détaillent les variants nucléotidiques simples (SNV), les insertions et les délétions (Indels), les variations du nombre de copies (CNV) et les fusions de gènes, ainsi que des signatures mutationnelles plus complexes tels que l'instabilité des microsatellites (MSI), la charge mutationnelle tumorale (TMB), le déficit de recombinaison homologue (HRD) ou la maladie résiduelle minimale (MRD). De plus, la plateforme SOPHiA DDM™ alimentée par l'Intelligence Artificielle détecte et identifie avec précision les altérations génomiques difficiles, telles que les mutations de CEBPA<sup>1</sup> ou de FLT3-ITD<sup>2</sup>, les mutations de saut d'exon14 de MET<sup>3</sup> ou les fusions génétiques rares.

---

1. CCAAT Enhancer Binding Protein Alpha

2. Une mutation très répandue qui se manifeste par une charge leucémique élevée et confère un mauvais pronostic aux patients

3. Un récepteur membranaire à activité tyrosine kinase dont le ligand est le facteur de croissance hépatocytaire

### 2.2.2 SOPHiA for RADIOMICS

Le logiciel **SOPHiA Radiomics** permet, dans un premier temps de lire des images médicales au format DICOM<sup>1</sup>. La Figure 1 représente les différents modes de vue possibles d'une image : axial, coronal, sagittal. En 3D (regroupement des trois modes), une image peut être vue comme une grille 3D de voxels d'une image. Cette grille possède une origine (0,0,0) ainsi qu'un repère ( $\vec{i}, \vec{j}, \vec{k}$ ). Lorsque l'on se déplace axialement, coronalement ou sagittalement sur cette grille, on dit que l'on change de « coupe ». L'extraction des caractéristiques effectuées sera issue de la segmentation en 3D.

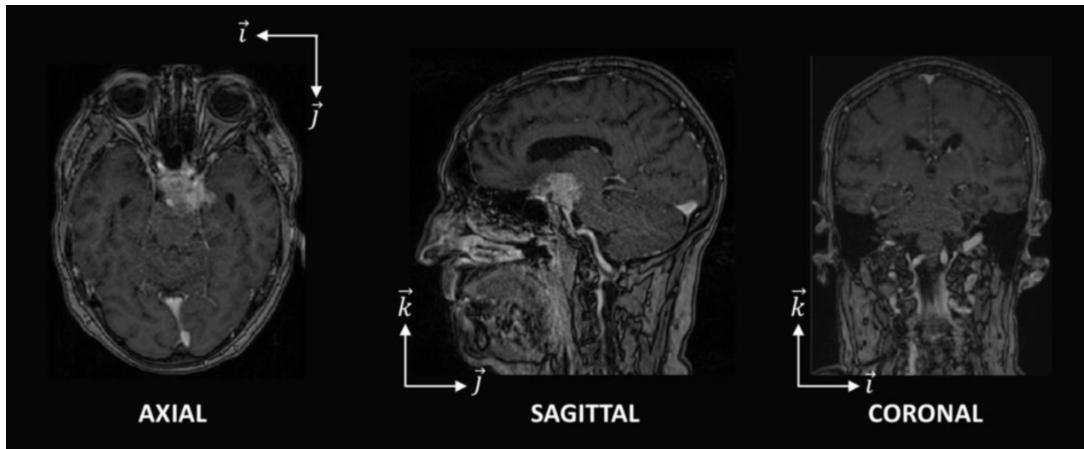


FIGURE 1 – Prises de vue axiale, sagittale et coronale d'un cerveau.

Dans un deuxième temps le logiciel permet de réaliser des segmentations de zone d'intérêt. Il possède de nombreux outils dont l'outil de segmentation. Le terme "segmentation" désigne un processus algorithmique utilisé pour isoler et extraire les voxels d'une zone d'intérêt, typiquement une tumeur. Dans le logiciel, la segmentation est réalisée grâce à un "modèle déformable" : un maillage 3D est initialisé et évolue dans l'espace jusqu'à s'aligner sur les contours de la forme voulue. Ce modèle contient un ensemble de paramètres dont les valeurs dépendent du type de tumeur à segmenter et de l'organe dans lequel elle se situe. De manière simplifiée, le principe d'une segmentation est de créer un maillage de départ et de le faire grossir à l'aide de forces afin qu'il prenne la forme de la tumeur, tel un ballon que l'on ferait gonfler dans une cavité afin qu'il épouse la forme. Les forces sont déterminées en partie par les intensités des voxels de l'image analysée, leurs gradients et leurs rapports avec les intensités voisines (indicateurs de texture). Une fois la segmentation réalisée, des indicateurs ("radiomics features") de la tumeur sont calculés (notamment à partir des formules de [Zwanenburg et al., 2020]). C'est à partir de l'information extraite des segmentations réalisées que l'équipe statistique dont je fais partie peut finaliser les analyses radiomiques.

La segmentation d'images alimentées par l'IA<sup>2</sup> et l'extraction de caractéristiques radiomiques transforment les images médicales 3D existantes en points de données inédits.

La plateforme SOPHiA DDM™ peut traiter et analyser les données provenant de tout type de technologie d'imagerie médicale tridimensionnelle, notamment les scanners de tomographie assistée par

1. Digital Imaging and Communications in Medicine. Ce format est une norme standard pour la gestion informatique des données issues de l'imagerie médicale.

2. Intelligence Artificielle

ordinateur (CT), de tomographie par émission de positons (PET), d'imagerie par résonance magnétique (IRM) et de tomographie par émission monophotonique (SPECT). L'entreprise a développé des algorithmes de segmentation alimentés par l'IA qui détectent les tumeurs dans les scanners, puis les segmentent et les reconstruisent en 3D, couvrant un large éventail de types de tumeurs majeures, notamment le cancer du poumon, du sein, du foie, du rein et du cerveau. L'extraction de caractéristiques radiomiques est effectuée sur les tumeurs segmentées, générant des points de données à travers des caractéristiques volumétriques, morphologiques, de premier ordre (c'est-à-dire l'hétérogénéité), de second ordre (c'est-à-dire la texture) et générées par l'apprentissage profond. Du fait de la multiplicité et de la variabilité de ses données extraites, les avis sont très vite partagés, ce qui a conduit à la standardisation de ceux-ci à travers IBSI (Image Biomarker Standardisation Initiative) [Zwanenburg et al., 2020] qui fournit une nomenclature et des définitions des biomarqueurs d'images normalisées reproductibles quelque soit l'étude, et mise à jour très régulièrement. La plateforme SOPHiA DDM™ dispose de capacités radiomiques pour la détection de la maladie, la discrimination des sous-types histologiques de la maladie, la prédiction de l'évolution de la tumeur et la prédiction de la progression de la maladie.

### 2.2.3 SOPHiA MULTIMODAL

En combinant des données de haute qualité au niveau individuel pour générer des perspectives multimodales, SOPHiA GENETICS exploite la puissance de l'IA avancée et des modèles d'apprentissage automatique. Aujourd'hui, la plateforme SOPHiA DDM™ permet l'analyse de données cliniques, biologiques, génomiques et radiomiques. À l'avenir, l'entreprise a pour intention de prendre en charge des modalités de données supplémentaires telles que la pathologie numérique<sup>1</sup>, la protéomique<sup>2</sup> et la métabolomique<sup>3</sup>. Les capacités de modélisation prédictive de la plateforme SOPHiA DDM™ comprennent le dépistage des maladies, la détection précoce des maladies, le diagnostic des maladies et la discrimination des sous-types, la prédiction de l'évolution des maladies, la prédiction de la réponse au traitement, la sélection et le suivi des traitements.

## 2.3 La branche RADIOMICS

Le département RADIOMICS a débuté son aventure il y a maintenant plus de trois ans sous la direction de Thierry Colin. Celui-ci est en pleine expansion puisqu'il compte aujourd'hui un peu plus d'une quarantaine d'employés. Ce département a pour but majeur le développement d'un logiciel dédié au traitement des données d'images médicales et à la personnalisation du suivi des patient - plus de détails section 2.2.2.

Les équipes Biostatistics, Image Processing et IT se structurent comme suit :

Business Management + Product Management + Subject Matter Expert			
	<p>@ Thierry Colin SVP Radiomics Business Area</p>		<p>@ Nayef Idrissi Global Product Manager</p>
			<p>@ Simon Duchene Subject Matter Expert (FR)</p>

1. Étude au microscope des cellules et tissus pathologiques (lame de verre)
2. Étudie de l'ensemble des protéines d'un organisme, d'un fluide biologique, d'un tissu, d'une cellule ou même d'un compartiment cellulaire.
3. Étude des métabolites issus de l'organisme ou provenant de l'environnement

Le département RADIOMICS est composé de plusieurs équipes :

1. Équipe Biostatistics : Modélisation statistique descriptive, explicative et prédictive.

	@Loic Ferrer	Manager Biostatistics Research
	@Guillaume Etchepare	Biostatistics Research Expert - DS4
	@Colombe Lopez	Junior Biostatistician - DS1

	@Jennifer Vargas	Junior Biostatistician - DS1
	@Mehdy Houkonnou	Junior Biostatistician - DS1

2. Équipe Image Processing : Algorithmes mathématiques de segmentation, traitement d'images, extraction de features radiomiques. (Image processing et Biostatistiques forment équipe "Data science").

	@Olivier Gallinato	Manager Data Engineering - Radiomics research
	@Floriane Gidel	Data Engineer - DS2
	@Antoine Huc	Junior Data Engineer - DS1
	@Juliette Busquet	Junior Data Engineer - DS1
	@Jason Siffre	Data Scientist - DS2
	@Lea Pfeiffer	Junior Data Scientist - DS1
	@Clement Hognon	PhD

3. Équipe IT (Information Technology) : Bases de données, architecture de code du logiciel et génie logiciel.

Radiomics Software Development Team			
	@Vivien Pianet	Manager software engineering	
	@Sasha Oudot	SDE II	
	@Maylis Dupouy	SDE II	
	@Yves Le Moigne	SDE II	@Thibault Cardaire
			SDE II

		
@Thibault Sabalcagaray SDE I	@Wafaa Namasse SDE II	@Adrien Danzon SDE II

Systems Operations

@Frederic Bruneteau Senior System Administrator

Ces équipes sont en constante collaboration et leurs travaux sont complémentaires.

### 2.3.1 Équipe Image Processing

Dirigée par Olivier Gallinato, l'équipe Image Processing développe des algorithmes de segmentation semi-automatiques pour l'extraction de caractéristiques radiomiques. Le principe d'une segmentation semi-automatique est d'initialiser manuellement une région d'intérêt dans l'image, qui est amenée à croître par la fusion de voxels jusqu'à ce que toute la lésion d'intérêt soit couverte. Une fois la segmentation réalisée, des indicateurs (ou features) radiomiques de la tumeur sont calculés en utilisant l'Initiative de Standardisation de Biomarqueurs d'Images [Zwanenburg et al., 2020]. L'IBSI fournit une nomenclature et des définitions de biomarqueurs d'images normalisées reproductibles quelque soit l'étude radiomique.

### 2.3.2 Équipe IT

L'équipe software est chargée de la gestion et du développement de l'application d'analyse radiomique appelé **SOPHiA Radiomics**. Elle se fait ainsi appeler l'équipe SG IT Radiomics et comprends des ingénieurs Front et Back End. Le logiciel est aujourd'hui commercialisé, principalement à destination de cliniciens au sein des hôpitaux ou des laboratoires pharmaceutiques. Le service SG IT est chargé du logiciel et de son interface. Il est à la fois la partie finale dans la chaîne de développement du produit et le premier visuel qu'aura le client. Ainsi, même si l'équipe comprend principalement des Ingénieurs Software, elle est en lien direct avec les managers produits responsables de la communication auprès des clients.

Finalement, vis-à-vis de son rôle très concret quant au produit, le service est fortement amené à échanger avec les nombreux autres services, que ce soit avec les équipes présentes à Bordeaux telles que les Data sciences, ou encore celles à l'international comme en Suisse, pour la gestion de la sécurité et de l'accès aux données.

### 2.3.3 Équipe Biostatistics

L'Équipe Biostatistics au nombre de quatre, dirigée par Loïc Ferrer, soit l'équipe dans laquelle j'ai effectué mon stage, effectue les analyses radiomiques à partir de données multimodales, dont

notamment les informations extraites à partir des segmentations. Ces analyses s'expriment principalement à travers l'application de modèles de Machine Learning pour prédire un diagnostic (par exemple le type histologique d'une tumeur) et/ ou le pronostic (progression à la première évaluation d'un patient) des maladies graves comme le cancer (poumons, seins etc...). Un intérêt spécial est donné sur l'interprétation des modèles et l'explication des prédictions issues de ceux-ci, afin d'assurer une utilisation en routine clinique.

### 3 Etat de l'art

Ce chapitre introduit les principales notions et les méthodes statistiques utilisées durant mon stage : les différents coefficient de corrélation, les algorithmes de machine learning, les techniques de validation des modèles, leurs indicateurs de performance prédictive et finalement les outils pour l'interprétabilité et l'explicabilité des modèles d'apprentissage supervisé. Ces techniques sont présentées dans le cadre d'un problème de classification, c'est-à-dire lorsque l'on cherche à prédire une réponse binaire.

#### 3.1 Coefficients de corrélation et tests statistiques

En analyse radiomique, les caractéristiques radiomiques sont calculées à partir de données extraites d'imagerie médicale. L'ensemble des formules pour les calculs des variables sont standardisées par l'IBSI [Zwanenburg et al., 2020]. Au travers de la prise de connaissance de ceux-ci il a été rapidement constaté qu'il existe des liens intrinsèques entre énormément de variables ( certaines sont le carré d'autres ou des combinaisons de certaines en forment de nouvelles ...). Les caractéristiques radiomiques sont donc énormément corrélées et l'utilisation des coefficients de corrélations nous apporte une solution simple et efficace dans la sélection de ces variables. En outre les tests statistiques sont utilisés pendant l'analyse explicative : l'association bi-variée (entre une caractéristique radiomique et la variable d'intérêt) dans le but d'identifier si la variable a un effet discriminant suivant le statut de la réponse à prédire.

##### 3.1.1 Corrélation de Pearson

Le coefficient de corrélation linéaire simple, dit de Bravais-Pearson (ou de Pearson), mesure la liaison linéaire existant entre deux variables quantitatives aléatoires. C'est une normalisation de la covariance par le produit des écarts-type des variables. En revanche, il n'est plus adapté lorsque les dépendances entre variables sont non linéaires et non monotones.

Soit  $X = (X_i)_{1 \leq i \leq n}$  et  $Y = (Y_i)_{1 \leq i \leq n}$  deux variables aléatoires, ayant respectivement comme moyenne  $\bar{X}$  et  $\bar{Y}$ , alors la corrélation de Pearson est définie comme suit :

$$r_{xy} = \frac{Cov(X, Y)}{\sigma_x \sigma_y} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X))^2]} \sqrt{E[(Y - E[Y])^2]}}$$

et estimé selon :

$$r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

**Propriétés :**

1. Il est compris entre -1 et 1. Il est égal à 1 lorsque la liaison entre  $X$  et  $Y$  est linéaire, positive et parfaite, et est égale à -1 lorsque la liaison est linéaire négative.
2. Si  $X$  et  $Y$  sont totalement indépendants alors  $r = 0$ . La réciproque n'est en général pas vraie.

3. Il peut être égal à zéro alors qu'il existe une liaison fonctionnelle entre les variables : c'est le cas des liaisons non monotones.

### 3.1.2 Corrélation de Spearman

Fondamentalement, le coefficient de Spearman  $\rho$  est aussi un cas particulier du coefficient de Pearson, calculé à partir des transformations des variables originelles. Ce n'est autre que la corrélation de Pearson évaluée non pas sur les observations elles-mêmes mais sur les rangs des observations des deux variables. Il fait référence aux notions de paires discordantes ou concordantes, *i.e* les deux paires d'observations  $(x_i, y_i)$  et  $(x_j, y_j)$ , on dira que ces paires sont :

- concordantes si  $x_i < x_j$  et  $y_i < y_j$  ou  $x_i > x_j$  et  $y_i > y_j$ ,
- discordantes si  $x_i < x_j$  et  $y_i > y_j$  ou  $x_i > x_j$  et  $y_i < y_j$ .

Son intérêt vient du fait que des paires concordantes ou discordantes peuvent être créées par des dépendances linéaires ou non entre les variables. Par exemple si  $y$  est une transformée monotone mais non linéaire de  $x$  alors cette dépendance sera parfaitement détectée par la corrélation de Spearman alors qu'elle peut ne pas l'être par la corrélation de Pearson (du moins pas aussi finement). On préférera donc Spearman lorsque l'on pense que les dépendances ne sont pas linéaires ou/et que les variables ne sont pas gaussiennes.

Le coefficient de corrélation de Spearman cumule les bonnes qualités, il permet de traiter des variables intrinsèquement ordinaires (un indice de satisfaction, une appréciation ou une note attribuée, etc.) et il est très robuste face aux points aberrants même lorsque l'effectif est faible. Néanmoins il possède des limites. Lorsque la liaison n'est pas monotone il n'est pas opérant ; Il est estimé comme suit :

Soit  $R_i = \text{Rang}(x_i)$  et  $S_i = \text{Rang}(y_i)$ , correspondant respectivement au rang de l'observation  $x_i$  dans  $X$  et  $y_i$  dans  $Y$ , alors le  $\rho$  de Spearman est ni plus ni moins de le coefficient de Pearson calculé sur les rangs :

$$\hat{\rho} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}} = 1 - \frac{6 \sum_{i=1}^n (R_i - S_i)^2}{n(n^2 - 1)}, \quad \bar{R} = \bar{S} = \frac{n+1}{2}$$

### 3.1.3 Corrélation de Kendall

Le coefficient de corrélation  $\tau$  de Kendall est défini pour mesurer l'association entre variables ordinaires, typiquement des classement (ou rangs) affectés par des juges, aussi entre deux variables dont l'une est continue et l'autre ordinaire. Son champ d'application couvre donc parfaitement celui du coefficient  $\rho$  de Spearman mais n'est pas à proprement parlé une variante du coefficient de Pearson. Il repose sur un principe différent et s'interprète donc différemment. Son principe repose également sur la notion de paires discordantes et concordantes et s'interprète comme le degré de correspondance entre 2 classements (ou 2 notations). En l'absence de rangs égaux dans les deux variables, il s'agit simplement de l'écart entre le nombre de paires concordantes  $n_c$  et discordantes

$n_d$  rapporté au nombre total de paires d'observations  $n$ , soit :

$$\hat{\tau}_b = \frac{n_c - n_d}{\frac{1}{2}n(n - 1)}$$

Étant fondée comme la corrélation de Spearman sur la concordance et la discordance des paires d'observations, son emploi est recommandé dans le cas de dépendance non linéaire entre deux variables et/ou pour des variables non gaussiennes.

### 3.1.4 Le test de Wilcoxon-Mann-Whitney

Le cadre général est le suivant : On dispose de deux échantillons,  $X = x_1, x_2, \dots, x_{n_x}$  et  $Y = y_1, y_2, \dots, y_{n_y}$  constitué chacun de réalisations indépendantes de deux variables aléatoires ayant des fonctions de répartitions inconnues  $F_x$  et  $F_y$ . La question est celle de savoir si les deux variables aléatoires possèdent la même distribution. Les deux tests suivants, de Mann-Whitney et de la somme des rangs de Wilcoxon, sont couramment utilisés pour répondre à la question qui nous intéresse ici. Par ailleurs, même si au premier abord leur approche est différente, ils sont en fait équivalents : on n'a donc pas à privilégier l'un plutôt que l'autre.

La logique qui sous-tend ce test est très simple : supposons que l'on mélange les deux ensembles d'observations, et que l'on classe par ordre croissant les valeurs de l'échantillon ainsi créé alors, si  $H_0$  est vraie, nous devrions observer une alternance régulière tout au long du support de l'échantillon joint des valeurs prises dans  $X$  d'une part et dans  $Y$  d'autre part. En revanche, l'apparition de zones de concentration d'observations issues de  $X$  ou de  $Y$  est défavorable à l'hypothèse nulle. Une illustration de mélange possible d'échantillons est la suivante :

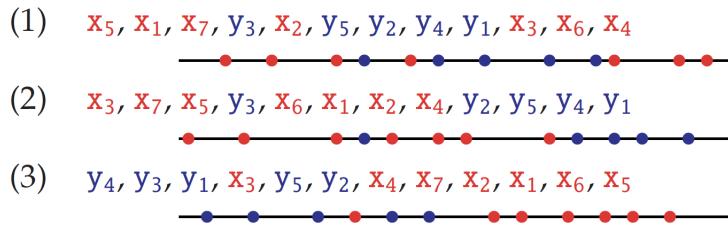


FIGURE 2 – Test de Mann-Whitney - Exemples de mélanges d'échantillons

Le test de Mann-Whitney est simplement le nombre de fois où un  $y$  précède un  $x$  dans l'échantillon mélangé et classé, soit :

$$U = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \mathbf{1}(x_i > y_j) \quad (1)$$

Si on note  $W_x$  (resp.  $W_y$ ) la somme des rangs des observations de  $X$  (resp. de  $Y$ ) dans l'échan-

tillon mélangé, on peut montrer que la statistique  $U$  de (1) est encore donnée par

$$U = \min(U_x, U_y), \quad \text{avec}$$

$$\begin{aligned} U_x &= n_x n_y + \frac{n_x(n_x + 1)}{2} - W_x, \quad \text{et} \\ U_y &= n_x n_y + \frac{n_y(n_y + 1)}{2} - W_y \end{aligned}$$

Lorsque  $n_x$  et  $n_y$  sont suffisamment grands,  $U$  suit une loi gaussienne de moyenne  $\frac{n_x n_y + 1}{2}$  et de variance  $\frac{n_x n_y (n_x + n_y + 1)}{12}$  et le test se ramène au test de la moyenne d'un échantillon gaussien.

### 3.1.5 Le test de Kruskal-Wallis

Le test de Kruskal-Wallis s'applique en présence d'au moins trois échantillons constitués d'individus indépendants et la question posée est celle de la similarité des distributions dont ils sont issus. Le plus fréquemment, les échantillons correspondent aux catégories d'une variable qualitative, par exemple aux découpages d'un échantillon par tranches d'âges, de revenus, de catégories socio-professionnelles, etc... Ce test va réaliser une analyse de la variance sur des transformées (scores) des observations initiales. Il généralise le test de Wilcoxon-Mann-Whitney, qui est utilisé pour comparer seulement deux groupes. Lorsque ce test entraîne des résultats significatifs alors au moins un des échantillons est différent des autres. Cependant ce test n'identifie pas où se trouve cette différence. Pour effectuer ce test, la méthode est de classer toutes les observations de tous les groupes ensemble, c'est-à-dire classer les données de 1 à  $N$  sans faire de groupes. Attribuer à toute valeur liée la moyenne des classements qu'ils auraient obtenus s'ils n'avaient pas été liés. La statistique de test est donnée par :

$$H = (N - 1) \frac{\sum_{i=1}^g n_i (\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^g \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2}, \quad \text{avec}$$

- $n_i$  est le nombre d'observations dans le groupe  $i$
- $r_{ij}$  est le rang (parmi toutes les observations) de l'observation  $j$  du groupe  $i$
- $g$  le nombre total d'échantillons (groupes) à comparer
- $N$  le nombre total d'observations sur l'ensemble des groupes
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$  est le rang moyen de toutes les observations du groupe  $i$ .
- $\bar{r} = \frac{1}{2}(N + 1)$  est la moyenne de tous les  $r_{ij}$

Et si les données ne contiennent pas de lien, on a alors une expression simplifiée de la statistique de test :

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^g n_i \bar{r}_{i\cdot}^2 - 3(N + 1)$$

La loi de  $H$  pour des effectifs plus importants ( $n_i \geq 5, \forall i$ ) est approchée par une loi du chi-deux à  $g - 1$  degrés de liberté.

## 3.2 Algorithmes de Machine Learning

Le Machine Learning (ML) utilise des algorithmes programmés qui apprennent et optimisent leurs paramètres à partir des données d'entrée, afin de prédire les valeurs de sortie avec la meilleure performance possible. De très nombreux algorithmes de ML ont été développés dans la littérature et dans les logiciels. Je vais présenter certains d'entre eux que j'ai utilisé au cours de mon stage. On peut notamment se focaliser sur 3 classes d'algorithmes : les modèles de régression paramétrique avec technique de régularisation, les algorithmes à base d'arbres de décision et les machines à vecteur support.

### 3.2.1 Méthodes Linéaires Généralisées avec Régularisation

La relation entre une variable réponse binaire et plusieurs variables explicatives est modélisée traditionnellement par la régression logistique où on prédit la probabilité  $p_i$  de survenue de l'évènement pour le sujet  $i$ . Les estimateurs basés sur la vraisemblance pour l'estimation des risques, sont instables et ont une grande variance quand  $n < p$  ( $n$  : nombre d'individus,  $p$  : nombre de variables explicatives), ou en cas de colinéarité entre les variables explicatives [Avalos, 2009]. Compte tenu du fait que dans les applications cliniques le nombre d'individus est assez faible, les méthodes de régularisation sont envisagées pour solutionner l'instabilité du modèle, éviter le surapprentissage<sup>1</sup> et améliorer la précision des prédictions.

**La régularisation Ridge** ajoute une contrainte sur les coefficients lors de la modélisation pour contrôler l'amplitude de leurs valeurs. La régression Ridge atteint ses meilleures performances de prédiction grâce à un compromis biais-variance. Cependant, on ne peut pas produire un modèle parcimonieux, car la régression conserve tous les prédicteurs. Soit  $S(\beta z_i) = \frac{1}{1+e^{\beta z_i}}$ , l'estimateur Ridge pour une régression logique est défini par :

$$\hat{\beta}_\lambda = \operatorname{argmax}_{\beta} \left( \sum_{i=1}^n y_i \log(S(\beta z_i)) + \sum_{i=1}^n (1 - y_i) \log(1 - S(\beta z_i)) - \lambda \sum_{j=1}^p \beta_j^2 \right)$$

où  $\lambda$  ( $\lambda \neq 0$ ) est un paramètre à fixer qui permet de contrôler l'impact de la pénalité,  $y_i$  est la variable réponse,  $z_{ij}$  sont les variables explicatives  $x_{ij}$  centrées et réduites et  $\beta$  est le vecteur des coefficients à estimer.

**La régularisation Lasso** peut être considérée comme une technique de sélection de variables, car elle réduit certains coefficients à 0. Le modèle essaie donc de conserver les caractéristiques essentielles pour la prédiction. Les limites de la régularisation LASSO sont rencontrées dans les problèmes de très grande dimension ( $n < p$ ), car lasso sélectionne au plus  $n$  variables avant de saturer le modèle. Les coefficients Lasso sont obtenus en pénalisant la log-vraisemblance par la valeur absolue des estimateurs :

---

1. Trop grande capacité à capturer des informations, et difficulté à généraliser les caractéristiques des données

$$\hat{\beta}_\lambda = \operatorname{argmax}_{\beta} \left( \sum_{i=1}^n y_i \log(S(\beta z_i)) + \sum_{i=1}^n (1 - y_i) \log(1 - S(\beta z_i)) - \lambda \sum_{j=1}^p |\beta_j| \right)$$

**La régularisation Elastic Net** contient à la fois les régularisations Lasso et Ridge. Elastic Net gère les problèmes  $n << p$  dans la sélection de variables par l'exclusion de variables non pertinentes (propriété conservée de la régularisation Lasso). En plus, elle partage le poids pour les groupes de variables prédictives corrélées (propriété conservée de la régularisation Ridge). On cherche à trouver les coefficients  $\beta_{\lambda,\alpha}$  qui résolvent :

$$\hat{\beta}_{\lambda,\alpha} = \operatorname{argmax}_{\beta} \left( \sum_{i=1}^n y_i \log(S(\beta z_i)) + \sum_{i=1}^n (1 - y_i) \log(1 - S(\beta z_i)) - \lambda [ \alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 ] \right)$$

Les paramètres  $\alpha$  et  $\lambda$  sont à définir. Si  $\alpha = 0$  on retrouve la régression Ridge, et si  $\alpha = 1$  on retrouve la régression Lasso. Tous les modèles régularisés sont basés sur des variables standardisées, étape considérée dans le package `glmnet` de R. Lorsque les coefficients finaux sont affichés, ils sont ramenés à leur échelle d'origine pour faciliter leur interprétation.

### 3.2.2 Méthodes à bases d'arbres de décision

Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. Les arbres de décision binaire évitent les hypothèses restrictives des modèles de régression paramétrique en appliquant de manière itérative des décisions binaires simples et en les combinant pour obtenir un meilleur résultat de prédiction. Cependant, il est connu que ces techniques sont sujets au surapprentissage parce que l'algorithme peut capter des relations non-linaires trop complexes entre les prédicteurs et la variable réponse. C'est pourquoi des méthodes de bagging ou de boosting ont été proposées. Les algorithmes Bagging créent un modèle à partir de l'agrégation des classificateurs construits indépendamment, alors que dans les algorithmes Boosting chaque classificateur est construit de manière séquentielle en tenant compte les résultats des classificateurs précédents [Mayr et al., 2014]. Les modèles présentés ci-dessous peuvent être appliqués à différents types de classifieur, ici nous parlerons de l'application avec les arbres de décisions.

#### C5-trees :

Une approche des arbres de classification est le modèle C5.0, qui est la version améliorée du C4.5 [Quinlan, 1993]. Ici, le critère de séparation des noeuds est basé sur l'Entropie [Shannon, 1948]. Soit  $c$  le nombre de classes et  $p_i$  la probabilité de tomber dans la classe  $i$  alors on a :

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Compte tenu de la mesure de pureté, l'algorithme doit encore décider de la caractéristique sur laquelle il faut effectuer le fractionnement. Pour ce faire, l'algorithme utilise l'entropie pour calculer le changement d'homogénéité résultant d'une division sur chaque caractéristique possible. Ce calcul est appelé gain d'information. Le gain d'information pour une caractéristique  $F$  est calculé comme étant la différence entre l'entropie dans le segment avant la division ( $S1$ ) et les partitions résultant de la division ( $S2$ ), c'est-à-dire :

$$gain(split) = Entropy(S1) - Entropy(S2).$$

Une complication est qu'après un fractionnement, les données sont divisées en plus d'une partition (deux dans le cas binaire). Par conséquent, la fonction de calcul de l' $Entropy(S2)$  doit prendre en compte l'entropie totale de toutes les partitions. Pour ce faire, on pondère l'entropie de chaque partition par la proportion d'individus tombant dans cette partition, ce qui peut être exprimé par la formule suivante :

$$Entropy(S2) = \sum_{i=1}^n w_i Entropy(P_i),$$

avec  $w_i$  la proportion d'individus tombant dans la partition  $P_i$  après la division.

Plus le gain d'information est élevé, plus une caractéristique est apte à créer des groupes homogènes après une partition sur cette caractéristique. D'où la caractéristique ayant le gain d'information le plus élevé est choisi pour prendre la décision. Les fractionnements avec des gains d'information plus importants sont plus attractifs que ceux avec des gains plus faibles. L'un des avantages de l'algorithme C5.0 est qu'il n'a pas d'opinion sur l'élagage ; il prend automatiquement de nombreuses décisions en utilisant des valeurs par défaut assez raisonnables. Sa stratégie globale consiste à post-élaguer l'arbre. Pour ce faire, il fait d'abord croître un grand arbre qui surajuste les données d'apprentissage. Ensuite, les noeuds et les branches qui ont peu d'effet sur les erreurs de classification sont supprimés. En outre dans le cas du traitement des valeurs manquantes, C5.0 permet soit d'estimer les valeurs manquantes en fonction d'autres attributs, soit de répartir statistiquement le cas parmi les résultats.

