Méthode de sélection de caractéristiques pronostiques et prédictives basée sur les forêts aléatoires pour le suivi thérapeutique des lésions tumorales par imagerie fonctionnelle TEP

Paul Desbordes

Doctorats : la nomenclature générique (47 modalités)

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité Informatique

Préparée au sein de l'Université de Rouen Normandie

Méthodes de sélection de caractéristiques radiomiques à valeur pronostique ou prédictive des lésions tumorales en imagerie TEP au ^{18}FDG basées sur les forêts aléatoires.

Présentée et soutenue par
Paul DESBORDES

Thèse soutenue publiquement le 02/05/2017 devant le jury composé de		
M. Philippe GIRAUD	PUPH - Hôpital George Pompidou, Paris	Président du Jury
Mme. Irène BUVAT	Directeur de recherche - Université Paris Sud, Orsay	Rapporteur
M. John LEE	Pr - Université catholique de Louvain	Rapporteur
M. Pierre VERA	PUPH - Université de Rouen Normandie	Examinateur
M. Sébastien VAUCLIN	Ingénieur de recherche - Dosisoft, Cachan	Examinateur
Mme. Su RUAN	Pr - Université de Rouen Normandie	Co-directrice
Mme. Isabelle GARDIN	Centre Henri Becquerel, Rouen	Co-directrice

Thèse dirigée par Isabelle GARDIN, Centre Henri Becquerel
et par Su RUAN, Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS)

THÈSE

En vue de l'obtention du grade de

DOCTEUR DE NORMANDIE UNIVERSITÉ

délivré par

L'UNIVERSITÉ DE ROUEN NORMANDIE

LABORATOIRE D'INFORMATIQUE DU TRAITEMENT DE L'INFORMATION ET DES SYSTÈMES

Mathématiques, **I**nformation et **I**ngénierie des **S**ystèmes

Discipline : INFORMATIQUE

Présentée et soutenue publiquement par

Paul DESBORDES

Méthodes de sélection de caractéristiques radiomiques à valeur pronostique ou prédictive des lésions tumorales en imagerie TEP au ^{18}FDG basées sur les forêts aléatoires.

Directrice de thèse : **Mme. Isabelle GARDIN**
Co-directrice de thèse : **Pr. Su RUAN**

Jury

M. Philippe GIRAUD,	PUPH, Hôpital George Pompidou, Paris	Président du Jury
Mme. Irène BUVAT,	Directeur de recherche, Université Paris Sud, Orsay	Rapporteur
M. John LEE,	Pr, Université catholique de Louvain	Rapporteur
M. Pierre VERA,	PUPH, Centre Henri Becquerel, Rouen	Examinateur
M. Sébastien VAUCLIN,	PhD, Société Dosisoft, Cachan	Examinateur
Mme. Isabelle GARDIN,	Centre Henri Becquerel, Rouen	Co-directrice
Mme. Su RUAN,	Pr, Université de Rouen	Co-directrice

Résumé

La radiomique propose de combiner des caractéristiques images avec celles issues de la clinique, de la génomique, de la protéomique, etc ... afin de mettre en place une médecine personnalisée dans la prise en charge du cancer. L'objectif est d'anticiper, à partir d'un examen initial, les chances de survie du patient ou la probabilité de la maladie de répondre à un traitement. En médecine, des méthodes statistiques classiques sont généralement utilisées comme l'analyse de Mann-Whitney pour les études prédictives et l'analyse des courbes de survie de Kaplan-Meier pour les études pronostiques. Cependant, l'augmentation du nombre de caractéristiques étudiées pose des problèmes pour l'utilisation de ces statistiques. C'est pour cela que nous nous sommes orientés vers l'utilisation des algorithmes d'apprentissage automatique et des méthodes de sélection de caractéristiques. Ces méthodes sont résistantes aux grandes dimensions, ainsi qu'aux relations non-linéaires entre caractéristiques. Nous avons proposé 2 méthodes de sélection des caractéristiques basées sur la méthode d'apprentissage automatique des forêts aléatoires. Nos méthodes ont permis la sélection de sous-ensembles de caractéristiques prédictives et pronostiques sur 2 bases de données (cancer de l'œsophage et du poumon). Nos algorithmes ont montré les meilleures performances de classification comparées aux méthodes statistiques classiques et aux autres méthodes de sélection des caractéristiques étudiées.

Radiomics proposes to combine image features with those extracted from other modalities (clinical, genomic, proteomic) to set up a personalized medicine in the management of cancer. From an initial exam, the objective is to anticipate the survival rate of the patient or the treatment response probability. In medicine, classical statistical methods are generally used, such as the Mann-Whitney analysis for predictive studies and analysis of Kaplan-Meier survival curves for prognostic studies. Thus, the increasing number of studied features limits the use of these statistics. We have focused our works on machine learning algorithms and features selection methods. These methods are resistant to large dimensions as well as non-linear relations between features. We proposed two features selection strategy based on random forests. Our methods allowed the selection of subsets of predictive and prognostic features on 2 databases (oesophagus and lung cancers). Our algorithms showed the best classification performances compared to classical statistical methods and other features selection strategies studied.

Remerciements

Je tiens à remercier en premier lieu les membres du jury de cette thèse pour l'intérêt porté à mes travaux. Je remercie tout particulièrement Madame Irène BUVAT et Monsieur John LEE, d'avoir accepté de rapporter ma thèse. Je remercie également Monsieur Philippe GIRAUD de m'avoir fait l'honneur de présider le jury. Je suis particulièrement reconnaissant pour l'ensemble des remarques formulées, critiques et conseils très enrichissants qui m'ont été prodigués.

Je souhaite exprimer ma reconnaissance à Madame Isabelle GARDIN, ainsi qu'à Madame Su RUAN qui ont encadré cette thèse. Je les remercie pour leur conseils très instructifs dans le domaine scientifique, et pour leur disponibilité notable, propice au bon déroulement de ma thèse. Je tiens également à les remercier pour leur aide précieuse m'ayant permis de parfaire la réalisation de ce mémoire.

Je remercie également l'entreprise Dosisoft pour avoir accompagnée ma thèse en commençant par Monsieur Hanna KAFROUNI pour m'avoir accepté au sein de son entreprise tout en me laissant une grande liberté dans mes recherches. Bien sûr, je remercie également Pascal PINEAU et Sébastien VAUCLIN pour leur aide et leur expertise concernant les images médicales.

Ma gratitude va ensuite vers l'ensemble des membres du Centre Henri Becquerel de Rouen avec qui j'ai eu la chances de travailler durant cette thèse. Je commencerai par remercier Monsieur Pierre VERA pour son expertise dans le domaine médical et son intérêt pour toutes les formes de recherches qui ont été pour moi une source de motivation. Ma gratitude va également à Messieurs Romain MODZELEWSKI et Pierrick GOUEL qui m'ont accompagné dans l'ensemble des étapes de ma thèse et qui ont su m'écouter lorsque des obstacles se présentaient. Ma reconnaissance va également à Monsieur Simon BERNARD, maitre de conférence de l'Université de Rouen, pour son expertise et ses conseils perti-

nents.

Je remercie tous les membres de l'équipe QuantIF, qui ont rendu ces années de thèse agréables et sympathiques : Maxime, Elyse, Amine, Hongmei et Chunfeng. J'espère avoir l'occasion de vous retrouver dans le futur. Je salut également mes anciens collègues et amis de la promotion IBIOM Hounsfield, Alban, Kévin et Mathieu.

Pour leur soutien infaillible et leurs encouragements réguliers, je remercie infiniment ma famille pour avoir fait de moi la personne que je suis aujourd'hui et m'avoir toujours tout donné, en particulier, mes grands parents, mon père Philippe, ma mère Christèle, mon frère Quentin, ma sœur Zoé et ma pacsou Sonia. Enfin, je remercie mes amis de toujours, qui sont devenus une véritable 2ème famille : Alexandre, Bertrand, Bicou, Camille, Clémence, Ewa, Julien et Mickaël.
Merci de m'avoir toujours soutenu, même aux moments les plus difficiles.

Pour Angie

Table des matières

Introduction générale

L'imagerie médicale, telle que la *TomoDensitoMétrie* (TDM), la *Tomographie par Émission de Positons* (TEP) au *2-[18]-Fluoro-2-désoxy-D-glucose* (FDG) et l'*Imagerie par Résonance Magnétique* (IRM), est une technique non invasive utilisée en routine clinique dans la prise en charge des patients atteints d'un cancer. Elle tend à jouer un rôle de plus en plus important dans la personnalisation du traitement s'appuyant sur les caractéristiques individuelles du patient et de l'image de la tumeur.

La médecine personnalisée s'est largement développée avec la génomique et la protéomique, qui sont des techniques invasives. Cependant, il existe une hétérogénéité spatiale et temporelle des caractéristiques phénotypiques tumorales [Marusyk et al. 2012] (capacité métastatique, survie à la thérapie, ...), rendant difficile les biopsies répétées, aussi bien dans l'espace que dans le temps, là où l'imagerie permet d'emblée de donner une représentation spatiale de la lésion et présente l'avantage d'être facile à répéter.

Deux idées sont sous-jacentes au concept de radiomique [Yip and Aerts 2016]. La première est que les caractéristiques tumorales cliniques, à l'échelle tissulaire, cellulaire et/ou génomique auraient un retentissement phénotypique en imagerie médicale. Cela revient à considérer que des caractéristiques de l'image sont fortement corrélées à des caractéristiques cliniques et biologiques. Le deuxième rationnel est que l'information portée par l'image serait complémentaire de celles provenant d'autres sources d'informations médicales permettant ainsi d'enrichir le nombre des caractéristiques de la tumeur.

L'interprétation médicale des images est généralement basée sur une simplement appréciation visuelle du contraste. Ce type d'interprétation est qualitatif et subjectif, même s'il a montré toute son efficacité dans la prise en charge des patients en cancérologie. Avec la radiomique, il devient nécessaire de quantifier ces interprétations pour des rai-

sons d'objectivité et de reproductibilité. Cependant, cela engendre un nombre et une complexité des caractéristiques d'images toujours croissants.

Comme la radiomique cherche à étudier un nombre important de caractéristiques, il est nécessaire d'avoir à disposition des outils statistiques suffisamment puissants pour une prédiction précise et robuste. Les algorithmes d'apprentissage automatique (ou "machine learning") sont particulièrement adaptés à cette problématique. Ces méthodes sont capables d'apprendre des modèles à partir d'observations, ce qui permet d'automatiser et d'améliorer le processus de prédiction [Parmar et al. 2015]. Les méthodes d'apprentissage automatique possèdent la capacité de gérer de nombreuses caractéristiques et de capter leurs relations non-linéaires. Elles conduisent alors à une meilleure puissance discriminante lors de l'analyse de plusieurs dizaines de caractéristiques en comparaison aux statistiques classiques [El Naqa et al. 2009].

Avec l'augmentation du nombre de caractéristiques, la radiomique place sa problématique dans un espace à grande dimension ce qui peut mener les méthodes statistiques classiques à des résultats faussés et biaisés. Afin de limiter la dimension, il peut être intéressant d'utiliser des stratégies de sélection de caractéristiques afin de réduire le nombre de caractéristiques étudiées. La sélection de caractéristiques est une méthode de réduction de la dimension du problème posé.

L'objectif de cette thèse a été de mettre en place plusieurs stratégies de sélection des caractéristiques radiomiques ayant une valeur pronostique ou prédictive de la réponse au traitement en cancérologie. Ces algorithmes ont été construits à partir de méthode d'apprentissage automatique, les forêts aléatoires. Ces méthodes ont ensuite été évaluées sur plusieurs bases de données que nous avons construites au préalable et dont nous avons extraits les caractéristiques cliniques des dossiers médicaux et des examens d'imagerie.

Le manuscrit de cette thèse se compose de 5 chapitres :

Le premier chapitre introduit le principe de l'imagerie TEP au FDG, ainsi que son intérêt médical en cancérologie. Nous terminerons en définissant les caractéristiques radiomiques proposées dans la littérature, en commençant par les caractéristiques du 1er ordre, les caractéristiques de forme et de texture.

Le chapitre 2 porte sur une revue de la littérature présentant le concept de radiomique et étudiant son apport en oncologie. Nous décrirons les différents paramètres pouvant modifier la valeur de ces caractéristiques, ainsi que leur robustesse pour établir la réponse à un traitement ou leur valeur pronostique.

Le chapitre 3 porte sur les principes des méthodes d'apprentissage automatique. Plusieurs méthodes d'apprentissage supervisée et non-supervisée seront présentées en détails. Puis, plusieurs méthodes de sélection des caractéristiques seront décrites. Enfin, les données de la littérature sur l'utilisation de ces méthodes en imagerie médicale seront présentées.

Le chapitre 4 porte sur la description des 2 méthodes de sélection de caractéristiques que nous avons développé en commençant par la description de l'algorithme des forêts aléatoires qui se place au cœur de nos méthodes. Puis, nous développerons notre première étape de sélection filtrante basée sur l'analyse des corrélations. Enfin, nous aborderons les différentes approches envisagées de sélection enveloppantes pour la sélection des sous-ensembles de caractéristiques.

Dans le chapitre 5, les performances de ces 2 méthodes seront évaluées sur 2 bases de données (cancer de l'œsophage et du poumon). L'influence des paramètres de ces méthodes sont évaluées sur la base de données du cancer de l'œsophage afin de les optimiser. Les méthodes sont comparées aux statistiques classiques généralement utilisées en imagerie médicales, ainsi qu'à d'autres méthodes de sélection basées sur des algorithmes d'apprentissage.

Chapitre 1

L'imagerie TEP au FDG : principe et caractéristiques

Sommaire

1.1 Le cancer

1.1.1 Les cellules cancéreuses

Les cellules cancéreuses ont pour origine une mutation du code génétique. Cette altération cause la dégénérescence du cycle cellulaire, régulant la durée de vie des cellules, ainsi que leur division. Les cellules cancéreuses présentent alors une multiplication exponentielle et anarchique créant des amas cellulaires. Ces regroupements sont appelés des tumeurs lorsqu'ils atteignent environ 100 000 cellules.

Cette croissance cellulaire est associée à une angiogenèse créant de nouveaux vaisseaux sanguins fournissant l'énergie nécessaire à toutes ces cellules hyperactives. En effet, comme la division cellulaire est un processus nécessitant une importante quantité d'énergie, ces cellules présentent une consommation de glucose supérieure à celle des cellules saines.

Les 2 grandes catégories de cancer sont les cancers solides et les cancers sanguins. Les tumeurs solides, comme les carcinomes ou les sarcomes, sont repérables par un amas localisé de cellules. Ils se distinguent des cancers des cellules sanguines, comme les leucémies ou les lymphomes, dont les cellules cancéreuses circulant dans le sang ou la lymphe sont dispersées dans l'organisme.

En 2015, le nombre de nouveaux cas de cancer en France métropolitaine a été estimé à 385 000 (211 000 hommes et 174 000 femmes) et le nombre de décès par cancer, à 149 500 (84 100 hommes et 65 400 femmes). Les cancers du poumon, de la prostate, du sein et du côlon-rectum sont les cancers les plus fréquents [1]. De ce fait, la prise en charge du cancer est un enjeu de santé publique.

1.1.2 Prise en charge de la maladie

Lors de la détection de symptômes ou d'un test de dépistage positif, les patients bénéficient de différents examens afin de confirmer ou d'infirmer le diagnostic de cancer. Ces examens portent sur différents types d'informations :

1. [Les cancers en France /Édition 2014, Institut National du Cancer], Janvier 2015, www.e-cancer.fr

— données cliniques (palpations, stade de la maladie, perte de poids, ...),

— données biologiques (résultats de la biopsie, bilan sanguin, ...),

— données génomiques et protéomiques.

Ces examens sont généralement accompagnés d'examens d'imagerie tel que l'échographie, la TDM, la TEP ou l'IRM. L'ensemble de ces données permet de préciser la nature de la tumeur et son stade d'évolution. Elles sont essentielles à la mise en place d'un traitement adapté.

Une fois le diagnostic posé, une réunion de concertation pluridisciplinaire est organisée afin de choisir et de planifier le meilleur traitement possible. Il en existe 3 grands types : la chimiothérapie, la radiothérapie et la chirurgie. Ils utilisent divers processus et agissent à différents niveaux. Ces traitements peuvent être associés, telle que la *RadioChimioThérapie* (RCT) qui combine une radiothérapie et une chimiothérapie. Le but est l'amplification de l'efficacité du traitement, parfois au détriment de l'augmentation des effets secondaires.