### **Gradient Boosting Machine (cas particulier des gradients boosting trees) :**

Le Gradient Boosting Machine (Boosting Trees) entraîne de nombreux modèles de manière progressive, additive et séquentielle, en utilisant des gradients pour minimiser la fonction de perte<sup>1</sup>, ainsi que pour identifier les déficiences prédictives des arbres construits à chaque étape. En particulier, GBM construit des modèles de régression en ajustant de manière séquentielle une fonction paramétrée simple aux *pseudo-résidus* actuels par moindres carrés. Les *pseudo-résidus* correspondent à la différence entre les valeurs observées et les probabilités prédites dans l'étape précédent pour chaque individu de la base d'apprentissage [Jerome, 2002] .

Soit  $\hat{y}_i^t$  la prédiction de l'observation  $i$  dans l'itération  $t$ , et  $f_t$  l'arbre construit dans l'itération  $t$ . Le modèle est entraîné de manière additive, en minimisant la fonction objectif suivante :

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2)$$

1. La fonction de perte indique dans quelle mesure les coefficients du modèle sont bons pour ajuster les données. Pour la classification, elle mesure la capacité du modèle prédictif pour classer les individus.

où  $l$  correspond à la fonction de perte ( $-(LogVraisemblance)$ ) pour les problèmes de classification),  $f_t(x_i)$  représente la prédiction de l'individu  $i$  par l'arbre de l'itération  $t$ ,  $\Omega$  une fonction en termes de  $log(odds)$  qui tient compte des résidus précédents.

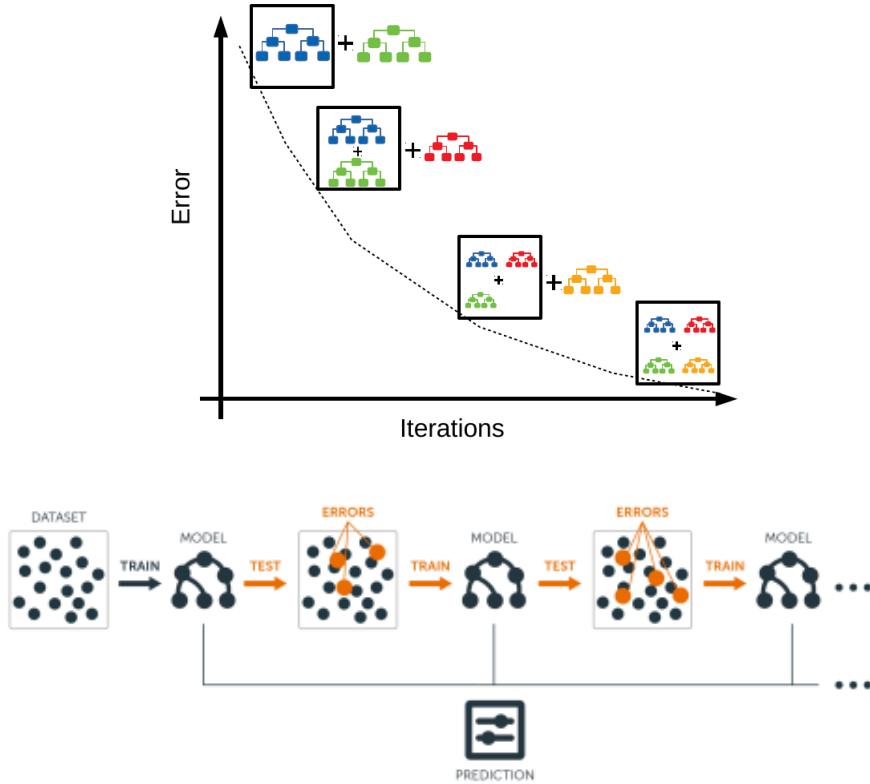


FIGURE 3 – Approche du Gradient Boosting

L'algorithme améliore les performances de manière itérative en considérant les observations caractérisées par des grands résidus calculés dans les itérations précédentes [Mayr et al., 2014]. Cet algorithme est implémenté sous R avec la librairie `gbm`.

#### XGBoost :

L'algorithme XGBoost (eXtreme Gradient Boosting) est une alternative qui reprend la théorie du Gradient Boosting pour obtenir des résultats supérieurs en utilisant moins de ressources informatiques. Cet algorithme d'apprentissage automatique est très efficace et largement utilisé en raison de sa grande précision de prédiction. Le facteur le plus important du succès de XGBoost est son adaptabilité dans tous les scénarios. Enfin, le système fonctionne plus de dix fois plus vite que les solutions courantes existantes sur une seule machine [Chen et al., 2019].

XGBoost n'a pas besoin de normaliser les données (dû à sa structure arborescente). L'idée principale de l'algorithme, comme pour GBM, est de combiner une série de classificateurs simples (les

arbres de décisions) avec une faible précision pour créer un classifieur puissant avec de meilleures performances de classification [Chen et al., 2019]. Sous R, cet algorithme peut être exécuté par la librairie `xgboost`.

### AdaBoost :

AdaBoost (Adaptive Boosting) est une technique de boosting très populaire qui vise à combiner plusieurs classificateurs faibles pour construire un classificateur fort. L'article original d'AdaBoost a été écrit par Yoav Freund et Robert Schapire [Freund and Schapire, 1996]. Un seul classificateur peut ne pas être en mesure de prédire avec précision la classe d'un objet, mais lorsque nous regroupons plusieurs classificateurs "faibles"<sup>1</sup> avec chacun apprenant progressivement des objets mal classés des autres, nous pouvons construire un modèle aussi fort. Le classificateur mentionné ici peut être n'importe lequel des classificateurs de base, des arbres de décision (souvent par défaut) à la régression logistique, etc.

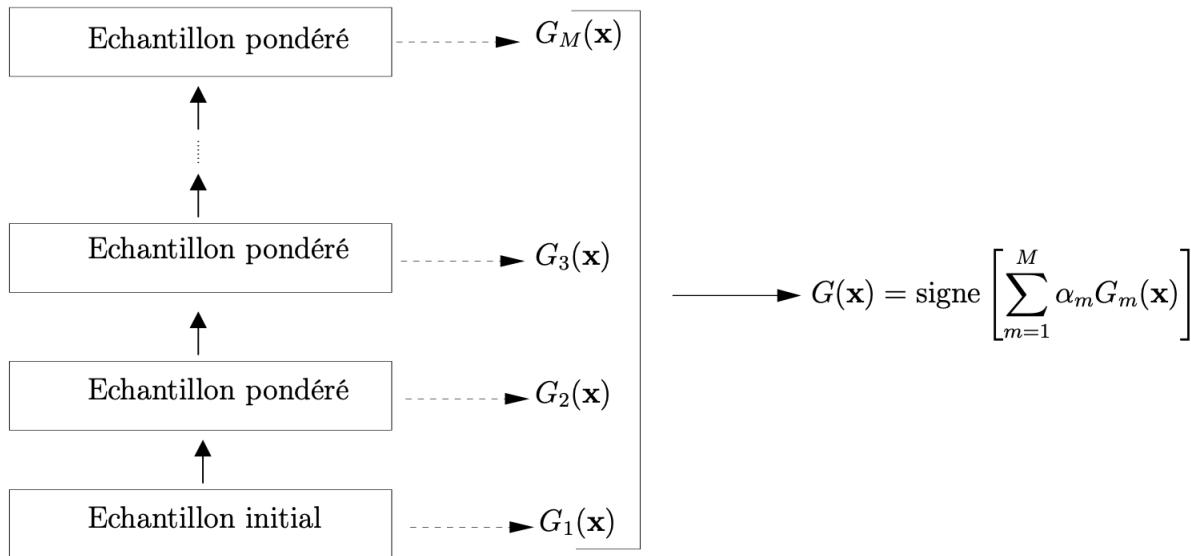


FIGURE 4 – Représentation de l'algorithme Adaboost [Trevor Hastie, 2009]

Plutôt que d'être un modèle en soi, AdaBoost peut être appliqué sur n'importe quel classificateur pour apprendre de ses lacunes et proposer un modèle plus précis. On désigne par  $g(x)$  une règle de classification "faible", l'idée consiste à appliquer la règle du boosting plusieurs fois en affectant "judicieusement" un poids différent aux observations à chaque itération (Voir figure 5). Les poids de chaque observation sont initialisés à  $\frac{1}{n}$  pour l'estimation du premier modèle. Ils sont ensuite mis à jour pour chaque itération. L'importance d'une observation  $w_i$  est inchangée si l'observation est bien classée, dans le cas inverse elle croît avec la qualité d'ajustement du modèle mesurée par  $\alpha_m$ . L'agrégation finale est une combinaison des règles  $g_1, \dots, g_m$  pondérée par les qualités d'ajustement de chaque modèle.

1. un classificateur qui fonctionne mieux que la devinette aléatoire, mais qui fonctionne toujours mal pour désigner des classes aux objets.

---

**Algorithm 1 AdaBoost**


---

**Entrée :**

- $\mathbf{x}$  l'observation à prévoir
- $d_n = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  l'échantillon
- Une règle faible
- $M$  le nombre d'itérations.

1. Initialiser les poids  $w_i = 1/n$ ,  $i = 1, \dots, n$

2. Pour  $m = 1$  à  $M$  :

- (a) Ajuster la règle faible sur l'échantillon  $d_n$  pondéré par les poids  $w_1, \dots, w_n$ , on note  $g_m(\mathbf{x})$  l'estimateur issu de cet ajustement
- (b) Calculer le taux d'erreur :

$$e_m = \frac{\sum_{i=1}^n w_i \mathbf{1}_{y_i \neq g_m(\mathbf{x}_i)}}{\sum_{i=1}^n w_i}.$$

- (c) Calculer :  $\alpha_m = \log((1 - e_m)/e_m)$

- (d) Réajuster les poids :

$$w_i = w_i \exp(\alpha_m \mathbf{1}_{y_i \neq g_m(\mathbf{x}_i)}), \quad i = 1, \dots, n$$

3. Sortie :  $\hat{g}_M(\mathbf{x}) = \sum_{m=1}^M \alpha_m g_m(\mathbf{x})$ .

FIGURE 5 – Etape de l'algorithme Adaboost [Freund and Schapire, 1996]

### 3.2.3 Machine à Vecteur Support (SVM)

La Machine à Vecteur Support (SVM) est un algorithme réputé pour sa précision et sa capacité à traiter des données de grande dimension. Les SVM appartiennent à la catégorie des méthodes à noyaux (kernel).

Ce modèle représente les observations de l'espace des prédicteurs X à travers d'un hyperplan qui sépare de manière optimale les observations de chaque classe [Cortes and Vapnik, 1995].

Dans la figure 6, les points situés sur les frontières sont appelés Vecteurs Support, et au milieu de la marge est situé l'hyperplan de séparation optimale. R propose la librairie e1071.

**Remarque :** Tous les algorithmes d'apprentissage étudiés peuvent également être obtenus sous R à l'aide de la librairie caret, en précisant dans la fonction train, l'algorithme d'intérêt, `method = c("gbm", "xgbTree", "svmLinear", "C5.0", etc...)`.

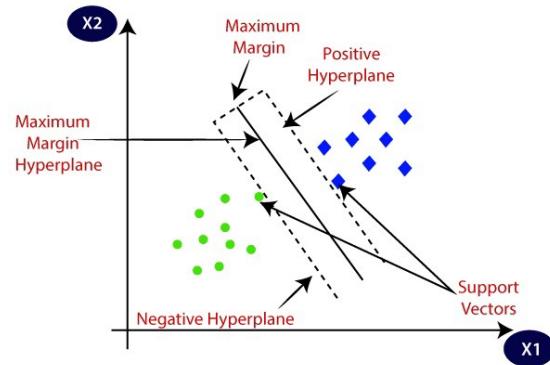


FIGURE 6 – Classification et séparation linéaire

## 3.3 Évaluation des capacités prédictives

### 3.3.1 La Validation croisée Leave-One-Out (LOOCV)

La Validation croisée Leave-One-Out est un cas particulier de validation croisée où le nombre de folds est égal au nombre d'instances dans l'ensemble de données. Ainsi, l'algorithme d'apprentissage est appliqué une fois pour chaque instance, en utilisant toutes les autres instances comme ensemble d'apprentissage.

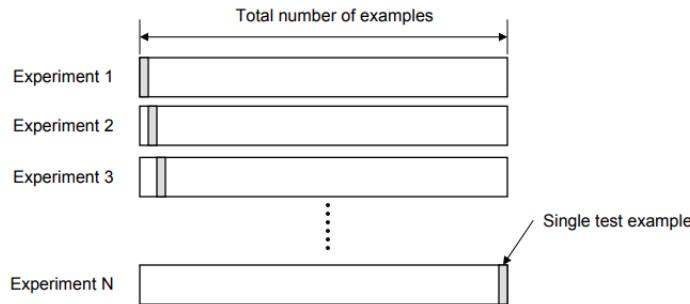


FIGURE 7 – La Validation croisée Leave-one-out

En effet, ici l'ensemble test est un seul élément, et l'ensemble d'apprentissage est l'ensemble de tous les autres éléments. Le modèle à chaque itération effectue la recherche de hyperparamètres optimaux suivant une métrique donnée (Sensibilité, PRAUC, AUC, F1-Score, etc...) sur l'ensemble

d'apprentissage. Ainsi les hyperparamètres optimaux sont choisis et le modèle entraîné est utilisé pour prédire l'ensemble test. Le modèle final est le modèle avec les hyperparamètres optimaux qui aura maximisé la métrique d'intérêt. La métrique d'intérêt étant séparément calculée pour chaque étape de la validation, les résultats sont ensuite moyennés afin d'estimer la métrique de performance globale.

Cette méthode est très utilisée dans les analyses où l'on dispose de très peu de patients/individus et que l'on veut néanmoins éviter de faire du sur-apprentissage, même si elle reste cependant légèrement biaisée car l'apprentissage se fait sur les données qui ont servi à la recherche des hyperparamètres optimaux. C'est dans cette optique que la Validation Croisée Emboîtée (Nested-Cross-Validation), qui est expliquée ci-dessous, est utilisée dans un but de validation des résultats prédictifs.

### 3.3.2 La Validation croisée Emboîtée

La Validation croisée Emboîtée ou Nested-Cross-Validation est une approche, de l'optimisation des hyperparamètres des modèles et de la sélection des modèles, qui tente de surmonter le problème du surapprentissage de l'ensemble de données d'apprentissage. Cette image issue de la documentation du package `mlr` illustre cette méthode.

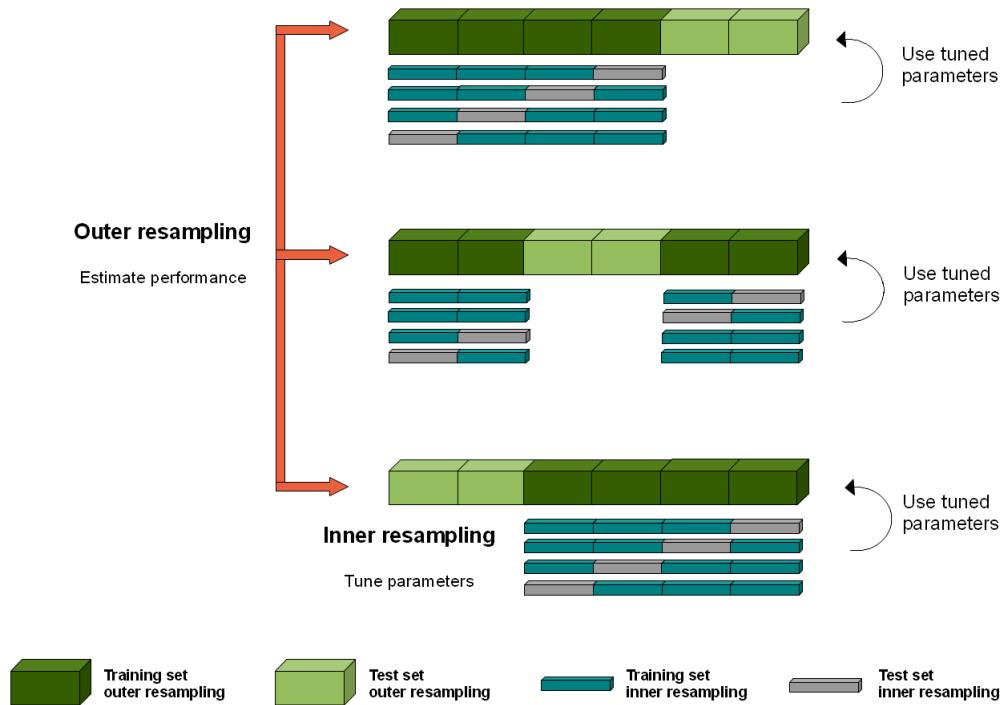


FIGURE 8 – La Validation croisée Emboîtée

Afin de surmonter le biais dans l'évaluation des performances, la sélection du modèle doit être

considérée comme une partie intégrante de la procédure d'ajustement du modèle, et doit être menée indépendamment dans chaque essai afin d'éviter le biais de sélection et parce qu'elle reflète la meilleure pratique dans l'utilisation opérationnelle [Cawley and Talbot, 2010].

La procédure est subdivisée en deux sous-validations croisées k-fold, une interne qui sert à traiter l'optimisation des hyperparamètres du modèle et une externe qui consiste en l'évaluation de la qualité du modèle. Ainsi, la procédure de validation croisée k-fold pour l'optimisation des hyperparamètres du modèle est imbriquée dans la procédure de validation croisée k-fold pour la sélection du modèle. L'utilisation de deux boucles de validation croisée permet également d'appeler la procédure "double validation croisée" ou "validation croisée imbriqué".

Typiquement, la procédure de validation croisée k-fold implique l'ajustement d'un modèle sur tous les folds sauf un et l'évaluation du modèle ajusté sur le fold retenu. Nous appellerons l'ensemble des folds utilisés pour former le modèle "ensemble de données de formation" et le fold retenu "ensemble de données de test". Chaque ensemble de données d'entraînement est ensuite fourni à une procédure d'optimisation des hyperparamètres, telle que la recherche sur grille ou la recherche aléatoire, qui trouve un ensemble optimal d'hyperparamètres pour le modèle. L'évaluation de chaque ensemble d'hyperparamètres est effectuée à l'aide de la validation croisée à k-fold qui divise l'ensemble de données d'entraînement fourni en k folds, et non l'ensemble de données original.

Néanmoins un inconvénient de la validation croisée imbriquée est l'augmentation considérable du nombre d'évaluations de modèles effectuées. Si  $n * k$  modèles sont ajustés et évalués dans le cadre d'une recherche traditionnelle d'hyperparamètres de validation croisée pour un modèle donné, ce nombre passe à  $k * n * k$  car la procédure est ensuite exécutée  $k$  fois de plus pour chaque fold de la boucle externe de la validation croisée imbriquée. Pour être concret, vous pouvez utiliser  $k = 5$  pour la recherche d'hyperparamètres et tester 100 combinaisons d'hyperparamètres de modèle. Une recherche traditionnelle d'hyperparamètres permettrait donc d'ajuster et d'évaluer  $5 * 100$  ou 500 modèles. La validation croisée imbriquée (nested) avec  $k = 10$  folds dans la boucle externe adapterait et évaluerait 5000 modèles.

### 3.4 Critères de performances prédictives

Les critères de performances prédictives permettent de comparer plusieurs algorithmes. Le choix du critère dépend de la problématique et du type de données. Nous présentons quelques critères utilisés dans cette étude, nous sommes dans le cadre d'une classification binaire.

**Matrice de Confusion :** Une matrice de confusion est utilisée pour avoir une image complète de la performance d'un modèle. En considérant deux classes la positive et la négative, aussi en considérant FN comme le nombre de faux négatifs, VP comme le nombre de vrais positifs, VN comme le nombre de vrais négatifs et FP comme le nombre de faux positifs, la matrice de confusion est définie de la manière suivante :

		Prédit	
		Positif	Négatif
Observé	Positif	VP	FN
	Négatif	FP	VN

TABLE 1 – Matrice de confusion

Les indicateurs suivants sont communément utilisés pour évaluer la performance des modèles de classification en utilisant la matrice de confusion ou en utilisant les probabilités d'appartenir à une classe. On y distingue donc :

**Précision globale :** C'est le taux d'observations bien classées dans l'échantillon total.

$$\text{ACC} = \frac{VP + VN}{VP + FP + VN + FN}$$

**Sensibilité :** C'est le taux de vrais positifs parmi les cas positifs.

$$\text{Sensibilité} = \frac{VP}{VP + FN}$$

**Spécificité :** C'est le taux de vrais négatifs parmi les cas négatifs.

$$\text{Spécificité} = \frac{VN}{VN + FP}$$

**Précision :** C'est le taux de vrais positifs parmi les cas prédis positifs.

$$\text{Précision} = \frac{VP}{VP + FP}$$

**Valeur prédictive négative :** C'est le taux de vrais négatifs parmi les cas prédis négatifs.

$$\text{VPN} = \frac{VN}{VN + FN}$$

**F-beta scores :** C'est un indicateur hybride utilisé pour les classes non-balancées.

$$\text{F}_\beta\text{Score} = \frac{(1 + \beta^2)VP}{(1 + \beta^2)VP + \beta^2FN + FP}$$

Les plus utilisés seront le F1score ( $\beta = 1$ ) et F2score ( $\beta = 2$ ).

**Brier-Score :** Le score de Brier est une fonction de score qui évalue l'exactitude des prédictions probabilistes. Pour les prédictions unidimensionnelles, il est strictement équivalent à l'erreur quadratique moyenne appliquée aux probabilités prédites.

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2,$$

avec  $p_i$  la probabilité prédite de la classe positive de l'observation réelle  $o_i$ .

### 3.5 Interprétabilité globale

Les algorithmes d'apprentissage utilisés comme outil d'aide à la décision peuvent nécessiter l'adhésion de l'humain, en particulier lorsque les domaines d'application sont critiques comme ici le cas de la médecine. La compréhension de l'algorithme automatique peut jouer un rôle fondamental dans la prise en main et la mise en place d'une collaboration homme – machine et limiter la résistance au changement numérique [John-Mathews, 2019]. Cette partie a pour but de nous aider à mieux comprendre le modèle sélectionné. Interprétabilité est le degré auquel un humain peut comprendre la cause d'une décision [Tim, 2017] ou c'est la mesure dans laquelle un humain peut prédire de manière cohérente le résultat du modèle [Been et al., 2016].

L'interprétation globale résume donc une interprétation de façon globale du modèle, plus explicitement, permet de savoir quelles sont les variables qui impactent le plus notre modèle et dans quelle mesure elles jouent un effet dans le calcul des prédictions. On y distingue comme stratégies :

#### 3.5.1 Importance des variables :

C'est une mesure qui détermine comme son nom l'indique l'importance d'une variable dans le modèle de machine learning, à quel point une variable est impliquée dans la construction du modèle de prédiction. L'idée du calcul de cette mesure repose sur un principe qui demeure intuitif. En se basant sur une métrique d'évaluation de performance prédictive ( AUC, FBeta-Score, etc.), on cherche à opposer les performances du modèle en prédiction avec et sans la variable à évaluer. Pour neutraliser l'effet de la variable, les valeurs de celle-ci sont aléatoirement perturbées. Ainsi, on casse le lien qu'elle peut entretenir avec la classe à prédire (et les autres variables par la même occasion). Le modèle est alors évalué à chaque permutation. Plus les performances sont dégradées plus la variable est importante.

#### 3.5.2 Graphique de dépendance partielle (PDP)

C'est une méthode d'interprétation globale qui donne une idée générale de la relation entre l'effet d'une variable et les prédictions faites par le modèle. Le PDP montre l'effet marginal d'une ou deux variables sur le résultat prédit du modèle. L'idée générale qui sous-tend la construction d'un PDP est de représenter graphiquement le changement de réponse moyen du modèle suite à une petite variation d'une ou deux variables. Si la variation concerne seulement une variable, on représente la dérivée partielle du classifieur en fonction de la variable en entrée. Si la variation concerne deux variables, une visualisation par "heat-map" peut être envisagée pour visualiser l'impact de l'interaction entre les deux variables sur la variable en output (dérivée partielle croisée). En bonus avec un PDP nous pouvons montrer le comportement non linéaire ou non monotone de la variable réponse.

### 3.5.3 Graphique d'Espérance Individuelle Conditionnelle (ICE)

C'est l'équivalent du PDP pour les individus. Le graphique d'ICE affiche une ligne par individus qui montre comment la prédiction de l'individu change lorsqu'une valeur de la variable change.

### 3.5.4 Effets Locaux Accumulés (ALE)

Les ALE décrivent comment les caractéristiques influencent en moyenne la prédiction avec le modèle. Les graphiques ALE sont une alternative plus rapide et non biaisée aux PDP à partir du moment où les variables sont corrélées. En effet, les PDP prennent comme hypothèse l'indépendance des variables (ce qui n'est pratiquement jamais évident sur un jeu de données), l'ALE lui est calculé à partir de la distribution conditionnelle des variables (ne tient pas compte de l'indépendance des variables) d'où sa caractéristique de "protocole robuste". L'ALE montre comment la prédiction change localement lorsque la variable varie. Il est plus utilisé dans un but de validation du graphique PDP.

### 3.5.5 Modèle de substitution global (Surrogate models)

Un modèle de substitution global est un modèle interprétable qui est formé pour se rapprocher des prédictions de notre modèle supposé complexe de base avec une "boîte noire". Nous remplaçons la boîte noire par un modèle plus simple : un arbre de décision, des modèles linéaires etc ... Sur ce modèle de substitution le  $R^2$  détermine son efficacité ou non. Si le  $R^2$  se rapproche de 1, la boîte noire peut être expliquée par le modèle de substitution, mais si il se rapproche de 0 cela signifie que le modèle de substitution n'approxime pas très bien la boîte noire.

## 3.6 Interprétabilité Locale

En complément de l'interprétabilité globale, qui explique au niveau populationnel l'estimation d'un modèle de prédiction, l'interprétabilité locale explique une prédiction selon un ensemble de variables. Ainsi, elle est utilisée pour justifier des prédictions individuelles. La stratégie utilisée dans cette étude est décrit ci-dessous.

### 3.6.1 Les modèles de substitution locale (LIME)

Les modèles de substitution locaux sont des modèles interprétables qui sont utilisés pour expliquer les prédictions individuelles des modèles d'apprentissage automatique de type boîte noire. L'algorithme de cette méthode crée un modèle autour d'une prédiction donnée afin de l'approximer localement. Contrairement au modèle de substitution globale, LIME se focalise sur l'apprentissage d'un modèle de substitution pour expliquer les prédictions individuelles. La procédure pour entraîner un modèle se déroule suivant les étapes suivantes :

- Sélection de l'individu d'intérêt pour lequel nous voulons avoir une explication de la prédiction de la boîte noire.
- Perturbation de l'ensemble des données et calcul des prédictions avec le modèle de la boîte noire pour ces nouveaux individus simulés.
- Pondération des nouveaux échantillons en fonction de leur proximité avec l'individu d'intérêt.
- Entraînement du modèle pondéré et interprétable sur l'ensemble de données avec les variations.
- Explication de la prédiction par interprétation du modèle local.

## 4 Robustesse des caractéristiques radiomiques

### 4.1 Contexte et motivation

Après l'extraction des données radiomiques issues des images médicales par l'équipe d'Image Processing, c'est au tour de l'équipe de Biostatistiques de prendre le relai à travers la réalisation d'une étude statistique : l'analyse radiomique. Les langages utilisés pour le traitement des données et l'analyse statistique sont R et Python.

Le but est de prédire un "outcome" ou "variable d'intérêt" (progression du statut d'une tumeur, analyse de survie de patients...) en utilisant les données radiomiques à l'aide de modèles d'apprentissage supervisé. Cette analyse statistique standardisée, se déroule suivant des étapes bien définies que sont :

1. Configuration de l'environnement et importation des données
2. Gestion des données brutes
3. Analyse explicative
4. Séparation du jeu de données
5. Sélection de variables ( Réduction de dimensionnalité )
6. Prédiction
7. Interprétabilité globale
8. Interprétabilité locale
9. Conclusion

La partie de la sélection de variables est une partie tout aussi complexe qu'importante dans l'analyse radiomique. Les caractéristiques radiomiques sont des caractéristiques quantitatives d'images médicales qui décrivent une maladie. Elles sont obtenues suivant un processus mettant en lien l'acquisition de l'image médicale, la délimitation ou segmentation de la tumeur qui peut être à la fois manuelle ou semi-automatique et l'extraction des caractéristiques. Les caractéristiques radiomiques extraites des segmentations sont nombreuses, diverses et variées, cependant, elles sont soumises à une standardisation d'un organisme : IBSI [Zwanenburg et al., 2020]. En effet, l'utilisation des caractéristiques radiomiques comme biomarqueurs de la réponse au traitement d'un patient exige que les caractéristiques calculées soient robustes et reproductibles. Cette préoccupation nous conduit au fait qu'aucune sélection de variables ne peut être effectuée si les variables en amont ne sont pas d'une solidité (robustesse) et d'une pertinence avérée. Ainsi l'équipe *Radiomics – Research* souhaite intégrer la sélection de variable robustesse avant l'utilisation d'une méthodologie de sélection de variables.

Une segmentation est la représentation géométrique d'une structure anatomique visible sur une image médicale. La problématique du processus de segmentation ou de la validation d'une segmentation est la disparité des résultats obtenus par un même spécialiste ou plusieurs, pour une même structure anatomique cible. En fonction de la qualité de l'image, des compétences radiologiques, de la spécialité et des objectifs de l'opérateur, voire du moment, les résultats de segmentation manuelle ou de validation d'une segmentation peuvent être très différents.

On y distingue ainsi trois types de variabilités :

1. Variabilité due aux logiciels : Suivant les versions et les logiciels utilisés les résultats des segmentation différent. Ainsi les features extraites peuvent être sujettes à des fluctuations.
2. Variabilité inter-opérateur : Suivant le clinicien, les délimitations des lésions tumorales peuvent différer.
3. Variabilité intra-opérateur : Le même clinicien à différents temps fournit des segmentations différentes.

On peut aussi rajouter à la liste des variabilités celle qui intervient sur l'acquisition des données brutes i.e la variabilité due aux centres ; Certaines cohortes de patients peuvent avoir la même pathologie mais peuvent provenir de différentes centres, ainsi les différentes méthodes de scanners/IRM peuvent influer sur les images obtenues donc sur les segmentations. Il n'y a pas de vérité en matière de segmentation. Il est donc indispensable de mesurer la variabilité inter-opérateur et d'évaluer l'impact sur le calcul des caractéristiques radiomiques, sur leur robustesse, puis sur les résultats des analyses prédictives, objectif ultime de la Radiomics. Or le processus de segmentation par un opérateur humain est souvent un travail extrêmement coûteux en temps et il est généralement réalisé par plusieurs experts uniquement sur des cohortes modestes afin d'étudier cette variabilité. L'automatisation de perturbations de segmentation est donc un outil très utile pour l'analyse de robustesse des caractéristiques radiomiques sur des cohortes larges.

## 4.2 Méthodologie d'évaluation de la robustesse des caractéristiques radiomiques

Comme énoncé précédemment pour quantifier la robustesse des caractéristiques radiomiques, des simulations de perturbations de segmentations ont été utilisées. Celle-ci a été développée par l'équipe Data Science - Radiomics Research de SOPHiA Genetics. L'objectif est de simuler des perturbations aléatoires de la segmentation d'une structure anatomique, validée par un expert, selon un coefficient Dice  $D$  [8] prédéfini.

Ces perturbations sont aléatoires et pourraient ne pas être représentatives de la variabilité inter-opérateur. Ainsi, une validation humaine des perturbations est réalisée afin de tenir compte des *a priori* introduits par l'oeil humain, par la connaissance anatomique ou par les objectifs cliniques. Ces perturbations peuvent être générées en grand nombre, sur un grand nombre de segmentations de structures anatomiques comparables, permettant ainsi une analyse plus fine de la robustesse des caractéristiques radiomiques associées à ces segmentations.

### 4.2.1 Représentation 3D d'une image médicale et d'une segmentation

Une image médicale comme une segmentation peuvent être représentées en format DICOM<sup>1</sup>, le standard en termes d'imagerie médicale. Il s'agit d'un standard international qui présente un gros avantage puisqu'il est valable sur toutes les machines. Il présente l'inconvénient d'être un empilement de structures 2D :

- D'images 2D composées de pixels et représentant des coupes 2D de l'image totale

---

1. Digital Imaging and Communication in Medicine

- Des contours 2D (lignes brisées fermées) représentant le contour 2D d'une structure visible sur la coupe 2D associée.

Pour pouvoir calculer les caractéristiques radiomiques en 3D (morphologies, textures...) le logiciel SOPHiA DDM for Radiomics transforme images et contours en structure 3D dont une image est la suivante :

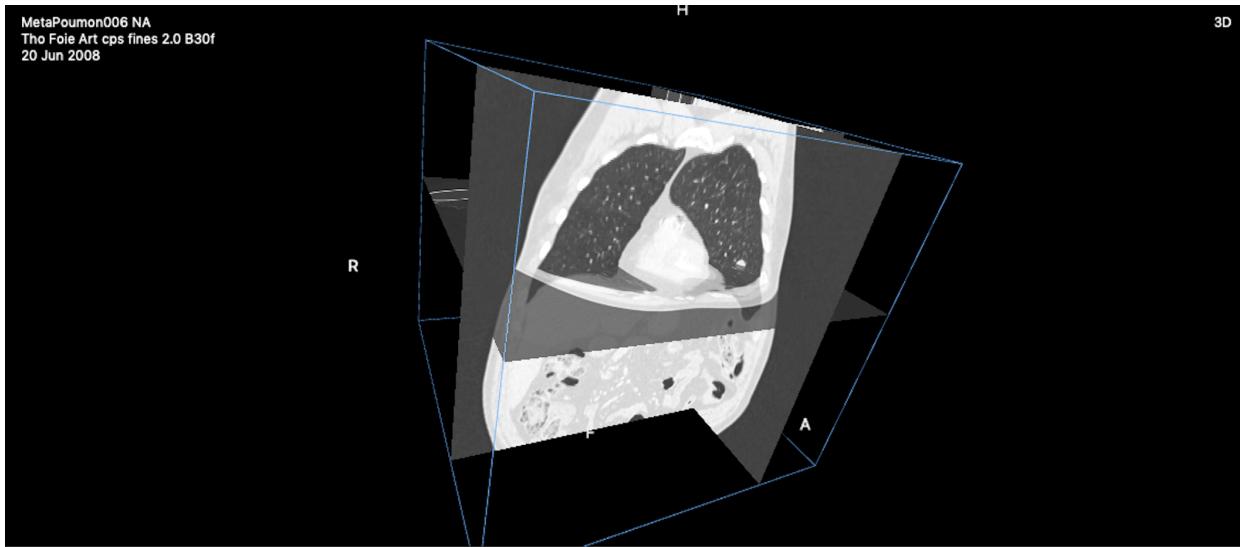


FIGURE 9 – Image 3D d'un parenchyme

Ainsi c'est sur cette représentation 3D que la segmentation de la tumeur aura lieu. La segmentation ici est semi-automatique. Un aperçu d'une segmentation réalisé en 3D mais aperçu sur des coupes en 2D est la suivante :

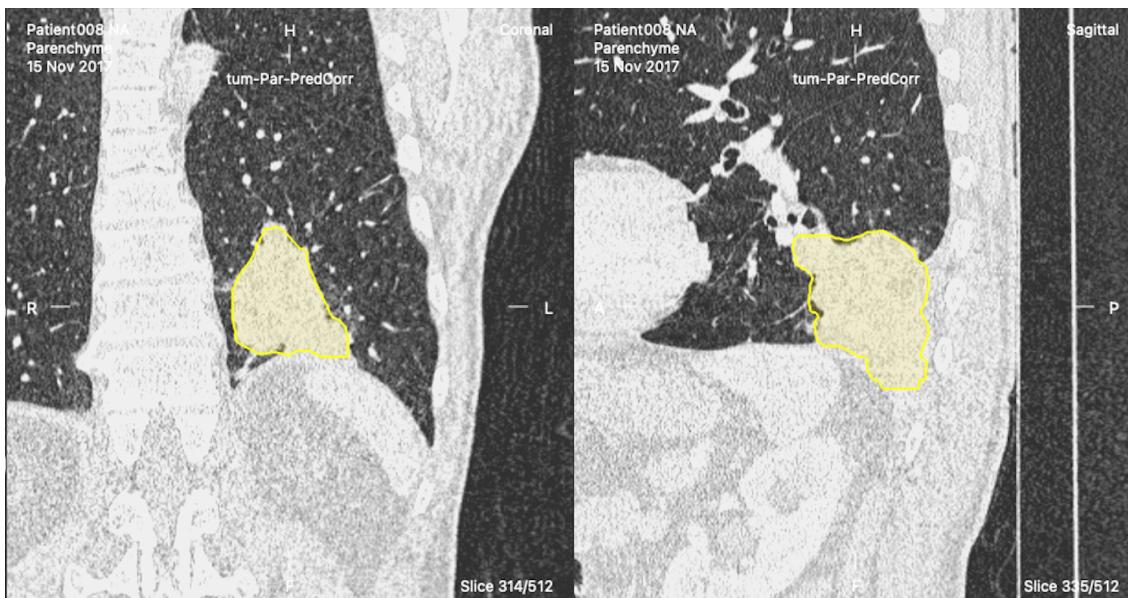


FIGURE 10 – Segmentation ou contour 2D d'une tumeur sur un parenchyme

La représentation 3D finale de la segmentation est un maillage voir figure 11

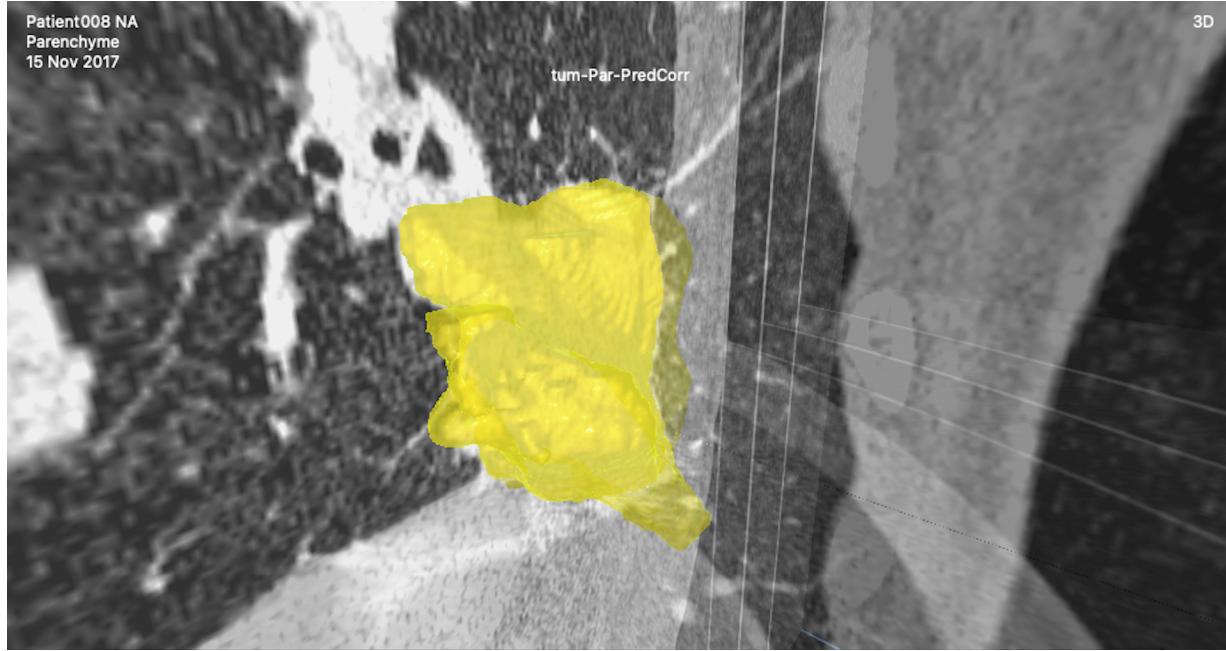


FIGURE 11 – Segmentation 3D d'une tumeur sur un parenchyme

Un maillage est une enveloppe surfacique fermée, immergée dans l'image 3D, constituée de triangles jointifs. Le maillage est donc constitué de faces triangulaires, reliés par leurs arêtes et leurs sommets. La qualité d'une segmentation est donc évaluée par la qualité de l'alignement de cette surface avec le bord de la structure anatomique cible.

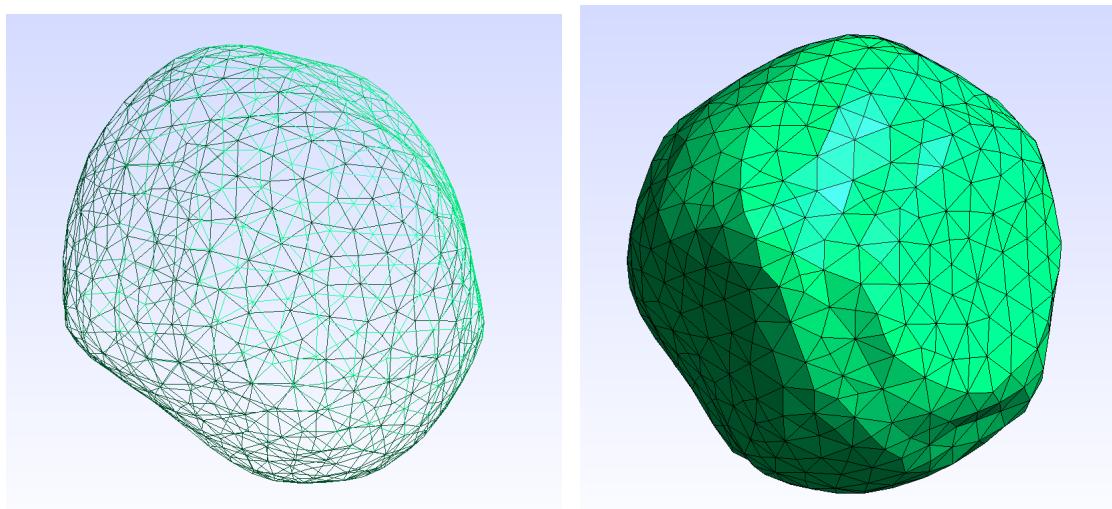


FIGURE 12 – Maillage d'une segmentation en 3D transparente et non transparente

Ce maillage est ensuite régularisé en utilisant un filtre moyen surfacique sur les positions de chaque sommet. On y perd du volume mais cette perte sera compensé à une autre étape du processus.

#### 4.2.2 Description brève du "Modèle déformable"

L'intérêt de la structure de maillage est de pouvoir être facilement déformable, en bougeant les sommets, tout en conservant la cohérence topologique. Certaines règles implémentées dans l'outil de gestion de maillage du logiciel SOPHiA DDM for Radiomics, qui ne sont pas détaillées ici, doivent toutefois être respectées : gestion des auto-collisions de maillage, gestion de la taille des arêtes par ajout ou suppression de sommets lorsqu'une arête devient trop grande ou trop petite.

La simulation de perturbation de maillage repose sur cet outil de déformation de maillage. Le mouvement appliqué à chaque sommet du maillage est itératif, de sorte qu'à chaque itération, la gestion des collisions et le contrôle des tailles d'arêtes soient effectués. Ce mouvement peut être généré par deux types de forces : des forces internes qui ne dépendent que de la structure du maillage (par exemple la régularisation par filtre moyen surfacique, déjà mentionnée) et des forces externes qui peuvent dépendre de l'information contenue dans l'image médicale 3D dans laquelle est immergé le maillage, ou toute autre image.

#### 4.2.3 Application à la perturbation de segmentation

Soit un Dice  $D$  fixé, le but est de générer une segmentation perturbée respectant :

$$\text{Dice}(\text{Originale}, \text{Perturbée}) = D$$

Le modèle déformable qui sera appliqué pour les perturbations réalisées ici est un modèle conçu spécifiquement par l'équipe d'Image Processing. La création d'une perturbation de segmentation se fait de façon itérative.

Dans un premier temps le maillage est déformé à travers des forces au niveau de chaque sommet. Après le déplacement de tous les sommets et la phase de gestion des collisions et de traitement des longueurs d'arêtes, une régularisation par filtre moyen surfacique est appliquée au maillage. Cette deuxième phase de régularisation permet à chaque itération d'assurer le lissage du maillage. Elle présente l'inconvénient d'introduire une perte de volume sur le maillage fermé qui est donc essentiellement convexe. Cette perte de volume est ensuite compensée par un paramètre de compensation défini empiriquement à partir du coefficient de Dice fixé  $D$ , et d'autres paramètres du maillage original.

Enfin, le processus itératif est stoppé lorsque le coefficient de Dice  $d$  mesuré à chaque itération entre la segmentation originale et sa perturbation courante vérifie  $d < D$ . Quand le critère d'arrêt est presque atteint, c'est-à-dire si la condition  $d < D + 0.005$  est vérifiée, alors le paramètre représentant le pas de temps fictif entre deux itérations est raffiné, afin d'assurer une convergence au plus proche du Dice cible.

Voici ci-dessus un exemple d'une perturbation de contour en 2D et 3D réalisé avec un critère d'arrêt de Dice  $D = 0.90$ . Le contour de couleur rouge étant celui de départ les le vert et jaune les contours perturbés simulés.

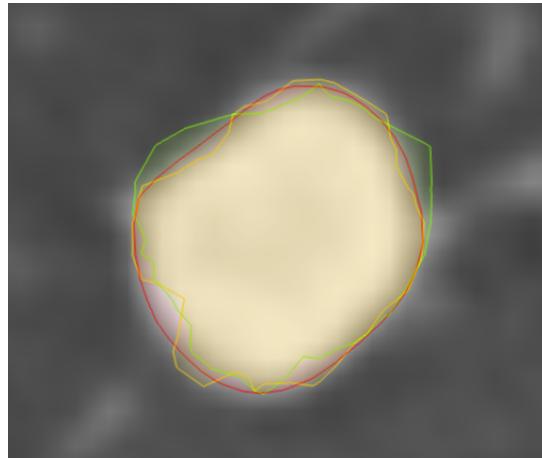


FIGURE 13 – Contour original (rouge) et ses perturbations en 2D avec un Dice = 0.90

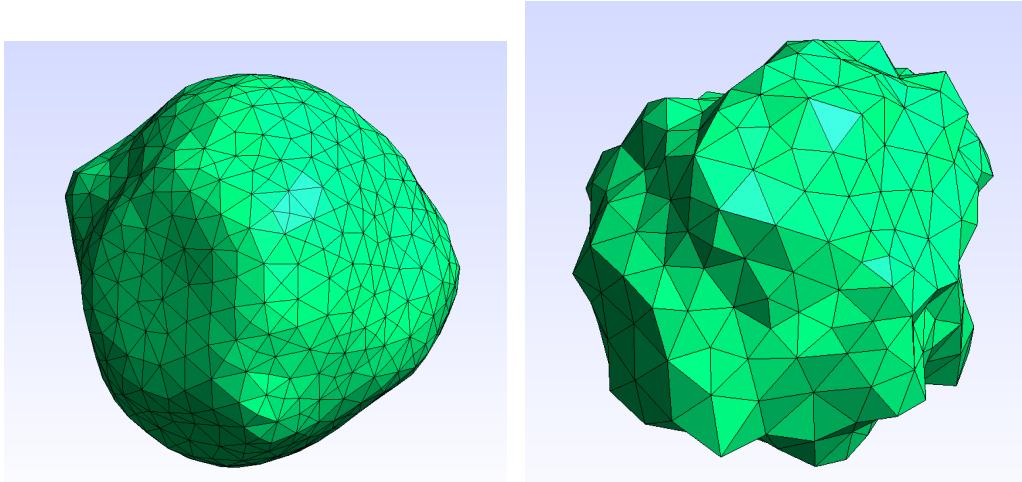


FIGURE 14 – De gauche à droite un maillage originale et une perturbation en 3D de Dice = 0.90

#### 4.2.4 Critère d'évaluation de la robustesse des caractéristiques radiomiques

L'objectif étant d'évaluer la robustesse des caractéristiques radiomiques, celles-ci sont calculées pour chaque maillage perturbé. Les caractéristiques ainsi que leur calcul tel qu'il est implémenté dans le logiciel SOPHiA DDM for Radiomics, sont définis par le standard de l'IBSI [Zwanenburg et al., 2020].

Ces contours perturbés ainsi générés seront utilisés pour caractériser cette variabilité (inter-opérateur). On considérera les caractéristiques radiomiques peu variables comme robustes et les autres comme non-robustes. Pour évaluer la robustesse des caractéristiques radiomiques nous allons utiliser le coefficient de corrélation intra-classe (ICC) [McGraw and Wong, 1996].

L'ICC est un indice de fiabilité largement utilisé dans les analyses de fiabilité test-retest, intra-opérateur et inter-opérateur où il y a la présence de deux ou plusieurs évaluateurs. Il peut prendre

des valeurs entre 0 et 1, 0 indiquant une fiabilité nulle entre les évaluateurs et 1 une fiabilité parfaite entre les évaluateurs. En termes simples, un ICC est utilisé pour déterminer si les éléments (ou les sujets) peuvent être évalués de manière fiable par différents évaluateurs. Il existe 10 formes d'ICCs selon [Koo and Li, 2016] et chaque forme implique des hypothèses distinctes dans leur calcul et conduira à des interprétations différentes, et dépendantes des trois facteurs suivants :

1. **Le modèle** : Anova à un facteur aléatoire, Anova à deux facteurs aléatoires, Anova à deux facteurs aléatoires et mixtes
2. **Le type de relation** : Cohérence (constancy) ou concordance (absolute agreement)
3. **Unités** : Un seul évaluateur ou la moyenne des évaluateurs

Voici un descriptif des trois différents facteurs :

#### 1. Les modèles :

Disposons-nous du même ensemble d'évaluateurs pour tous les sujets ? Disposons-nous d'un échantillon de correcteurs choisis au hasard dans une population plus large ou d'un échantillon spécifique de correcteurs ? Telles sont les questions qui peuvent nous guider dans le choix du modèle à utiliser.

- **Anova à un facteur aléatoire** : Dans ce modèle, chaque sujet est évalué par un ensemble différent d'évaluateurs qui ont été choisis au hasard dans une population plus large d'évaluateurs possibles. En pratique, ce modèle est rarement utilisé dans l'analyse de la fiabilité clinique car la majorité des études de fiabilité impliquent généralement le même ensemble d'évaluateurs pour mesurer tous les sujets. Une exception serait les études multicentriques pour lesquelles la distance physique entre les centres empêche le même groupe d'évaluateurs d'évaluer tous les sujets. Dans ce cas, un groupe d'évaluateurs peut évaluer un sous-groupe de sujets dans un centre et un autre groupe d'évaluateurs peut évaluer un sous-groupe de sujets dans un autre centre, et donc le modèle à un facteur aléatoire doit être utilisé dans ce cas.
  - **Anova à deux facteurs aléatoires** : Si nous sélectionnons aléatoirement nos évaluateurs à partir d'une population plus large d'évaluateurs présentant des caractéristiques similaires, le modèle à effets aléatoires à deux voies est le modèle de choix. En d'autres termes, nous choisissons le modèle à effets aléatoires à deux voies si nous envisageons de généraliser nos résultats de fiabilité à tous les évaluateurs qui possèdent les mêmes caractéristiques que les évaluateurs sélectionnés dans l'étude de fiabilité. Ce modèle est approprié pour évaluer les méthodes d'évaluation clinique basées sur les évaluateurs (par exemple, l'amplitude passive des mouvements) qui sont conçues pour une utilisation clinique de routine par des cliniciens ayant des caractéristiques spécifiques (par exemple, des années d'expérience) comme indiqué dans l'étude de fiabilité.
  - **Anova à deux facteurs aléatoires et mixtes** : Nous devons utiliser le modèle à effets mixtes à deux voies si les évaluateurs sélectionnés sont les seuls évaluateurs d'intérêt. Avec ce modèle, les résultats ne représentent que la fiabilité des évaluateurs spécifiques impliqués dans l'expérience de fiabilité. Ils ne peuvent pas être généralisés à d'autres évaluateurs, même si ceux-ci ont des caractéristiques similaires à celles des évaluateurs sélectionnés dans l'expérience de fiabilité. Par conséquent, le modèle à effets mixtes à deux voies est moins couramment utilisé dans l'analyse de la fiabilité inter-juges.
2. **Les types de relation** : Pour les modèles à deux facteurs aléatoires et à deux facteurs aléatoires et mixtes , il existe 2 définitions de la CCI : "concordance" et "cohérence". Le choix

de la définition de l'ICC dépend de l'importance que l'on accorde à l'accord absolu ou à la cohérence entre les évaluateurs.

- **Cohérence** : Nous nous intéressons aux différences systématiques entre les notes attribuées par les juges (par exemple, les juges ont-ils attribué une note faible ou élevée à des sujets similaires ?)
- **Concordance** : Nous sommes intéressés par les différences absolues entre les évaluations des juges (par exemple, quelle est la différence absolue entre les évaluations du juge A et du juge B).
- 3. **Les unités** : Cette sélection dépend de la manière dont le protocole de mesure sera mené dans l'application réelle. Par exemple, si nous prévoyons d'utiliser la valeur moyenne de 3 évaluateurs comme base d'évaluation, la conception expérimentale de l'étude de fiabilité doit impliquer 3 évaluateurs, et le type "moyenne de k évaluateurs" doit être sélectionné. Inversement, si nous prévoyons d'utiliser la mesure d'un seul évaluateur comme base de la mesure réelle, le type "évaluateur unique" doit être sélectionné même si l'expérience de fiabilité implique 2 évaluateurs ou plus.
  - **Un seul évaluateur** : Nous sommes uniquement intéressés par l'utilisation des évaluations d'un seul évaluateur comme base de mesure.
  - **La moyenne des évaluateurs** : Nous sommes intéressés par l'utilisation de la moyenne des évaluations de tous les juges comme base de mesure.

Dans notre situation l'ICC nous permettra de évaluer la variabilité des caractéristiques radiomiques suivant les différentes perturbations effectuées : à quel point les caractéristiques radiomiques extraites des différentes perturbations sont - elles différentes entre elles ? On choisit cet indicateur car il est basé sur un calcul de variance soit la mesure de la dispersion des valeurs d'un échantillon. Ainsi nous avons comme seul facteur aléatoire les différentes perturbations des segmentations de chaque patient. Sachant que les perturbations sont indépendantes d'un patient à un autre et d'une perturbation à une autre (par définition), nous avons décider d'utiliser le modèle d'analyse de variance à un facteur aléatoire soit l'ICC(1,1). Durant toute la suite du rapport sauf par précision lors d'une ambiguïté lorsque l'on parlera de l'ICC on fera référence à l'ICC(1,1).

#### 4.2.5 Modélisation

Nous allons donc appliquer le modèle d'analyse de variance à un facteur aléatoire et calculer pour chaque caractéristique radiomique son coefficient de corrélation intra-classe pour ensuite déduire parmi celles-ci les robustes des moins robustes grâce à un seuil fixé arbitrairement. Dans ce modèle, chaque patient est évalué par un ensemble différent de déformations de tranche (coupe suivant la 3D de l'organe), choisi au hasard dans une population plus large d'évaluateurs possibles (multitudes de possibilités suivant le coefficient de DICE sélectionné).

On considère  $p$  le nombre de patients dans cette étude et  $n$  le nombre de perturbations par patients.  $Y_{i,j}$  la valeur de la caractéristique radiomique extraite de la  $j$ -ème perturbation pour le  $i$ -ème patient et  $\alpha_i$  l'effet aléatoire de la perturbation du  $i$ -ème patient.

On obtient alors le modèle suivant :

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, p \quad \text{et} \quad j = 1, \dots, n \quad (3)$$

où  $\mu$  est une constante et  $\epsilon_{ij}$ ,  $\alpha_i$  des variables aléatoires *i.i.d* telles que

$$\alpha_i \sim \mathcal{N}(0, \sigma_a^2) \quad (4)$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (5)$$

$$\epsilon_{i,j} \perp\!\!\!\perp \alpha_i, \forall i \quad (6)$$

Ainsi on peut en déduire les calculs suivants :

$$\begin{aligned} \text{La variance des } Y_{i,j} : \quad & Var(Y_{ij}) = Var(\mu + \alpha_i + \epsilon_{ij}) \\ & = Var(\alpha_i + \epsilon_{ij}) \\ & = Var(\alpha_i) + Var(\epsilon_{ij}) \quad \bigg) \downarrow_1 \\ & Var(Y_{ij}) = \sigma_a^2 + \sigma^2 \end{aligned}$$

$$\begin{aligned} \text{L'espérance des } Y_{i,j} : \quad & \mathbb{E}[Y_{ij}] = \mathbb{E}[\mu + \alpha_i + \epsilon_{ij}] \\ & = \mathbb{E}[\mu] + \mathbb{E}[\alpha_i] + \mathbb{E}[\epsilon_{ij}] \quad \bigg) \mathbb{E}[\alpha_i] = \mathbb{E}[\epsilon_{ij}] = 0 \\ & \mathbb{E}[Y_{ij}] = \mu \end{aligned}$$

La corrélation entre  $Y_{ij}$  et  $Y_{kl}$ ,

$$Cor(Y_{ij}, Y_{kl}) = \begin{cases} 0 & \text{si } i \neq k \\ \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} & \text{si } i = k \text{ et } j \neq l \\ 1 & \text{si } i = k \text{ et } j = l \end{cases}$$

---

1.  $\epsilon_{i,j} \perp\!\!\!\perp \alpha_i, \forall i$

$$\begin{aligned}
\text{Si } i \neq k : Cor(Y_{ij}, Y_{kl}) &= \frac{Cov(Y_{ij}, Y_{kl})}{\sqrt{Var(Y_{ij})} \sqrt{Var(Y_{kl})}} \\
&= \frac{Cov(\mu + \alpha_i + \epsilon_{ij}, \mu + \alpha_k + \epsilon_{kl})}{\sqrt{Var(Y_{ij})^2}} \\
&= \frac{Cov(\alpha_i + \epsilon_{ij}, \alpha_k + \epsilon_{kl})}{Var(Y_{ij})} \\
&= \frac{Cov(\alpha_i, \alpha_k) + Cov(\alpha_i, \epsilon_{kl}) + Cov(\epsilon_{ij}, \alpha_k) + Cov(\epsilon_{ij}, \epsilon_{kl})}{\sigma_a^2 + \sigma^2} \\
&= \frac{0}{\sigma_a^2 + \sigma^2} \quad \left. \right]_1 \\
Cor(Y_{ij}, Y_{kl}) &= 0
\end{aligned}$$

$$\begin{aligned}
\text{Si } i = k \text{ et } j = l : Cor(Y_{ij}, Y_{ij}) &= \frac{Var(Y_{ij})}{\sqrt{Var(Y_{ij})^2}} \\
&= \frac{Var(Y_{ij})}{Var(Y_{ij})}
\end{aligned}$$

$$Cor(Y_{ij}, Y_{ij}) = 1$$

$$\begin{aligned}
\text{Si } i = k \text{ et } j \neq l : Cor(Y_{ij}, Y_{il}) &= \frac{Cov(Y_{ij}, Y_{il})}{\sqrt{Var(Y_{ij})} \sqrt{Var(Y_{il})}} \\
&= \frac{Cov(\mu + \alpha_i + \epsilon_{ij}, \mu + \alpha_l + \epsilon_{il})}{\sqrt{Var(Y_{ij})^2}} \\
&= \frac{Cov(\alpha_i + \epsilon_{ij}, \alpha_l + \epsilon_{il})}{Var(Y_{ij})} \\
&= \frac{Cov(\alpha_i, \alpha_l) + Cov(\alpha_i, \epsilon_{il}) + Cov(\epsilon_{ij}, \alpha_l) + Cov(\epsilon_{ij}, \epsilon_{il})}{\sigma_a^2 + \sigma^2} \\
&= \frac{Var(\alpha_i) + Cov(\epsilon_{ij}, \epsilon_{il})}{\sigma_a^2 + \sigma^2} \quad \left. \right]_2 \\
Cor(Y_{ij}, Y_{il}) &= \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} \quad \left. \right]_3
\end{aligned}$$

Ainsi les perturbations d'un même patient sont corrélées tandis que celles de patients différents sont indépendantes. On définit ainsi le coefficient de corrélation intra-classe aussi par :

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} \tag{7}$$

- 
1.  $Cov(\epsilon_{ij}, \epsilon_{kl}) = Cov(\alpha_i, \alpha_j) = Cov(\alpha_i, \epsilon_{kl}) = 0, \forall i \neq k \text{ et } j \neq l$
  2.  $Cov(\alpha_i, \epsilon_{il}) = 0$
  3.  $Cov(\epsilon_{ij}, \epsilon_{il}) = 0$

On obtient ainsi la table de l'Anova suivante :

Source de variation	Df	Sommes des carrés	Carrés moyens	Expressions
Entre différents patients	p-1	$SS_B$	$MS_B$	$\sigma^2 + n\sigma_a^2$
Entre les perturbations (même patient)	p(n-1)	$SS_W$	$MS_W$	$\sigma^2$

On pose  $\bar{Y}_i = \sum_{j=1}^n \frac{Y_{ij}}{n}$ ,  $\bar{Y}_{..} = \sum_{i=1}^p \sum_{j=1}^n \frac{Y_{ij}}{an}$ , alors :

$$SS_B = n \sum_{i=1}^p (\bar{Y}_i - \bar{Y}_{..})^2$$

$$SS_W = \sum_{i=1}^p \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i..})^2$$

Ce qui nous conduit à l'estimation des carrés moyens suivant :

$$\begin{aligned} \mathbb{E}[MS_W] &= \frac{1}{p(n-1)} \mathbb{E}\left[\sum_{i=1}^p \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i..})^2\right] \\ &= \frac{1}{p(n-1)} \mathbb{E}\left[\sum_{i=1}^p \sum_{j=1}^n (\epsilon_{ij} - \bar{\epsilon}_{i..})^2\right] \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p \mathbb{E}\left[\sum_{j=1}^n (\epsilon_{ij} - \bar{\epsilon}_{i..})^2\right] \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p \mathbb{E}\left[\sum_{j=1}^n (\epsilon_{ij}^2 - 2\epsilon_{ij}\bar{\epsilon}_{i..} + \bar{\epsilon}_{i..}^2)\right] \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p \mathbb{E}\left[\sum_{j=1}^n (\epsilon_{ij}^2) - 2n\bar{\epsilon}_{i..}^2 + n\bar{\epsilon}_{i..}^2\right] \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p \sum_{j=1}^n (\mathbb{E}[\epsilon_{ij}^2] - n\mathbb{E}[\bar{\epsilon}_{i..}^2]) \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p \sum_{j=1}^n (n\sigma^2 - n\frac{\sigma^2}{n}) \\ &= \frac{1}{p(n-1)} \sum_{i=1}^p (n-1)\sigma^2 \\ \mathbb{E}[MS_W] &= \sigma^2 \end{aligned}$$

$$\begin{aligned}
\mathbb{E}[MS_B] &= \frac{1}{a-1} \mathbb{E}\left[n \sum_{i=1}^p (\bar{Y}_i - \bar{Y}_{..})^2\right] \\
&= \frac{n}{a-1} \mathbb{E}\left[\sum_{i=1}^p ((\alpha_i - \bar{\alpha}_{..}) + (\bar{\epsilon}_i - \bar{\epsilon}_{..}))^2\right] \\
&= \frac{n}{a-1} \mathbb{E}\left[\sum_{i=1}^p ((\alpha_i - \bar{\alpha}_{..})^2)\right] + \mathbb{E}\left[\sum_{i=1}^p (\bar{\epsilon}_i - \bar{\epsilon}_{..})^2\right] \\
&= \frac{n}{a-1} ((a-1)\sigma_a^2 + (a-1)\frac{\sigma_a^2}{n}) \\
\mathbb{E}[MS_B] &= \sigma^2 + n\sigma_a^2
\end{aligned}$$

Alors on obtient une réécriture de l'ICC suivante :

$$ICC = \frac{MS_B - MS_W}{MS_B + (n-1)MS_W} \quad (8)$$

#### 4.2.6 Méthodologie d'évaluation de la robustesse

Dans la suite, pour chaque caractéristique radiomique nous calculerons donc l'ICC, à l'aide de la formule décrite précédemment(7). Ensuite nous fixerons un seuil arbitraire (ici 0.90), et nous considérons comme caractéristiques radiomiques robustes, soit celles qui ont une valeur de l'ICC supérieure au seuil soit celles qui ont la borne inférieure de l'intervalle de confiance de la valeur de l'ICC supérieure au seuil. Ainsi on obtient en amont une base constituée de caractéristiques radiomiques robustes donc fiables. Sur cette nouvelle base on pourra appliquer des méthodes de sélection de variables diverses.