Pour aider au choix du traitement optimal et le personnaliser, la recherche de caractéristiques permettant l'anticipation de l'évolution de la maladie est l'un des thèmes majeurs en recherche contre le cancer. On distingue les caractéristiques prédictives renseignant sur l'efficacité du traitement, des caractéristiques pronostiques renseignant sur la survie du patient (*"Overall Survival" ou survie globale* (OS) et *"Progression Free Disease" ou survie sans récidive* (PFS)). Ces caractéristiques peuvent venir des différentes modalités citées précédemment : données cliniques, biologiques, génétiques ou protéomiques, ainsi que des caractéristiques extraites des images médicales.

Une fois le traitement débuté, le patient est suivi afin d'évaluer l'efficacité du traitement. De nouveau, l'imagerie médicale joue un rôle fondamental pour le suivi de l'évolution de la maladie en cours du traitement et à distance, afin de contrôler l'apparition d'éventuelles récidives.

1.1.3 Les critères WHO, RECIST et PERCIST

En 1979, afin de normaliser l'évaluation de la réponse thérapeutique, l'*Organisation Mondiale de la Santé* (OMS) (ou *"World Health Organization"* (WHO)) a proposé des re-

commandations pour évaluer la réponse au traitement des tumeurs solides [Miller et al. 1981]. Ces critères permettent la séparation des patients en 4 groupes : *Réponse Complète* (RC), *Réponse Partielle* (RP), *Maladie Stable* (MS) et *Maladie Progressive* (MP). Cette séparation se fait en fonction de plusieurs facteurs notamment des critères anatomiques d'évolution de la tumeur en fin de traitement. Cependant, les critères WHO sont discutables. Tout d'abord, la notion de lésion non-mesurable est utilisée alors que ce critère est subjectif. De plus, ces critères sont peu robustes aux biais de mesure et ils ne spécifient pas, par exemple, le nombre maximal/minimal de lésions à prendre en compte lors de l'évaluation.

Les critères *"Response Evaluation Criteria in Solid Tumors"* (RECIST) ont été proposés pour dépasser ces limites. Ils s'appuient également sur les évolutions technologiques en imagerie (RECIST 1.0 [Therasse et al. 2005] et RECIST 1.1 [Eisenhauer et al. 2009]). Ces critères s'appuient sur l'imagerie anatomique TDM en suivant l'évolution du plus grand diamètre des lésions au cours du temps. Cette mesure est faite en supposant que la forme des tumeurs est elliptique. Un patient est considéré comme ayant une RC si toutes ses lésions ont disparu et que tous les ganglions lymphatiques atteints ont un diamètre inférieur à 10 mm. Un patient est défini comme ayant une RP si le diamètre des lésions a diminué de 30 % en moyenne, MP si le diamètre a augmenté d'au moins 20 % et MS dans les autres cas.

L'une des nouveautés de RECIST 1.1 par rapport à la version 1.0 est l'utilisation "raisonnable" des données qualitatives issues de l'imagerie fonctionnelle TEP au FDG. Les informations fournies par la TEP peuvent être utilisées pour la prise en compte de l'apparition de nouvelles lésions. L'imagerie fonctionnelle présente en effet un intérêt particulier en cancérologie car elle permet la mise en évidence d'anomalies métaboliques engendrées par la maladie. Ces informations sont complémentaires de celles apportées par l'imagerie TDM anatomique. En effet, la TEP au FDG permet la visualisation du métabolisme glucidique des cellules. Cela est particulièrement intéressant en oncologie car les cellules cancéreuses présentent une consommation plus élevée en glucose que les cellules saines. Par ailleurs, les évolutions de la maladie ont un retentissement physiologique plus précoce qu'anatomique permettant ainsi *a priori* de modifier le traitement plus ra-

pidement en imagerie TEP que TDM. De plus, les critères RECIST 1.1 ne sont pas toujours adaptés à toutes les localisations ou traitements, c'est notamment le cas pour les lésions non-mesurables.

En 2009, ont été proposé dans la littérature [Wahl et al. 2009] les critères *"PET Response Criteria in Solid Tumors"* (PERCIST). Au contraire des critères morphologiques précédents, ces critères placent la TEP au centre du suivi thérapeutique. Wahl et al. proposent l'utilisation des informations métaboliques quantitatives présentes dans l'imagerie TEP à partir d'un indice d'intensité de fixation du radiotraceur correspondant à la moyenne des intensités au sein d'une zone de 1 mL autour de l'intensité maximale appelée "peak". Cet indice est calculé, puis comparé entre 2 examens successifs permettant ainsi de séparer les patients en 4 catégories. Un patient est considéré comme en RC si toutes les lésions ont disparu et que l'intensité de fixation sur l'imagerie post-thérapeutique de la lésion est inférieure à celle d'une zone de référence saine (foie ou aorte). Un patient est défini comme étant en RP si la variation du "peak" entre 2 examens est de -30 % pour la zone de plus forte intensité, MP si le "peak" est de +30 % et MS dans les autres cas.

Enfin, l'un des éléments important du protocole PERCIST est la définition d'un protocole d'acquisition optimisé pour les études multicentriques [Boellaard et al. 2008] dans le but d'obtenir des résultats reproductibles. En effet, ce protocole, mis à jour en 2015 [Boellaard et al. 2015], limite la variabilité des images lors de leur acquisition.

Une littérature abondante a montré l'intérêt de l'imagerie TEP en oncologie, aussi bien pour la pose du diagnostic, que pour l'évaluation du stade de la maladie et le suivi thérapeutique [Ben-Haim and Ell 2008]. De ce fait, nous nous sommes intéressés à l'imagerie TEP au FDG en oncologie, en particulier à sa valeur pronostique et prédictive. Dans les sections suivantes, nous allons présenter le principe de l'imagerie médicale TEP en oncologie, ainsi que les caractéristiques que l'on peut en extraire. L'intérêt de l'imagerie TEP en oncologie couvre aussi bien les tumeurs solides qu'hématologiques. Les données de la littérature étant très abondantes, nous nous focaliserons principalement sur le cancer de l'œsophage et du poumon, en raison de l'intérêt clinique de notre équipe de recherche pour ces 2 tumeurs [Vera et al. 2014] et [Lemarignier et al. 2014].

1.2 Principe de l'imagerie TEP au FDG

1.2.1 Principe de l'imagerie TEP

La TEP (ou *Positron Emission Tomography* (PET), Figure 1.1) est une méthode d'imagerie permettant la visualisation de la distribution spatiale d'un vecteur associé à un émetteur de positons. L'association des 2 forme un radiopharmaceutique. Le but est de mettre en évidence la fonction ciblée par le vecteur. Ce dispositif d'imagerie donne une information fonctionnelle des tissus biologiques contrairement aux méthodes d'imagerie anatomique comme la TDM.

FIGURE 1.1 – Exemple d'un dispositif d'imagerie TEP/TDM (Biograph Horizon, Siemens).

La TEP est un système d'imagerie photonique en coïncidence basée sur la désintégration par émission de positons e^+ (antiparticule de l'électron e^-, de masse égale mais de charge opposée).

Un émetteur de positons est un élément présentant un excès de protons dans son noyau. Deux voies de désintégration sont possibles. La première est la capture électronique. La deuxième est l'émission β^+ (Équation 1.1) créant un noyau fils Y ayant le même nombre de masse, mais un proton de moins que le noyau père X. L'apparition de Y est accompagné de l'émission d'un positon e^+ et d'un neutrino ν_e.

$$ {}^{A}_{Z}X \rightarrow {}^{A}_{Z-1}Y + e^+ + \nu_e \tag{1.1} $$

Le positon réalise un court parcours (au maximum 3,8 mm dans l'eau pour le fluor 18) avant de rencontrer une particule sœur : l'électron. Cette rencontre matière-antimatière engendre une réaction d'annihilation donnant naissance à 2 photons γ (Figure 1.2). Ces derniers sont émis dans une même direction, un sens opposé à 180° l'un de l'autre avec une énergie de 511 keV chacun.

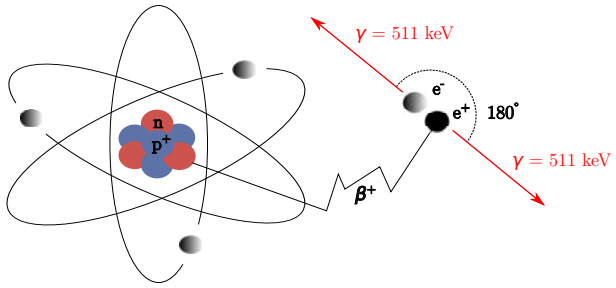


FIGURE 1.2 – Après une courte distance, le positon e^+ obtenu par émission β^+ est annihilé avec un électron e^- donnant naissance à deux photons γ émis dans une même direction, en sens opposé à 180° l'un de l'autre et avec une énergie de 511 keV chacun.

Les photons γ résultant de l'annihilation vont alors interagir avec la matière qu'ils traversent, tissus biologiques et détecteurs, ayant comme effets possibles une diffusion avec perte d'énergie des photons et une atténuation responsable d'une diminution du nombre de photons incidents.

Le fait que deux photons soient émis simultanément après annihilation est une information permettant l'estimation du lieu de la désintégration. En mode 3D, cette détection est effectuée par plusieurs couronnes de détecteurs. On nomme ligne de réponse, une ligne dans l'espace tridimensionnelle correspondant au trajet de 2 photons détectés en coïncidence, c'est-à-dire simultanément dans une fenêtre temporelle de quelques ns. On distingue alors plusieurs cas :

— Les coïncidences vraies (Figure 1.3a) lorsque les photons provenant d'une même annihilation et n'ayant subi aucune diffusion Compton sont détectés.

— Les coïncidences fortuites (Figure 1.3b) lorsque 2 photons ne provenant pas de la même annihilation sont détectés créant une ligne de réponse fausse.

— Les coïncidences diffusées (Figure 1.3c) lorsque 2 photons provenant d'une même annihilation sont détectés mais qu'au moins l'un d'entre eux a subi une diffusion Compton altérant l'interprétation de la ligne de réponse.

— Les coïncidences multiples (Figure 1.3d) lorsque plus de 2 photons sont détectés

au cours d'une même fenêtre de temps.

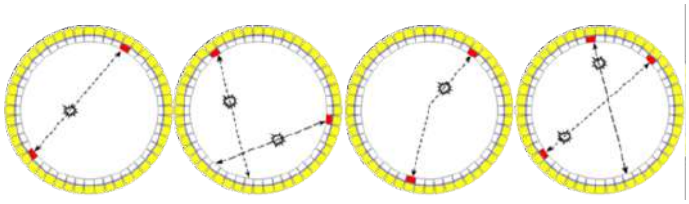


FIGURE 1.3 – Types de coïncidences enregistrées par le système de détection : (a) coïncidence vraie, (b) coïncidence diffusée, (c) coïncidence fortuite et (d) coïncidence multiple.

Il est intéressant de noter que la ligne de réponse n'est qu'une approximation du lieu d'émission du positon du fait de son parcours avant l'annihilation. Cette approximation est une limite intrinsèque à la méthode de détection en coïncidence. Sur les dispositifs TEP récents, on trouve un système électronique de mesure de la durée séparant l'enregistrement des deux évènements, nommée le temps de vol. Cela permet d'avoir une estimation de la position de l'annihilation et ainsi d'améliorer la résolution spatiale du système de détection.

L'étape qui consiste à passer de ce système de détection à la distribution en 3D de l'activité mesurée est appelée reconstruction tomographique. A partir d'un signal $f(x, y)$ d'une coupe 2D d'un objet contenant de la radioactivité, un ensemble de projections mono-dimensionnelles sont obtenues (Figure 1.4). Une projection $p(\theta, u)$ correspond à l'intégrale sur v du signal étant donné un angle θ et une position u. Une projection, en chaque point u, correspond ainsi à la somme de toutes les lignes de réponse rencontrées le long de l'axe v. Enfin, les projections sont rangées dans un sinogramme, ou chaque ligne correspond à une projection à un angle donné.

A partir de là, l'objectif est de reconstruire le signal tridimensionnel correspondant à un problème inverse. Il existe 2 familles de méthodes de reconstruction d'images : les méthodes analytiques et les méthodes itératives. Cependant, en imagerie TEP, les méthodes analytiques ne sont pas utilisées. En fin de traitement, une image numérique médicale est obtenue.

En dépit des progrès techniques, la résolution des images TEP reste en deçà de celles des images obtenues par les systèmes d'imageries anatomiques, comme la TDM et l'IRM. Le problème du bruit dans les images reconstruites, et surtout leur résolution spatiale li-

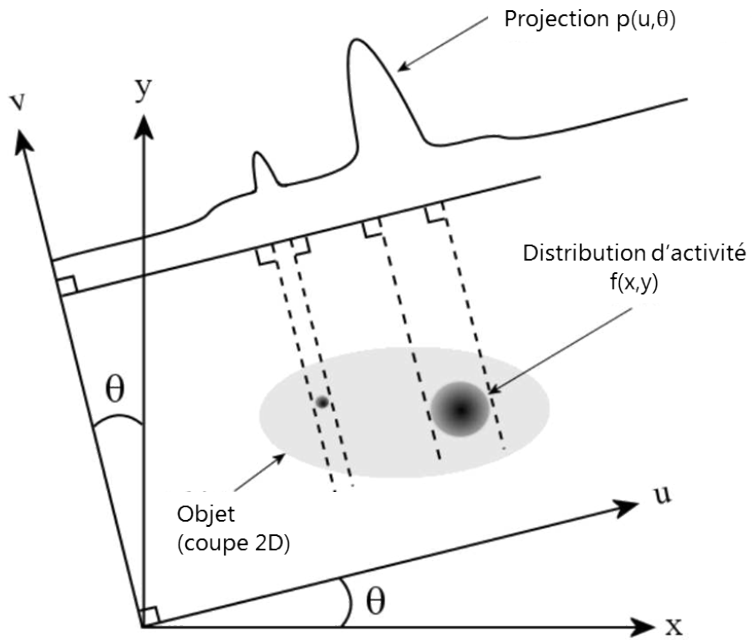


FIGURE 1.4 – Principe d'obtention des projections. Le signal $f(x, y)$ est projeté selon un angle θ sur un profil. Le nouveau signal $p(\theta, u)$ correspond ainsi à une projection de la distribution d'activité $f(x, y)$ sous l'angle d'incidence θ.

mitée, constituent des limites majeures à la quantification en TEP. L'effet de volume partiel qui est dû à la fois à la faible résolution spatiale du système d'imagerie et à l'échantillonnage des données, est une source de biais d'autant plus critique dans le contexte du suivi thérapeutique qu'il dépend de nombreux paramètres, ce qui rend sa correction complexe.

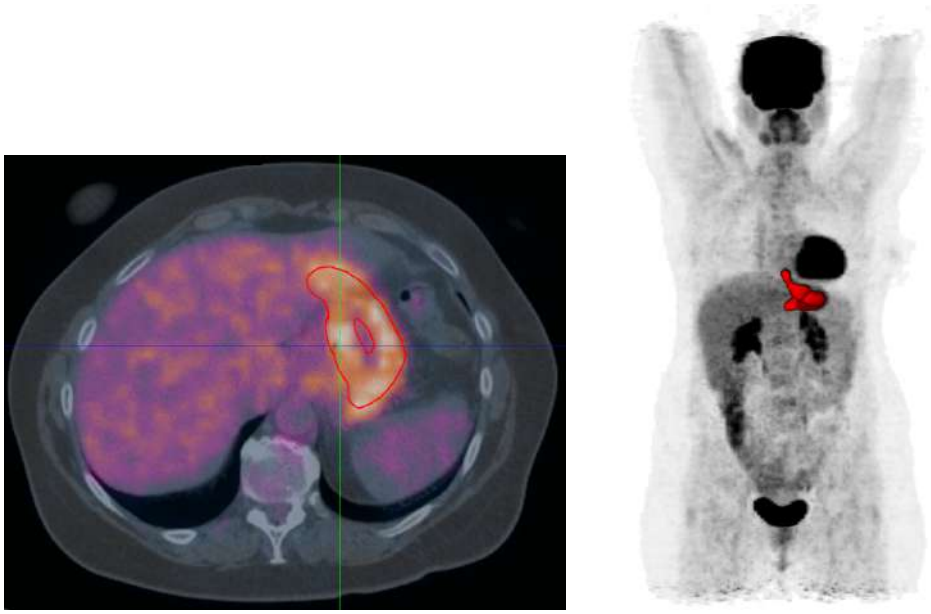
Les dispositifs d'imagerie TEP sont associés à des TDM permettant la visualisation anatomique et fonctionnelle en un seul examen et de réaliser une correction d'atténuation. Par ailleurs, plusieurs systèmes existent, matériels et logiciels, pour corriger des biais de mesures et ainsi améliorer la quantification de la radioactivité. Plus récemment, des dispositifs associant l'imagerie TEP à l'imagerie IRM ont été développés menant à l'apparition des dispositifs hybrides TEP/IRM.