## 5 Application : Prédition de la réponse au traitement chez des patients atteints de cancer du poumon en stade avancé

### 5.1 Contexte

Le cancer du poumon est l'un des cancers les plus répandus en France, il attaque toutes catégories d'âge et de sexe. Chez les patients en stade avancé, le pronostic est mauvais mais des solutions thérapeutiques existent. Durant mon stage, mon application s'est réalisée chez des patients avec un cancer du poumon non à petites cellules en stade 3 ou 4, c'est-à-dire en stade très avancé. Ces données proviennent d'une cohorte observationnelle d'un hôpital français. Les patients ne présentaient pas de mutations du gène EGFR ou ALK, ils ont été traités par immunothérapie avec Pembrolizumab. Des images issues de CT-scans ont été collectées pour chaque patient avant l'initiation du traitement. Ces patients ont ensuite été suivis longitudinalement, notamment pour évaluer leur réponse au traitement.

### 5.2 Objectifs

Les patients ont eu une première évaluation de la réponse au traitement deux mois après l'initiation du pembrolizumab. Cette évaluation s'est basée sur l'évolution de la taille tumorale ou l'apparition de nouvelles lésions depuis l'image pré-traitement : c'est le critère RECIST. Ce critère permet une évaluation unidimensionnelle simple des tumeurs des patients dans le but de déceler une progression au niveau des tumeurs soit par une progression du diamètre des lésions cibles<sup>1</sup> de 20% ou plus<sup>2</sup>, soit parce qu'il y a une augmentation indiscutable des lésions non-cibles<sup>3</sup> ou soit encore parce que des nouvelles lésions<sup>4</sup> sont indiscutablement apparues.

L'objectif clinique principal de cette étude était d'identifier les patients qui ont progressé à la première évaluation, c'est-à-dire ceux chez qui le traitement n'a pas eu d'effet (augmentation de la somme des lésions cibles de plus de 20%, ou apparition de nouvelles lésions tumorales). En effet, les cliniciens souhaitaient pouvoir identifier ces patients afin de leur proposer une alternative thérapeutique. Cette étude a été réalisée en deux temps. Une première analyse a été réalisée en développant une signature prédictive du statut à la première évaluation à partir des caractéristiques radiomiques pré-traitement, en ignorant l'identification des caractéristiques radiomiques non robustes. Dans un second temps, à l'aide de perturbations de segmentations initiales, nous avons caractérisé les caractéristiques radiomiques non-robustes et une signature radiomique robuste a ainsi été développée.

### 5.3 Données

Notre étude a été réalisée à partir de données collectées chez 57 patients traités pour un cancer du poumon en stade avancé. Parmi eux, 31 avaient progressé à la première évaluation post-traitement, et 26 qui n'avaient pas progressé (stabilité de la maladie ou réponse au traitement).

Les segmentations des lésions tumorales pré-traitement ont été réalisées par l'équipe Image Processing à l'aide du logiciel SOPHiA RADOMICS. Pour chaque patient, les caractéristiques radiomiques

---

1. Les lésions cibles sont des lésions mesurant en pratique plus de 1cm de diamètre.

2. dans ce cas, pour éviter toute ambiguïté de mesure sur de petites lésions, il est aussi nécessaire que la progression soit supérieure à 5 mm en valeur absolue

3. Les lésions non-cibles sont toutes les lésions qui ne peuvent pas être considérées comme cibles

4. Les nouvelles lésions sont des tumeurs qui sont indiscutablement apparues, de façon non équivoque

extraites de ces segmentations ont été utilisées afin de prédire le statut (progression ou non) à la première évaluation. Ainsi, environ 201 biomarqueurs quantitatifs (classés en cinq classes différentes à savoir *Morphological features*, *First order features*, *Histogram features*, *Original features* et *Textures features*) ont été extraits.

Parmi eux, les caractéristiques morphologiques *Morphological features* décrivent les aspects géométriques d'une région d'intérêt (ROI), comme par exemple la surface et le volume. Les statistiques de premier ordre *First order features* décrivent la distribution des intensités des voxels dans la ROI. Les caractéristiques de l'histogramme ou *Histogram features* sont les statistiques calculées à partir d'un histogramme d'intensité généré en discrétisant la distribution d'intensité originale de l'image à l'intérieur du ROI. Les caractéristiques originales ou *Original features* sont l'ensemble des statistiques calculées à partir de l'image médicale brute (c'est une classe propre à l'équipe Image Processing). Enfin les caractéristiques de textures ou *Textures features* regroupe toutes les statistiques calculées à partir de diverses matrices qui caractérisent les niveaux de gris des zones du ROI i.e (GLCM<sup>1</sup>, GLSZM<sup>2</sup>, GLRLM<sup>3</sup>, NGTDM<sup>4</sup> et GLDM<sup>5</sup>).[Zwanenburg et al., 2019]

Afin d'identifier les caractéristiques radiomiques robustes et de développer une signature prédictive fiable, une deuxième base de données a été étudiée. Celle-ci comportait les caractéristiques radiomiques issues des perturbations des segmentations originales de ces patients. Cette deuxième base se décompose en plusieurs versions selon le Dice utilisé (0.99, 0.96, 0.93, 0.90, 0.87) pour les perturbations. Pour chaque Dice, entre 10 et 20 perturbations ont été réalisées pour chaque patient.

#### 5.4 Environnement, importation et gestion des données brutes

L'environnement de travail utilisé est **RStudio** version 1.4.1106. L'ensemble des algorithmes a été réalisé sur un ordinateur dont les caractéristiques sont les suivantes :

Type	MacBook Pro (16 pouces, 2019)
Processeur	2,6 GHz Intel Core i7 6 coeurs
Mémoire	16 Go 2667 MHz DDR4
Graphisme	AMD Radeon Pro 5300M 4 Go

TABLE 2 – Caractéristiques de l'ordinateur

Notre jeu de données ne comprenait pas de valeurs manquantes. Notamment, il n'y avait pas de données manquantes au niveau des variables explicatives car toutes les caractéristiques radiomiques ont été extraites pour chaque patient. Afin de réaliser l'analyse statistique, les variables redondantes (colinéaires ou anti-colinéaires) et les variables non-informatives (avec variance nulle) ont été identifiées puis supprimées du reste de l'analyse. La détection de valeurs aberrantes ou d'individus extrêmes a été ensuite réalisée à l'aide de statistiques descriptives et d'une analyse factorielle (analyse en composantes principales).

- 
1. A Gray Level Co-occurrence Matrix
  2. A Gray Level Size Zone
  3. A Gray Level Run Length Matrix
  4. A Neighbouring Gray Tone Difference Matrix
  5. Gray Level Dependence Matrix

## 5.5 Analyses descriptive et explicative

Dans cette partie, nous allons présenter quelques analyses que l'on a faites avant de passer à la prédiction. Cette analyse descriptive et explicative nous a permis d'avoir une première impression sur le jeu de donnée. Il s'agit notamment de la distribution de variable d'intérêt, l'analyse des associations bi-variées à travers des tests statistiques non paramétriques, l'analyse des liaisons entre paires de variables et enfin l'analyse de la distribution des variables explicatives.

### 5.5.1 Distribution de la variable réponse

Parmi les 57 patients du jeu de données, 26 (45.6%) ont progressé à la première évaluation tandis que 31 (54.4%) n'ont pas progressé (Figure 15). La variable réponse est donc assez équilibrée.

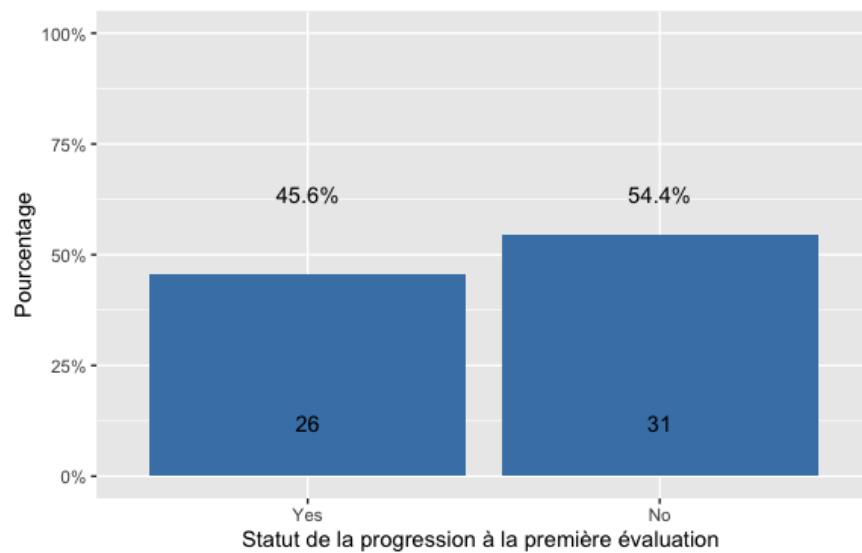


FIGURE 15 – Distribution de la variable réponse : Progression à la première évaluation

### 5.5.2 Associations bi-variées

Le jeu de données est composé uniquement de données numériques, excepté la variable réponse qui est un facteur à deux niveaux ("Yes" pour la progression et "No" pour la non progression). L'interaction entre toutes les variables et la progression des patients est présentée ci-dessous. Les p-valeurs sont calculées à partir du test de la somme des rangs de Wilcoxon et du test exact de la somme des rangs de Wilcoxon avec le test  $H_0$  : la distribution de la caractéristique est similaire selon les deux statuts possibles à la première évaluation. Nous calculons ces interactions entre les caractéristiques et calculons les p-valeurs et les p-valeurs ajustées (selon l'ajustement de Benjamin-Hochberg), voir un extrait dans le tableau 3 ci-dessous.

Variables	Yes (N=26)	No (N=31)	Total (N=57)	p.v <sup>1</sup>	p.v ajustées
Volume				0.214	0.956
Moyenne (SD <sup>2</sup> )	229.059 (346.533)	113.974 (133.496)	166.469 (257.878)		
IC <sup>3</sup>	4.040 - 1560.000	0.939 - 517.000	0.939 - 1560.000		
Homogeneity				0.631	0.956
Moyenne (SD)	0.111 (0.050)	0.111 (0.036)	0.111 (0.042)		
IC	0.049 - 0.276	0.048 - 0.183	0.048 - 0.276		
Area				0.067	0.892
Moyenne (SD)	274.658 (265.849)	177.934 (199.226)	222.054 (234.898)		
IC	24.400 - 1220.000	5.470 - 807.000	5.470 - 1220.000		

TABLE 3 – Extrait des associations entre les variables et la variable d'intérêt

Les tests réalisés nous ont permis de mettre en avant les variables significatives i.e les variables dont les distributions sont significativement différentes suivant le statut de la progression à la première évaluation. Ces variables discriminent la variable d'intérêt de part leur distribution. Ce sont majoritairement des caractéristiques morphologiques et de textures.

Une liste de ses variables est affichée, voir tableau 4 ci-dessous :

	Variables	p-values
1	Asphericity	0.002
2	Compactness.1	0.002
3	Compactness.2	0.002
4	Least.axis.length..cm.	0.031
5	Major.axis.length..cm.	0.009
6	Maximum.3D.diameter..cm.	0.001
7	Minor.axis.length..cm.	0.002
8	Spherical.disproportion	0.002
9	Sphericity	0.002
10	Volume.density...aligned.bounding.box	0.022
11	Volume.density...convex.hull	0.010
12	Volume.density...enclosing.ellipsoid	0.006
13	Volume.density...oriented.bounding.box	0.009

TABLE 4 – Les variables ayant les p-valeurs les plus significatives i.e p-valeurs < 0.05

Un exemple de distribution soit celle de la variable *Compactness 1 ou Compacité 1* qui est

- 
- 1. p-valeurs
  - 2. Variance
  - 3. Intervalle de confiance

une caractéristique radiomique appartenant à la classe des caractéristiques radiomiques de formes (Shapes features). Elle est l'une des features qui permettent de quantifier la déviation du volume du ROI<sup>1</sup> par rapport à un sphéroïde représentatif. Elle est une mesure de la compacité ou de la forme sphérique du volume.

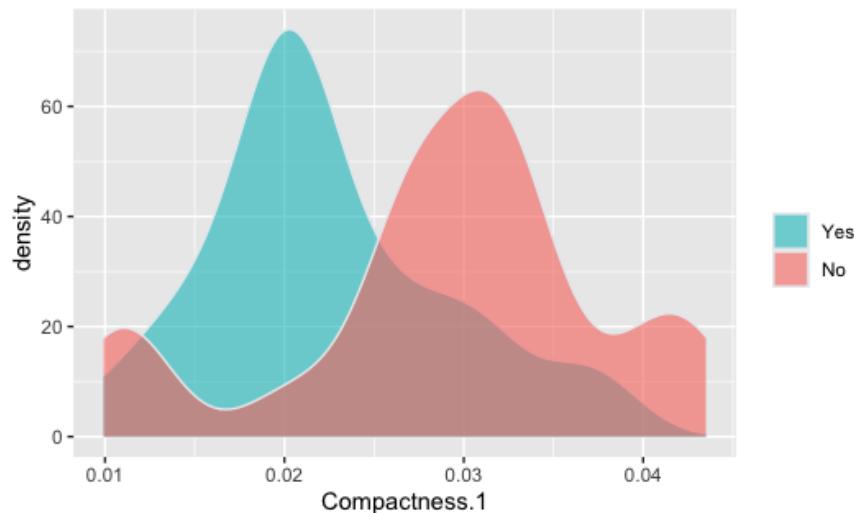


FIGURE 16 – Distribution de la variable Compactness 1 suivant la variable d'intérêt

### 5.5.3 Diagramme de corrélation

Dans cette partie nous affichons la matrice de corrélation triangulaire supérieure, avec comme méthode d'ordonnancement la classification ascendante hiérarchique (les variables sont classées suivant le clustering de variable réalisée au préalable). La matrice est créée grâce à la fonction `corrplot` avec l'argument `order` pour le classement des variables. La matrice de corrélation, la figure 17 nous montre une corrélation forte entre les paires de variables à la fois positive que négative. Sur la figure ci-dessous plus la couleur est bleu plus la corrélation entre les variables est positive (+1) et plus elle est rouge plus la corrélation entre les variables est négative (-1).

En effet, ce graphe que l'on peut agrandir (zoom) nous sert juste de motivation pour réajuster le nombre de variables en fonction de la corrélation entre paires.

---

1. Region of Interest

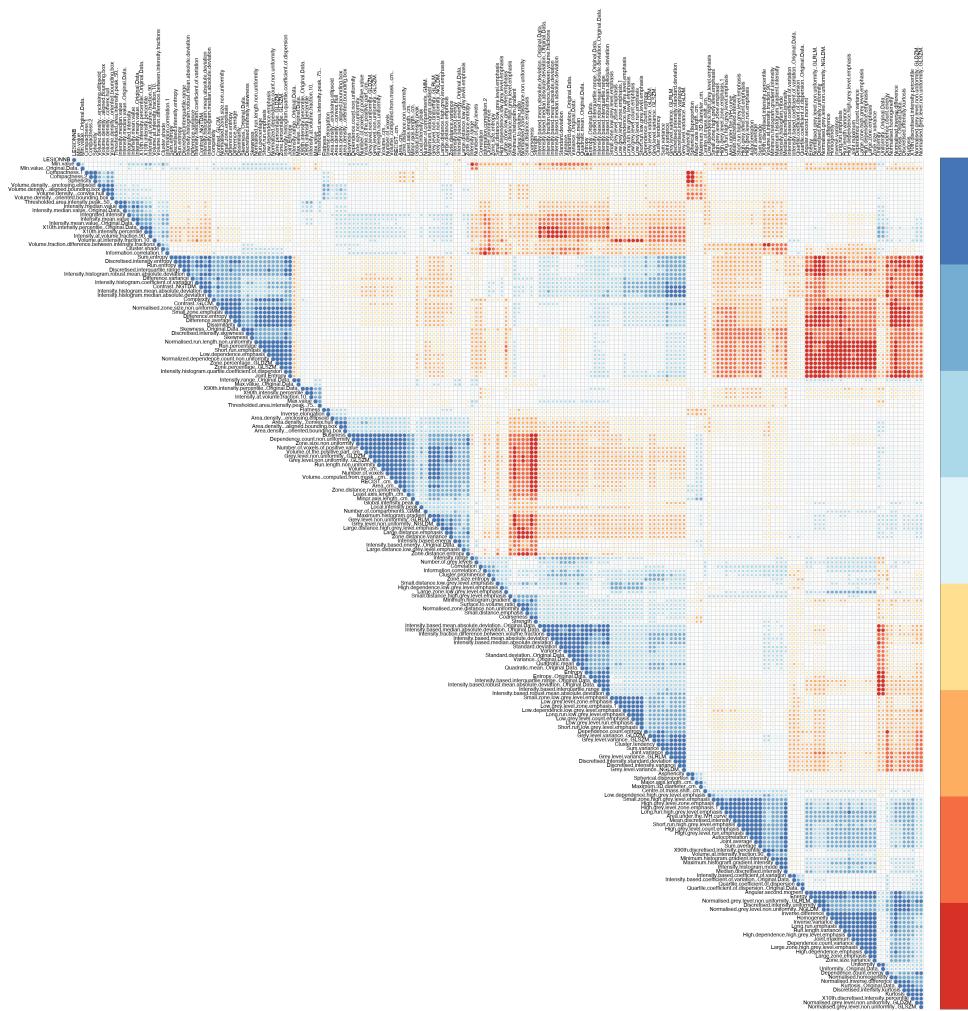


FIGURE 17 – Matrice de corrélation (bleu = 1, rouge = -1, blanc = 0)

#### 5.5.4 Distribution des variables explicatives

Les variables explicatives sont les caractéristiques radiomiques extraites des segmentations, ayant des distributions inconnues. Des tests ont été réalisé dans le but d'en détecter la nature. Cette détection aidera dans le choix de la sélection de variable suivant les coefficients de corrélations (kendall, spearman ou pearson). Notamment si les variables sont gaussiennes ou non (distribution cible). Pour ce faire nous avons utilisé le test de Shapiro-Wilks à 5% avec comme hypothèse nulle,  $H_0$  : l'échantillon  $x_1, \dots, x_n$  est issu d'une population normalement distribuée. En effet nous avons obtenu que 31 variables sur un total de 201 variables sont gaussiennes ce qui donne une proportion d'environ 15% de variables gaussiennes. Le critère alors utilisé pour la sélection de variables est donc le coefficient de "Kendall" qui omet l'hypothèse de normalité des variables et de linéarité entre paires

- . Une distribution d'une variable gaussienne celle de la variable *Compactness 1* voir figure 16.

L'analyse explicative ainsi faite, nous n'allons pas faire de séparation en base d'entraînement et de test des données due à la faible quantité de ceux-ci. En outre, cette partie intervient uniquement dans le cas où l'on disposerait d'une base de données avec suffisamment d'individus. Dans ce cas, le jeu de données sera séparé en deux parties : une partie d'entraînement et une partie test. La partie d'entraînement sera d'environ 70% de la base totale et permettra le développement et l'entraînement des algorithmes d'apprentissage automatique et la partie test d'environ 30% permettra l'évaluation non biaisée des performances prédictives des algorithmes sélectionnés a posteriori suite à la phase d'entraînement. Cette séparation de la base de données sera effectuée de manière aléatoire, et en ajoutant si besoin une stratification pour conserver la même proportion dans chaque base. Néanmoins, dans le cas où l'on se trouverait en possession d'une base de données avec peu d'individus (souvent une trentaine), d'autres méthodes pourront être utilisées pour fournir des performances prédictives et entraîner les algorithmes de la manière les moins biaisés que possible.

## 5.6 Prédiction de la progression à la première évaluation

### 5.6.1 Modélisation

Dans cette partie, dans ce premier volet, nous construisons un modèle capable de prédire au mieux la progression à la première évaluation des patients atteints du cancer. Pour cela nous utilisons un panel d'algorithmes d'apprentissage supervisée (ML<sup>1</sup>) que nous optimisons suivant leurs hyperparamètres respectifs et suivant différentes métriques. Les modèles utilisés sont listés ci-dessous :

1. Régression Logistique avec régularisations Lasso, Ridge et Elastic-net
2. Les k-voisins les plus proches
3. Modèle bayésien naïf
4. Analyse Discriminante linéaire
5. Les Machines à Vecteur Support (noyaux linéaire, radial, polynomial)
6. Les méthodes à bases d'arbres de décision :
  - CART
  - Forêts aléatoires
  - C5trees
  - Gradient Boosting Machine (GBM)
  - Extrême Gradient Boosting (Xgboost)
  - AdaBoost

Notre jeu de données étant de taille de 57 individus, nous avons privilégié la validation croisée Leave-one-out (LOOCV). La méthode recommandée quand cela est possible pour l'optimisation des algorithmes via la sélection des meilleurs hyperparamètres se vaudrait être une Cross validation 10-fold répétée 30 fois sur la base d'entraînement.

L'optimisation ainsi faite, il nous faut quantifier les performances des algorithmes. En effet, les critères de performances prédictives sont l'ensemble des métriques de performance qui permettent

---

1. Machine Learning

d'évaluer ou de comparer les algorithmes de machines learning. Certaines métriques utilisent un seuil de décision binaire et sont donc définies à partir de la matrice de confusion obtenue. C'est le cas pour l'accuracy, la sensibilité, la spécificité, la précision et le F1 score. En revanche, d'autres métriques se basent sur les probabilités prédictives d'appartenir à une classe et sont invariantes selon le seuil de décision. C'est le cas pour l'AUC, le PRAUC et le BrierScore [Martin, 2016].

Comme performance prédictives nous avons :

1. Sensibilité
2. Spécificité
3. Précision globale
4. Précision
5. Valeur prédictive négative
6. Fbeta-Scores (F1score, F2score...)
7. AUC
8. PrAUC positif
9. PrAUC negatif
10. Brier-Score

Dans notre étude les hyperparamètres des algorithmes ont été optimisés par maximisation du F1-score.

Dans la plupart des études du fait de la difficulté de l'acquisition de données de patients, nous sommes sujet à la présence d'un jeu de données comportant un nombre de variables (un minimum de 200 variables) largement supérieure au nombre d'individus. Étant donc dans un cadre de "grande dimension", l'optimisation des algorithmes de machine learning sur le jeu de données doit passer par une réduction de dimension (sélection de variables pour éviter les effets potentiels de la malédiction de la grande dimension qui se produisent lorsque le nombre de variables prédictives est élevé par rapport au nombre de patients étudiés). Il est habituel de distinguer trois types de méthodes de sélection de variables : les méthodes "filtres" , "envelopantes" et "embarquées" [Wu et al., 2013]. Dans notre cas, une méthode "filtres" sera utilisée pour réduire la dimensionnalité des données. Plusieurs autres méthodes existent telles que (a sélection de variable LASSO, forêts aléatoires par importance de variables, vsurf, par clustering, etc. Sur l'ensemble de ces techniques un travail post-étude a été effectué dans un but comparatif et suivant nos objectifs la sélection choisie est celle basée sur les corrélations.

Cette méthode de sélection se décompose en deux étapes et a pour but la suppression des variables les plus corrélées et expliquant le moins la variable réponse.

En effet, soit un seuil  $s$  de corrélation fixé, on sélectionne les paires de variables dont la valeur absolue du coefficient de corrélation de Kendall est supérieure au seuil  $s$ . Le choix du coefficient de corrélation de Kendall est fait car les caractéristiques radiomiques respectent rarement les critères pour l'utilisation des autres coefficients de corrélations (Pearson et Spearman) comme la normalité des variables et la linéarité entre paires.

Ensuite parmi ces paires l'on élimine la variable qui explique le moins la variable d'intérêt grâce au critère de l'AUC ( Soit  $X_1$  et  $X_2$  deux caractéristiques radiomiques telles que  $|cor(X_1; X_2)| \geq s$  on garde  $X_1$  si  $AUC(X_1; Y) \geq AUC(X_2; Y)$  sinon on conserve  $X_2$ ,  $Y$  étant la variable d'intérêt).

L'utilisation du critère de l'*AUC* permet de garder la variable qui discrimine le mieux la classe d'intérêt.

La méthodologie de sélection nécessitant un seuil nous avons testé un seuil<sup>1</sup> à 0.90 et à un seuil à 0.80. Le meilleur seuil a été gardé selon les performances prédictives obtenues. La sélection de variables des corrélations a ainsi conservé un total de 117 variables pour le seuil à 0.90 et 59 variables pour le seuil à 0.80 sur les 201 variables initialement disponibles .

L'apprentissage de nos nombreux algorithmes ainsi réalisé nous avons conservés les 5 meilleures algorithmes il s'agit de l'AdaBoost, les arbres de décisions, les Forêts aléatoires, le Gradient Boosting Machine et le modèle C5trees. Les résultats des performances prédictives de ces algorithmes sont présentés ci-dessous sous formes de tableaux.

	AUC	PRAuc	Accuracy	Sensibilité	Spécificité	Précision	NPV <sup>2</sup>	F1score	BrierScore
AdaBoost	0.76	0.61	0.75	0.88	0.65	0.68	0.87	0.77	0.21
CART	0.76	0.44	0.75	0.88	0.65	0.68	0.87	0.77	0.19
GBM	0.75	0.60	0.75	0.73	0.77	0.73	0.77	0.73	0.25
C5trees	0.70	0.52	0.70	0.65	0.74	0.68	0.72	0.67	0.25
Random Forest	0.68	0.59	0.68	0.65	0.71	0.65	0.71	0.65	0.23

TABLE 5 – Tableau des performances prédictives de la sélection de variable Kendall à 0.90

	AUC	PRAuc	Accuracy	Sensibilité	Spécificité	Précision	NPV	F1score	BrierScore
GBM	0.72	0.61	0.72	0.77	0.68	0.67	0.78	0.71	0.26
C5trees	0.68	0.61	0.68	0.69	0.68	0.64	0.72	0.67	0.24
AdaBoost	0.65	0.50	0.65	0.69	0.61	0.60	0.70	0.64	0.20
CART	0.65	0.42	0.65	0.69	0.61	0.60	0.70	0.64	0.25
Random Forest	0.65	0.61	0.65	0.62	0.68	0.62	0.68	0.62	0.23

TABLE 6 – Tableau des performances prédictives de la sélection de variable Kendall à 0.80

Les meilleures performances sont associées à la sélection de variables via corrélations de Kendall au seuil de 0.90 .

Le meilleur modèle semblerait être le modèle AdaBoost issu de la sélection des corrélations de Kendall au seuil 0.90. C'est le modèle qui maximise la métrique choisi le "F1score" avec une valeur de 0.77. Cependant le nombre d'itération de cet algorithme de boosting, utilisant les arbres de décisions comme classifieur faible, étant de 1 il se ramenait à un simple arbre de décision qui n'est pas facile à généraliser du fait qu'il s'imbrique très souvent des valeurs de l'ensemble d'apprentissage. Le modèle finalement choisi est le Gradient Boosting Machine.

Les performances au niveau de la prédiction des individus peuvent se caractériser sous la forme d'une pseudo matrice de confusion. Ce n'est pas réellement une matrice de confusion car la stratégie utilisée reste le LOOCV : chaque individu est prédit avec un modèle différent optimisé par l'ensemble

1. Ces seuils sont arbitraires

des individus de l'ensemble d'apprentissage. Cependant cette matrice nous donne un aperçu des prédictions des individus considérés comme ensemble test à chaque itération.

		Observé	
		Positive(Oui)	Negative(Non)
Prédit	Positive	19	7
	Negative	7	24

TABLE 7 – Matrice de confusion pour le modèle GBM

### 5.6.2 Interprétabilité globale

Afin d'interpréter de façon globale quelles sont les covariables qui impactent la prédiction de la progression à la première évaluation, et dans quelles mesures elles jouent un rôle dans le calcul des prédictions, les outils suivants ont été calculés en ré-estimant le modèle de prédiction à partir du jeu de données complet, c'est-à-dire ajusté sur toute la population.

#### 1. Importance des variables

On cherche à déterminer les caractéristiques radiomiques qui contribuent le plus dans le modèle de prédiction. On réalise alors l'analyse de l'importance de features sur les données d'apprentissage utilisées dans notre modèle boîte noire (modèle GBM). Grâce à la fonction `variable_importance` du package `DALEX` on peut mesurer la perte de capacité discriminante d'après le RMSE<sup>1</sup>, au fur et à mesure qu'on permute les valeurs de chacune des variables. Les variables associées aux pertes de discrimination les plus fortes sont celles qui peuvent être considérées comme les plus essentielles dans l'estimation des prédictions des modèles.