Pour une description plus détaillé du fonctionnement physique de l'imagerie TEP ainsi que de la reconstruction tomographique, nous recommandons la lecture de [Das B K. and Das 2015].

1.2.2 Le Fluoro-2-désoxy-D-glucose

Le principal radiopharmaceutique utilisé en imagerie TEP oncologique est le FDG. Il permet d'étudier le métabolisme glucidique des cellules. Il s'agit d'un analogue du glucose dans lequel un atome de fluor 18 émetteur de positons est fixé, de période physique de 110 minutes. Une fois injecté, les cellules assimilent le FDG par le mécanisme de glycolyse. Il est alors phosphorylé en FDG-6P par une enzyme cytosolique : l'hexokinase. Or, il n'existe pas dans les cellules humaines d'enzyme pour transformer le FDG-6P. De ce fait, ce dernier reste piégé dans le cytosol.

L'imagerie TEP permet de détecter la présence du FDG dans l'organisme. La fixation n'est pas spécifique aux cellules cancéreuses, mais est présente dans l'ensemble de l'organisme. Cependant, comme les cellules tumorales possèdent un métabolisme glucidique plus important que les cellules saines, la concentration en FDG est supérieure dans ces zones (Figure 1.5a). La fixation en FDG est différente d'un patient à l'autre en fonction de son métabolisme, du type de lésion, mais également au sein d'une même tumeur créant des hétérogénéités de fixation. C'est le cas, par exemple, dans les zones de nécrose présentent généralement dans les tumeurs volumineuses, comme l'illustre la Figure 1.5.



(a) (b)

FIGURE 1.5 – (a) Coupe transverse TEP au FDG et (b) *"Maximum Intensity Projection"* (MIP) d'un patient présentant une tumeur de l'œsophage. La tumeur apparait colorée sur le MIP. Les autres organes présentant également une fixation normale du FDG sont le cerveau et le cœur dû à leur activé permanente, les reins et la vessie pour leur rôle de filtration.

Il existe d'autres radiopharmaceutiques utilisés en imagerie TEP oncologique. On peut rapidement citer la Fluoro-L Thymidine, témoignant de la prolifération cellulaire, ainsi que la Fluoro-Misonidazole qui est un traceur de l'hypoxie.

1.2.3 Le SUV

L'image TEP reflète la concentration du radiotraceur dans les tissus biologiques. Cependant, celle-ci, ainsi que le contraste, évolue au cours du temps en fonction de la pharmacocinétique du traceur. Afin d'obtenir un contraste suffisant entre les tissus sains et pathologiques, il est nécessaire d'attendre un certain temps après l'injection (i.e. 60 min pour le FDG). L'intensité du signal reçu dépend de plusieurs paramètres comme :

— l'activité de produit radioactif injecté,

— le temps d'attente entre l'injection et l'acquisition,

— le volume du patient, sa glycémie, son état de repos.

Par ailleurs, de nombreux phénomènes physiques viennent biaiser la mesure de l'activité présente dans les tissus biologiques. Pour comparer différents examens TEP, il a été proposé de normaliser les mesures d'intensité des images TEP. Cette intensité est nommée *"Standardized Uptake Value"* (SUV) et a été proposée pour la première par Kenney et al. [Kenney et al. 1941]. Si le radiotraceur est distribué uniformément dans l'ensemble de l'organisme, la moyenne des SUV est égale à 1. Tout écart à 1 traduit une répartition non-uniforme du radiotraceur dans le volume dans lequel il est distribué.

Le *"Standardized Uptake Value Body Weight"* (SUV_{BW}) est le type de SUV le plus utilisé dans la littérature. Il normalise la concentration d'activité estimée par l'imagerie TEP par l'activité injectée et le poids du patient. Comme la masse volumique du corps humain est proche de celle de l'eau, le poids du patient en grammes est assimilé à son volume en millilitres. Le SUV_{BW} n'a donc pas d'unité. L'Équation 1.2 permet le calcul de la valeur du SUV_{BW}.

$$SUV_{BW} = \frac{\text{Concentration d'activité } \left(\frac{\text{kBq}}{\text{mL}}\right)}{\text{Activité injectée (kBq)/Poids du patient (g)}} \tag{1.2}$$

Afin de proposer la méthode de normalisation la plus appropriée physiologiquement,

certains auteurs ont proposés l'utilisation d'autres expressions mathématiques du SUV, comme la normalisation par la masse maigre des patients *"Lean Body Mass"* (LBM) ou par la surface corporelle *"Body Surface Area"* (BSA). Le SUV_{LBM} est calculé à partir de l'Équation 1.2 où le poids du patient est remplacé par sa masse maigre en grammes. Il existe plusieurs formules pour calculer la LBM, plusieurs d'entre elles sont recensées dans l'article de Erselcan et al., en fonction de l'âge, du poids et du sexe du patient [Erselcan et al. 2002]. Tout comme le SUV_{BW}, le SUV_{LBM} est sans unité. Enfin, le SUV_{BSA} est calculée en remplaçant le poids du patient (Équation 1.2) par sa surface corporelle. Cette dernière est calculée grâce à l'Équation 1.3 [Du Bois and Du Bois 1916]. Le facteur $0,00718$ est un coefficient en $m^2/kg.cm$, ainsi la BSA est exprimée en m^2 et le SUV_{BSA} en m^2/mL.

$$BSA\ (m^2) = 0,00718 \times \text{poids}\ (kg)^{0,425} \times \text{taille}\ (cm)^{0,725} \tag{1.3}$$

Le SUV_{BW} est celui qui est généralement utilisé dans la littérature. Par la suite, il sera simplement nommé SUV.

1.2.4 Apport clinique du SUV_{max}

Grâce à l'imagerie fonctionnelle TEP, les médecins peuvent de manière rapide et visuelle détecter les zones d'hyperfixation du radiotraceur correspondant à une valeur de SUV élevée. Dès la fin des années 90, l'analyse qualitative subjective a été remplacée par une analyse quantitative à l'aide de caractéristiques calculées sur ces images. Une des premières caractéristiques à avoir été étudiée est le SUV_{max} qui est la valeur maximum du SUV dans la lésion car elle est la caractéristique la plus facile à déterminer.

Dans le Tableau 1.1 sont référencés des articles de la littérature abordant l'étude de la valeur prédictive et pronostique de plusieurs caractéristiques, dont le SUV_{max}, dans le cancer du poumon non à grandes cellules. Lowe et al. ont proposé l'utilisation d'un seuil de SUV_{max} de 2,5 pour différencier le nodule pulmonaire malin du nodule bénin [Lowe et al. 1998]. L'idée sous-jacente est que plus un nodule fixe le FDG, plus son SUV est élevé et plus le risque de malignité est important. Ce seuil est un compromis entre une sensibilité souhaitée relativement élevée et une spécificité acceptable. Dans une étude

prospective de Bryant et al. [Bryant and Cerfolio 2006], sur une cohorte de 585 patients atteints d'un cancer pulmonaire, 24 % des nodules étaient malins lorsque le SUV_{max} était compris entre 0 et 2,5. Ce pourcentage passait à 80 % si le SUV_{max} était mesuré entre 2,6 et 4 et à 96 % lorsque le SUV_{max} était supérieur à 4,1.

La revue de la littérature de Van de Wiele et al. [Van De Wiele et al. 2013] recense des articles abordant, entre autre, l'intérêt du SUV_{max} appliqué au cancer des poumons, mais également aux cancers *Oto-rhino-laryngologie* (ORL) et de l'œsophage. Il y est rapporté que le SUV_{max} et son évolution au cours du traitement (ΔSUV_{max}) présentent une valeur pronostique et prédictive de la réponse au traitement. Dans les Tableaux 1.2 et 1.3 sont regroupés plusieurs articles de la littérature abordant l'étude de la valeur prédictive et pronostique, respectivement, de plusieurs caractéristiques dans le cancer de l'œsophage. L'évolution précoce du SUV_{max}, notamment durant les 2 premières semaines de traitement, est présentée comme étant particulièrement importante [Zhu et al. 2012]. La valeur du SUV_{max} s'est avérée être efficace pour l'identification des groupes de patients ayant des pronostics différents. D'après [Teyton et al. 2008], un seuil de SUV_{max} de 9 est défini comme optimal sur une étude pronostique d'une cohorte de 52 patients atteints d'un cancer l'œsophage. Les patients présentant une fixation intense du FDG ($SUV_{max} > 9$) avaient une médiane de survie de 13 mois, alors qu'elle était de 26 mois pour les autres patients. Ce résultat est en accord avec l'analyse rétrospective de Al-Taan et al. [Al-Taan et al. 2014] réalisée sur 22 études. Il en ressort que des différences existent dans la valeur du seuil optimal. Par exemple, Al-Taan et al. le situe plutôt autour de 11.

Ces données de la littérature montrent, aussi bien pour les cancers pulmonaires que de l'œsophage, qu'il n'existe pas actuellement de consensus sur les caractéristiques prédictives et/ou pronostiques, ni sur la valeur des seuils à appliquer.

TABLEAU 1.1 – Études prédictives et pronostiques utilisant des statistiques classiques des caractéristiques images (1^{er} et $2^{ème}$ ordres) extraites d'un examen TEP initial au FDG chez des patients présentant un cancer du poumon non à petites cellules. **Adénocarcinome* (ADC), *Carcinome épidermoïde* (SCC), ***"Metabolically Tumour Volume"* (MTV) *"Total Lesion Glycolysis"* (TLG), *"Tumor Length"* (TL).

Référence	Recrut.	Nb de patients	Histologie*	Stade TNM	Statistiques utilisées	Caractéristiques étudiés	Résultats**
[Lowe et al. 1998]	Rétrospec.	89	?	?	?	SUV_{max}	SUV_{max} → tissus sains et pathologiques
[Bryant and Cerfolio 2006]	Prospec.	585	40% ADC 34% SCC	?	Test exact de Fisher, test du chi^2, test de Wilcoxon	SUV_{max}	SUV_{max} → tissus sains et pathologiques
[Yan et al. 2011]	Rétrospec.	120	49% ADC 51% SCC	III à IV	Kaplan Meier, test de log-rank, analyse de Cox, courbes ROC	SUV_{max} & MTV	SUV_{max} (>12) & MTV (>34 mL) → OS et SUV_{max} (>12) & MTV (>34 mL) → PFS
[Lee et al. 2012]	Rétrospec.	61	?	I à IV	Kaplan Meier, test de log-rank, analyse de Cox	SUV_{max} & MTV	MTV → OS & PFS
[Zhang et al. 2013]	Rétrospec.	328	39% ADC 28% SCC	II à IV	Kaplan Meier, test de log-rank, analyse de Cox, courbes ROC	SUV_{max}, SUV_{moy}, MTV & TLG	SUV_{max} (> 2,22), SUV_{moy} (>1,27), MTV (>66) cm^3 & TLG (>205) → OS
[Kim et al. 2012]	Rétrospec.	91	56% ADC 34% SCC	I à III	test du chi^2, courbes ROC, test de Mann–Whitney U, test de Kruskal–Wallis, Kaplan Meier, test de log-rank, analyse de Cox	MTV & TLG	MTV & TLG → OS & PFS

[Soussan et al. 2013]	Rétrospec.	32	38% ADC 50% SCC	III	Kaplan Meier, test de log-rank, analyse de Cox	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG, Volume & Diametre TDM	aucun
[Chen et al. 2013]	Rétrospec.	187	66% ADC 22% SCC	I à IV	Kaplan Meier, test de log-rank, analyse de Cox	SUV_{max}, MTV, TLG	SUV_{max} (>15), MTV (>68 cm^3) & TLG (> 655) → PFS
[Cook et al. 2013]	Rétrospec.	53	40% ADC 45% SCC	I à III	Analyse univariée, courbes ROC	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG & 4 GLDM (Coarseness, Contrast, Busyness, Complexity)	Coarseness, Contrast & Busyness (GLDM) → RC
[Cook et al. 2013]	Rétrospec.	53	40% ADC 45% SCC	I à III	Kaplan Meier, test de log-rank, analyse de Cox, courbes ROC	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG & 4 GLDM (Coarseness, Contrast, Busyness, Complexity)	Coarseness, Contrast & Busyness (GLDM) → PFS et Coarseness (GLDM) → OS
[Vera et al. 2014]	Prospec.	52	42% ADC 52% SCC	I à III	ANOVA, test du chi^2, régression logistique, courbes ROC	SUV_{max}, SUV_{moy}, SUV_{peak} & MTV	SUV_{moy} & MTV → OS
[Pyka et al. 2015]	Rétrospec.	45	40% ADC 53% SCC	I à III	Kaplan Meier, test de log-rank, analyse de Cox, courbes ROC	SUV_{max}, SUV_{moy}, MTV, COV, diamètre TDM, et 5 textures relatives à 64 intensités (Entropie, Corrélation, Contraste (GLCM), Busyness, Coarsenes (GLDM))	Entropie (GLCM) → PFS
[Fledelius et al. 2017]	Rétrospec.	50	?	?	Régression linéaire, courbes ROC	SUL_{max}, SUL_{peak}, TLG	TLG → RC

TABLEAU 1.2 – Études prédictives utilisant des statistiques classiques des caractéristiques images (1er et 2ème ordres) extraites d'un examen TEP initial au FDG chez des patients présentant un cancer de l'œsophage localement avancé, traité par RCT avec une chirurgie optionnelle.

Référence	Recrut.	Nb de patients	Histologie	Localisation	Stade TNM	Statistiques utilisées	Caractéristiques étudiés	Résultats
[Swisher et al. 2004]	Rétrospec.	83	88% ADC 12% SCC	1% sup 6% moy 93% inf	0 à IV	?	SUV_{max}	aucun
[Levine et al. 2006]	Prospec.	64	81% ADC 14% SCC	6% sup 11% moy 83% inf	I à IV	test t de Student, chi^2 & test de Fisher	SUV_{max}	SUV_{max} (>4) → RC
[Kato et al. 2007]	Rétrospec.	27	SCC	?	I à IV	test t de Student, chi^2 & test de Fisher, ANOVA, régression logistique	"SUV", Longueur tumorale	"SUV" (>8,2) & TL → RC
[Rizk et al. 2009]	Rétrospec.	189	ADC	?	II à IV	test t de Student, chi^2 & test de Fisher	SUV_{max}	SUV_{max} (>4,5) → RC
[Tixier et al. 2011]	Rétrospec.	41	24 % ADC 76 % SCC	24 % sup 37 % moy 39 % inf	I à IV	courbes ROC, Kruskal-Wallis	SUV_{max}, SUV_{moy}, SUV_{peak}, 7 carac. du 1er ordre, 31 carac. de texture relatives à 64 intensités	SUV_{peak}, Entropie (GLCM), Coarsness (GLDM), GLNUz et ZLNU (GLSZM) → RC
[Hatt et al. 2011c]	Rétrospec.	50	26 % ADC 74 % SCC	26 % sup 40 % moy 34 % inf	I à IV	courbes ROC, Kruskal-Wallis	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG & TL	MTV, TLG & TL → RC

[Jayachandran et al. 2012]	Rétrospec.	37	73% ADC 27% SCC	8% sup 16% moy 76% inf	I à IV	test logrank, Analyse de Cox	SUV_{max}, MTV, TLG	aucun
[Palie et al. 2013]	Prospec.	48	SCC	30 % sup 47 % moy 23 % inf	II à IV	Régression logistique, courbes ROC, ANOVA	SUV_{max}, MTV, TLG	SUV_{max}, MTV, TLG → RC
[Blom et al. 2013]	Rétrospec.	73	73 % ADC 24 % SCC	53 % sup 22 % moy 25 % inf	II à IV	test de Mann-Whitney U, Kruskal-Wallis	SUV_{max}, MTV, Diamètre tumoral	MTV → RC
[Nakajo et al. 2017]	Rétrospec.	52	SCC	13 % sup 23 % moy 16 % inf	II à IV	test de Mann-Whitney U, test de Kruskal-Wallis, courbes ROC, chi^2	SUV_{max}, SUV_{moy}, MTV, TLG et 6 textures relatives à 64 intensités GLCM (entropie, homogénéité et dissimilarité) et GLSZM (GLNU, ZLNU et ZP)	MTV, TLG, GLNUz et ZLNU → RC

TABLEAU 1.3 – Études pronostiques utilisant des statistiques classiques des caractéristiques images (1er et 2ème odre) extraites d'un examen TEP initial au FDG chez des patients présentant un cancer de l'œsophage localement avancé, traité par RCT avec une chirurgie optionnelle.