---

1. Root mean square error

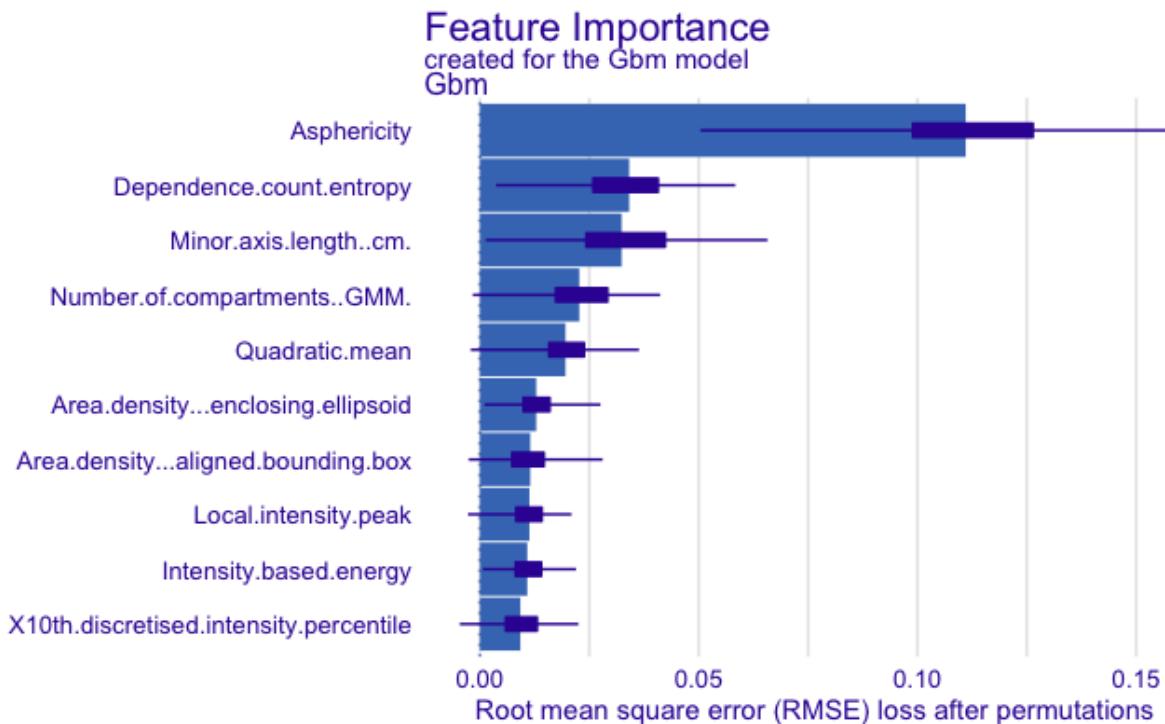


FIGURE 18 – Top 10 des variables les plus importantes pour la prédiction

On peut déduire de la figure 18 que les trois variables les plus influentes sont : *Asphericity*, *Dependence count entropy* et *Minor axis length*. Plus une barre est grande, plus la variable contribue à la prédiction car la perte de précision associée à cette variable est grande. En plus, le boxplot montre la répartition des valeurs de perte ( $RMSE_{X_{orig}} - RMSE_{X_{perm}}$ ) obtenues par les permutations (ici 99), il peut être interprété comme un intervalle de confiance<sup>1</sup>. Si on rompt le lien entre la variable *Asphericity* et la variable à prédire la perte du RMSE estimé est d'environ 0.11 (la plus élevée obtenue), avec un  $IC_{95\%} = [0.05, 0.16]$ , et ainsi de suite pour les autres variables.

Ainsi nous avons deux caractéristiques de forme (*Asphericity*, *Minor axis length* et une caractéristique de texture (*Dependence count entropy*). L'*Asphericity* ou l'asphérité décrit dans quelle mesure la ROI<sup>2</sup> s'écarte d'une sphère parfaite, les volumes parfaitement sphériques ayant une asphérité de 0. Aussi le *Minor axis length* ou la longueur du petit axe de la ROI donne une mesure de la distance à laquelle le volume s'étend le long du deuxième axe le plus grand. La *Dependence count entropy* est l'entropie de la matrice de dépendance des niveaux de gris voisins (NGLDM : Cette matrice vise à capturer la grossièreté de la texture globale et est invariante par rotation. En résumé elle témoigne de la proximité de voxels de même niveau de gris). [Zwanenburg et al., 2020].

1. Cet intervalle de confiance est construit autour de la médiane. Sachant que IQR : intervalle interquartile, l'intervalle est défini donc par : médiane  $+/- 1.58 * IQR / \sqrt{n}$

2. Region of Interest

## 2. Relations entre caractéristiques radiomiques et prédictions

Après avoir identifié les features radiomiques qui jouent un rôle important dans la prédiction de la progression à la première évaluation, l'étape suivante consiste à comprendre l'effet de chacune de ces variables sur les prédictions obtenues par le modèle GBM. Comme on a vu dans les sections 3.5.2, 3.5.3, 3.5.4, le PDP, l'ICE et l'ALE permettent de caractériser ces liens. Le paquet R utilisé pour ces visualisations a été `iml`, avec chaque méthode spécifiée dans l'argument `method = c("pdp", "pdp+ice", "ale")`.

La figure 19 montre l'impact de la variable *Asphericity* sur les prédictions faites par le modèle GBM. La variable décrit dans quelle mesure la ROI<sup>1</sup> s'écarte d'une sphère parfaite. Plus la ROI s'écarte de la forme d'une sphère ( valeur s'éloignant de 0), plus la valeur de la variable augmente et la probabilité d'un patient de progresser augmente. A titre d'illustration, pour cette variable en particulier, d'après le PDP, un individu avec une Asphéricité égal à 1 a une probabilité marginale estimée à environ 61% de progresser ( variable d'intérêt = "Yes"). Pour le graphique ICE, on observe que toutes les courbes individuelles suivent la même tendance), néanmoins les trajectoires sont assez différentes (point de départ). La ligne jaune représente le comportement moyen (PDP). Enfin, pour l'ALE on observe qu'une augmentation de la variable traduit une augmentation de la probabilité prédite par rapport à la probabilité moyenne.

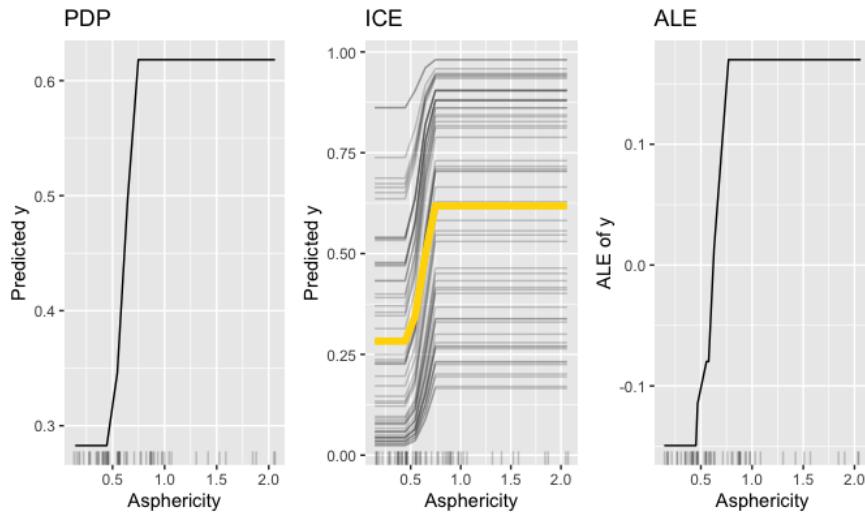


FIGURE 19 – De gauche à droite : PDP, ICE et ALE pour la variable Asphéricité

Les figures 20 et 21 représentent respectivement l'influence des variables *Dependence count entropy* et *Minor axis length* sur les prédictions faites par le modèle. Les informations apportées par ces deux figures sont similaires à celle précédentes ainsi que les analyses qui en ressortent.

1. Region of Interest

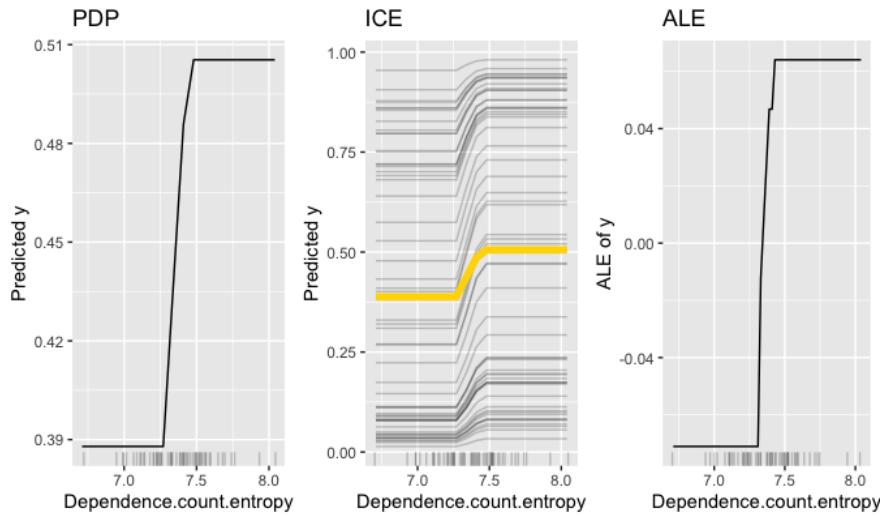


FIGURE 20 – De gauche à droite : PDP, ICE et ALE pour la variable Dependence count entropy

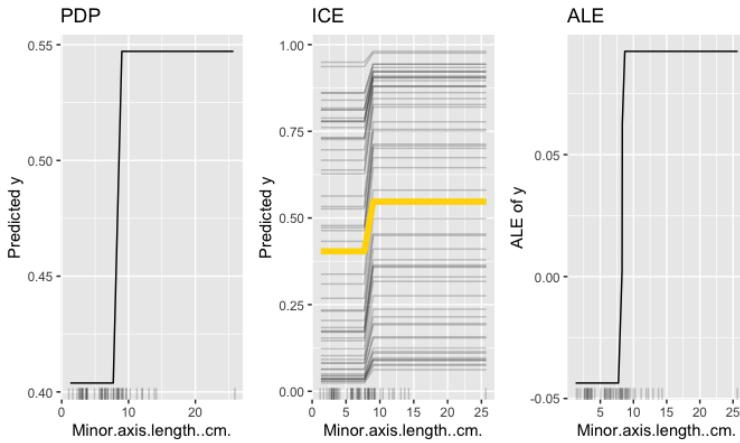


FIGURE 21 – De gauche à droite : PDP, ICE et ALE pour la variable Minor axis length

### 3. Modèle de substitution global

Le modèle Gradient Boosting Machine n'est pas un modèle facile à interpréter. Pour cela on a utilisé un modèle de substitution interprétable qui tente d'expliquer les prédictions issues du modèle. La technique choisie a été un arbre de régression conditionnel, obtenu à partir de `cmtree` du package `partykit`.

Le modèle de substitution construit à partir de tous les variables donne un  $R^2$  de 0.36 cela nous amènerait à conclure que les approximations du modèle GBM sont pas suffisantes pour remplacer le modèle GBM initial. Cependant à partir des trois variables les plus importantes du modèle GBM, le modèle de substitution capture 76% de la variabilité des prédictions. Ce modèle de substitution peut remplacer le modèle GBM initial. La modélisation des probabilités issues du modèle GBM à partir des variables importantes est visible sur la figure 22 .

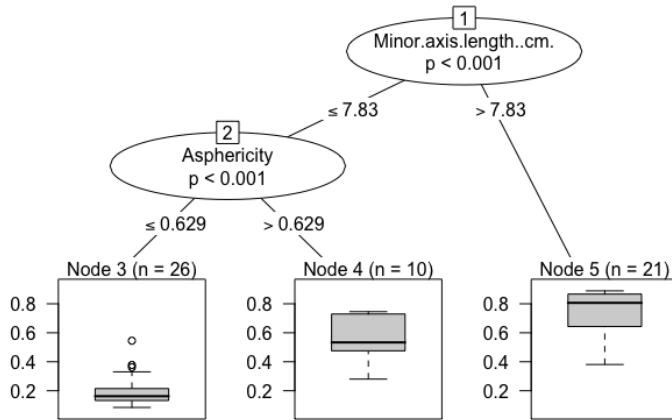


FIGURE 22 – Arbre de décision obtenu comme modèle de substitution global

La distribution des noeuds terminaux nous montre que le modèle de substitution prédit un nombre élevé de patients qui ont progressé à la première évaluation lorsque le *Minor axis length* est supérieure à 7.83. Traduisant ainsi un fort étirement du volume sur le long du deuxième axe le plus grand de la tumeur. Aussi lorsque la variable *Minor axis length* est inférieure à 7.83 et la variable *Asphericity* est supérieure à 0.629, la probabilité pour qu'un patient progresse est légèrement élevé (médiane à 0.5).

Tandis que lorsque la variable *Minor axis length* est inférieure à 7.83 et la variable *Asphericity* est inférieure à 0.629 la probabilité des patients à progresser est très basse.

Un  $R^2$  de 0.78 peut être vu comme une bonne approximation mais pas suffisante pour remplacer le modèle initiale par le modèle surrogate.

Les visualisations étudiées aident à comprendre le modèle de machine learning dans une perspective globale : en identifiant les variables ayant le plus grand impact global sur la variable réponse. Cependant, ces outils ne sont pas adaptés à l'explication d'une nouvelle observation. On va donc appliquer les techniques d'interprétabilité locale afin de comprendre quelles sont les caractéristiques les plus influentes pour déterminer la prédiction d'un nouvel individu.

### 5.6.3 Interprétabilité locale

Afin d'illustrer l'utilisation des méthodes d'interprétabilité locale, nous avons généré une observation synthétique à l'aide de la fonction `syn` et du package `synthpop` de R. [Beata Nowok and Dibben, 2016]. Il s'agit d'un patient atteint du cancer du poumon ayant une probabilité prédite de ne pas progresser à la première évaluation de 0.85. Les méthodes d'interprétabilité locale expliquées dans la section 3.7.1 permettront de comprendre comme cette probabilité prédite a été obtenue.

#### 1. Modèle de substitution locale avec LIME

Le modèle local LIME cherche à expliquer la prédiction d'un individu à travers d'un modèle de facile interprétation. Pour ce faire, LIME génère des observations aléatoires autour de l'individu à expliquer, en tenant compte la distance entre cette observation et les observations créées. LIME ajuste un modèle de régression pondéré dans la localité d'appartenance de l'individu, en considérant comme variables prédictives les variables de départ discrétissées. En général, LIME utilise des régressions linéaires comme Ridge ou Lasso. Le modèle obtenu permettra d'expliquer en détail la prédiction de l'observation d'intérêt.

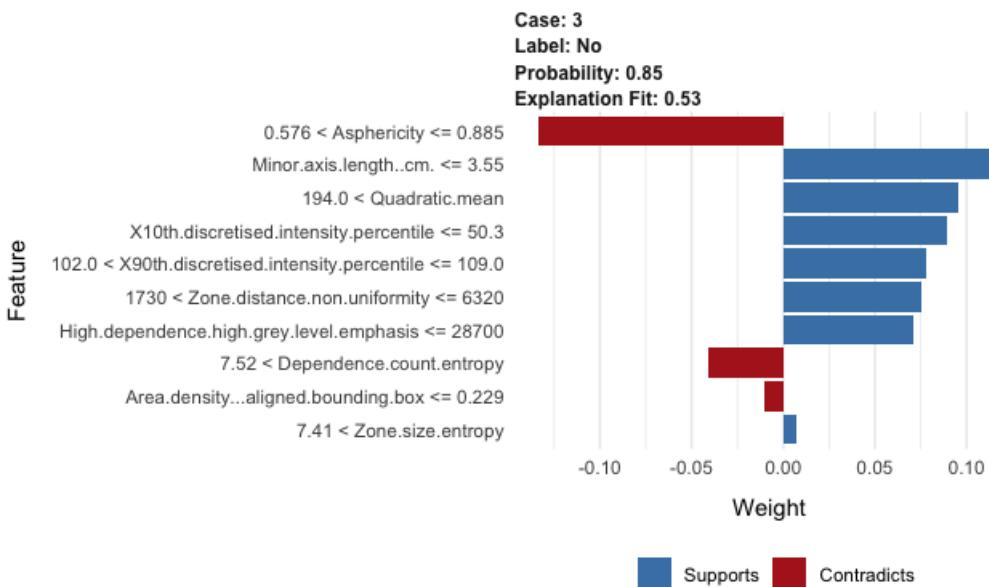


FIGURE 23 – Interprétabilité locale LIME pour une prédiction individuelle

La figure 23 expose les résultats du modèle local obtenu. Dans la visualisation, Probability correspond à la probabilité prédictée par le modèle GBM, et Explanation Fit correspond au  $R^2$  du modèle simple. Les 10 features les plus importantes qui expliquent le comportement de la région locale sont représentées par des barres : les Supports correspondent à une augmentation de la probabilité prédictée et les Contradits à une diminution de la probabilité prédictée pour l'individu d'intérêt.

Les probabilités prédictées par le modèle GBM et par le modèle local sont très proches. Cependant, on observe que le  $R^2$  n'est pas suffisamment élevé (tableau 8), comportement qui peut être dû à la définition du voisinage ou possiblement à la violation de l'hypothèse de linéarité dans la région évaluée. Il faut donc appliquer prudemment LIME, et vérifier que toutes les sorties soient cohérentes.

Prédiction GBM	Prédiction Surrogate Local	$R^2$ Surrogate Local
0.85	0.85	0.53

TABLE 8 – Prédictions obtenues par le modèle complexe et le modèle simple

## 5.7 Impact de la sélection des variables robustes

La sélection des variables robustes est réalisée à partir de perturbations des segmentations des patients suivant plusieurs Dices (0.99, 0.96, 0.93, 0.90, 0.87) et grâce au critère de sélection l'ICC fixé à deux seuils possibles : 0.90 ou 0.80. Dans cette partie nous allons mettre en évidence l'impact de cette sélection sur les performances prédictives pour en ressortir la meilleure configuration qui permet d'avoir à la fois une sélection robuste des variables et des bonnes performances prédictives.

### 5.7.1 Modélisation de l'ICC(1,1)

Pour réaliser la sélection de variables robustes nous avons utilisé les données des segmentations perturbées. Sur R, nous avons la possibilité de calculer l'ICC soit avec la fonction `icc` du package `irr` soit en utilisant un modèle linéaire à effets aléatoires avec la fonction `lme` du package `nlme`. Le calcul de l'ICC se fera ainsi donc pour chaque caractéristique radiomique.

Pour la première stratégie il faut créer une table de données avec en colonnes les différentes perturbations et en lignes les différents patients. Un exemple de la table de données pour l'utilisation de la fonction `icc` est la suivante :

	Perturbation_1	Perturbation_2	Perturbation_3	Perturbation_4	Perturbation_5
Patient_1	274.00	295.00	264.00	358.00	262.00
Patient_2	244.00	267.00	271.00	254.00	267.00
Patient_3	119.00	141.00	111.00	150.00	138.00
Patient_4	384.00	371.00	395.00	356.00	396.00
Patient_5	107.00	87.10	79.40	98.40	91.80
Patient_6	485.00	488.00	486.00	508.00	499.00

TABLE 9 – Perturbations de la variable "Grey level variance"

La fonction calcule et nous donne en sortie les estimations du modèle Anova à effets aléatoires choisi. Dans notre cas c'est l'ICC(1,1) qui a été utilisé. La fonction estime ainsi la valeur de l'ICC et son intervalle de confiance associé.

### 5.7.2 Sélection de la configuration des différents DICEs pour les perturbations et seuil de l'ICC la sélection de variables robustes

En utilisant plusieurs perturbations de segmentations différentes nous avons déduit la configuration la plus optimale pour la sélection de variables. Pour cela nous avons calculé pour chaque Dice donné à 0.99, 0.96, 0.93, 0.90 et 0.87, les coefficients ICCs de chaque variable. Une fois l'ensemble des coefficients calculé, un seuil a été fixé et les variables dont l'ICC est inférieur à celui-ci sont considérées comme non robustes et les autres comme robustes. Afin la bonne configuration du Dice, nous avons calculé le nombre de variables robustes pour toutes les différentes configurations du Dice selon les seuils d'ICCs suivants : 0.75, 0.80, 0.85, 0.90, 0.95, 0.99.

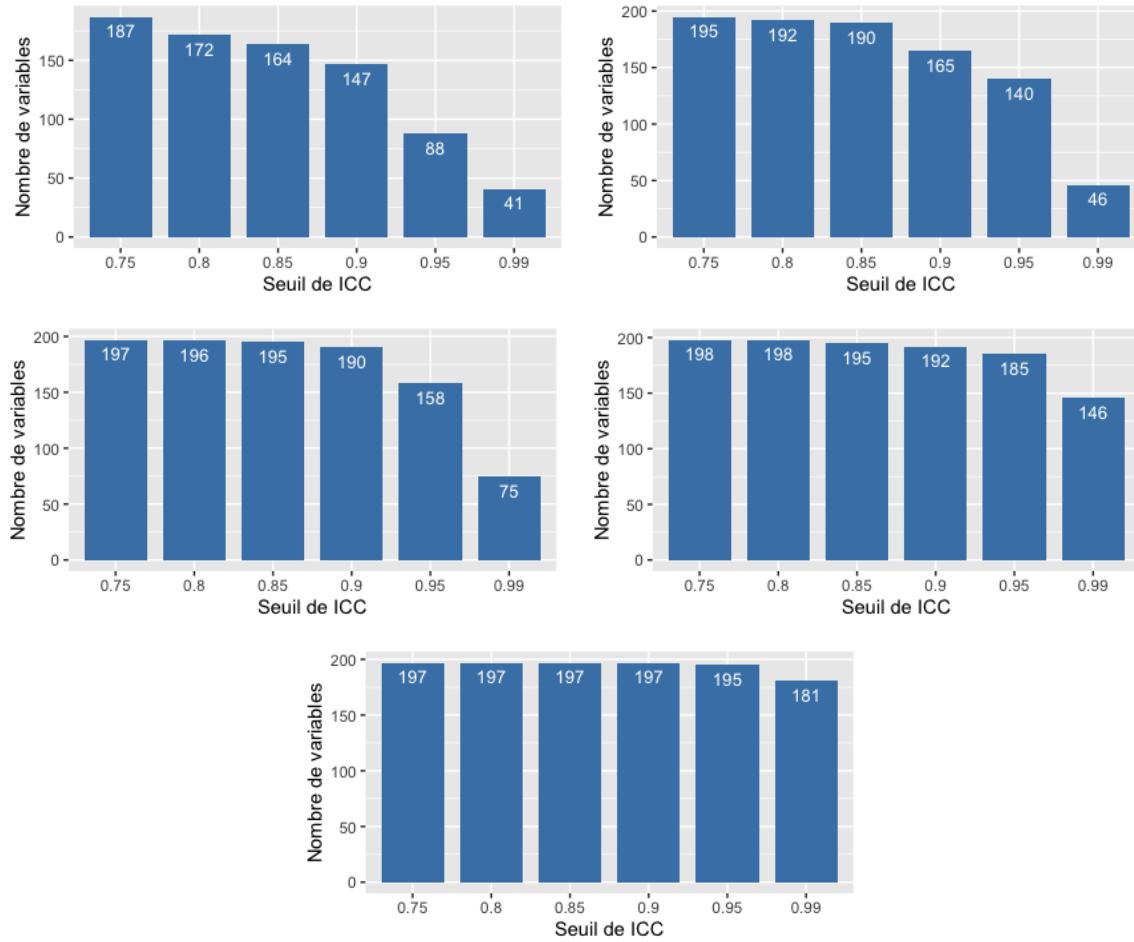


FIGURE 24 – Le nombre de variables robustes en fonction du seuil de L'ICC suivant les Dices dans l'ordre : 0.87, 0.90, 0.93, 0.96, 0.99

La figure 24 nous montre que plus le DICE est élevé plus le nombre de variables robustes est élevé et vice-versa. C'est tout à fait logique car plus le Dice est élevé, plus la segmentation perturbée se rapproche de la perturbation originale. Cela implique donc une diminution de la variabilité interopérateur et les caractéristiques radiométriques sont moins variables.

La configuration du Dice à 0.99 est éliminée car elle ne permet pas la détection de variables robustes (la segmentation originale et la segmentation perturbée sont quasi-identitique dans ce cas). C'est aussi le cas de la configuration du Dice 0.87, qui nous donne après visualisation une trop forte dissimilarité de la segmentation perturbée avec la segmentation originale, elle ne détecte que très peu de variables robustes. Cette situation s'éloigne donc de l'objectif de la simulation.

Nous avons donc choisi la configuration du Dice à 0.90 car elle élimine un assez gros nombre de variables sans trop virer d'information et aussi c'est ce qui est utilisé dans la littérature [Owens et al., 2018].

La bonne configuration du Dice trouvée il nous faut définir un seuil de l'ICC pour la sélection des variables robustes. Notre choix se portera entre les seuils utilisés ci-dessus. Dans la littérature les seuils fixés sont variables soit à 0.80 [Belli et al., 2018], [Lu et al., 2016], soit à 0.85 [Leijenaar et al., 2015] et soit à 0.9 [Lee et al., 2021]. Pour sélectionner le meilleur seuil nous avons utilisé la même méthodologie que dans la section 5.6.1 afin de prédire le statut de chaque patient à la première évaluation après initiation du traitement. Cette stratégie nous a permis de comparer les performances des algorithmes avec la sélection de variables robuste suivant les différents seuils.

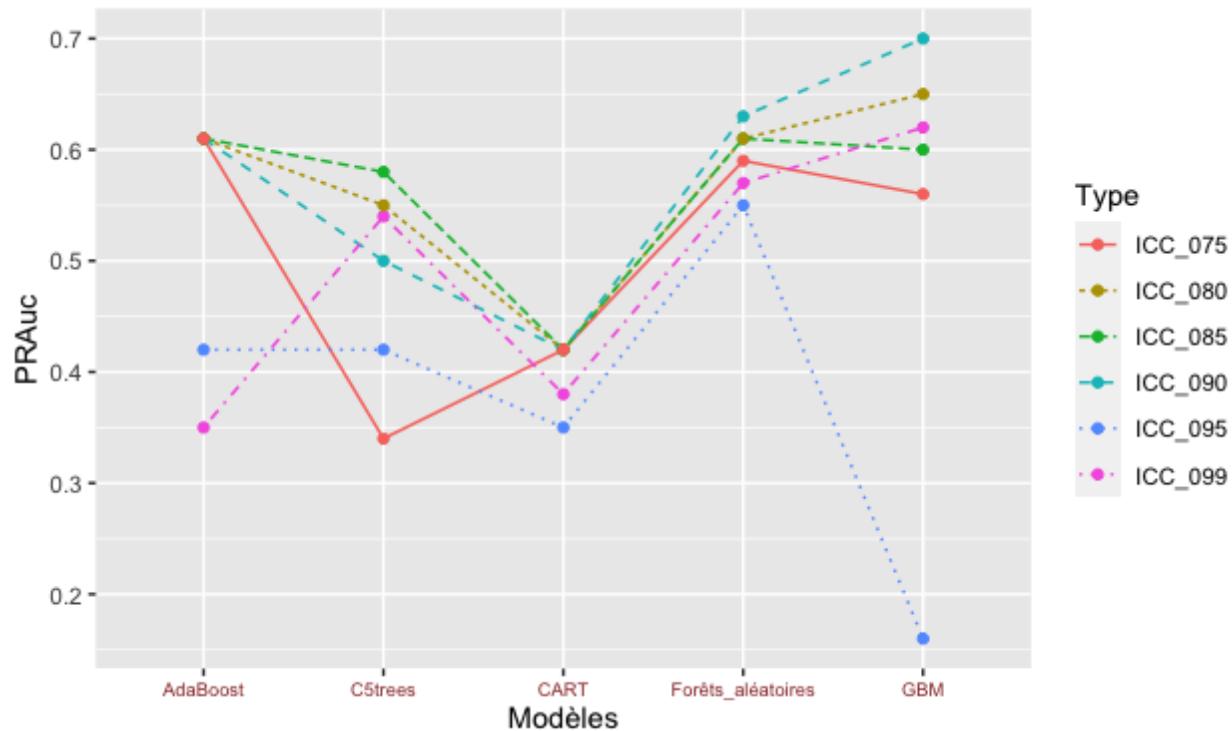


FIGURE 25 – Comparaison du PRAuc suivant les différents seuils de l'ICC pour le Dice 0.90

La figure 25 nous montre les performances au niveau du PRAuc<sup>1</sup> des 5 meilleurs algorithmes suivant les différents seuils de l'ICC. En prenant en compte donc la littérature [Duron et al., 2019] et les résultats des performances, le seuil que nous avons donc choisi pour la sélection de variables robustes dans notre étude est fixé à 0.90 (ratio performances prédictives/nombres de variables sélectionnées). Ce seuil est bien celui qui maximise l'algorithme sélectionné (GBM) par l'analyse de prédiction (section 5.6.1).

1. généralisation du F1score pour tous les seuils de décisions possibles

### 5.7.3 Sélection des variables robustes à partir de l'ICC

La configuration pour les perturbations étant trouvée il faut à présent choisir la stratégie de sélection des variables robustes grâce à l'ICC. Pour cela deux stratégies ont été utilisées, la première est l'utilisation de la valeur de l'estimation de l'ICC [Granzier et al., 2020] et la deuxième est l'utilisation de la borne inférieure estimée de l'intervalle de confiance à 95% de la valeur de l'ICC [Zwanenburg et al., 2019].

#### I - Sélection des caractéristiques robustes selon l'ICC estimé :

Une fois l'ensemble des coefficients calculés pour chaque variable, un seuil est fixé (ici on utilise 0.90) et l'ensemble des variables dont la valeur de l'ICC est inférieure à 0.90 sont considérés comme non robustes et dans le cas contraires elle sont dites robustes. Les résultats du nombre de variables pour différents seuils de l'ICC pour la configuration choisie (Dice = 0.90) sont visibles figure 24.

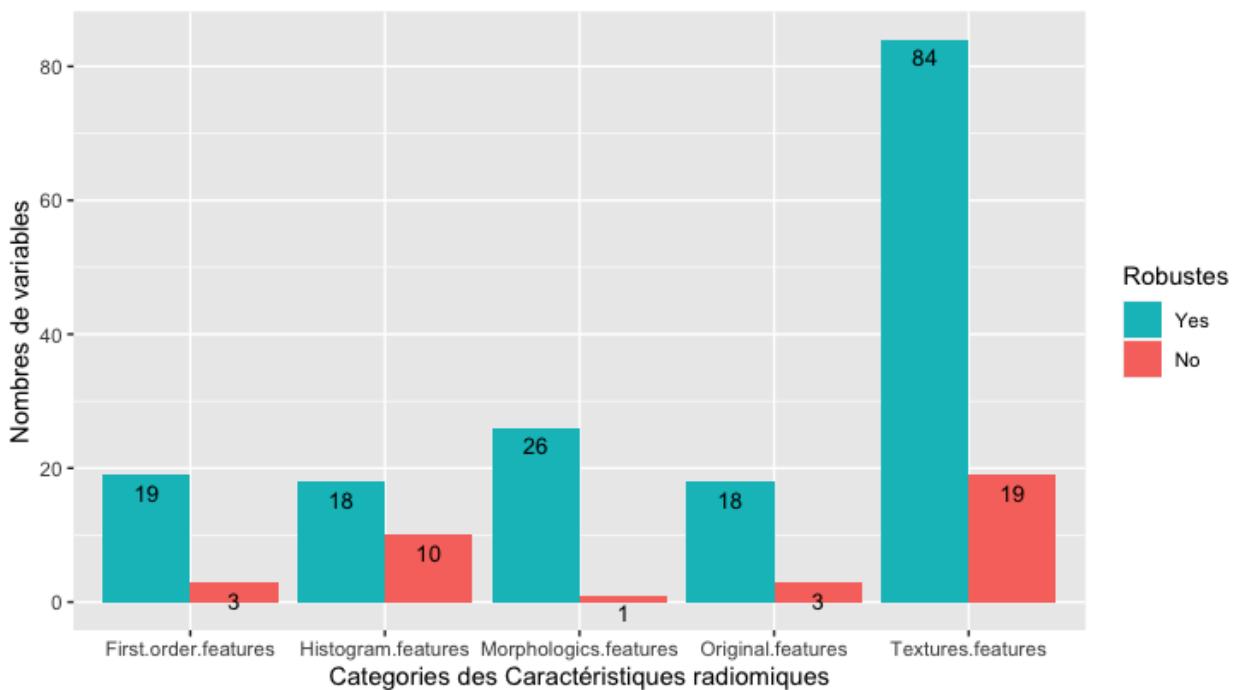


FIGURE 26 – Nombre de variables robustes par classes

La figure 26 résume par classes de caractéristiques radiomiques le nombre de variables robustes sélectionnées ou non, on peut remarquer que dans toutes les catégories de caractéristiques radiomiques il y a la présence de variables non robustes. La classe des statistiques des histogrammes est la classe la plus impactée par cette sélection. Cependant les catégories des caractéristiques de textures et des caractéristiques morphologiques sont celles qui contiennent le plus de caractéristiques robustes. La liste détaillée par classes de caractéristiques des variables robustes sélectionnées/ non est présentée sur la figure 50. Les caractéristiques de couleurs vertes sont celles qui sont retenues comme robustes et celles de couleurs rouges sont celles qui ne sont pas robustes. La suite des graphiques par classes est visible en section 8

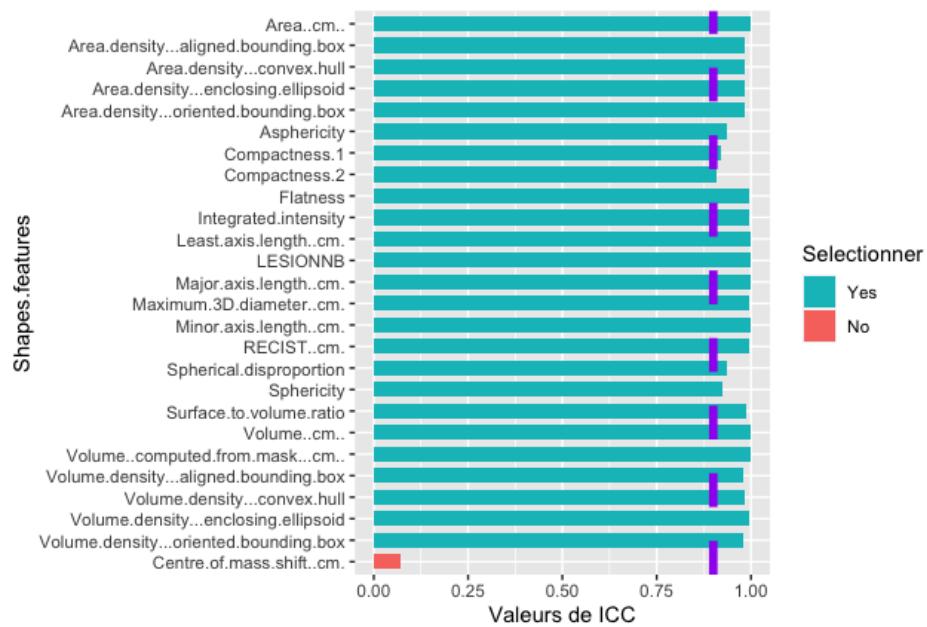


FIGURE 27 – Shapes features selection

Le récapitulatif des variables non sélectionnées avec la valeur de leur ICC est visible dans le tableau ci-dessous :

TABLE 10 – Variables non robustes

Variables	ICC	Variables	ICC
1 Intensity.based.coefficient.of.variation	6.2e-09	19 Small.zone.high.grey.level.emphasis	0.87
2 Intensity.based.coefficient.of.variation..Original.Data.	0.02	20 Small.zone.low.grey.level.emphasis	0.87
3 Centre.of.mass.shift..cm.	0.07	21 Sum.average	0.87
4 Quartile.coefficient.of.dispersion..Original.Data.	0.26	22 X90th.discretised.intensity.percentile	0.87
5 Number.of.compartments..GMM.	0.74	23 Autocorrelation	0.88
6 Maximum.histogram.gradient.intensity	0.79	24 High.dependence.low.grey.level.emphasis	0.88
7 Thresholded.area.intensity.peak..50..	0.79	25 High.grey.level.count.emphasis	0.88
8 Volume.at.intensity.fraction.10.	0.79	26 High.grey.level.run.emphasis	0.88
9 Minimum.histogram.gradient.intensity	0.82	27 Median.discretised.intensity	0.88
10 Local.intensity.peak	0.83	28 Low.dependence.low.grey.level.emphasis	0.89
11 Intensity.histogram.mode	0.85	29 Small.distance.low.grey.level.emphasis	0.89
12 Thresholded.area.intensity.peak..75..	0.86	30 Volume.at.intensity.fraction.90.	0.89
13 Area.under.the.IVH.curve	0.87	31 Volume.fraction.difference.between.intensity.fractions	0.89
14 High.grey.level.zone.emphasis	0.87	32 Max.value	0.895
15 High.grey.level.zone.emphasis.1	0.87	33 Low.grey.level.zone.emphasis	0.897
16 Joint.average	0.87	34 Low.grey.level.zone.emphasis.1	0.897
17 Mean.discretised.intensity	0.87	35 Long.run.high.grey.level.emphasis	0.898
18 Short.run.high.grey.level.emphasis	0.87	36 Min.value..Original.Data.	0.73

Parmi les variables non sélectionnées comme robustes, nous avons des variables d'un peu toutes

les classes de caractéristiques radiomiques et principalement de la classe des caractéristiques de l'histogramme des intensités. (voir figure 26). Concernant les variables non robustes, avant l'étude nos services en soupçonnaient quelques unes à cause de leurs méthodologies de calcul. Parmi celles-ci le *Max value* et *Min value Orignal Data*, ces variables caractérisent respectivement la valeur maximale et minimale des zones très foncées et très claires de la RIO. Il suffit donc qu'une perturbation de segmentations déborde légèrement sur une zone très foncées ou très claires pour que ces valeurs fluctuent drastiquement. Ce débordement explique aussi donc la présence d'autant de caractéristiques radiomiques issues de l'histogramme des intensités comme variables non robustes à savoir *Intensity.based.coefficient.of.variation*, *X90th.discretised.intensity.percentile etc...*

## II - Sélection des caractéristiques robustes selon la borne inférieure estimée de l'intervalle de confiance à 95% de l'ICC

Ici la stratégie est la même que la partie précédente à l'exception que le seuil de l'ICC pour la sélection est fixé non pas par la valeur de l'ICC de chaque variable mais par la valeur de la borne inférieure de l'intervalle de confiance de celle-ci. Avant de se fixer ce seuil arbitraire nous avons calculé le nombre de variables sélectionnées suivant différents seuils : 0.75 , 0.80, 0.85, 0.90, 0.95, 0.99. Les résultats se présentent sous forme de graphique en barre suivant :

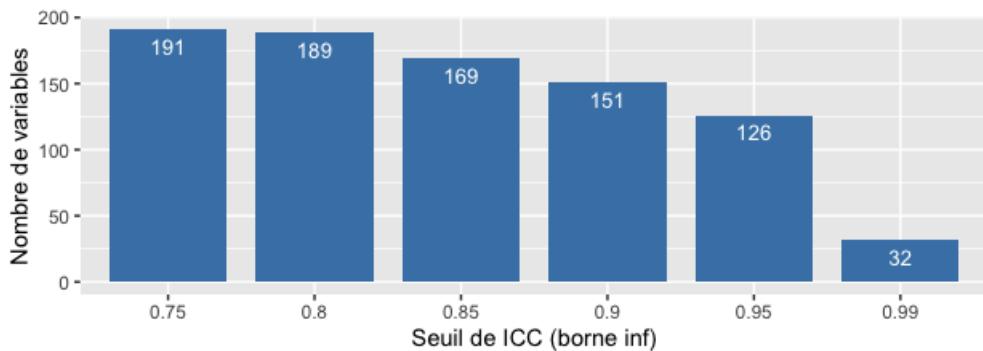


FIGURE 28 – Le nombre de variables sélectionnées en fonction du seuil de l'ICC (borne inférieure de l'IC à 95%)

Dans la continuité de la section précédente le seuil est fixé à 0.90, on considère donc comme robuste l'ensemble des variables dont la borne inférieure de l'intervalle de confiance de la valeur de l'ICC est supérieure à 0.90 . La figure 29 résume par classes de caractéristiques radiomiques le nombre de variables robustes sélectionnées ou non, on note que le nombre de variables non robustes à augmenter comparativement à la méthode de sélection précédente.

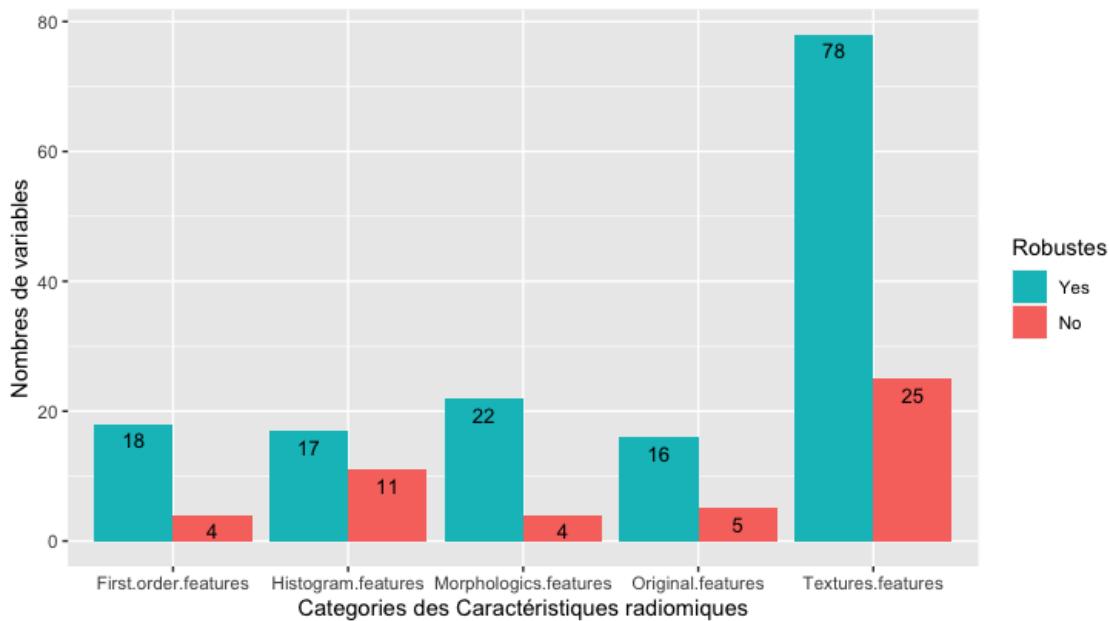


FIGURE 29 – Nombre de variables robustes par classes

Les résultats ci-dessous présentent respectivement l'ensemble des variables sélectionnées suivant leurs classes de variables en graphique en barre et l'ensemble des variables non-sélectionnées sous forme de tableau. La suite des graphiques se trouve en section 8.

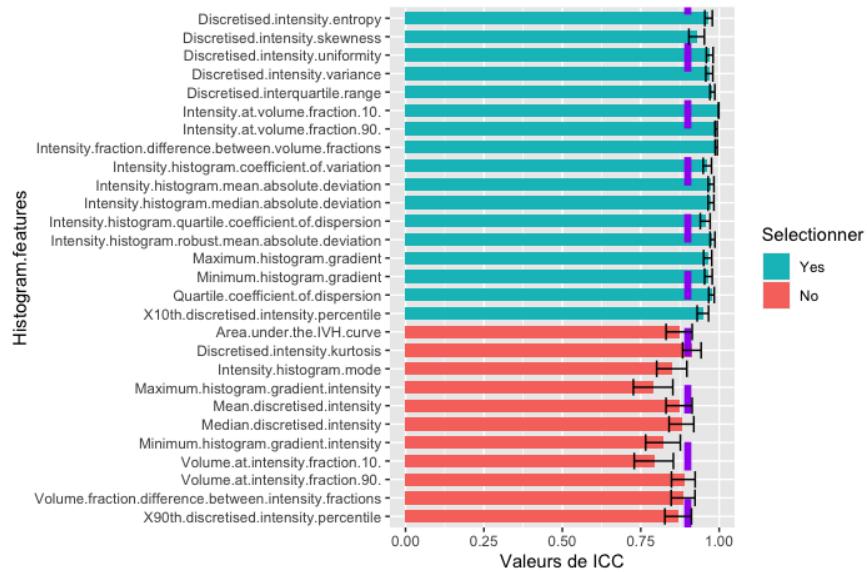


FIGURE 30 – Histogram features selection

Le récapitulatif des variables non sélectionnées avec la valeur de leur ICC est visible dans le

tableau ci-dessous :

	Variables	ICC	Inf	Sup
1	Intensity.based.coefficient.of.variation	6.2e-09	5.1e-10	0.01
2	Intensity.based.coefficient.of.variation..Original.Data.	0.02	0.002	0.06
3	Centre.of.mass.shift..cm.	0.07	0.04	0.12
4	Quartile.coefficient.of.dispersion..Original.Data.	0.26	0.19	0.36
5	Number.of.compartments..GMM.	0.74	0.66	0.81
6	Min.value..Original.Data.	0.74	0.67	0.81
7	Thresholded.area.intensity.peak..50..	0.79	0.73	0.85
8	Maximum.histogram.gradient.intensity	0.79	0.73	0.85
9	Volume.at.intensity.fraction.10..	0.79	0.73	0.85
10	Minimum.histogram.gradient.intensity	0.82	0.77	0.88
11	Local.intensity.peak	0.83	0.77	0.88
12	Intensity.histogram.mode	0.85	0.80	0.90
13	Thresholded.area.intensity.peak..75..	0.86	0.81	0.90
14	Small.zone.low.grey.level.emphasis	0.87	0.82	0.91
15	X90th.discretised.intensity.percentile	0.87	0.83	0.91
16	Sum.average	0.87	0.83	0.91
17	Joint.average	0.87	0.83	0.91
18	High.grey.level.zone.emphasis.1	0.87	0.83	0.91
19	High.grey.level.zone.emphasis	0.87	0.83	0.91
20	Short.run.high.grey.level.emphasis	0.87	0.83	0.91
21	Small.zone.high.grey.level.emphasis	0.87	0.83	0.91
22	Mean.discretised.intensity	0.87	0.83	0.91
23	Area.under.the.IVH.curve	0.87	0.83	0.91
24	Autocorrelation	0.88	0.83	0.92
25	High.grey.level.run.emphasis	0.88	0.83	0.92
26	High.grey.level.count.emphasis	0.88	0.83	0.92
27	High.dependence.low.grey.level.emphasis	0.88	0.84	0.92
28	Median.discretised.intensity	0.88	0.84	0.92
29	Volume.fraction.difference.between.intensity.fractions	0.89	0.85	0.92
30	Volume.at.intensity.fraction.90..	0.89	0.85	0.92
31	Low.dependence.low.grey.level.emphasis	0.89	0.85	0.92
32	Small.distance.low.grey.level.emphasis	0.89	0.85	0.93
33	Max.value	0.90	0.86	0.93
34	Low.grey.level.zone.emphasis	0.90	0.86	0.93
35	Low.grey.level.zone.emphasis.1	0.90	0.86	0.93
36	Long.run.high.grey.level.emphasis	0.90	0.86	0.93
37	Large.distance.low.grey.level.emphasis	0.90	0.86	0.93
38	Low.dependence.high.grey.level.emphasis	0.90	0.87	0.94
39	Compactness.2	0.91	0.87	0.94
40	Max.value..Original.Data.	0.91	0.88	0.94
41	Intensity.range..Original.Data.	0.91	0.88	0.94
42	Kurtosis	0.91	0.88	0.94
43	Discretised.intensity.kurtosis	0.92	0.88	0.94
44	Compactness.1	0.92	0.89	0.95
45	Long.run.low.grey.level.emphasis	0.92	0.89	0.95
46	Short.run.low.grey.level.emphasis	0.92	0.89	0.95
47	Low.grey.level.run.emphasis	0.92	0.89	0.95
48	Low.grey.level.count.emphasis	0.92	0.89	0.95
49	Sphericity	0.92	0.899	0.95

Le nombre de variable non-robustes est passé de 36 à 49 soit une augmentation d'environ 36%. Cette augmentation est due au fait que le nombre de caractéristiques indéterminées est en corrélation avec le nombre de perturbations, car l'intervalle de confiance à 95 % de l'ICC se rétrécit avec l'augmentation des perturbations répétées. Il est donc possible d'augmenter le nombre de caractéristiques robustes et non robustes en augmentant le nombre de perturbations, mais avec des rendements décroissants. Le critère de sélection avec la valeur de l'ICC est moins rigoureux que la comparaison par rapport à l'intervalle de confiance et peut conduire à l'inclusion de caractéristiques qui ont une probabilité raisonnable (entre 2,5 et 50%) de ne pas répondre au critère. Cela est particulièrement le cas si les intervalles de confiance sont larges et se superposent à des valeurs  $ICC < 0.90$  (robustesse médiocre). Néanmoins on retrouve comme classes avec le plus de variables robustes la classe des caractéristiques morphologiques, la classe des caractéristiques de textures (comme avec la sélection par la valeur de l'ICC), mais aussi la classe des statistiques de premier ordre.

**Remarque :** Plusieurs caractéristiques de par leur non normalité (distribution non gaussienne) se sont avérées non robustes, c'est le cas de la variable *Center of mass shift*. Cependant suite à des transformations de ces variables notamment en utilisant la fonction "logarithme népérien" celles-ci deviennent robustes ( $ICC > 0.90$ ). Cela est dû aux hypothèses gaussiennes du modèle Anova à un facteur aléatoire. Parmi l'ensemble de ses variables seule la variable centre de masse shift apporte une amélioration des modèles de prédiction, elle sera donc conservée lors de la sélection de variables robustes.

## 5.8 Comparaison des différentes méthodes de sélections

Notre sélection de variables s'est faite à partir des deux bases de données des variables robustes (sélection suivant la valeur de l'ICC et sélection suivant IC à 95% de la valeur de l'ICC) en utilisant la même méthodologie de sélection qu'en section 5.6.1. Par la suite nous allons comparer suivant leurs performances prédictives respectives chacune de ses bases pour en déduire laquelle est la plus efficace.

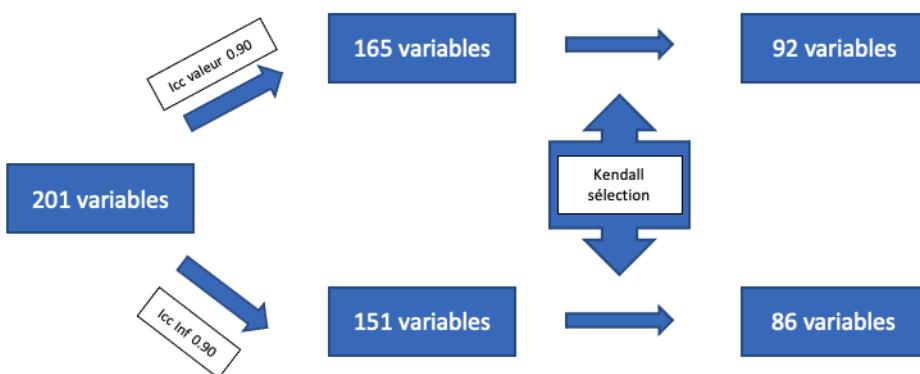


FIGURE 31 – Bilan des sélections de variables avec l'ICC

Les résultats des performances prédictives obtenues avec sélection de variables, soit selon l'ICC à 0.90, soit selon la borne inférieure de l'intervalle de confiance à 95% de l'ICC, sont résumés dans les Tables 11 et 12, respectivement.

	AUC	PRAuc	Accuracy	Sensitivity	Specificity	Precision	NPV	F1score	BrierScore
AdaBoost	0.75	0.61	0.74	0.85	0.65	0.67	0.83	0.75	0.17
C5trees	0.66	0.50	0.67	0.62	0.71	0.64	0.69	0.63	0.28
GBM	0.72	0.70	0.72	0.69	0.74	0.69	0.74	0.69	0.28
CART	0.73	0.42	0.72	0.85	0.61	0.65	0.83	0.73	0.22
Foret	0.68	0.62	0.68	0.65	0.71	0.65	0.71	0.65	0.23

TABLE 11 – Résultats de la sélection de variable suivant ICC à 0.90

	AUC	PRAuc	Accuracy	Sensitivity	Specificity	Precision	NPV	F1score	BrierScore
AdaBoost	0.75	0.61	0.74	0.85	0.65	0.67	0.83	0.75	0.17
C5trees	0.68	0.52	0.68	0.65	0.71	0.65	0.71	0.65	0.27
GBM	0.72	0.63	0.72	0.73	0.71	0.68	0.76	0.70	0.28
CART	0.73	0.42	0.72	0.85	0.61	0.65	0.83	0.73	0.22
Foret	0.68	0.63	0.68	0.65	0.71	0.65	0.71	0.65	0.22

TABLE 12 – Résultats de la sélection de variable suivant la borne inférieure de l'intervalle de confiance de l'ICC à 0.90

Nous notons que les performances prédictives issues des deux stratégies de sélection de variables robustes sont très proches. La métrique d'optimisation des algorithmes étant le F1-score, nous avons observé de plus près la différence entre ces deux résultats et les résultats issues de la prédiction dans la section 5.6.1 soit celle sans la sélection de variables robustes à travers les figures ci-dessous. Nous nous sommes focalisés sur le PRAuc qui généralise le F1-score suivant tous les seuils de décision possibles.

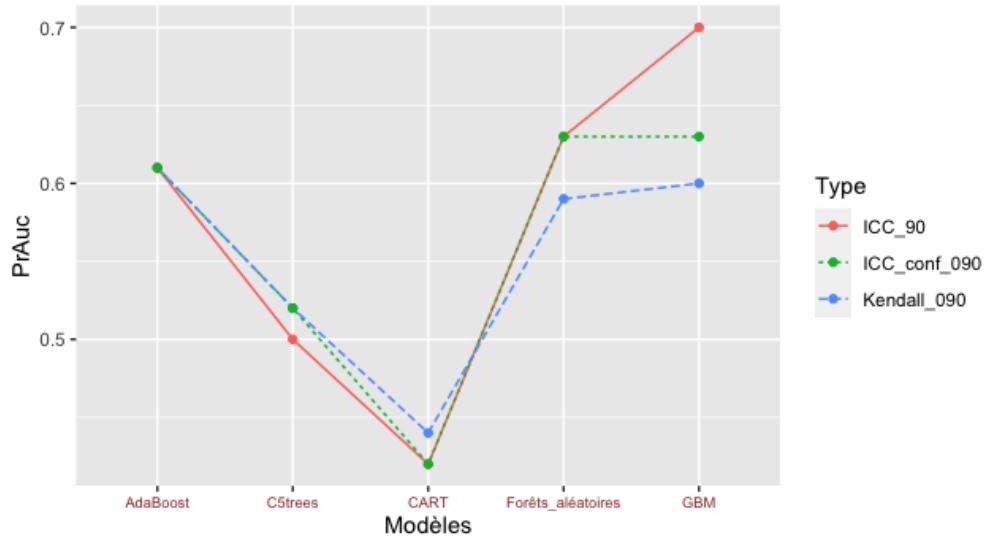


FIGURE 32 – Comparaison du PRAuc selon les trois méthodes de sélection

La figure 32 nous montre les performances des 5 meilleurs algorithmes issus des différentes bases (avec et sans les variables robustes). Au niveau du modèle GBM, la méthodologie de sélection incluant la sélection des variables robustes suivant la valeur de l'ICC semble se distinguer des deux autres sélections de variables (GBM avec ICC 0.90 a les meilleures performances). Dans l'ensemble, les différents modes de sélection semblent avoir des performances assez proches.

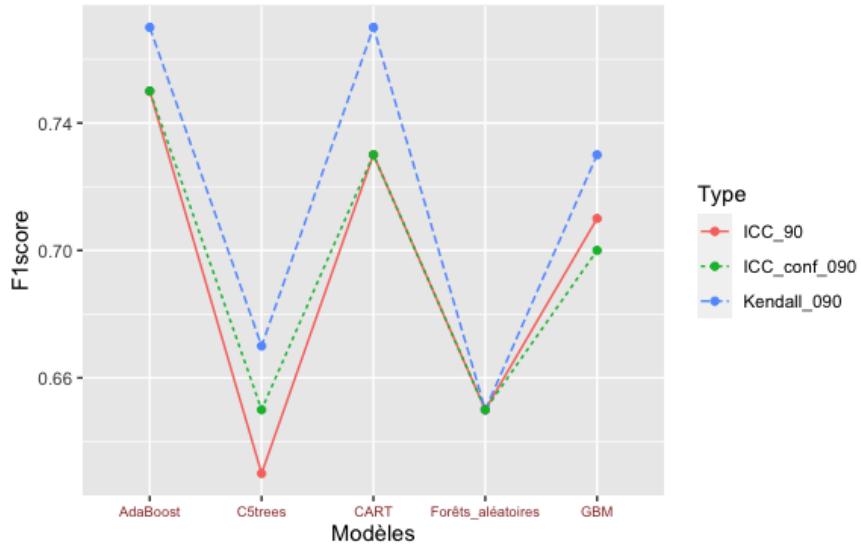


FIGURE 33 – Comparaison du F1score selon les trois méthodes de sélection

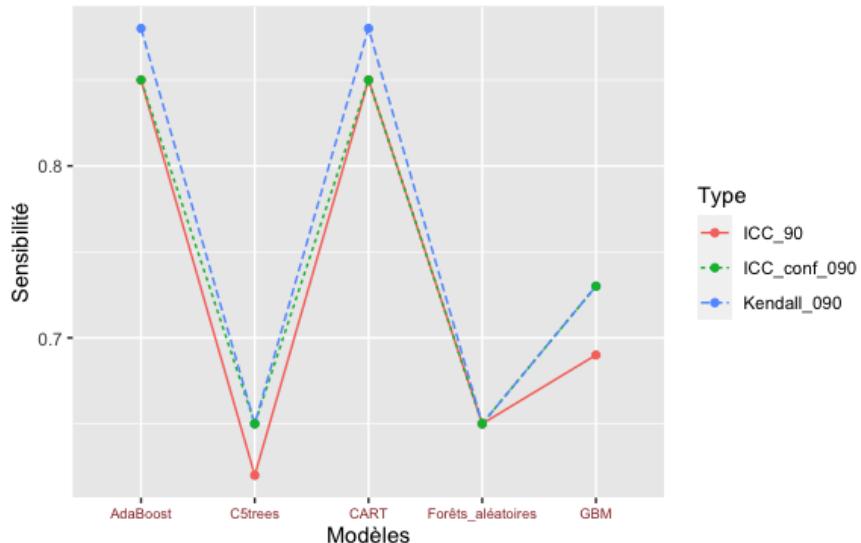


FIGURE 34 – Comparaison de la Sensibilité selon les trois méthodes de sélection

À travers ces deux précédents graphiques, la sélection de variable sans la sélection de variables robustes tend à avoir les meilleures performances, néanmoins elle ne plafonne pas de très loin les

sélections intégrant la sélection de variables robustes. Exemple pour le modèle sélectionné soit le modèle GBM nous sommes à un F1score de 0.73 comparé à 0.69 pour la sélection suivant la valeur de l'ICC et 0.70 pour la sélection suivant la borne inférieure de l'intervalle de confiance de la valeur de l'ICC. Les résultats sont sensiblement égaux, sachant aussi que l'on n'a pas à disposition un jeu de données avec beaucoup d'individus. Il n'y a pas de différence flagrante entre les performances de la sélection avec et sans variables robustes.

En effet l'objectif était de pouvoir créer des signatures robustes, signatures modélisées à l'aide des caractéristiques radiomiques robustes. À travers les résultats ci-dessus nous retrouvons que les performances prédictives des algorithmes sur la base de toutes les variables et celles sur la base des caractéristiques radiomiques robustes sont similaires à peu près. La sélection avec l'ICC 0.90 a donné les meilleures performances au niveau du PRAuc avec le meilleur modèle sélectionné le GBM. Nous avons, à l'aide de variables robustes uniquement, put obtenir des performances prédictives aussi bien que celles obtenues sans le processus de sélection de variables robustes.

Les résultats obtenus nous montrent donc que l'utilisation des variables robustes peut améliorer les résultats des algorithmes prédictifs ou à défaut faire aussi mieux que l'utilisation de toutes les variables.

## 5.9 Interprétabilité des modèles

Les modèles issues des modélisations incorporant la sélection des caractéristiques radiomiques robustes ainsi trouvés, nous allons réalisé dans cette partie leur interprétabilité pour établir une cohérence avec les résultats précédents de la section 5.6.2.

### 5.9.1 Interprétabilité Globale

#### 1. Importance des variables

La figure 35 nous montre le top des variables qui contribuent le plus dans le modèle de prédiction issues des variables robustes (respectivement de gauche à droite, issues de la sélection à l'aide de la valeur de l'ICC et issues de la sélection à l'aide de l'IC de l'ICC à 0.90).

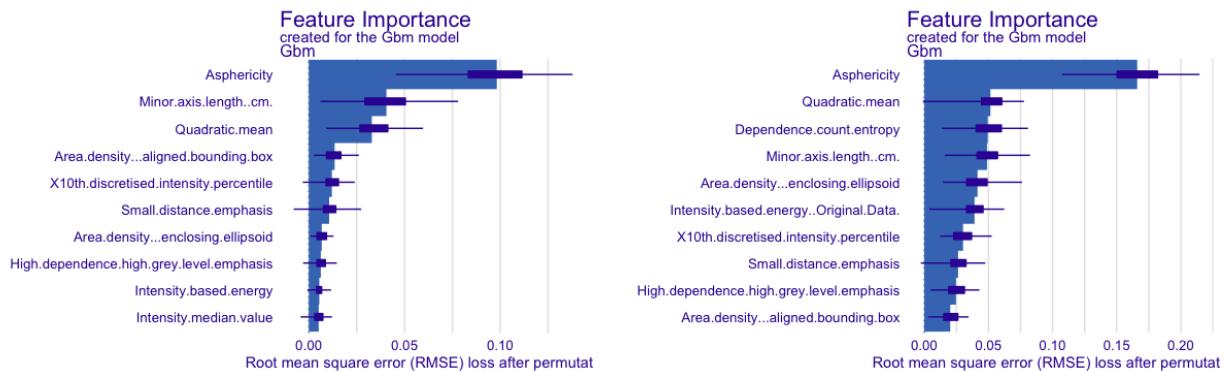


FIGURE 35 – De gauche à droite le top 10 des variables importantes du modèle avec l'ICC 0.90 et avec l'IC de l'ICC 0.90

Sur l'ensemble des variables importantes on note une certaine cohérence avec celle sélectionnée auparavant. Cependant nous pouvons voir, parmi les trois variables les plus importantes, la montée de la caractéristique *Quadratic mean* dans les deux modèles, c'est une caractéristique radiomique de la classe des statistiques de premier ordre (*First order features*). La *quadratic mean* ou *Root mean square intensity feature* c'est la racine carrée de la moyenne des carrés des intensité des voxels inclus dans le masque d'intensité du ROI [Zwanenburg et al., 2020].

## 2. Relations entre caractéristiques radiomiques et prédictions

Les graphiques suivants nous montrent l'impact des top 3 variables importantes sur les prédictions faites par le modèle GBM modélisé avec la sélection des variables robustes utilisant d'une part l'ICC (figure 36) et d'autre part la valeur de l'IC de l'ICC (figure 37).