Référence	Recrut.	Nb de patients	Histologie	Localisation	Stade TNM	Statistiques utilisées	Caractéristiques étudiés	Résultats
[Teyton et al. 2008]	Prospec.	52	23% ADC 77% SCC	19% sup 38% moy 43% inf	?	Kaplan-Meier, test de logrank et analyse de Cox	SUV_{max}	SUV_{max} (>9) → OS
[Rizk et al. 2009]	Rétrospec.	189	ADC	?	II à IV	test t de Student, chi^2 & test de Fisher	SUV_{max}	aucun
[Hyun et al. 2010]	Rétrospec.	151	3 % ADC 97 % SCC	16 % sup 49 % moy 35 % inf	I à IV	courbes ROC, Kaplan-Meier, Analyse de Cox	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG, TL	MTV → OS
[Hatt et al. 2011b]	Rétrospec.	45	27 % ADC 73 % SCC	24 % sup 38 % moy 38 % inf	I à IV	courbes ROC, Kaplan-Meier, Analyse de Cox	SUV_{max}, SUV_{moy}, SUV_{peak}, MTV, TLG, TL	MTV (<14cm^3, >85cm^3), TL et TLG (>180) → OS
[Jayachandran et al. 2012]	Rétrospec.	37	73% ADC 27% SCC	8% sup 16% moy 76% inf	I à IV	Kaplan-Meier, Test logrank, Analyse de Cox	SUV_{max}, MTV, TLG	aucun
[Lemarignier et al. 2014]	Rétrospec.	67	100 % SCC	24 % sup 48 % moy 28 % inf	II à IV	Kaplan-Meier, Test logrank, Analyse de Cox	SUV_{max}, SUV_{moy}, MTV, TLG	MTV (<14 cm^3 → OS
[Hatt et al. 2015]	Rétrospec.	112	56 % ADC 44 % SCC	NA	I à IV	courbes ROC, Kaplan-Meier, Analyse de Cox	SUV_{max}/SUV_{moy}/SD, 4 caract de texture (2 GLCM et 2 GLSZM)	Dissimilarity (GLCM) → OS, mieux que MTV
[Nakajo et al. 2017]	Rétrospec.	52	SCC	13 % sup 23 % moy 16 % inf	II à IV	courbes ROC, Kaplan-Meier, Analyse de Cox	SUV_{max}, SUV_{moy}, MTV, TLG et 6 textures relatives à 64 intensités GLCM (entropie, homogénéité et dissimilarité) et GLSZM (GLNU, ZLNU et ZP)	aucun

1.3 Les caractéristiques en imagerie TEP au FDG

1.3.1 Introduction

Comme nous venons de le voir, le SUV_{max} constitue une caractéristique très importante. Cependant, certains auteurs ont fait la remarque qu'il ne caractérise la tumeur qu'à partir d'un seul voxel. Par ailleurs, il est particulièrement sensible au bruit [Buvat 2007]. Pour améliorer cette robustesse, des auteurs ont proposé l'utilisation de valeurs moyennes du SUV sur la lésion. Cependant, ces caractéristiques nécessitent la segmentation au préalable de la lésion afin d'obtenir un volume tumoral. En cancérologie, ce volume est une caractéristique importante intervenant dans la stadification de la maladie. En particulier, il a été utilisé en TDM dans les critères RECIST décrits précédemment [Eisenhauer et al. 2009].

Si l'on considère l'image de la lésion et les intensités des voxels, d'autres caractéristiques correspondant à des statistiques du 1^{er} ordre peuvent également être calculées. Par ailleurs, il a été montré que les tumeurs présentant une hétérogénéité avaient tendance à résister aux traitements et étaient de moins bons pronostics [Jackson et al. 2007]. C'est pourquoi certains auteurs se sont intéressés à l'hétérogénéité de fixation du FDG [El Naqa et al. 2009].

En 2003, Miller et al. [Miller et al. 2003] ont caractérisé des tumeurs du col de l'utérus à l'aide d'un score dépendant, entre autres, de l'hétérogénéité évaluée par une approche visuelle. Plus tard, Kalff et al. [Kalff et al. 2006] ont proposé des critères basés sur la fixation tumorale du FDG afin de caractériser la réponse thérapeutique de carcinomes rectaux. De ce fait, des caractéristiques statistiques ont été proposées dans la littérature [El Naqa et al. 2009]. Ces caractéristiques statistiques, bien connus dans le domaine du traitement de l'image, peuvent être classées en statistiques du 1^{er} ordre, du $2^{ème}$ ordre et d'ordres supérieurs en fonction du nombre de voxels intervenant dans les relations étudiées. La différence de base est que les statistiques du 1^{er} ordre estiment les propriétés des valeurs individuelles des voxels, comme la moyenne et la variance, en ignorant l'interaction spatiale entre voxels. Les statistiques d'ordres supérieurs estiment les propriétés de deux ou

plusieurs voxels se situant à des emplacements spécifiques les uns par rapport aux autres.

De même, les lésions sphériques avec des bords bien nets sont considérées comme moins agressives que des tumeurs diffuses et infiltrantes. De ce fait, des caractéristiques de forme ont été proposées dans la littérature [El Naqa et al. 2009]. Dans le même esprit, on retrouve dans la littérature portant sur le traitement de l'image, l'utilisation de la dimension fractale [Mandelbrot 1994] ou des ondelettes [Meyer and Salinger 1992]. Ces caractéristiques ont été utilisées en imagerie TEP oncologique [Breki et al. 2016] [Miwa et al. 2014] [Takeshita et al. 2016]. Cependant, ces caractéristiques n'ont pas été étudiées dans le cadre de cette thèse car nous avons voulu concentrer notre attention sur les caractéristiques du 1er ordre, de forme et de texture.

Dans la section suivante nous allons présenter les différentes caractéristiques rencontrées dans la littérature en imagerie TEP oncologique en donnant leur définition et leur intérêt clinique.

1.3.2 Caractéristiques statistiques du 1er ordre

1.3.2.1 Volume tumoral et caractéristiques dérivées du SUV

Les caractéristiques statistiques du 1er ordre qualifient les intensités des voxels d'un *"Volume Of Interest"* (VOI) dans sa globalité. En imagerie TEP au FDG, comme la lésion cancéreuse est hyperfixante, il est possible de l'isoler du reste de l'image pour définir le volume métabolique tumoral (MTV).

A l'intérieur de celui-ci, il est possible d'extraire plusieurs caractéristiques autres que le SUV_{max} (Figure 1.6). il s'agit du :

— SUV_{moy} correspondant à la moyenne des SUV des voxels d'un VOI. Il reflète l'activité métabolique générale de la tumeur.

— SUV_{peak} [Wahl et al. 2009] correspondant à la moyenne maximale des SUV calculée dans un VOI d'un millilitre. C'est un compromis entre le SUV_{max} et le SUV_{moy}.

— SUV_{somme} correspondant à la somme des SUV dans un VOI.

— TLG [Larson et al. 1999] correspondant au produit du SUV_{moy} d'un VOI par son MTV.

— TL correspondant à la mesure du plus grand diamètre tumoral.

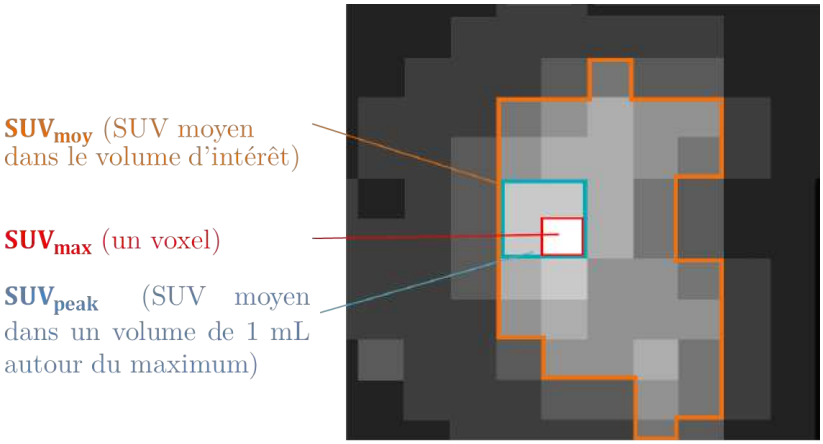


FIGURE 1.6 – Représentation des dérivés du SUV sur une coupe transverse d'une lésion.

1.3.2.2 Histogramme de fréquences des intensités et histogramme intensité-volume

Comme dans toutes statistiques du 1er ordre, il est possible de représenter les intensités, ou SUV, sous forme d'histogrammes et de calculer des statistiques à partir de cette distribution.

L'histogramme des intensités est une fonction h qui donne pour chaque niveau d'intensité (i) le nombre de voxels ayant cette intensité dans une image ou un VOI I de taille $\omega = X \times Y \times Z$ (Équation 1.4) :

$$h(i) = \sum_{x,y,z}^{X,Y,Z} \delta(I(x,y,z),i) \qquad \text{avec } i = 0,1,\ldots,G-1 \tag{1.4}$$

Où l'image I peut être considérée comme une fonction de trois variables x, y et z avec $x \in [1,X]$, $y \in [1,Y]$ et $z \in [1,Y]$. $I(x,y,z)$ peut prendre des valeurs discrètes telles que $i = 0,1,\ldots,G-1$, où G est le nombre de niveaux de gris dans l'image. $\delta(j,i)$ est la fonction de Kronecker telle que Équation 1.5 :

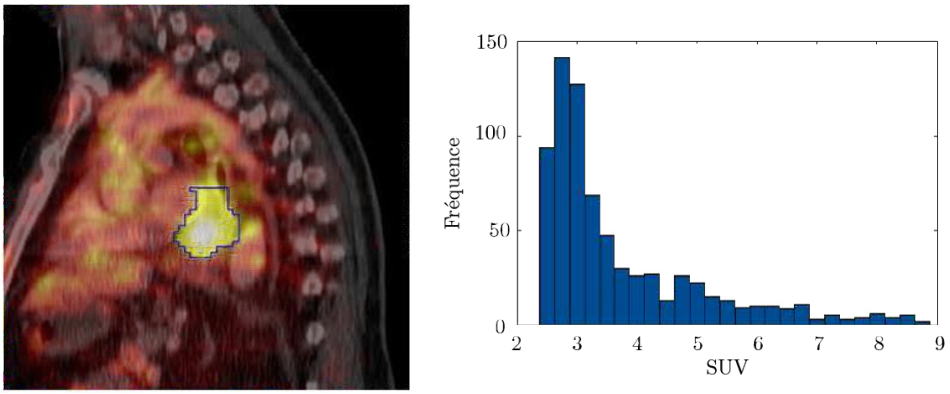
$$\delta(j,i) = \begin{cases} 1 \text{ si } j = i \\ 0 \text{ sinon} \end{cases} \tag{1.5}$$

La densité de probabilité d'occurrence des niveaux d'intensité $p(i)$ est obtenue en divisant les valeurs $h(i)$ par le nombre total de voxels dans l'image (Équation 1.6).

$$p(i) = \frac{1}{\omega} \times h(i) \qquad \text{avec } i = 0, 1, \ldots, G-1 \tag{1.6}$$

Si les niveaux de gris ne sont pas des valeurs discrètes, comme l'imagerie TEP, l'histogramme analyse des intervalle d'intensités.

L'histogramme de fréquences des SUV est un résumé simple et concis de l'information statistique de 1er ordre contenue dans l'image (Figure 1.7). La forme de l'histogramme est révélatrice de l'homogénéité des SUV dans le VOI. L'histogramme d'une image peu contrastée est distribué sur une étroite plage d'intensités alors qu'une image hétérogène correspond à une large plage d'intensités.



(a) (b)

FIGURE 1.7 – Coupe sagittale TEP/TDM d'une tumeur pulmonaire segmentée en bleu (a) et histogramme associé des SUV (b) d'après [Tan et al. 2013a].

Plusieurs caractéristiques peuvent être calculées à partir de ces histogrammes. Elles sont définies dans le Tableau 3.1 [Herlidou et al. 1999]. Elles sont simples à calculer à partir du moment où le VOI a été défini.

Par ailleurs, l'*Histogramme Intensité-Volume* (HIV) a été proposé par El Naqa et al. [El Naqa et al. 2009]. Il représente la distribution d'activité dans la tumeur en traçant la fraction volumique en fonction du SUV (Figure 1.8), c'est-à-dire la fraction des voxels possédant un SUV supérieur ou égal à x. Cet histogramme est construit de manière comparable à l'histogramme dose volume utilisé largement en radiothérapie. El Naqa et al. ont défini plusieurs caractéristiques à extraire de cet histogramme :

— I_x % représentant le minimum de SUV des x % du volume ;

— V_x % représentant la fraction de volume dont le SUV est au moins égal à x % de

TABLEAU 1.4 – Principales caractéristiques statistiques du 1er ordre.

Caractéristiques	Formules
Intensité moyenne	$\mu = \sum_{i=0}^{G-1} i.p(i)$
Variance	$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2.p(i)$
Asymétrie (skewness)	$\mu_3 = \frac{1}{\sigma^3} \sum_{i=0}^{G-1} (i-\mu)^3.p(i)$
Aplatissement (ou kurtosis)	$\mu_4 = \frac{1}{\sigma^4} \sum_{i=0}^{G-1} (i-\mu)^4.p(i)$
Énergie	$E = \sum_{i=0}^{G-1} [p(i)]^2$
Entropie	$H = \sum_{i=0}^{G-1} p(i) log_2[p(i)]$
COV	$COV = \frac{\sigma}{\mu}$

SUV_{max} ;

— Des soustractions entre des valeurs particulières de I_x % (respectivement V_x %), afin de rendre compte de l'hétérogénéité de la distribution dans le volume étudié.

1.3.2.3 Apports cliniques

Les caractéristiques statistiques du 1er ordre ont pour avantage d'être faciles à mesurer par rapport à ceux d'ordres supérieurs. En particulier, le SUV_{peak} mis en avant dans les critères PERCIST présentés précédemment (voir sous-section 1.1.3, page 7). En effet, la variation du ΔSUV_{peak} est déterminante pour séparer les patients en différents groupes de répondeurs. La revue de la littérature de Van de Wiele [Van De Wiele et al. 2013] recense plusieurs publications étudiant de manière rétrospective le pouvoir prédictif et pronostique de plusieurs caractéristiques statistiques. Des caractéristiques du 1er ordre, comme le SUV_{moy}, le SUV_{peak} ou encore le MTV sont présentées dans de nombreux articles comme ayant un intérêt prédictif ou pronostique de l'OS et de la PFS.

Dans les Tableaux 1.2 et 1.3 (page 22) sont regroupés les données de la littérature concernant le cancer de l'œsophage. Tixier et al. ont étudié le pouvoir prédictif de la réponse au traitement de plusieurs caractéristiques du 1er ordre [Tixier et al. 2011]. Pour cela, une cohorte de 41 patients présentant un cancer de l'œsophage, traités par RCT a

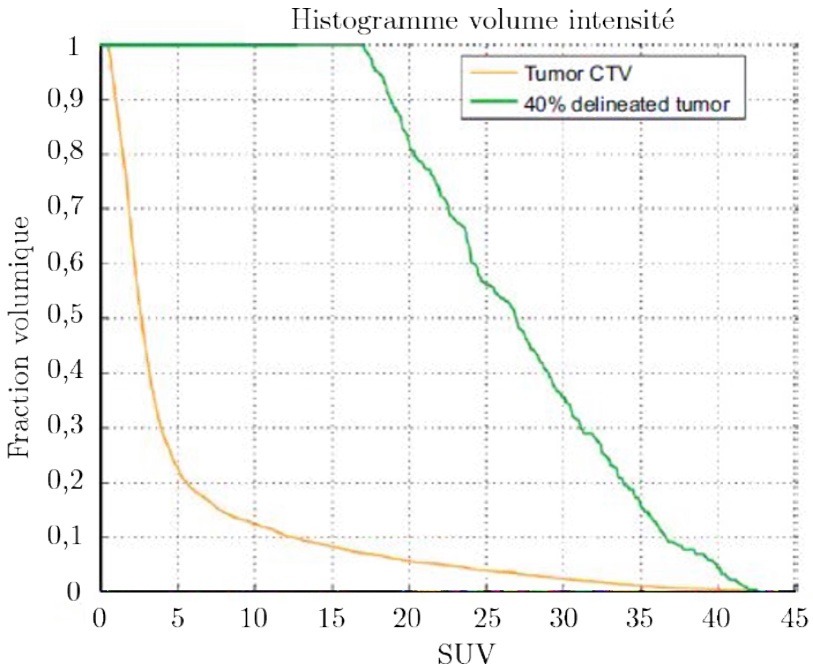


FIGURE 1.8 – Exemple d'HIV sur les données TEP d'une patiente atteinte d'un cancer du col de l'utérus : en marron, la courbe du *"Volume Tumoral Clinique"* (CTV) et en vert celle du volume TEP segmenté par un seuillage à 40 % du SUV_{max} tumoral [El Naqa et al. 2009].