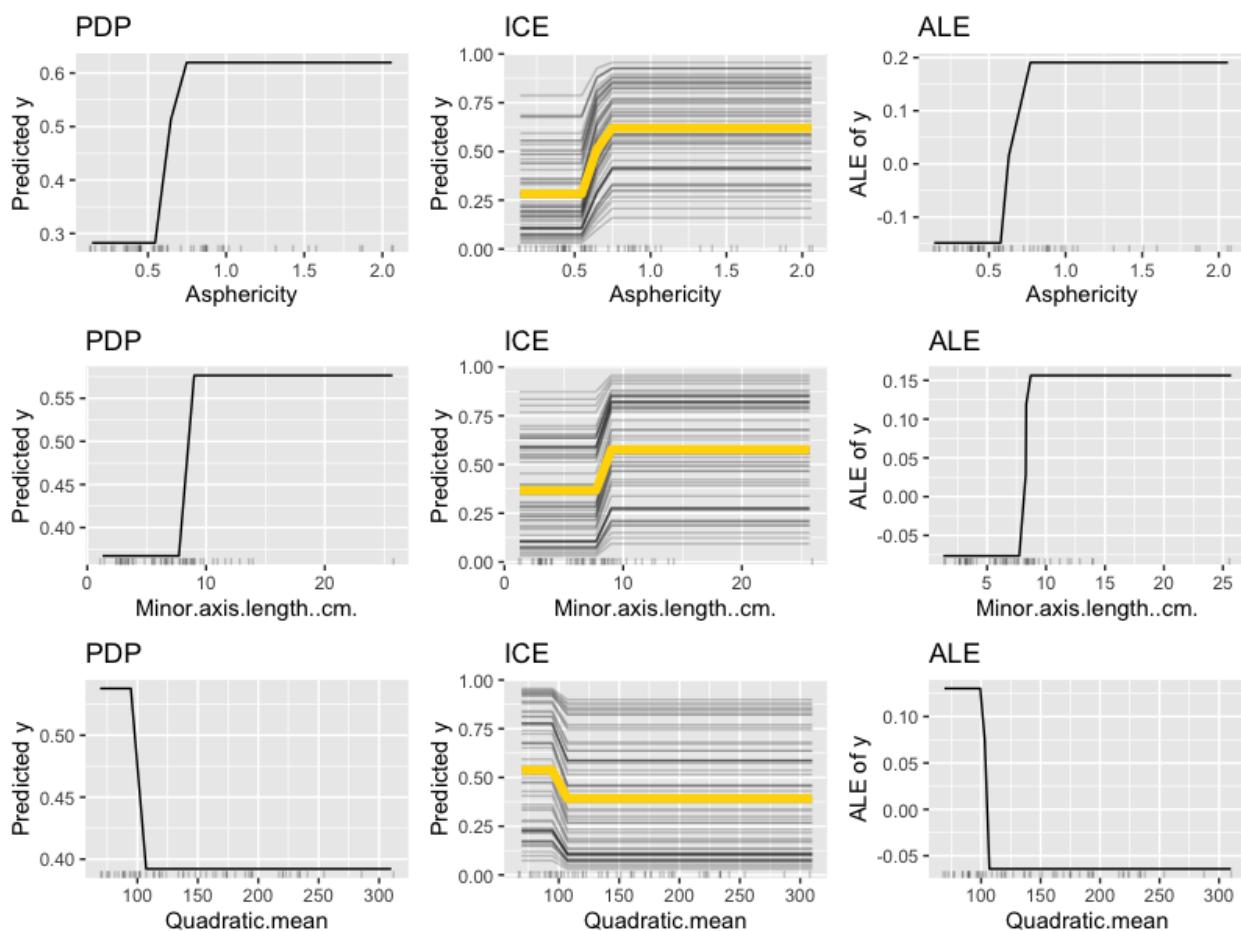


FIGURE 36 – De gauche à droite : PDP, ICE et ALE pour le top 3 des variables importantes : ICC 0.90

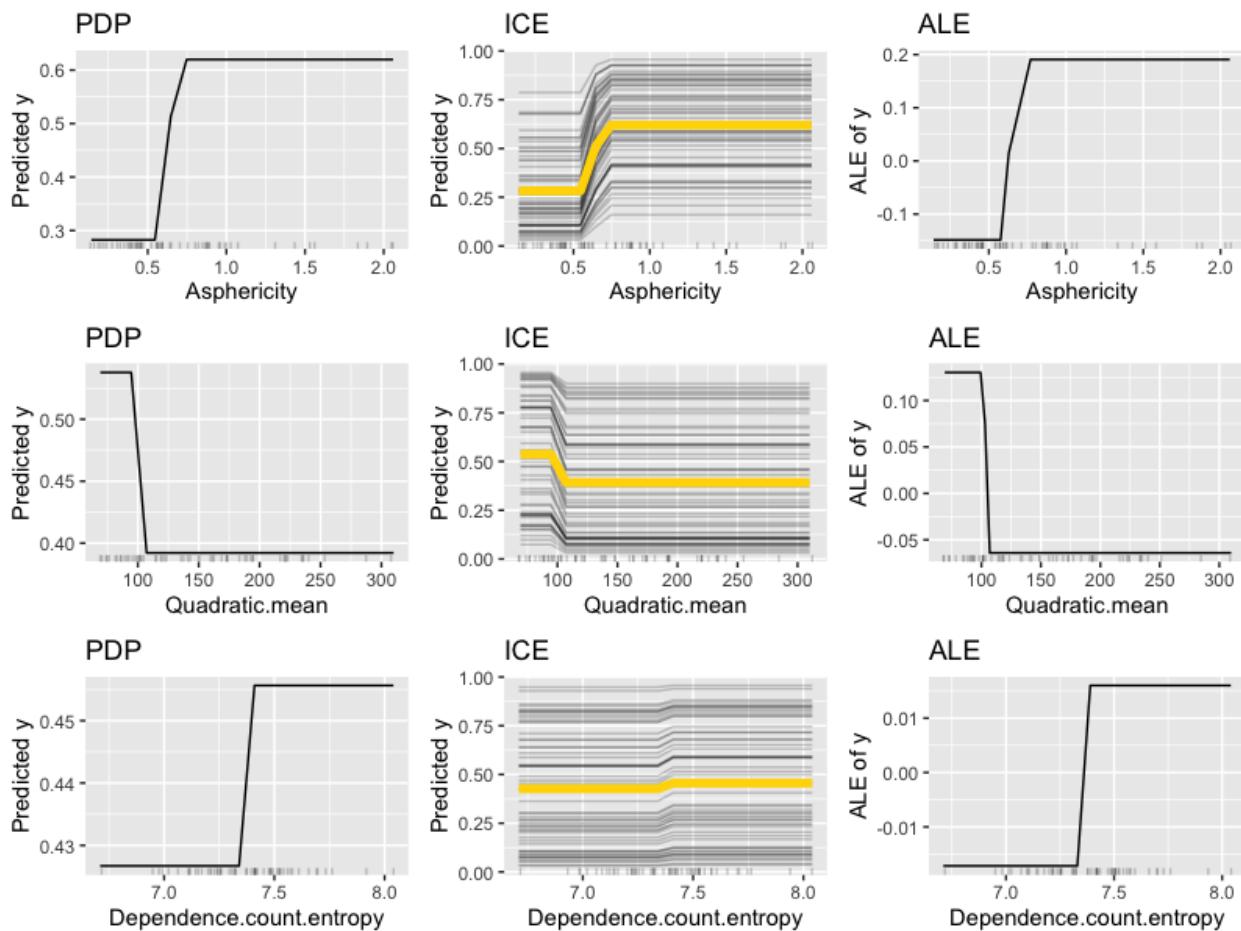


FIGURE 37 – De gauche à droite : PDP, ICE et ALE pour le top 3 des variables importantes : IC de l'ICC 0.90

Les variables qui apparaissent et leurs interprétations ne diffèrent pas avec les résultats précédents voir section 5.6.2. Cependant la nouvelle variable top 3, *Quadratic mean*, d'après le PDP, un individu avec une moyenne quadratique supérieure ou égal à 100 environ a une probabilité marginale estimée à environ près de 60% de progresser (variable d'intérêt = "Yes"). Pour le graphique ICE, on observe que toutes les courbes individuelles suivent la même tendance, néanmoins les trajectoires sont assez différentes (point de départ), la ligne jaune représentant le comportement moyen (PDP). Enfin, pour l'ALE on observe qu'une augmentation de la variable traduit une diminution de la probabilité prédictive par rapport à la probabilité moyenne.

### Modèle de substitution global

#### 1. Première base : ICC 090

Le modèle de substitution construit à partir des variables robustes sélectionnées suivant l'ICC 0.90 donne un  $R^2$  de 0.49 cela nous amènerait à conclure que les approximations du modèle

GBM sont pas suffisantes pour remplacer le modèle GBM initial. Cependant à partir des trois variables les plus importantes du modèle GBM, le modèle de substitution capture 88% de la variabilité des prédictions. Ce modèle de substitution peut remplacer le modèle GBM initial. La modélisation des probabilités issues du modèle GBM à partir des variables importantes est visible sur la figure 38

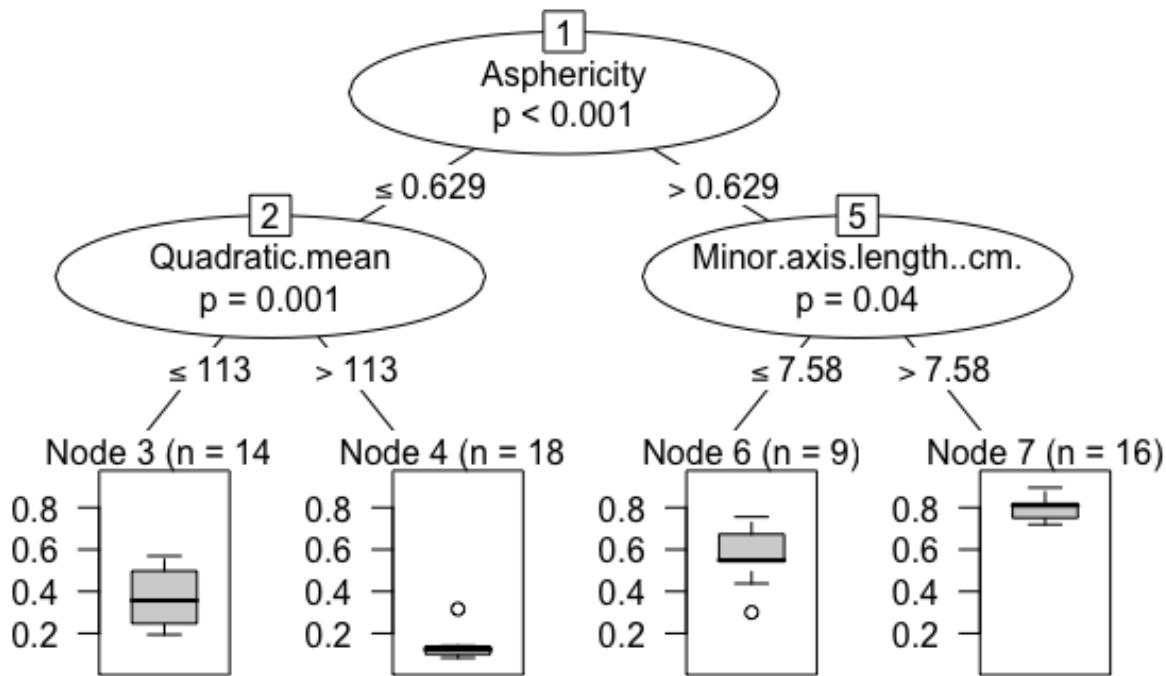


FIGURE 38 – Arbre de décision obtenu comme modèle de substitution global

## 2. Deuxième base : IC de l'ICC 0.90

Dans ce cas, le modèle de substitution construit à partir des variables robustes donne un  $R^2$  de 0.34, les approximations du modèle GBM ne sont toujours pas suffisantes pour remplacer le modèle GBM initial. Cependant à partir des trois variables les plus importantes du modèle GBM, le modèle de substitution capture 80% de la variabilité des prédictions. Ce modèle de substitution peut remplacer le modèle GBM initial. La modélisation des probabilités issues du modèle GBM à partir des variables importantes est visible sur la figure 39

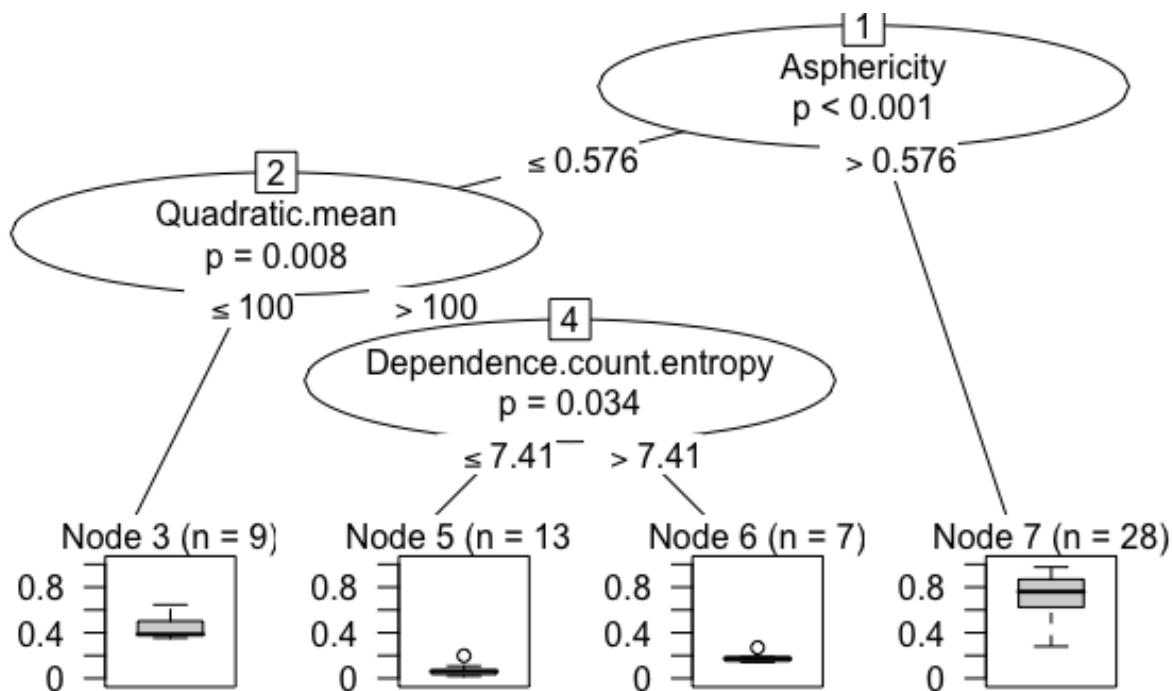


FIGURE 39 – Arbre de décision obtenu comme modèle de substitution global

Les résultats issues des modèles de substitution coïncident avec les analyses de PDP. Nous notons toutefois que l'on arrive à mieux substituer le modèle issu de l'utilisation de la valeur de l'ICC ( $R^2 = 0.88$ ) comparativement aux deux autres modèles ( $R^2 = 0.80$  et  $0.76$ ).

L'interprétabilité nous indique qu'il y a de la cohérence au sein des modèles estimés à partir des différentes sélections des caractéristiques radiomiques robustes. Cette méthodologie révèle donc une son importance au sein de l'analyse radiomique permettant la production de signature robuste et fiable.

## 6 Travaux supplémentaires

Durant mon stage, j'ai eu la chance de réaliser des travaux supplémentaires autant pour mon équipe que pour l'équipe d'Image Processing. Ci-dessous je vais exposer brièvement chacun des travaux menés au cours de mon stage.

- Recherche d'un marqueur simple à partir de caractéristiques radiomiques 2D pour prédire un score de la présence d'une tumeur dans un parenchyme segmenté afin de pondérer la sélection aléatoire de coupe lors de l'entraînement par apprentissage profond : cette expérience m'a permis d'approfondir mes connaissances sur l'optimisation des algorithmes de machine learning et par la même occasion me familiariser avec mon environnement de travail qu'est l'analyse radiomique.
- Étude descriptive et comparative des différentes techniques de sélection de variables pour un jeu de données d'analyse radiomique : ce travail a permis d'identifier la meilleure stratégie de sélection de variables dans une analyse radiomique.
- Création d'un tutoriel reproductible sous R. À partir d'un jeu de données, chacune des étapes de la sélection des variables robustes, en passant par le calcul de l'ICC suivant un seuil puis par la prédiction d'une variable d'intérêt, à l'interprétabilité du modèle optimal.
- Validation de la méthode de segmentation à travers des analyses notamment du Biais relatif entre les segmentations originales et les segmentations perturbées : ce travail nous a permis d'observer à quel point les caractéristiques radiomiques issues des segmentations perturbées sont éloignées par rapport à la segmentation originale.

## 7 Conclusion

### Principaux résultats

De nos jours l'analyse radiomique est un domaine en plein essor, d'autant plus qu'elle évolue avec l'avancée des nouvelles techniques d'intelligence artificielle. Les modèles de machine learning deviennent de plus en plus puissants et accomplissent des innovations qui ne sont plus négligeables notamment dans le domaine de l'aide à la décision pour des applications médicales. Pour fournir une utilisation de modèle fiable et robuste les caractéristiques radiomiques se doivent d'être fiables et robustes. Cependant ces caractéristiques ou features radiomiques sont soumises à des variabilités suite à leur mode d'acquisition (extraites de la segmentation d'une zone d'intérêt) [Owens et al., 2018]. Ces variabilités peuvent être dues à différents logiciels, à différents opérateurs et à différentes segmentations pour un opérateur donné [Lee et al., 2021].

Pendant mon stage j'ai pu quantifier la variabilité inter-opérateur (plusieurs cliniciens segmentent différemment une ROI) à travers la détection de variables robustes et son impact sur les prédictions estimées. Dans l'impossibilité de faire recours à plusieurs cliniciens pour effectuer une expérience sur les segmentations, une simulation approchant de façon utopique cette situation a été réalisée. Cette simulation a fait recours à la génération de segmentations perturbées ayant avec une segmentation originale un coefficient de similarité de Dice pouvant être fixé entre 0.87 et 0.90 : le coefficient finalement utilisé est le Dice à 0.90 car c'est la configuration qui avait les meilleures performances prédictives avec les algorithmes d'apprentissage supervisés par rapport au nombre de variables sélectionnées et aussi parce que c'est ce qui est le plus souvent utilisé dans la littérature [Granzier et al., 2020]. Ainsi, grâce à ces perturbations nous avons pu quantifier la variabilité intra-opérateur des features radiomiques à l'aide du coefficient de corrélation intra-classe ICC.

Dans une première partie j'ai donc effectué une analyse radiomique standard sur une cohorte de patients atteints de cancer du poumon au stade IV sous Pembrolizumab (Section 5.6.1). L'objectif était de prédire la progression à la première évaluation après initiation du traitement pour chaque patient. Plus précisément, l'objectif était de détecter les patients ayant progressé (pour ces patients une autre alternative de traitement est possible). Les données étant déséquilibrées (variables/individus), la sélection de variable des corrélations de Kendall a été réalisée. Nous avons donc comparé une multitude de modèles d'apprentissage supervisé avec une méthode d'optimisation qu'est le LOOCV avec la maximisation du F1score. Le modèle GBM était le plus performant avec un F1score à 0.75. Ensuite, j'ai réalisé une interprétabilité de ce modèle afin de mieux comprendre son fonctionnement sur les données, cela a permis de déceler trois variables importantes : *Asphericity*, *Dependance count entropy* et *Minor axis length*, soit deux features radiomiques morphologiques et une caractéristique de texture.

Un autre travail a consisté en la détection des caractéristiques radiomiques robustes et non-robustes (caractéristiques radiomiques qui varient suivant la méthode d'acquisition). J'ai utilisé le coefficient de corrélation inter-classe ICC(1,1) pour ce faire. Pour chaque caractéristique radiomique, j'ai calculé les valeurs des ICCs et une caractéristique est dite robuste si sa valeur ou la valeur de la borne inférieure de son intervalle de confiance est supérieure à 0.90 (seuil arbitraire fixé suivant les

performances obtenues des algorithmes et aussi d'après la littérature [Duron et al., 2019]). Ensuite, au travers de ces deux nouvelles bases de caractéristiques radiomiques robustes nous avons repris l'étude précédente (pour chacune des études la méthodologie de sélection des variables fût la sélection de variables des corrélations de Kendall) en vue de comparer les résultats d'une étude comprenant toutes les variables et les résultats d'une étude comprenant uniquement les variables robustes. Le but étant de pouvoir quantifier l'impact de la sélection des variables robustes sur les prédictions.

La comparaison des prédictions estimées des modèles d'apprentissage supervisé prenant en compte la sélection de variables robustes et celles ne les prenant pas en compte nous indique une différence soit moindre (la différence en valeur absolue entre le F1score étant de  $\pm 0.05$ ) soit supérieure (le modèle avec la sélection de variable robuste utilisant l'ICC performe avec un PRAuc supérieure de +0.1) ; en fonction de la métrique utilisée. Autrement dit avoir des modèles basés sur des variables robustes et avoir des modèles basés sur toutes les variables apportent à priori les mêmes voire de meilleures résultats (selon l'étude réalisée). Cette étude permet donc d'affecter à la sélection de variables robustes une place dans le processus de l'analyse radiomique, d'autant plus quelle permet une réduction plus importante de la dimensionnalité (92 ou 86 variables utilisées au lieu de 117 variables). Le modèle finalement choisi en fonction des performances est le modèle avec la sélection utilisant l'ICC qui a donné comme variables importantes les trois caractéristiques radiomiques suivantes :

- Asphéricité ou *Asphericity*
- La longueur du petit axe de la ROI ou *Minor axis length*
- La moyenne quadratique des intensités des voxels inclus dans le masque d'intensité du RIO ou *Quadratic mean*

#### **Limites rencontrées :**

Ce stage n'a pas toujours été une succession de réussites, j'ai également été confronté à quelques difficultés.. La majeure difficulté a été l'absence d'un jeu de données important. En effet, ce jeu de données ne permettait pas l'établissement d'un modèle d'apprentissage supervisé totalement non-biaisé, efficace et ne permettait pas également d'observer des différences significatives entre les performances prédictives obtenues sans/avec sélection des caractéristiques radiomiques robustes ! Ensuite, l'étude de la robustesse des caractéristiques radiomiques ici ne met en jeu qu'une seule catégorie de variabilité ; la variabilité inter-opérateur. Cette variabilité ne pouvant pas être matériellement caractérisée, elle a été simulée. Cette simulation est issue d'un processus itératif créé par l'équipe Image Processing. Même si elle nous permet de caractériser une certaine forme de robustesse de variables, elle est tout de même réalisée sous certaines hypothèses et contraintes : les simulations de perturbations de segmentation ne peuvent pas être considérées totalement comme l'approche de la segmentation faites par différents cliniciens. Les simulations sont totalement indépendantes contrairement à des cliniciens qui en voyant une RIO ont des aprioris sur leurs segmentations. Néanmoins même si cette technique de simulation de perturbation a déjà été utilisé dans la littérature Duron et al. [2019], nous avons effectué des validations visuelles de toutes les perturbations.

De plus, ces simulations ont un coût computationnel élevé. En outre, les modèles suivant lesquels l'ICC est calculé dépendent d'hypothèses gaussiennes qui ne sont pas toujours respectées par toutes les features : c'était le cas de la variable *Center of mass shift* qui était considérée comme non robuste alors qu'elle apportait une certaine amélioration au niveau des algorithmes prédictifs et est importante pour les cliniciens. En effet le "déplacement du centre de masse" ou *Center of mass shift*

est la différence entre le centre morphologique et le centre pondéré par l'intensité des voxels.

D'un autre côté nous avions cherché à caractériser les incertitudes au niveau des prédictions à travers la modélisation d'intervalles de confiance sur les valeurs des métriques obtenues et sur les probabilités des individus estimées. Cependant ne disposant pas d'assez de données et en utilisant la méthode du LOOCV, les méthodes traditionnelles de Bootstrap et de Bootstrap-Leave-one-out n'ont guère fonctionné.

### **Travaux Futurs :**

L'objectif à court terme est l'application de cette stratégie de sélection des caractéristiques robustes dans DeepLungIV : cohorte observationnelle de 4000 patients atteints de cancer de poumon en stade IV et traités par chimiothérapie et/ou immunothérapie. L'objectif principal de l'étude sera de prédire la probabilité de progression à la première évaluation (réalisée 2 mois après l'initiation du traitement). Des objectifs secondaires seront notamment la prédiction de la PFS (survie sans progression) et OS (survie globale). Toutes ces prédictions se baseront sur les données collectées en pré-traitement chez les patients (caractéristiques radiomiques issues d'images médicales, données cliniques, données biologiques, données génomiques, etc.)

Ensuite, il faudra à travers DeepLungIV ou d'une autre base améliorer et perfectionner la démarche pour la détection des variables robustes non seulement suivant la variabilité inter-opérateur mais aussi suivant la variabilité intra-opérateur, afin d'intégrer cette sélection dans tous les projets majeurs concernés à l'équipe Biostatistiques. Le but est de pouvoir réaliser des modèles robustes basés sur des caractéristiques invariantes suivant le mode d'acquisition : cela permettra entre-autre la rédaction d'articles. Cette stratégie pourra donc être appliquée à plusieurs types d'analyse ( différentes pathologies à savoir cancer du rein, du cerveau, du sein etc.).

A moyen terme il s'agira de pouvoir inclure l'incertitude due aux segmentations pour chaque probabilité à travers le calcul des intervalles de prédictions des prédictions individuelles dans l'ensemble des analyses.

A long terme, l'idée est de pouvoir implémenter cette stratégie dans le logiciel SOPHiA RADIO-MICS, afin que ces outils plus robustes soient destinés aux professionnels de la santé pour appuyer leur diagnostic médical et proposer des traitements adaptés à chaque patient.

## Références

- Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Number 4. Springer, 2 edition, 2009.
- Yoav Freund and Robert E. Schapire. Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning*, pages 148–156, 1996. ISSN 0706-652X, 1205-7533. doi : 10.1.1.133.1040. URL [http://www.public.asu.edu/\\$sim\\$jye02/CLASSES/Fall-2005/PAPERS/boosting-icml.pdf](http://www.public.asu.edu/$sim$jye02/CLASSES/Fall-2005/PAPERS/boosting-icml.pdf).
- George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. *Machine Learning Proceedings 1994*, pages 121–129, 1994. doi : 10.1016/b978-1-55860-335-6.50023-4.
- E. J. Limkin, R. Sun, L. Dercle, E. I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, E. Deutsch, and C. Ferté. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals of Oncology*, 28(6) :1191–1206, 2017. ISSN 15698041. doi : 10.1093/annonc/mdx034.
- Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, and et al. The image biomarker standardization initiative : Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2) :328–338, May 2020. ISSN 1527-1315. doi : 10.1148/radiol.2020191145. URL <http://dx.doi.org/10.1148/radiol.2020191145>.
- R. W.Y. Granzier, N. M.H. Verbakel, A. Ibrahim, J. E. van Timmeren, T. J.A. van Nijnatten, R. T.H. Leijenaar, M. B.I. Lobbes, M. L. Smidt, and H. C. Woodruff. MRI-based radiomics in breast cancer : feature robustness with respect to inter-observer segmentation variability. *Scientific Reports*, 10(1), 2020.
- Constance A. Owens, Christine B. Peterson, Chad Tang, Eugene J. Koay, Wen Yu, Dennis S. Mackin, Jing Li, Mohammad R. Salehpour, David T. Fuentes, Laurence E. Court, and Jinzhong Yang. Lung tumor segmentation methods : Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS ONE*, 13(10) :1–22, 2018.
- Alex Zwanenburg, Stefan Leger, Linda Agolli, Karoline Pilz, Esther G.C. Troost, Christian Richter, and Steffen Löck. Assessing robustness of radiomic features by image perturbation. *Scientific Reports*, 9 :1–10, 2019.
- Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Jean-claude Sadik, Isabelle Thomassin-naggara, Laure Fournier, Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Daniel Balvay, Saskia Vande Perre, and Afef Bouchouicha. Gray-level discretization impacts reproducible MRI radiomics texture features To cite this version : HAL Id : hal-02066691 Gray-level discretization impacts reproducible MRI radiomics texture features. 2019.

- E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. La-combe, and J. Verweij. New response evaluation criteria in solid tumours : Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2) :228–247, 2009. URL <http://dx.doi.org/10.1016/j.ejca.2008.10.026>.
- Lee R. Dice. Measures of the Amount of Ecologic Association Between Species Author ( s ) : Lee R . Dice Published by : Ecological Society of America. *Ecology*, 26(3) :297–302, 1945.
- Marta Avalos. Sélection de variables avec lasso dans la régression logistique conditionnelle. *41èmes Journées de Statistique, SFdS*, ..., pages 1–7, 2009.
- Andreas Mayr, H. Binder, O. Gefeller, and M. Schmid. The evolution of boosting algorithms : From machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6) :419–427, 2014. doi : 10.3414/ME13-01-0122.
- J R Quinlan. J. Ross Quinlan C4.5 Programs for Machine Learning . 5(3) :302, 1993.
- C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) : 379–423, 1948. doi : 10.1002/j.1538-7305.1948.tb01338.x.
- Friedman H. Jerome. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38 (4) :367–378, 2002.
- Minghua Chen, Qunying Liu, Shuheng Chen, Yicen Liu, Chang Hua Zhang, and Ruihua Liu. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. *IEEE Access*, 7 :13149–13158, 2019.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. pages 273–297, 1995.
- Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70) :2079–2107, 2010. URL <http://jmlr.org/papers/v11/cawley10a.html>.
- Jean-Marie John-Mathews. *Interprétabilité en Machine Learning*. Univiserté de Paris Saclay, 2019.
- Miller Tim. *Explanation in artificial intelligence : Insights from the social sciences*. arXiv Preprint arXiv :1706.07269, 2017.
- Kim Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. *Examples are not enough : learn to criticize. Criticism for interpretability*. Advances in Neural Information Processing Systems, 2016.
- McGraw and Wong. *Forming inferences about some intraclass correlation coefficients*. International series of monographs on physics. Psychol Methods, 1996.

Terry K. Koo and Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15 :155–163, 2016.

D. P. Martin. Handling Class Imbalance with R and Caret - An Introduction. 2016.

Bo Wu, Mingzhou Zhou, Xipeng Shen, Yaoqing Gao, Raul Silvera, and Graham Yiu. Simple profile rectifications go a long way statistically exploring and alleviating the effects of sampling errors for program optimizations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7920 LNCS(97) :654–678, 2013. doi : 10.1007/978-3-642-39038-8\_27.

Gillian M. Raab Beata Nowok and Chris Dibben. Synthpop : Bespoke creation of synthetic data in R. *Journal of Statistical Software*, pages 1–26, 2016.

Maria Luisa Belli, Martina Mori, Sara Broggi, Giovanni Mauro Cattaneo, Valentino Bettinardi, Italo Dell’Oca, Federico Fallanca, Paolo Passoni, Emilia Giovanna Vanoli, Riccardo Calandrino, Nadia Di Muzio, Maria Picchio, and Claudio Fiorino. Quantifying the robustness of [18F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Physica Medica*, 49(February) :105–111, 2018.

Lijun Lu, Wenbing Lv, Jun Jiang, Jianhua Ma, Qianjin Feng, Arman Rahmim, and Wufan Chen. Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma : Impact of Segmentation and Discretization. *Molecular Imaging and Biology*, 18, 2016.

Ralph T.H. Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter J.C. Van Elmpt, Esther G.C. Troost, Ronald Boellaard, Hugo J.W.L. Aerts, Robert J. Gillies, and Philippe Lambin. The effect of SUV discretization in quantitative FDG-PET Radiomics : The need for standardized methodology in tumor texture analysis. *Scientific Reports*, 5(August) :1–10, 2015.