été étudiée. Plusieurs caractéristiques ont été extraites des examens TEP au FDG. Après un test statistique de Kruskal–Wallis et l'utilisation de courbes ROC, le SUV_{max}, le SUV_{moy} et le SUV_{peak} sont apparus comme étant prédictifs de la réponse au traitement en séparant les patients ayant une RC et les autres ($p = 0,034$, $p = 0,044$ et $0,012$, respectivement). Cependant, seul le SUV_{peak} a permis la séparation des patients en 3 catégories (RC, MS et non-répondeur) de manière significative ($p = 0,045$).

En 2009, El Naqa et al. ont proposé l'utilisation, dans des études prédictives, des caractéristiques extraites des histogrammes des fréquences des intensités et intensité-volume pour caractériser l'hétérogénéité tumorale [El Naqa et al. 2009]. Deux cohortes de patients ont été étudiées, une première de 14 patients présentant un cancer du col de l'utérus et une deuxième de 9 patients présentant un cancer ORL. Après une analyse combinant ces caractéristiques par régression logistique, les caractéristiques V_{10-90} et V_{90} apparaissent prédictifs pour ces 2 cancers montrant ainsi l'intérêt de l'étude de l'hétérogénéité de la fixation du FDG.

1.3.3 Caractéristiques de forme

1.3.3.1 Définition des caractéristiques

En oncologie, les lésions sphériques avec des bords bien nets sont considérés comme moins agressives que des tumeurs diffuses et infiltrantes. De ce fait, des outils capables d'étudier la forme du volume de l'objet étudié présentent un intérêt potentiel.

Voici quelques exemples de caractéristiques de forme rencontrées dans la littérature.

— La sphéricité ("sphericity") de la lésion [Hofheinz et al. 2014] qui mesure le caractère circulaire défini selon l'Équation 1.7.

$$\text{sphericity} = \frac{1}{36\pi}\frac{S_{VOI}^3}{V_{VOI}^2} \tag{1.7}$$

Où S_{VOI} et V_{VOI} sont respectivement la surface et le volume du VOI étudié. Le facteur $\frac{1}{36\pi}$ s'assure que la sphéricité soit égale à 1 pour une sphère.

— La solidité ("solidity") est une caractéristique mesurant la convexité d'un objet. Elle est définie comme étant la proportion de voxels de l'objet contenue dans une enveloppe convexe ajustée au volume de l'objet défini selon l'Équation 1.8 [Chang et al. 2001].

$$\text{solidity} = A_{VOI}/H_{hull} \tag{1.8}$$

Où A_s est la zone du VOI présent à l'intérieur de l'enveloppe et H_{hull} est le volume de l'enveloppe.

— L'ampleur ("extent") est similaire à la caractéristique précédente en remplaçant l'enveloppe convexe par une boite englobante.

— Le nombre d'Euler caractérise la différence entre le nombre de voxels connexes d'une région et le nombre de trous symbolisés par les régions moins fixantes, comme l'illustre la Figure 1.9.

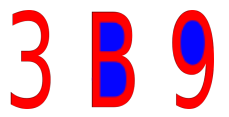


FIGURE 1.9 – Illustration de formes présentant différents nombres d'Euler, où les voxels connexes sont en rouge et les trous en bleu. Le "3" a un nombre d'Euler de 1, le "B" de -1 et le "9" de 0.

1.3.3.2 Apports cliniques

Dans l'article d'El Naqa et al. [El Naqa et al. 2009], les auteurs ont également étudié l'association de plusieurs caractéristiques de forme pour la prédiction de la réponse au traitement. Après une analyse combinant ces caractéristiques par régression logistique, les auteurs ont proposé un modèle prédictif applicable aux cancers ORL combinant le V_{90} (caractéristique du 1er ordre) à une caractéristique de forme : "l'extent". Le score généré par ce modèle a présenté un coefficient de corrélation de Spearman (ρ) de 0,87 avec la réponse au traitement et une *"Area Under ROC Curves"* (AUC) de 1.

Hofheinz et al. [Hofheinz et al. 2014] ont entre autres étudié l'intérêt pronostique de la sphéricité des lésions en ORL sur une cohorte de 37 patients. Les volumes tumoraux ont été segmentés automatiquement et la sphéricité extraite. L'analyse statistique univariée de Cox a montré que la sphéricité était prédictive de la PFS et de l'OS. La sphéricité calculée sur l'examen TEP initial permettrait l'amélioration de la prédiction de la progression tumorale.

1.3.4 Caractéristiques statistiques du 2ème ordre et supérieur

1.3.4.1 Définition des matrices d'analyse de texture

Les caractéristiques statistiques du 1er ordre présentées précédemment n'apportent aucune information concernant la disposition des voxels les uns par rapport aux autres. C'est pourquoi il peut être intéressant d'utiliser des méthodes d'ordre supérieur pour réaliser une analyse approfondie étudiant les relations respectivement entre les voxels, tels que les caractéristiques de texture. On parle de caractéristiques du 2ème ordre et supérieur, car contrairement à celles décrites précédemment, ces dernières étudient respectivement les relations entre les éléments voisins 2 à 2 ou plus.

La texture d'une image correspond à la représentation mathématique de l'aspect vi-

suelle d'une surface, ce qui permet d'exprimer la fréquence de répétition d'un motif ou de l'homogénéité des intensités comme illustré Figure 1.10.

FIGURE 1.10 – Illustration de différentes textures de l'image.

Il existe quatre principales matrices de texture proposées dans la littérature :

— Les matrices de cooccurrences (*"Gray Level Cooccurrence Matrix"* (GLCM)), introduites par Haralick [Haralick et al. 1973], contiennent l'ensemble des statistiques du 2ème ordre de l'image qui caractérisent les relations d'intensité entre les couples de voxels voisins.

— La matrice de différence des niveaux de gris (*"Gray Level Difference Matrix"* (GLDM)) a été proposée par Amadasun [Amadasun and King 1989]. Elle décrit les différences d'intensité entre voisins et contient des statistiques d'ordre supérieur à la matrice précédente.

— Les matrices de longueurs de plages homogènes (*"Gray Level Run Length Matrix"* (GLRLM)), introduites par Galloway [Galloway 1975], caractérisent la longueur des plages de même intensité dans une direction donnée.

— La matrice de longueur de zones homogènes (GLSZM) a été proposée par Thibault [Thibault et al. 2009]. Elle donne la longueur des zones ayant la même intensité dans toutes les directions simultanément.

Dans le Tableau 1.5 sont regroupées les caractéristiques que l'on peut extraire des différentes matrices. La construction des matrices et les expressions mathématiques des différentes caractéristiques sont données en Annexes (page 173).

1.3.4.2 Apports cliniques

Tixier et al. [Tixier et al. 2011] ont étudié le pouvoir prédictif de la réponse au traitement de 38 caractéristiques de texture en plus des caractéristiques du 1er ordre sur une cohorte de 41 patient atteints d'un cancer de l'œsophage traité par RCT (Tableau

TABLEAU 1.5 – Principales caractéristiques statistiques du 2ème ordre et supérieur

Matrices	**Caractéristiques du 2ème ordre et plus**
Matrices de cooccurrences (GLCM)	Variance, Énergie, Entropie, Corrélation, Dissimilarité, Contraste, Homogénéité, Moment différentiel inverse (IDM), "Cluster shade", "Cluster tendency"
Matrice des différences des niveaux de gris (GLDM)	"Coarseness", "Contrast", "Busyness", "Complexity", "Strength"
Matrices des longueurs de plages homogènes (GLRLM)	"Short Run Emphasis" SRE, "Long Run Emphasis" (LRE), "Low Gray level Run Emphasis" (LGRE), "High Gray-level Run Emphasis" (HGRE), "Short Run Low Gray-level Emphasis" (SRLGE), "Long Run Low Gray-level Emphasis" (LRLGE), "Short Run High Gray-level Emphasis" (SRHGE), "Long Run High Gray-level Emphasis" (LRHGE), "Run Percentage" (RPr), "Gray Level Non-Uniformity" (GLNUr), "Run Length Non-Uniformity" (RLNU)
Matrice des longueurs de zones homogènes (GLSZM)	"Short Zone Emphasis" (SZE), "Long Zone Emphasis" (LZE), "Low Gray level Zone Emphasis" (LGZE), "High Gray-level Zone Emphasis" (HGZE), "Short Zone Low Gray-level Emphasis (SZLGE), "Long Zone Low Gray-level Emphasis" (LZLGE), "Short Zone High Gray-level Emphasis" (SZHGE), "Long Zone HighGray-level Emphasis" (LZHGE), "Zone Percentage" (ZP), "Gray Level Non-Uniformity" (GLNUz), "Zone Length Non-Uniformity" (ZLNU)

1.2, page 20). Les caractéristiques ont été extraites des examens TEP au FDG, puis analysées par un test statistique de Kruskal–Wallis, ainsi que par des courbes ROC. Leurs résultats montrent que les caractéristiques de textures permettent de mieux séparer les patients que celles du 1er ordre (voir Figure 1.11). En effet, l'entropie locale issue de la matrice GLCM, la "coarseness" issue de la matrice GLDM, ainsi que ZLNU et GLNU de la matrice GLSZM présentent toutes les 4 un intérêt prédictif de la réponse au traitement ($p = 0,0006$, $p = 0,0002$, $p = 0,0002$ et $p = 0,0002$, respectivement).

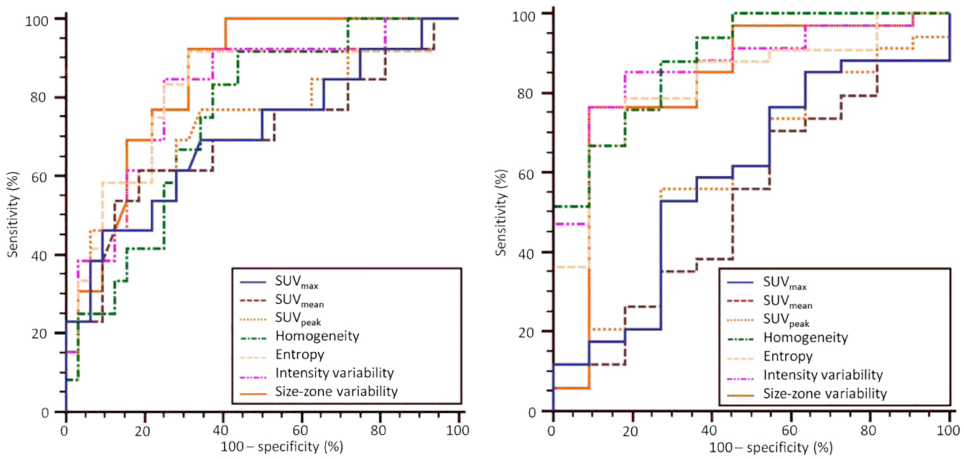


FIGURE 1.11 – Courbes ROC pour le SUV_{max}, le SUV_{moy}, le SUV_{peak}, l'homogénéité, l'entropie, GLNUz et ZLNU (a) des répondeurs complets et (b) des répondeurs partiels ou non-répondeurs d'après [Tixier et al. 2011].

L'intérêt prédictif de ces caractéristiques a été retrouvé dans 2 articles successifs de Tan et al. [Tan et al. 2013a] [Tan et al. 2013b]. Dans ces articles, les auteurs ont étudié une cohorte de 20 patients atteints d'un cancer de l'œsophage traités par RCT associée à une chirurgie. Les patients ont effectué un examen TDM/TEP au FDG avant (pre-RCT) et après traitement (post-RCT). Dans [Tan et al. 2013a], 192 caractéristiques du 1er ordre, de texture et de forme ont été étudiées par tumeur. Leurs résultats montrent notamment que 3 caractéristiques issues de la matrice GLCM extraites de l'imagerie TEP post-RCT sont prédictifs de la réponse au traitement : l'inertie (AUC = 0,850), la corrélation (AUC = 0,800) et le "cluster tendency" (AUC = 0,780). Ces caractéristiques présentent de meilleures performances que la variation du SUV_{max} entre les 2 examens TEP (AUC = 0,790) ou que la "skewness" pre-RCT (AUC = 0,760). Dans [Tan et al. 2013b], 41 caractéristiques du 1er ordre, de texture (GLCM) et des différences d'histogrammes ont été extraites des images

TDM/TEP au FDG. Leurs résultats montrent notamment que 5 caractéristiques de texture issues de la matrice GLCM extraites de l'imagerie TEP post-RCT présentaient un intérêt prédictif de la réponse au traitement : l'inertie, l'énergie, l'entropie, la corrélation ainsi que le moment différentiel inverse ($p = 0,01$, $p = 0,02$, $p = 0,04$, $p = 0,02$ et $p = 0,01$, respectivement). De même, les caractéristiques issues des différences d'histogrammes donnent des résultats intéressants.

Dans [Cook et al. 2013], les auteurs ont étudié le pouvoir prédictif de la réponse au traitement et pronostique des caractéristiques de texture issues de la matrice GLDM sur une cohorte de 53 patients présentant un cancer des poumons et traités par RCT. Leurs résultats montrent que certaines caractéristiques de texture apparaissent prédictives pour séparer les patients répondeurs, des non-répondeurs au traitement. L'analyse statistique de Kaplan-Meier révèle que les patients présentant une valeur élevée de la "coarseness" dans la tumeur primaire ont des temps plus faibles d'OS et de PFS ($p = 0,015$). De même, les temps de PFS étaient plus longs pour les patients présentant de fortes valeurs de contraste et de "busyness" ($p = 0,003$ et $p = 0,002$, respectivement).

1.4 Conclusion

Dans cette partie, il a été présenté l'intérêt de l'imagerie fonctionnelle TEP au FDG comme valeur pronostique et prédictive de la réponse au traitement en oncologie. Le développement des outils informatiques ces dernières années a permis l'extraction d'une importante quantité de caractéristiques décrivant les images TEP.

La 1ère caractéristique qui a permis d'établir ce type de lien est la valeur du SUV_{max} avant le traitement. Progressivement, de nombreuses autres caractéristiques ont été proposées dans la littérature basées sur les statistiques du 1er ordre, du 2ème ordre et supérieur, ainsi que sur les caractéristiques de forme. Certaines ont montré des résultats intéressants aussi bien en pronostique qu'en prédictif.

Dans la littérature, il a également été proposé de s'intéresser aux variations du SUV_{max} et du MTV en cours de traitement. Si l'on comprend tout l'intérêt d'une telle approche, elle présente 2 inconvénients. Le 1er est la réalisation d'un 2ème examen en cours de trai-

tement, là où l'on aimerait bien avoir une information en amont. Le 2ème est la manipulation de 2 fois plus de caractéristiques. De ce fait, dans notre travail, nous nous sommes limités à l'étude des caractéristiques obtenues sur l'imagerie TEP pré-thérapeutique.

L'objectif de cette thèse a donc été de proposer et d'évaluer des algorithmes de sélection des caractéristiques radiomiques ayant une valeur pronostique ou prédictive de la réponse au traitement en cancérologie basées sur une méthode d'apprentissage automatique, les forêts aléatoires. La prédiction précoce permettrait la mise en place d'une médecine personnalisée.

En plus de cette augmentation du nombre de caractéristiques issues des images médicales, des chercheurs ont récemment proposé de les associer à des données biologiques multimodales (clinique, protéomique, génomique, ...). Dans le chapitre suivant est développé ce concept, nommé la radiomique. Par ailleurs, des auteurs ont étudié l'influence de plusieurs paramètres sur la valeur des caractéristiques de l'image et leur possible retentissement sur la valeur pronostique et prédictive des caractéristiques. Nous en ferons également une revue de la littérature.