Joonsang Lee, Angela Steinmann, Yao Ding, Hannah Lee, Constance Owens, Jihong Wang, Jinzhong Yang, David Followill, Rachel Ger, Dennis MacKin, and Laurence E. Court. Radiomics feature robustness as measured using an MRI phantom. *Scientific Reports*, 11 :1–14, 2021.

## 8 Annexes

Les graphiques de sélection des variables robustes suivant la valeur de l'ICC :

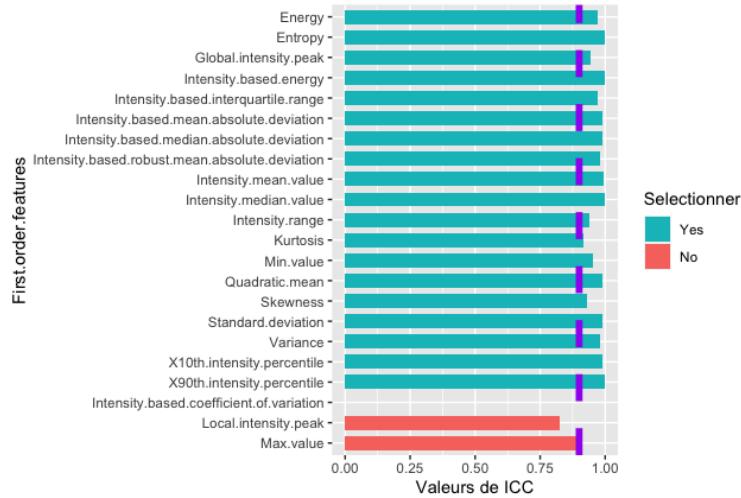


FIGURE 40 – First orders features selection

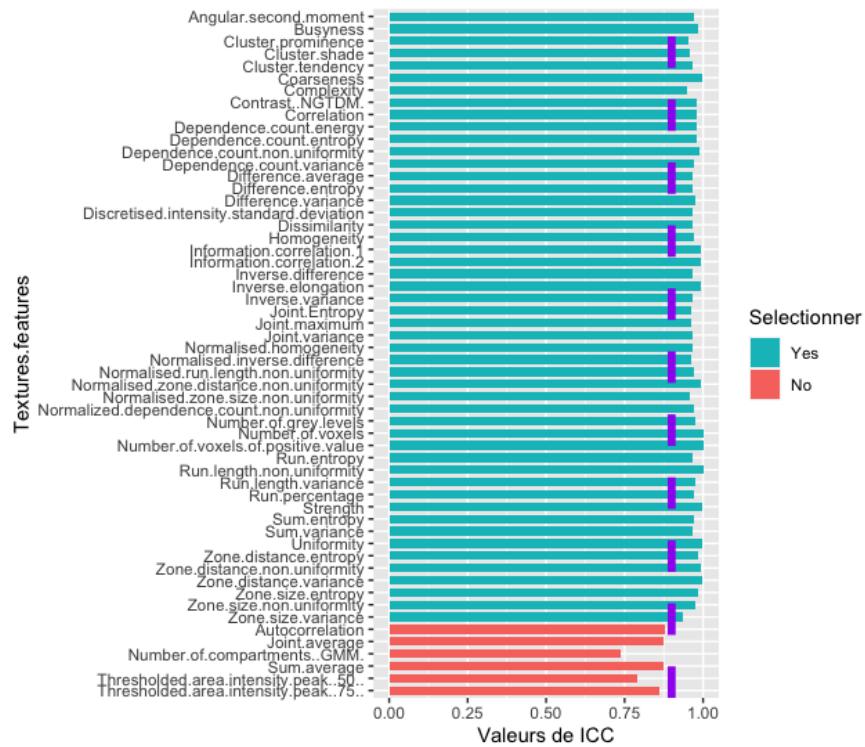


FIGURE 41 – Textures features selection

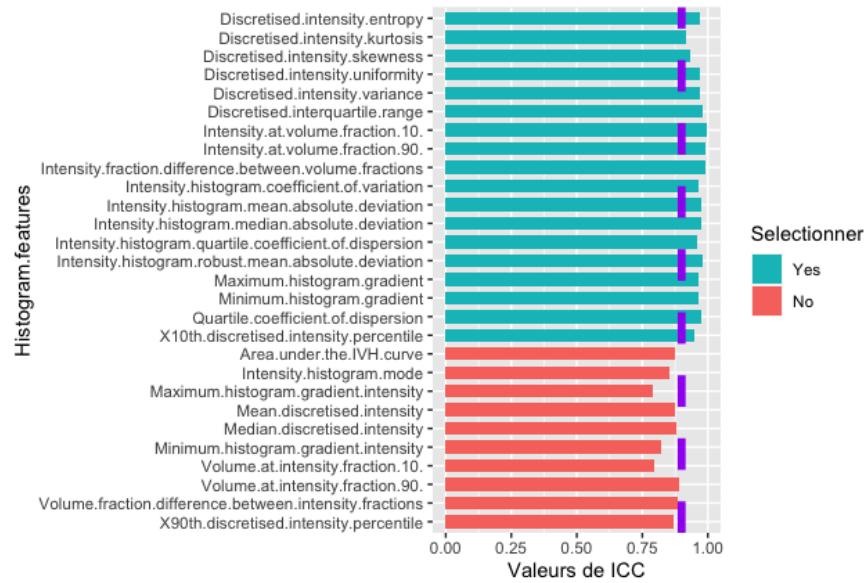


FIGURE 42 – Histogram features selection

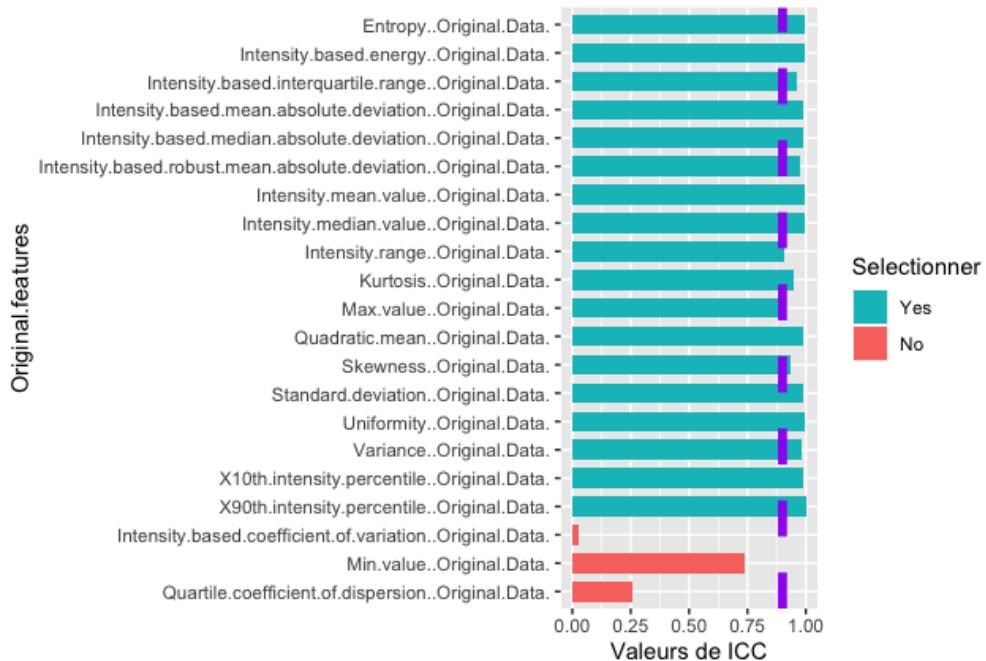


FIGURE 43 – Original features selection

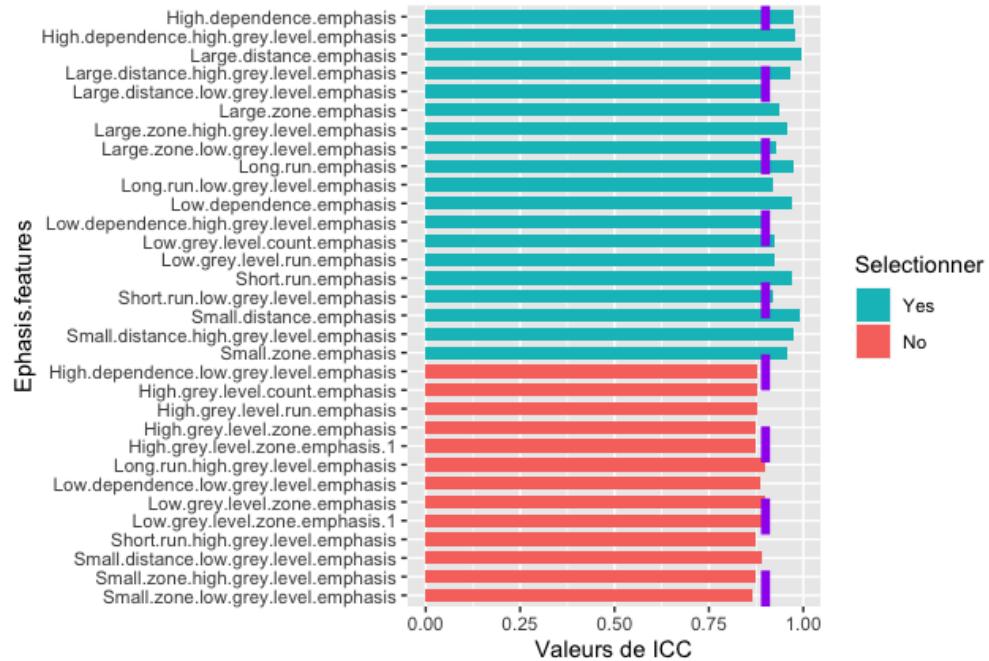


FIGURE 44 – Emphasis features selection

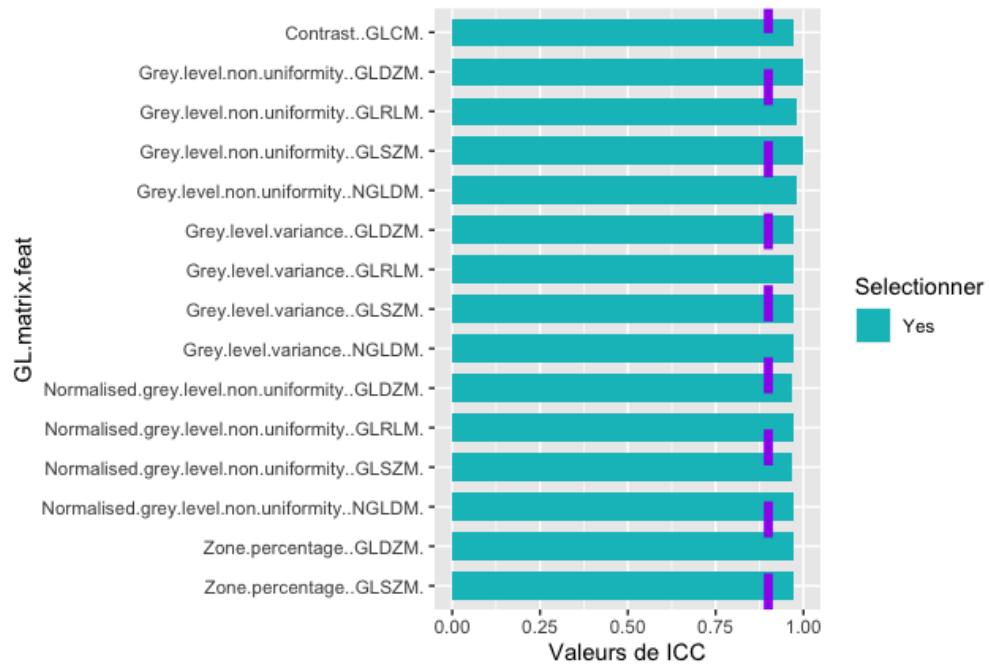


FIGURE 45 – GL matrix features selection

Les graphiques de sélection des variables robustes suivant la valeur de la borne inférieure de l'intervalle de confiance de l'ICC :

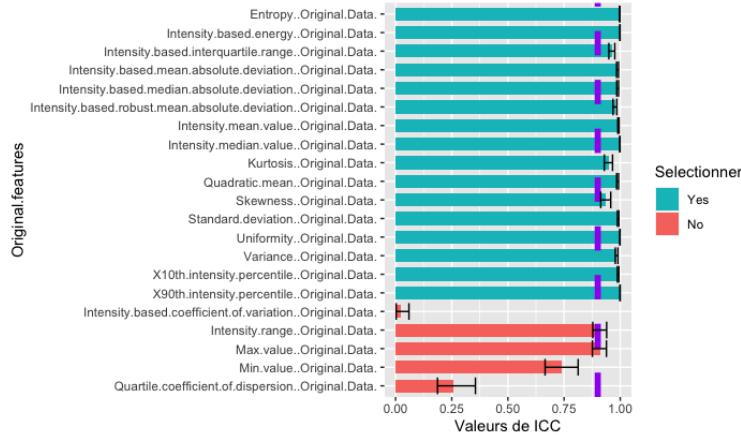


FIGURE 46 – Original features selection

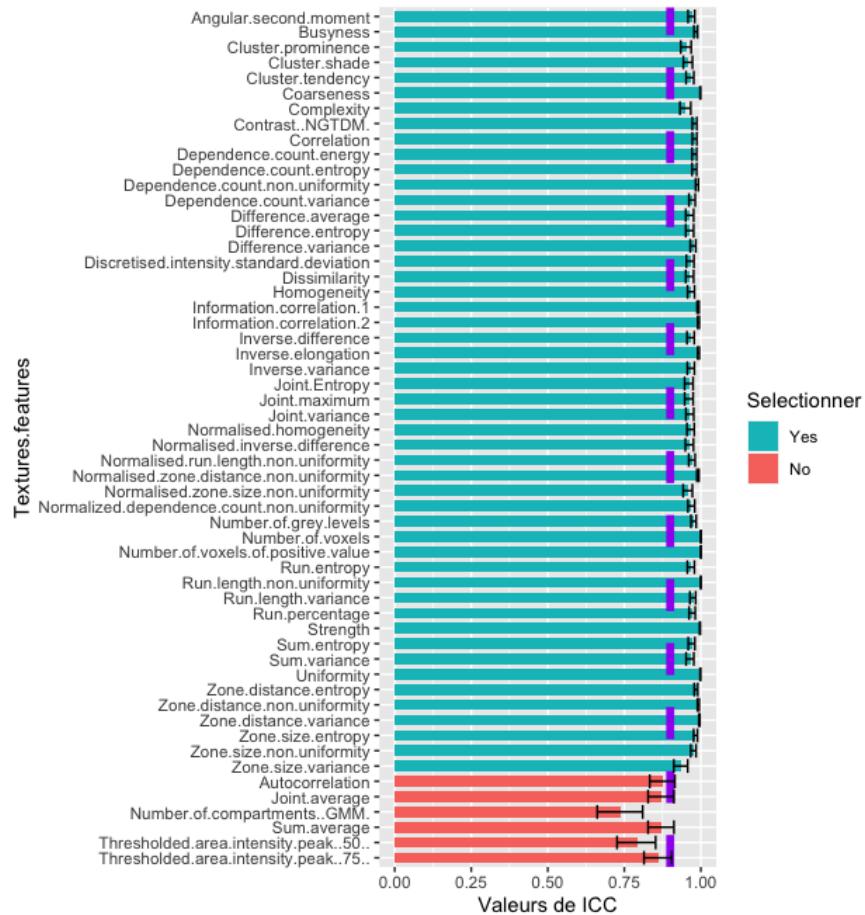


FIGURE 47 – Textures features selection

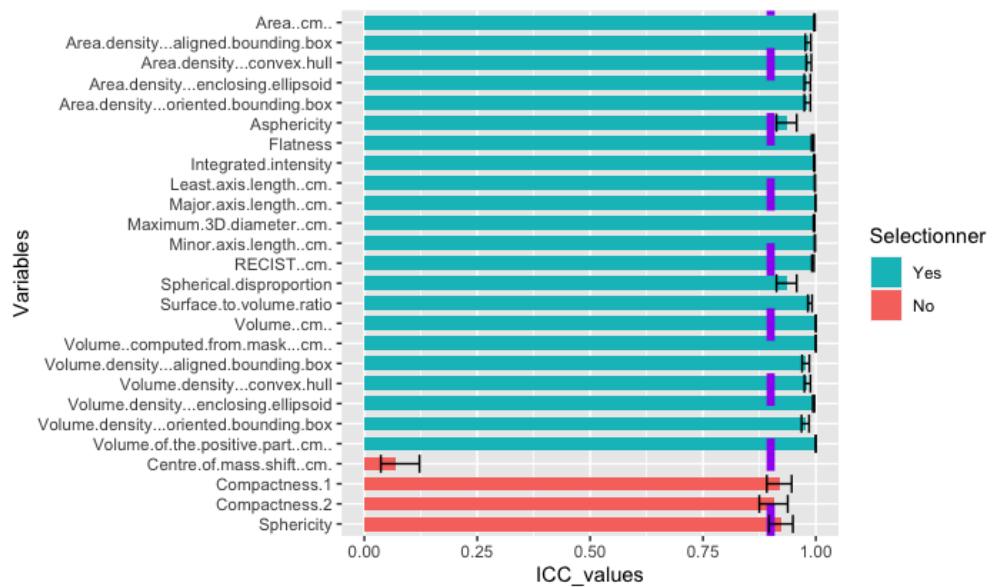


FIGURE 48 – Shapes features selection

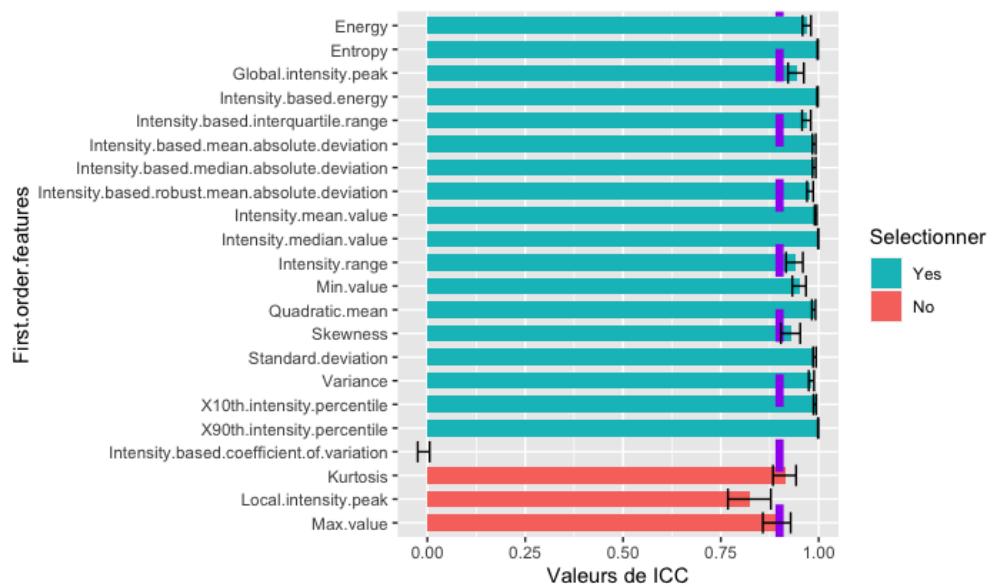


FIGURE 49 – First orders features selection

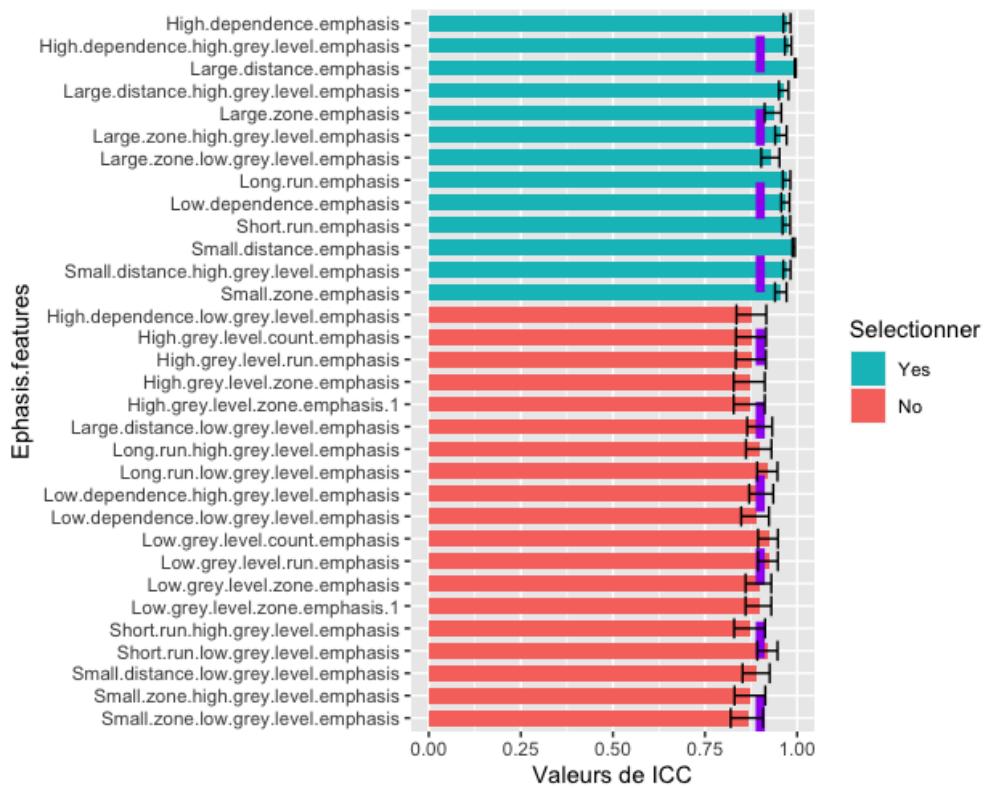


FIGURE 50 – Emphasis features selection

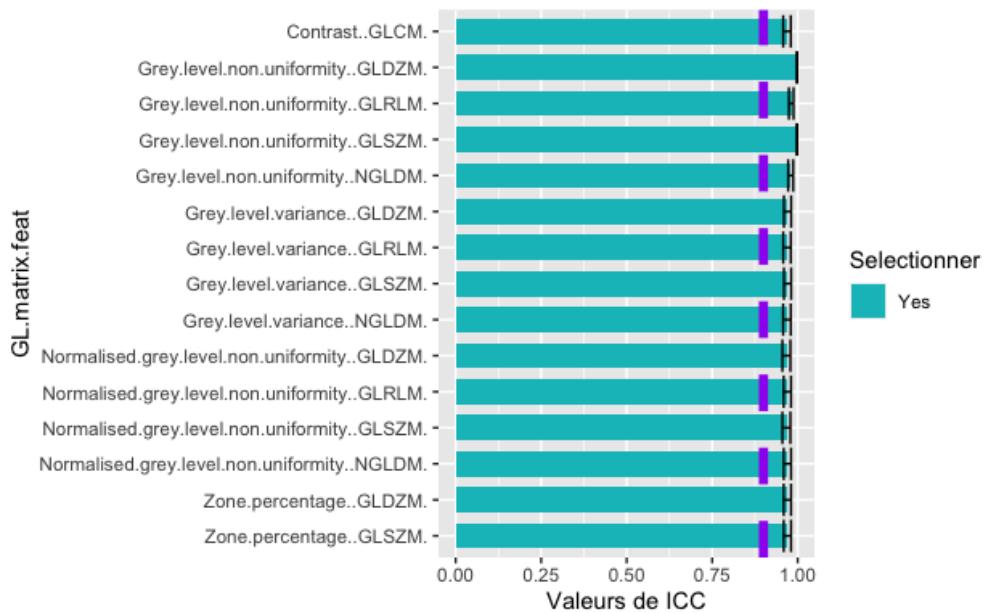


FIGURE 51 – GL matrix features selection

Table des features :

	Caractéristiques
1	Area..cm..
2	Area.density...aligned.bounding.box
3	Area.density...convex.hull
4	Area.density...enclosing.ellipsoid
5	Area.density...oriented.bounding.box
6	Asphericity
7	Centre.of.mass.shift..cm.
8	Compactness.1
9	Compactness.2
10	Flatness
11	Integrated.intensity
12	Least.axis.length..cm.
13	Major.axis.length..cm.
14	Maximum.3D.diameter..cm.
15	Minor.axis.length..cm.
16	Spherical.disproportion
17	Sphericity
18	Surface.to.volume.ratio
19	Volume..cm..
20	Volume..computed.from.mask...cm..
21	Volume.density...aligned.bounding.box
22	Volume.density...convex.hull
23	Volume.density...enclosing.ellipsoid
24	Volume.density...oriented.bounding.box
25	Volume.of.the.positive.part..cm..

TABLE 13 – Morphological Features

	Caractéristiques
1	X10th.intensity.percentile
2	X90th.intensity.percentile
3	Energy
4	Entropy
5	Global.intensity.peak
6	Intensity.mean.value
7	Intensity.median.value
8	Intensity.range
9	Intensity.based.coefficient.of.variation
10	Intensity.based.energy
11	Intensity.based.interquartile.range
12	Intensity.based.mean.absolute.deviation
13	Intensity.based.median.absolute.deviation
14	Intensity.based.robust.mean.absolute.deviation
15	Kurtosis
16	Local.intensity.peak
17	Max.value
18	Min.value
19	Quadratic.mean
20	Skewness
21	Standard.deviation
22	Variance

TABLE 14 – First order features

	Caractéristiques
1	X10th.discretised.intensity.percentile
2	X90th.discretised.intensity.percentile
3	Area.under.the.IVH.curve
4	Discretised.intensity.entropy
5	Discretised.intensity.kurtosis
6	Discretised.intensity.skewness
7	Discretised.intensity.uniformity
8	Discretised.intensity.variance
9	Discretised.interquartile.range
10	Intensity.at.volume.fraction.10.
11	Intensity.at.volume.fraction.90.
12	Intensity.fraction.difference.between.volume.fractions
13	Intensity.histogram.coefficient.of.variation
14	Intensity.histogram.mean.absolute.deviation
15	Intensity.histogram.median.absolute.deviation
16	Intensity.histogram.mode
17	Intensity.histogram.quartile.coefficient.of.dispersion
18	Intensity.histogram.robust.mean.absolute.deviation
19	Maximum.histogram.gradient
20	Maximum.histogram.gradient.intensity
21	Mean.discretised.intensity
22	Median.discretised.intensity
23	Minimum.histogram.gradient
24	Minimum.histogram.gradient.intensity
25	Quartile.coefficient.of.dispersion
26	Volume.at.intensity.fraction.10.
27	Volume.at.intensity.fraction.90.
28	Volume.fraction.difference.between.intensity.fractions

TABLE 15 – Histogram features

	Caractéristiques
1	X10th.intensity.percentile..Original.Data.
2	X90th.intensity.percentile..Original.Data.
3	Entropy..Original.Data.
4	Intensity.mean.value..Original.Data.
5	Intensity.median.value..Original.Data.
6	Intensity.range..Original.Data.
7	Intensity.based.coefficient.of.variation..Original.Data.
8	Intensity.based.energy..Original.Data.
9	Intensity.based.interquartile.range..Original.Data.
10	Intensity.based.mean.absolute.deviation..Original.Data.
11	Intensity.based.median.absolute.deviation..Original.Data.
12	Intensity.based.robust.mean.absolute.deviation..Original.Data.
13	Kurtosis..Original.Data.
14	Max.value..Original.Data.
15	Min.value..Original.Data.
16	Quadratic.mean..Original.Data.
17	Quartile.coefficient.of.dispersion..Original.Data.
18	Skewness..Original.Data.
19	Standard.deviation..Original.Data.
20	Uniformity..Original.Data.
21	Variance..Original.Data.

TABLE 16 – Original data Features

	Caractéristiques
1	High.dependence.emphasis
2	High.dependence.high.grey.level.emphasis
3	High.dependence.low.grey.level.emphasis
4	High.grey.level.count.emphasis
5	High.grey.level.run.emphasis
6	High.grey.level.zone.emphasis
7	High.grey.level.zone.emphasis.1
8	Large.distance.emphasis
9	Large.distance.high.grey.level.emphasis
10	Large.distance.low.grey.level.emphasis
11	Large.zone.emphasis
12	Large.zone.high.grey.level.emphasis
13	Large.zone.low.grey.level.emphasis
14	Long.run.emphasis
15	Long.run.high.grey.level.emphasis
16	Long.run.low.grey.level.emphasis
17	Low.dependence.emphasis
18	Low.dependence.high.grey.level.emphasis
19	Low.dependence.low.grey.level.emphasis
20	Low.grey.level.count.emphasis
21	Low.grey.level.run.emphasis
22	Low.grey.level.zone.emphasis
23	Low.grey.level.zone.emphasis.1
24	Short.run.emphasis
25	Short.run.high.grey.level.emphasis
26	Short.run.low.grey.level.emphasis
27	Small.distance.emphasis
28	Small.distance.high.grey.level.emphasis
29	Small.distance.low.grey.level.emphasis
30	Small.zone.emphasis
31	Small.zone.high.grey.level.emphasis
32	Small.zone.low.grey.level.emphasis

TABLE 17 – Emphasis features

	Caractéristiques
1	Contrast..GLCM.
2	Grey.level.non.uniformity..GLDZM.
3	Grey.level.non.uniformity..GLRLM.
4	Grey.level.non.uniformity..GLSZM.
5	Grey.level.non.uniformity..NGLDM.
6	Grey.level.variance..GLDZM.
7	Grey.level.variance..GLRLM.
8	Grey.level.variance..GLSZM.
9	Grey.level.variance..NGLDM.
10	Normalised.grey.level.non.uniformity..GLDZM.
11	Normalised.grey.level.non.uniformity..GLRLM.
12	Normalised.grey.level.non.uniformity..GLSZM.
13	Normalised.grey.level.non.uniformity..NGLDM.
14	Zone.percentage..GLDZM.
15	Zone.percentage..GLSZM.

TABLE 18 – GLMatrix features

	Caractéristiques
1	Angular.second.moment
2	Autocorrelation
3	Busyness
4	Cluster.prominence
5	Cluster.shade
6	Cluster.tendency
7	Coarseness
8	Complexity
9	Contrast..NGTDM.
10	Correlation
11	Dependence.count.energy
12	Dependence.count.entropy
13	Dependence.count.non.uniformity
14	Dependence.count.variance
15	Difference.average
16	Difference.entropy
17	Difference.variance
18	Discretised.intensity.standard.deviation
19	Dissimilarity
20	Homogeneity
21	Information.correlation.1
22	Information.correlation.2
23	Inverse.difference
24	Inverse.elongation
25	Inverse.variance

TABLE 19 – Textures features 1

	Caractéristiques
26	Joint.Entropy
27	Joint.average
28	Joint.maximum
29	Joint.variance
30	Normalised.homogeneity
31	Normalised.inverse.difference
32	Normalised.run.length.non.uniformity
33	Normalised.zone.distance.non.uniformity
34	Normalised.zone.size.non.uniformity
35	Normalized.dependence.count.non.uniformity
36	Number.of.compartments..GMM.
37	Number.of.grey.levels
38	Number.of voxels
39	Number.of.voxels.of.positive.value
40	RECIST..cm.
41	Run.entropy
42	Run.length.non.uniformity
43	Run.length.variance
44	Run.percentage
45	Strength
46	Sum.average
47	Sum.entropy
48	Sum.variance
49	Thresholded.area.intensity.peak..50..
50	Thresholded.area.intensity.peak..75..
51	Uniformity
52	Zone.distance.entropy
53	Zone.distance.non.uniformity
54	Zone.distance.variance
55	Zone.size.entropy
56	Zone.size.non.uniformity
57	Zone.size.variance

TABLE 20 – Textures features 2