Chapitre 2

La radiomique en imagerie TEP oncologique et étude des caractéristiques

Sommaire

2.1 La radiomique

2.1.1 Le concept

L'imagerie médicale est une technique non-invasive utilisée en routine clinique dans la prise en charge des patients atteints d'un cancer. Elle tend à jouer un rôle de plus en plus important dans la personnalisation du traitement s'appuyant sur les caractéristiques individuelles du patient et de l'image de la tumeur.

La médecine personnalisée s'est largement développée avec la génomique et la protéomique, qui sont des techniques invasives. Cependant, il existe une hétérogénéité spatiale et temporelle des caractéristiques tumorales [Marusyk et al. 2012], rendant difficile les biopsies répétées, aussi bien dans l'espace que dans le temps, là où l'imagerie permet d'emblée de donner une représentation spatiale de la lésion et présente l'avantage d'être facile à répéter.

C'est dans ce contexte que la notion de la radiomique s'est développée. La radiomique est une analyse quantitative, consistant en l'extraction d'un grand nombre de données numériques de l'image médicale afin d'obtenir des informations prédictives et pronostiques concernant les patients pris en charge pour une pathologie cancéreuse [Bourgier et al. 2015]. Ces données d'imagerie médicale sont possiblement associées à des données dites "omiques" issues des techniques biologiques, apportant diverses informations concernant les gènes (génomique), les protéines (protéomique) ou autres sujets d'étude, reflet de la physiologie et de la physiopathologie du patient et de la tumeur.

Deux idées sont sous-jacentes au concept de radiomique. La première est que les caractéristiques tumorales cliniques, à l'échelle tissulaire, cellulaire et/ou génomique auraient un retentissement phénotypique en imagerie médicale, comme l'illustre la Figure 2.1 d'après la revue de la littérature de Yip et al. [Yip and Aerts 2016]. Cela revient à considérer que des caractéristiques de l'image sont fortement corrélées à des caractéristiques cliniques et biologiques. Le deuxième rationnel est que l'information portée par l'image serait complémentaire de celle provenant d'autres sources d'informations, comme l'illustre la Figure 2.2 d'après Lambin et al. [Lambin et al. 2012], permettant ainsi d'enrichir le

nombre des caractéristiques tumorales.

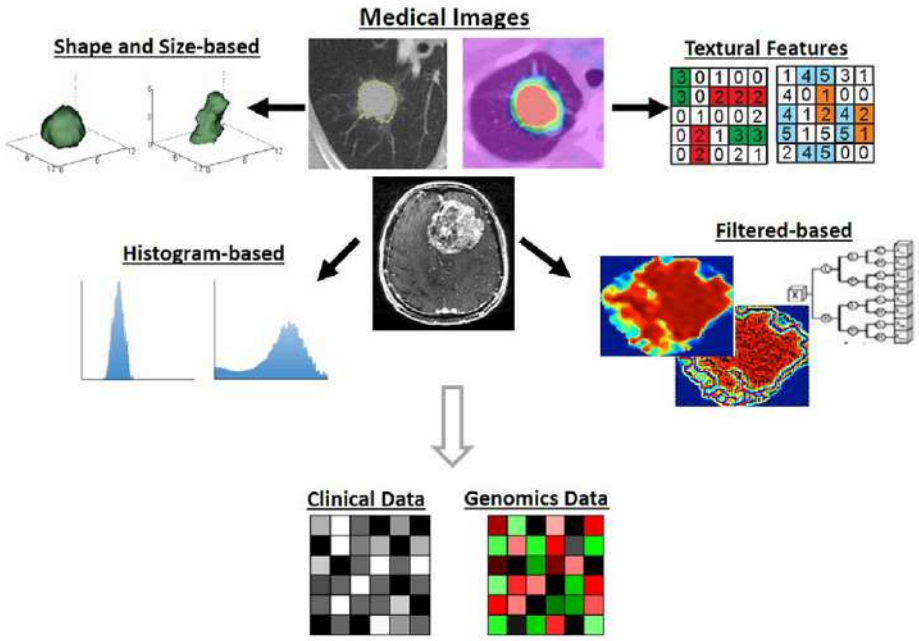


FIGURE 2.1 – Principe de la radiomique et de l'extraction de caractéristiques de l'image tumorale d'après [Yip and Aerts 2016].

L'interprétation médicale des images est généralement basée sur une simple interprétation visuelle du contraste. Même si ce type d'interprétation a montré toute son efficacité dans la prise en charge des patients en cancérologie, elle reste qualitative et subjective, là où en radiomique il devient nécessaire de quantifier les caractéristiques pour des raisons d'objectivité, de reproductibilité, du nombre toujours croissant des caractéristiques étudiées et de la complexité de la modélisation de certains index.

2.1.2 Apport en oncologie

La revue de la littérature de Yip et al. [Yip and Aerts 2016] montre que de nombreux travaux étayent l'intérêt clinique de la radiomique en cancérologie, aussi bien sur des images IRM, TDM que TEP au FDG. Ces données concernent les tumeurs solides, et ce pour de nombreuses localisations (pulmonaires, ORL, œsophage, sein, col de l'utérus, etc ...), ainsi que les tumeurs hématologiques. Elles concernent potentiellement tous les types de traitements (chimiothérapie, radiothérapie, traitements concomitants).

Plusieurs caractéristiques radiomiques ont montré leur utilité dans la stadification de la maladie par leur capacité à différentier de manière significatives les stades précoces et

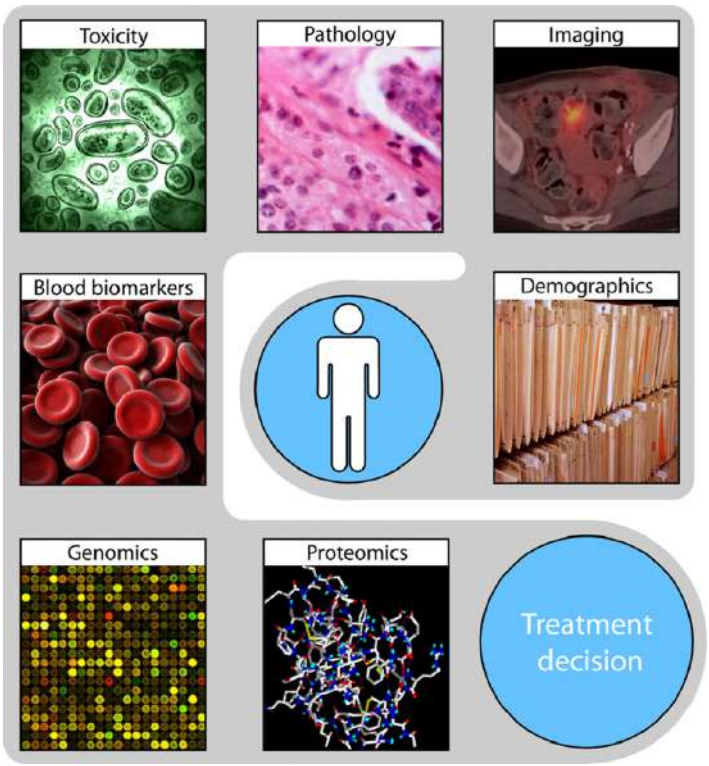


FIGURE 2.2 – Exemple d'informations complémentaires, dont les informations radiomiques de l'image contribuant à un traitement personnalisé, d'après [Lambin et al. 2012]

avancés de la maladie. Cette stadification précoce peut aider à mieux identifier les patients, pour ensuite sélectionner le meilleur traitement. Dans [Dong et al. 2013], des tumeurs de l'œsophage de 40 patients ont été évaluées par la classification de l'"American Joint Committee on Cancer" (AJCC) ("American Joint Committee on Cancer"). Des caractéristiques TEP comme le SUV_{max}, l'entropie et l'énergie issues de la matrice de cooccurrence se sont révélées être significativement corrélées avec les stades T et N ("Tumor, Nodes, Metastasis Stage" (Stade TNM), voir Tableau A.1 en Annexes, page 171). L'entropie, par exemple, avec un seuil de 4,7 a permis de séparer les tumeurs ayant un stade inférieur ou supérieur à IIb. Dans une étude plus récente, Mu et al. [Mu et al. 2015] ont classé en 2 groupes (stade précoce (stade I et II), stade avancé (stade III et IV)) 42 patientes atteints de cancer du col de l'utérus à l'aide de l'imagerie TEP. La caractéristique RPr issue de la matrice de longueurs de plages homogènes a été jugée la plus pertinente dans la prédiction du stade de la maladie.

Les caractéristiques radiomiques ont également montré leur utilité dans la discrimination tissulaire entre tissus sains et tissus pathologiques. Petkovska et al. [Petkovska et al. 2006] ont montré que l'utilisation des caractéristiques de la matrice de cooccurrence is-

sues du TDM donne de meilleurs résultats dans l'identifier les nodules malins et bénins que l'inspection visuelle réalisée par 3 radiologues experts. Dans [Xu et al. 2014] est proposé une méthode de diagnostic assisté par ordinateur combinant des caractéristiques de texture TDM et TEP au FDG pour différencier les lésions malignes et bénignes dans divers sites tumoraux. La fixation du FDG au sein des lésions malignes a été observées comme étant plus hétérogène que celle des tissus bénins. L'utilisation de ces caractéristiques a mené à une précision supérieure à 75 %, en comparaison avec le diagnostic histologique des lésions.

Plusieurs études ont mis en évidence une forte relation entre les caractéristiques des images et la génétique sous-jacente de la tumeur. Segal et al. [Segal et al. 2007] ont étudié l'intérêt des caractéristiques radiomiques sur une base de données de 28 patients présentant des carcinomes hépatocellulaires. Une analyse qualitative des examens TDM a permis la sélection de 138 caractéristiques. Parallèlement, une analyse de l'expression génique de la tumeur a permis d'identifier des groupes de gènes, nommés "modules biologiques" correspondant à une fonction biologique définie. Les corrélations entre ces 2 types de caractéristiques ont été étudiées. Les auteurs ont constaté que certaines des caractéristiques étaient représentatives de l'expression génique de la tumeur et avaient une valeur pronostique. Un travail similaire a été effectué chez des patients atteints de glioblastome par Diehn et al. [Diehn et al. 2008] à partir d'images IRM. Les auteurs ont mis en évidence que le module biologique de l'hypoxie, comprenant 5 gènes, était corrélé avec la caractéristique d'augmentation de la prise de contraste. Comme ces gènes sont impliqués dans le processus de l'hypoxie, ce trait pourrait permettre la sélection des patients candidats à un traitement par médicaments anti-angiogéniques.

Enfin, les caractéristiques radiomiques permettent la prédiction de la réponse au traitement et la survie du patient. Cook et al. [Cook et al. 2013] ont comparé le pouvoir prédictif du SUV_{max} ainsi que de 4 caractéristiques de texture de la matrice de différence des niveaux de gris sur une base de donnée de 53 patients atteints d'un cancer du poumon. Ils ont constaté que la "coarseness", la "busyness" et le contraste issus de cette matrice pouvaient mieux différencier les patients répondeurs de ceux non-répondeurs au traitement par CRT que les dérivés du SUV. De plus, la "coarseness" a été trouvée comme étant un

facteur pronostique de la survie globale des patients. De même, Mi et al. [Mi et al. 2015] ont étudié 79 caractéristiques cliniques et de l'imagerie TEP de 25 patients atteints d'un cancer du poumon et traités par RCT. Leurs résultats ont montré que la combinaison du SUV_{max} avec deux autres caractéristiques issues de la matrice GLSZM était l'association de caractéristiques apportant les meilleurs résultats prédictifs. Par ailleurs, de nombreux autres exemples de la littérature ont déjà été cités dans le Chapitre 1 (voir Tableau 1.1, 1.2 et 1.3, page 18).

2.2 Étude des caractéristiques

2.2.1 La problématique

La multiplication du nombre de caractéristiques, dont certaines correspondent à une expression mathématique relativement élaborée, présente potentiellement un réel intérêt dans le cadre de la médecine personnalisée comme l'atteste l'étude précédente de la littérature. Cependant, cela est au détriment de la compréhension de la signification de ce que représentent visuellement ces caractéristiques. Ainsi, autant un médecin appréhende très bien la signification d'un volume tumoral de petite ou de grande taille, autant il lui est complètement étranger la signification d'une valeur élevée d'une caractéristique tel que le "busyness" de la matrice GLDM.

Ceci peut constituer un frein au développement de la radiomique car les médecins souhaitent de manière légitime, avoir une compréhension médicale de ce qu'ils observent ou mesurent. Dans ce contexte, il est important d'avoir une meilleure compréhension de la signification des caractéristiques de texture, ce d'autant qu'il existe des différences de définition de ces caractéristiques selon les auteurs et les publications.

Par ailleurs, les variations des images TEP induites par les différentes méthodes d'acquisition ou de reconstruction peuvent avoir un impact sur les caractéristiques des images [Boellaard et al. 2008]. Il est important de connaitre la sensibilité de chaque caractéristique et de contrôler si ces variations altèrent les résultats des études prédictives et pronostiques.

Dans cette section nous aborderons les différents paramètres pouvant modifier la valeur des caractéristiques de texture. Nous commencerons par aborder la problématique des différents protocoles d'acquisition des images TEP, des différences de définitions des caractéristiques, ainsi que des différentes méthodes de ré-échantillonnage des niveaux de gris. Puis, nous parlerons des corrélations qui peuvent exister entre les caractéristiques, de l'influence que peuvent avoir les paramètres de reconstruction de l'image sur elles ainsi que des différents problèmes liés au volume tumoral et à sa segmentation. Enfin, nous terminerons en évoquant les problèmes engendrés par l'utilisation des études statistiques classiques.

2.2.2 Influence de l'acquisition des images

Dans la revue de la littérature de Boellaard et al. [Boellaard et al. 2004], les auteurs ont recensé de nombreux facteurs concernant la préparation des patients et l'acquisition des examens pouvant affecter les mesures du SUV, alors qu'il existe des variations de protocole entre les différents centres d'imagerie [Boellaard et al. 2008].

Il est essentiel que la distribution du SUV dans la lésion soit reproductible. Plusieurs facteurs peuvent influencer cela. On peut citer la mesure de l'activité initiale, le temps d'attente entre l'injection et l'acquisition. D'après Boellaard et al. [Boellaard et al. 2008], une erreur de 8 minutes induit une erreur de mesure sur le SUV de 5 %. Pour limiter ces variations, il est recommandé de prendre un temps entre injection et acquisition de 60 ± 5-10 minutes.

Le modèle du tomographe ainsi que les protocoles d'acquisition et de reconstruction peuvent être à l'origine de réponses différentes dans la mesure des SUV ainsi que dans la résolution spatiale. Si l'effet de chacun d'eux sur la valeur du SUV obtenue est faible en moyenne (< 15 %), leur accumulation peut entraîner une grande variabilité des mesures [Boellaard 2009]. Dans cette revue de la littérature, Boellaard et al. [Boellaard 2009] montrent que l'utilisation de protocole d'acquisition commun rend la mesure du SUV très reproductible. De même, en se basant sur les résultats d'une série d'études, Weber et al. [Weber 2010] ont ainsi relevé que l'utilisation d'un même protocole permettait de réduire l'écart-type des mesures répétées du SUV à environ 10 %.

En médecine personnalisée, là où il convient de déterminer des caractéristiques radiomiques à valeur prédictive, se pose le problème de la robustesse de ces caractéristiques vis à vis de l'acquisition et de la reconstruction sur des machines différentes avec des protocoles de reconstruction différents.

Plusieurs études ont étudié la reproductibilité des caractéristiques d'images face aux variations de protocole d'acquisition. D'après De Langen et al. [de Langen et al. 2012], le SUV_{moy} est une caractéristique extrêmement robuste, meilleure que celle du SUV_{max}. De Langen et al. ont montré également que la répétabilité était meilleure pour des volumes avec une forte fixation du FDG. van Velden et al. [van Velden et al. 2016] ont étudié la répétabilité de 105 caractéristiques radiomiques sur une base de données de 11 patients atteints d'un cancer du poumon. Ces patients ont effectués 2 examens TEP à 3 jours d'intervalle. Leurs résultats montrent que 98 % des caractéristiques possèdent une répétabilité similaire à celles du SUV_{moy}. Ces caractéristiques ne sont donc pas extrêmement sensibles aux différents paramètres d'acquisition.

2.2.3 Différence de définition des caractéristiques

Il existe des différences de définition des caractéristiques radiomiques [Orlhac et al. 2014] dues à une absence de consensus. Il arrive que 2 caractéristiques de texture avec des définitions différentes soient désignées sous le même nom. C'est le cas par exemple pour l'homogénéité extraite de la matrice GLCM. Cette dernière est utilisée dans l'article de Tixier et al. [Tixier et al. 2011], ainsi que dans celui de Willaime et al. [Willaime et al. 2013]. Cependant, la formule utilisée pour cette caractéristique diffère entre ces 2 articles. Il en est de même pour la "size-zone variability" qui est définie différemment dans 2 articles de Tixier et al. [Tixier et al. 2011] et [Tixier et al. 2012].

Les méthodes de calcul peuvent également être différentes. Certaines caractéristiques de texture peuvent être extraites d'une seule coupe transversale de la tumeur (texture 2D) ou du volume tumoral entier (texture 3D) [Ng et al. 2013], [Fave et al. 2015]. De même, certains auteurs ont considéré individuellement les différentes directions des matrices GLCM, alors que d'autres ont préféré extraire les caractéristiques de la matrice moyenne des 13 directions [Hatt et al. 2015].

L'ensemble de ces constatations montre qu'il serait souhaitable d'harmoniser les définitions [Buvat et al. 2015] et de bien définir les modes de calcul des caractéristiques. Ainsi, nos matrices de texture sont calculées en 3D. Lorsqu'une matrice est fonction de la direction et de la distance, comme c'est le cas pour la GLCM, les caractéristiques sont calculées sur la matrice moyenne des matrices issues des 13 directions possibles en 3D avec une distance de 1. Enfin, les expressions mathématiques des caractéristiques étudiées (voir Annexes, page 173) sont identiques à celles utilisées par Orlhac et al. [Orlhac et al. 2015].

2.2.4 Influence du ré-échantillonnage de l'image

Afin de calculer les différentes matrices de texture, il est nécessaire de ré-échantillonner les images TEP. Le but du ré-échantillonnage est de regrouper le très grand nombre de valeurs du SUV en un nombre limité d'intensité. Ainsi, un examen TEP qui pouvait présenter initialement plusieurs centaines de différents niveaux de gris, se retrouve après ré-échantillonnage à quelques dizaines d'intensité en fonction de la méthode utilisée. Cette étape permet de réduire le temps de calcul des matrices de textures. Elle permet également la réduction du bruit dans l'image et donc dans le calcul des caractéristiques [Tixier et al. 2012]. Par contre, cela se fait en contrepartie d'une perte d'information.

Il existe 2 approches pour discrétiser les images : la première est une approche par ré-échantillonnage relatif (R_r) et la deuxième par ré-échantillonnage absolu (R_a).

Une première méthode de ré-échantillonnage relatif a été proposée par Tixier et al. [Tixier et al. 2011]. Elle consiste à fixer au préalable le même nombre D d'intensité pour chaque examen. Cette approche est relative car chaque examen est ré-échantillonné avec D niveaux de gris. Comme la gamme de SUV utilisés est différente d'un patient à l'autre le pas de ré-échantillonnage est différent. Les nouveaux niveaux de gris $R_{r1}(i)$ sont réaffectés pour chaque valeur de SUV(i) de la manière suivante (Équation 2.1) [Tixier et al. 2011] :

$$R_{r1}(i) = D \times \frac{\mathrm{SUV}(i) - \mathrm{SUV}_{\min}}{\mathrm{SUV}_{\max} - \mathrm{SUV}_{\min} + 1} \tag{2.1}$$

Une autre formule de ré-échantillonnage relatif R_{r2}, légèrement différente, a été proposée par Orlhac et al. [Orlhac et al. 2014] (Équation 2.2) :

$$R_{r2}(i) = D \times \frac{\mathrm{SUV}(i) - \mathrm{SUV}_{\mathrm{min}}}{\mathrm{SUV}_{\mathrm{max}} - \mathrm{SUV}_{\mathrm{min}}} \tag{2.2}$$

L'impact du paramètre D sur les caractéristiques de texture a été étudié. Orlhac et al. [Orlhac et al. 2014] ont montré que les caractéristiques de texture ré-échantillonnées par R_{r2} sont donc peu sensibles au nombre de niveaux de gris utilisés lorsque celui-ci est supérieur ou égal à 32. Hatt et al. [Hatt et al. 2015] ont conclu que ré-échantillonner avec R_{r1} et plus de 64 niveaux de gris n'apportait pas d'informations complémentaires. C'est pourquoi ils conseillent de limiter le nombre de niveaux de gris à 64.

Le deuxième type de méthode de ré-échantillonnage est l'approche absolue. Plutôt que de fixer le nombre d'intensité D dans l'image ré-échantillonnée, il est préféré de fixer le pas B de ré-échantillonnage, aussi appelée résolution d'intensité [Leijenaar et al. 2015]. Cette méthode est considérée comme étant absolu car le pas est identique pour l'ensemble des examens. Ce type de ré-échantillonnage est réalisé de la manière suivante d'après Leijenar et al. [Leijenaar et al. 2015] R_{a1} (Équation 2.3) :

$$R_{a1}(i) = \left\lfloor \frac{\mathrm{SUV}(i)}{B} \right\rfloor - min\left(\left\lfloor \frac{\mathrm{SUV}(i)}{B} \right\rfloor \right) + 1 \tag{2.3}$$

où $[min(\mathrm{SUV}(i)/B) + 1]$, assure que les intensités du VOI commencent à 1. B est fixé par l'utilisateur en fonction du $\mathrm{SUV}_{\mathrm{min}}$ global ($\mathrm{SUV}_{\mathrm{minG}}$), du $\mathrm{SUV}_{\mathrm{max}}$ global ($\mathrm{SUV}_{\mathrm{maxG}}$) dans l'ensemble des VOI de la cohorte étudiée, ainsi que du nombre maximal de niveaux de gris voulu dans l'étude. Il peut donc être défini comme ($\mathrm{SUV}_{\mathrm{maxG}}$ - $\mathrm{SUV}_{\mathrm{minG}}$)/$D$.

Orlhac et al. [Orlhac et al. 2015] utilise une expression, légèrement différente, en fixant au préalable le $\mathrm{SUV}_{\mathrm{minG}}$ égale à 0. L'équation de ré-échantillonnage devient alors R_{a2} (Équation 2.4) :

$$R_{a2}(i) = \frac{\mathrm{SUV}(i)}{B} \tag{2.4}$$

Leijenaar et al. [Leijenaar et al. 2015] ont comparé les stratégies de ré-échantillonnage R_{r2} et R_{a1}. Les auteurs ont montré que l'utilisation d'un ré-échantillonnage des images

TEP avec un pas d'intensité B identique sur l'ensemble d'une cohorte de patients est plus appropriée. Ils précisent que l'utilisation de cette méthode limite les biais inter- et intra-patients. Ces conclusions sont en accords avec les résultats obtenus par Orlhac et al. [Orlhac et al. 2015] sur base de données de 45 patients présentant un cancer pulmonaire et une autre issue d'une étude de fantômes. Des images TEP ont été extraites 7 caractéristiques de texture obtenues à partir des images ré-échantillonnées par l'approche R_{a2} et R_{r2} : l'homogénéité et l'entropie de la matrice GLCM, SRE, LRE et RLNU de la matrice GLRLM et LGZE et HGZE de la matrice GLSZM. Les coefficients de corrélation de rang de Spearman ont été calculés entre les caractéristiques de texture, le MTV et le SUV_{max}. Les auteurs ont tout d'abord montré qu'il existait une importante corrélation entre les caractéristiques de texture et le MTV lorsqu'un ré-échantillonnage relatif R_{r2} était utilisé. Cette corrélation n'existe pas avec R_{a2}, mais il apparait alors une corrélation entre les caractéristiques de texture et le SUV_{max} qui n'existait pas avant. Cette corrélation n'est pas contre-intuitive car la gamme de SUV limitée par le SUV_{max} est la base de l'hétérogénéité. Ainsi, cette gamme de SUV disparait avec l'utilisation du ré-échantillonnage relatif. Les auteurs concluent que les caractéristiques de texture ré-échantillonnées par une méthode absolue sont plus pertinentes pour rendre compte des informations concernant la distribution d'amplitude et spatiale des SUV dans les images TEP. Enfin, l'utilisation d'un ré-échantillonnage absolu permet de générer des caractéristiques possédant un meilleur pouvoir discriminant dans la séparation des tumeurs et des tissus sains [Orlhac et al. 2015]. Par exemple, l'entropie basée sur un ré-échantillonnage relatif R_{r2} n'a pas permis la distinction entre la tumeur et le tissu sain ($p = 0,762$), alors que le même indice basé sur un ré-échantillonnage absolu R_{a2} a permis de différencier ces deux types de tissus ($p < 0,0001$). R_{a2} permet également une amélioration de la séparation des patients en fonction de leur type histologique (ADC et SCC).

Par ailleurs, la valeur de B doit être un compromis entre le nombre de niveaux de gris voulu et le SUV_{maxG} de la cohorte de patient étudié. Orlhac et al. [Orlhac et al. 2015] ont étudié plusieurs valeurs de SUV_{maxG} (5, 20 et 25) menant à des valeurs de B allant de 0,23 à 0,39 pour une gamme de niveaux de gris de 64 avec leur cohorte de patients. Les auteurs ont étudié la capacité de plusieurs caractéristiques de texture à séparer les différentes his-

tologies de 45 patients (ADC et SCC). Leurs conclusions restent inchangées malgré les variations de B, ce qui montre que le choix du pas de ré-échantillonnage n'influe pas fortement le pouvoir discriminant des caractéristiques. Dans [Leijenaar et al. 2015], la cohorte étudiée présentait un SUV_{maxG} autour de 30, ce qui a mené à étudier des caractéristiques ré-échantillonnées avec B égal à 0,5 afin de respecter une gamme de niveaux de gris de 64 dans notre cohorte.

En conclusion, il semble préférable d'utiliser un ré-échantillonnage absolu qui limite la corrélation entre les caractéristiques de texture et le volume [Orlhac et al. 2015]. De plus, cette méthode est conseillée par les auteurs dans le cadre de l'étude d'une ou plusieurs cohortes de patient [Leijenaar et al. 2015], [Orlhac et al. 2015]. Comme nous utilisons déjà les mêmes définitions des caractéristiques proposées par Orlhac et al. nous nous sommes également tournés vers leur formule de ré-échantillonnage R_{a2} (voir Équation 2.4) [Orlhac et al. 2015]. Le pas de ré-échantillonnage B utilisé dépendra du SUV_{maxG} de notre cohorte de patient.

2.2.5 Corrélations entre les caractéristiques

Le nombre de caractéristiques extraites des images TEP n'est pas nécessairement synonyme d'augmentation de la quantité d'informations apportées. Tixier et al. [Tixier et al. 2011] ont étudié les corrélations de 38 caractéristiques de texture sur une base de données de 41 patients atteints d'un cancer de l'œsophage. Pour cela, les auteurs se sont servis du coefficient de corrélation de Pearson (r). Ils ont montré que les caractéristiques extraites de la matrice GLRLM et celles de la matrice GLSZM sont fortement corrélées ($|r| \geq 0,9$), ce qui n'est pas surprenant, compte tenu du fait que l'une représente les longueurs des plages et l'autre des zones homogènes. Dans le contexte de l'image 3D, la matrice GLRLM ne présente pas d'intérêt particulier et n'a pas été retenue dans nos travaux.

Orlhac et al. [Orlhac et al. 2014] ont étudié les corrélations entre les caractéristiques en s'intéressant à 3 bases de données (72 lésions issues d'un cancer colo-rectal, 24 d'un cancer des poumons et 54 d'un cancer du sein). A partir des images TEP, 41 caractéristiques dont 31 de texture ont été extraites et leur coefficient de corrélation de Pearson a été étudié 2 à 2. Une corrélation a été considérée comme forte entre 2 caractéristiques si

$|r|$ était supérieur à 0,80 et p inférieur à 5 %. Les auteurs ont d'abord montré que certaines caractéristiques de texture étaient fortement corrélées avec le MTV dans les 3 types de tumeurs solides étudiées et ils ont associé les caractéristiques corrélées au sein de groupe de corrélation respectant des critères ($|r| \geq 0,8$ et $p \leq 5$ %).

Enfin, comme abordé dans la partie précédente, le type de ré-échantillonnage utilisé influe sur les corrélations obtenues [Orlhac et al. 2015] [Orlhac et al. 2017]. Ainsi, l'utilisation d'une approche relative a plutôt tendance à corréler les caractéristiques de texture avec le MTV (Équation 2.1 et 2.2) alors qu'une approche absolue révèle plutôt des corrélations avec le SUV_{max} (Équation 2.3 et 2.4).

2.2.6 Influence de la reconstruction de l'image

Plusieurs étapes du processus de reconstruction des images TEP modifient les valeurs des caractéristiques : l'algorithme de reconstruction tomographique et ses paramètres (nombre d'itérations et de sous-ensembles), la taille de la matrice de reconstruction, ainsi que la largeur à mi-hauteur du filtre gaussien en post-reconstruction.

Galavis et al. [Galavis et al. 2010] ont étudié les variations de 50 caractéristiques radiomiques extraites des images TEP face à 2 algorithmes de reconstruction *"Ordered Subset Expectation Maximisation"* (OSEM) et *"Iterative-Vue Point"* (ITER). A partir d'une cohorte de 20 patients atteints de différents cancers solides (carcinome du poumon, de l'épiglotte et de l'œsophage), les auteurs ont extrait 8 caractéristiques du 1er ordre, 23 de la matrice GLCM, 11 de la matrice GLRLM, 5 de la matrice de niveau de gris voisin *"Neighboring Gray Level Dependence Matrix"* (NGLDM) et 3 de la matrice GLDM. Les auteurs ont testé différents paramètres : le nombre d'itérations (2 et 4), la taille de la matrice de reconstruction (128×128 et 256×256) et la largeur à mi-hauteur du filtre gaussien post-reconstruction (3 et 5 mm). Au final, à partir d'une même acquisition, 10 images ont été reconstruites avec ces différents paramètres. Pour évaluer les variations, chaque caractéristique a été extraite puis moyennée sur ces 10 reconstructions. L'écart relatif par rapport à cette moyenne a été calculé pour chaque caractéristique de chaque image reconstruite (en %). En prenant en compte l'ensemble des paramètres testés, les résultats montrent que seulement 4 caractéristiques sur 50 présentent une faible variation (< 5 %) : l'entropie et l'énergie du 1er

ordre, le coefficient de corrélation maximal (matrice GLCM) et LGRE (matrice GLRLM). Six caractéristiques présentent une variation modérée (< 30 %) : l'entropie et la somme des entropies (matrice GLCM), HGRE et GLNUr (matrice GLRLM), le *"Small Number Emphasis"* (SNE) (matrice NGLDM) et l'entropie-GL (NGLDM). Enfin, les dernières 40 caractéristiques présentent d'importantes variations.

Plus récemment, une étude similaire a été réalisé par Yan et al. sur une cohorte de 20 patients atteints d'un cancer pulmonaire [Yan et al. 2015]. La robustesse face aux différents paramètres de l'algorithme OSEM a été étudié pour 64 caractéristiques de l'image TEP (3 dérivés du SUV, 6 caractéristiques du 1er ordre et 55 caractéristiques de texture). Les images ont été au préalable ré-échantillonnées selon la formule R_{r2} [Orlhac et al. 2014]. L'influence des paramètres suivants a été étudiée : le nombre d'itérations (1, 2 et 3), la taille de la matrice de reconstruction (128 × 128 et 256 × 256) et la largeur du filtre gaussien post-reconstruction (2,5, 3,5, 4,5 et 5,5 mm). Les éventuelles variations engendrées par l'utilisation des méthodes de correction, le *"Time Of Flight"* (TOF) et la *"Point-Spread Function"* (PSF), ont été étudiées. Ces variations ont été calculées en mesurant le COV, c'est-à-dire le ratio entre l'écart-type et la moyenne (en %), pour chaque caractéristique. Une bonne robustesse est défini par un COV inférieur à 5 %. Concernant les 64 caractéristiques étudiées, 7 seulement présentent une robustesse supérieure à celle du SUV_{max}, considérée comme celle de référence : l'entropie (1er ordre), la différence d'entropie, la différence inverse et le moment différentiel inverse (matrice GLCM), le LGRE et le HGRE (matrice GLRLM) et le LGZE (matrice GLSZM). D'autres caractéristiques ont présenté une robustesse plus faible que celle du SUV_{max} mais supérieure à celle du SUV_{moy} et donc considérée comme acceptable. On retrouve dans ce cas l'énergie (1er ordre), l'entropie et la somme des entropies (matrice GLCM), le HGZE (GLSZM) et le SNE (NGLDM). Les résultats montrent que les paramètres engendrant le plus de variations sont, dans l'ordre d'importance, la taille de la matrice de reconstruction avec 36/64 caractéristiques présentent de fortes variations, suivi de la largeur du filtre gaussien (8/64), puis du nombre d'itérations (3/64). Ni le TOF, ni la PSF n'ont augmenté la variation des caractéristiques étudiées.

Doumou et al. [Doumou et al. 2015] ont constaté que les caractéristiques radiomiques

étaient insensibles aux variations induites par la largeur à mi-hauteur du filtre gaussien en post-reconstruction.

A noter que ces données de la littérature portent sur des méthodes de ré-échantillonnage relatif et que l'on a peu d'information sur les méthodes de ré-échantillonnage absolu. Par ailleurs, nous avons sélectionné une cohorte de patients présentant des images TEP obtenues avec les mêmes paramètres de reconstruction afin d'éviter ce type de problème.

2.2.7 Influence de la méthode de segmentation du MTV

La délinéation du VOI est une étape cruciale afin d'extraire les caractéristiques radiomiques de l'image TEP. Il existe de nombreuses méthodes de segmentation des examens TEP : des méthodes manuelles, semi-automatiques ou automatiques. Il n'existe pas actuellement de consensus sur la meilleure méthode de segmentation [Zaidi and El Naqa 2010]. De ce fait, plusieurs auteurs ont étudié l'influence de la définition du MTV sur la reproductibilité des caractéristiques radiomiques.

Leijenaar et al. [Leijenaar et al. 2013] ont étudié la reproductibilité de 100 caractéristiques radiomiques chez 11 patients atteints d'un cancer du poumon. Les auteurs ont constaté que la majorité des caractéristiques des images TEP étaient peu sensibles à la variation inter-utilisateur et intra-utilisateur de la définition du MTV.

Doumou et al. [Doumou et al. 2015] ont étudié l'impact de l'utilisation de différentes méthodes de segmentation automatique de la tumeur par seuillage simple (de 45 à 60 % du SUV_{max}) sur les caractéristiques TEP au FDG. Les auteurs ont conclu que pour ce type de segmentation, l'influence sur les caractéristiques radiomiques de la variation du seuil était faible.

Orlhac et al. [Orlhac et al. 2014] ont montré, dans un 1^{er} temps, que 2 méthodes de segmentation différentes (seuillage fixe à 40 % du SUV_{max} et segmentation adaptative [Nestle et al. 2005]) donnent des volumes tumoraux significativement différents. Dans un $2^{ème}$ temps, la robustesse de 41 caractéristiques de l'image TEP extraites de 3 cohortes de patients présentant 3 cancers différents (colorectal, poumons et seins) a été étudiée : 10 caractéristiques du 1^{er} ordre et 31 de texture issues des matrices GLCM, GLRLM, GLDM et GLSZM. Comme pour [Yan et al. 2015], la robustesse du SUV_{max} est proposée comme

étant celle de référence et celle du SUV_{moy} comme acceptable. Seulement, 5 caractéristiques de texture obtiennent une meilleure robustesse que le SUV_{moy}. Il s'agit de l'entropie (matrice GLCM), le SRE, le LRE (matrice GLRLM) et le RPr (matrice GLRLM) et le SZE (matrice GLSZM). Puis, 12 caractéristiques présentent une robustesse comprise entre celle du SUV_{moy} et celle du MTV : l'homogénéité (matrice GLCM), le HGRE (matrice GLSZM) et le ZP (matrice GLSZM). Enfin, 19 caractéristiques ont montré d'importantes variations. Orlhac et al. [Orlhac et al. 2015] ont montré que cette sensibilité à la définition du volume tumoral est indépendante au type de tumeur aux vues des résultats identiques entre les 3 types de cancers étudiés.

Concernant notre travail, afin de limiter les variations inter- et intra-utilisateur nous avons préféré nous baser sur l'utilisation d'une méthode de segmentation développée au sein de notre laboratoire basée sur une méthode de seuillage adaptatif [Vauclin et al. 2009]. L'utilisation de la même méthode objective sur l'ensemble de la cohorte permet la réduction des biais de segmentation.

2.2.8 Influence de la taille du MTV

Brooks et al. [Brooks and Grigsby 2014] ont étudié la variation d'une caractéristique de texture, l'entropie (matrice GLCM), face aux variations du nombre de voxels dans le VOI des images TEP. Pour cela, une base de données de 70 patients atteints d'un cancer du col de l'utérus a été étudiée. Les résultats montrent que l'entropie est fortement corrélée avec le MTV si ce dernier est inférieur à 700 voxels (45 cm^3 pour leurs images). Ainsi, en dessous de ce seuil, les auteurs ont montré que cette caractéristique était 5 fois plus sensible aux changements de volume et donc reflétait plus le volume tumoral que l'hétérogénéité de fixation. D'après les auteurs, il n'est donc pas pertinent d'étudier cette caractéristique pour une tumeur possédant un volume inférieur à 700 voxels, ce qui en oncologie correspond déjà a un volume conséquent et élimine de fait un nombre important de patients. Enfin, il est important de noter que le volume exprimé en voxels est plus pertinent pour illustrer l'influence de la taille du MTV sur les caractéristiques de texture.

Cette analyse a été critiquée par Hatt et al. [Hatt et al. 2015] dû au fait qu'elle n'a considéré qu'une seule caractéristique calculée sur une GLCM 2D moyennée sur l'en-

semble des directions et 150 niveaux de gris. Ces auteurs ont alors étudié l'influence du volume sur 4 caractéristiques de texture en se basant sur une base de données de 555 patients présentant différents cancers (poumons, seins, ...). Les coefficients de corrélation de Spearman entre le MTV et les caractéristiques de texture ont été calculés. En augmentant progressivement le seuil limite du MTV de 3 à 60 cm^3, les auteurs ont trouvé que la corrélation entre les caractéristiques et le MTV avait tendance à diminuer. En effet, la dissimilarité est passée de $|\rho| = 0,80$ pour un seuil de 3 cm^3 à 0,21 pour 60 cm^3. L'évolution est identique pour l'ensemble des caractéristiques de texture étudiées, excepté pour le HGZE qui conserve une faible corrélation tout au long de l'étude ($|\rho| = 0,10$). La plus importante diminution du coefficient de corrélation de Spearman se fait entre un seuil de 3 et de 10 cm^3. Les auteurs recommandent de signaler la corrélation des caractéristiques avec le MTV pour chaque étude afin d'indiquer les informations indépendantes ou redondantes.

Nous comprenons tout à fait la problématique du calcul des caractéristiques de texture dans les lésions de faible volume énoncé par Brooks et al. [Brooks and Grigsby 2014]. Cependant, il est complexe dans notre problématique d'obtenir une base de donnée contenant un nombre de patients suffisant ayant une lésion de volume supérieur à 700 voxels ($\sim$ 45 cm^3). C'est pourquoi nous avons suivi les recommandations de Hatt et al. [Hatt et al. 2015] spécifiant de conserver les observations de faible volume tout en contrôlant les corrélations entre caractéristiques de texture et le MTV.

2.2.9 Analyses statistiques

Comme présenté précédemment, le nombre de caractéristiques extraites de l'imagerie TEP est croissant (caractéristiques du 1^{er} ordre, de forme ou de texture). Dans les Tableaux 1.1, 1.2 et 1.3 (Section 1.2.4 - Apport clinique du SUV_{max}, pages 18) sont présentés plusieurs articles étudiant les pouvoirs prédictifs et pronostiques de différentes caractéristiques pour les cancers du poumon et de l'œsophage ainsi que les tests statistiques utilisés. La méthodologie statistique varie en fonction de la problématique posée. Un résumé des différents tests statistiques classiques est donné Figure 2.3.

S'il est nécessaire de déterminer la valeur de seuil ("cut-off" en anglais) d'une caracté-

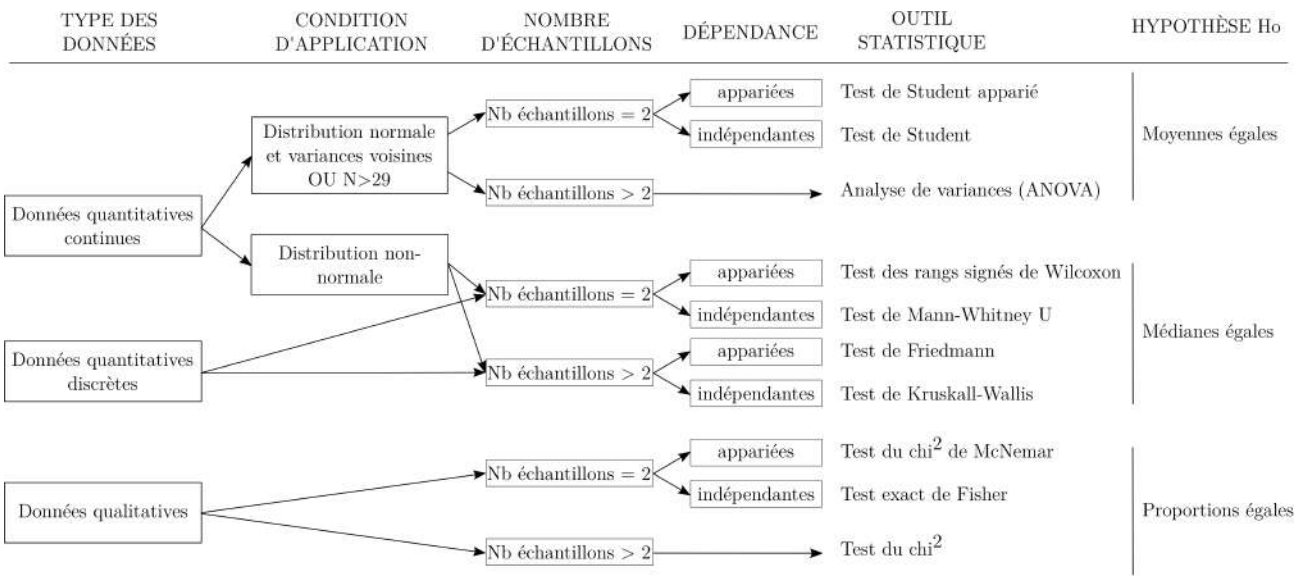


FIGURE 2.3 – Guide pour choisir un test statistique approprié en fonction de la situation.

ristique quantitative continue pour séparer une population en groupes à risque, l'analyse des courbes ROC [Berghmans et al. 2008] est utilisée. Plusieurs valeurs de seuil sont testées afin de trouver celle qui donne la relation la plus significative par rapport à la vérité terrain (par exemple répondeurs contre non-répondeurs). Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs (fraction des patients répondeurs qui sont effectivement détectés) en fonction du taux de faux positifs (fraction des patients non-répondeurs qui sont incorrectement détectés).

Dans le cas d'analyse de survie des patients, on retrouve généralement l'utilisation de l'analyse univariée de Kaplan-Meier [Kaplan and Meier 1958]. Le taux de survie en fonction du temps pour un ou plusieurs groupes de patients est représenté sur un graphe. On obtient alors une courbe de survie par groupe. Afin d'évaluer si ces courbes sont significativement différentes entre elles, on utilise des tests comme le test de log-rank [Mantel 1966]. Ce test non-paramétrique permet la comparaison des distributions de survie 2 à 2. Plus la valeur de p est faible, plus la différence entre les courbes est significative. Ensuite, peut-être réalisé une analyse multivariée de Cox [Cox and Oakes 1984] combinant des caractéristiques indépendantes parmi celles qui ont une valeur pronostique lors de l'analyse univariée. L'analyse de Cox est une méthode semi-paramétrique et permet de générer un modèle à partir du sous-ensemble des caractéristiques ayant une valeur pronostique. Les modèles de survie étudient le temps écoulé avant qu'un événement ne

survienne (survie et décès). La probabilité du devenir du patient (décès ou tout autre événement d'intérêt) est appelé le danger ("hazard", noté H). Un *"Hazard-Ratio"* (HR) peut alors être obtenu grâce à cette méthode en rapportant le "hazard" d'un patient sur celui d'un autre patient ne présentant aucun facteur de risque à l'instant t.

Plusieurs tests non-paramétriques existent pour comparer les différents groupes de patients entre eux (Figure 2.3). Le test de Kruskal-Wallis [Kruskal and Wallis 1952] permet la comparaison des rangs de plus de 2 échantillons indépendants contenant des données quantitatives discrètes ou quantitatives continues présentant une distribution non-normale. Le but est de déterminer si les échantillons proviennent d'une même population ou si au moins un échantillon provient d'une population différente. L'hypothèse nulle (H_0) posée est que l'ensemble des médianes des échantillons sont égales. Si on rejette l'hypothèse nulle, alors au moins un échantillon est différent. Le test de Friedmann est utilisé en cas de présence de plus de 2 échantillons appariés. En cas de l'étude de 2 échantillons, les tests des rangs signés de Wilcoxon et les tests de Mann-Whitney U [Wilcoxon 1945] (aussi appelé somme des rangs de Wilcoxon) sont utilisés, respectivement, pour des données appariées ou indépendantes.

Pour l'étude de données quantitatives présentant une distribution gaussienne avec des variances voisines, ou avec un effectif supérieur ou égal à 30 par groupe, il existe 3 tests paramétriques. On peut utiliser l'un des 2 tests de Student (apparié ou non) si 2 échantillons sont étudiés, ou une analyse de variances ANOVA si l'on s'intéresse à plus d'échantillons. Ces tests s'intéressent à l'égalité des moyennes des différents groupes. Enfin, pour l'étude des données qualitatives, ce sont les proportions qui sont comparés dans l'hypothèse nulle. On retrouve le test du chi^2, le test du chi^2 de McNemar et le test exact de Fisher qui sont non-paramétriques.

D'après Chalkidou et al. [Chalkidou et al. 2015], l'utilisation de ces méthodes doit être rigoureuse pour éviter les erreurs d'interprétation. Les risques sont l'augmentation de l'erreur statistique de type I proportionnellement à l'augmentation du nombre d'hypothèses testées, l'influence des corrélations entre caractéristiques sur les résultats statistiques ou l'absence de cohorte de validation. D'autres facteurs d'instabilité augmentant la probabilité qu'un résultat soit faux sont répertoriés dans [Ioannidis 2005] comme la

petite taille des bases de données.

L'utilisation de l'approche de recherche de cut-off montre quelques limites. En effet, Hilsenbeck et al. [Hilsenbeck et al. 1992] ont montré que plus le nombre de seuillages testés augmente, plus la probabilité d'obtenir par erreur un résultat statistiquement significatif augmente. Chalkidou et al. [Chalkidou et al. 2015] conseillent l'utilisation d'une correction selon la formule de Altman [Altman et al. 1994]. La combinaison de ces analyses multiples augmente également l'erreur statistique de type I pour les différents tests réalisés. Pour corriger cette erreur, la méthode de Benjamini-Hochberg [Hochberg and Benjamin 1990] est présenté comme étant la plus pertinente d'après Chalkidou et al.

Comme présenté précédemment, de nombreuses corrélations ont été mises en évidence entre certaines caractéristiques de texture et celles du 1er ordre [Tixier et al. 2011], [Orlhac et al. 2014]. Cette colinéarité entre les caractéristiques de texture peut conduire au phénomène connu sous le nom "bouncing betas" [Kiers and Smilde 2007]. L'instabilité des poids de coefficients de régression dans un modèle multivarié lorsqu'il existe une multi-colinéarité entre les variables, ainsi que les petits changements dans les données conduisent à des coefficients de régression très différents.

Sur la base de ces observations, Chalkidou et al. ont réalisé une revue de la littérature portant sur l'utilisation des caractéristiques de texture des images TEP ou TDM pour la recherche d'intérêt prédictif de la réponse au traitement ou pronostique [Chalkidou et al. 2015]. Sur les 15 articles identifiés, une probabilité moyenne d'erreur de type I de 76 % (gamme = 34-99 %) a été estimé. Enfin, seulement 3/15 des études ont utilisé un ensemble de données de validation pourtant indispensable pour éviter les mauvaises conclusions. De ce fait, les résultats obtenus dans la majorité des articles portant sur l'analyse de caractéristiques basées sur des recherches de "cut-off" ou tests statistiques sont à prendre avec circonspection.

2.3 Conclusion

Dans cette section, nous avons présenté le développement de la radiomique cherchant à développer la personnalisation du traitement des patients. Ce concept cherche à

mettre à profit la complémentarité des informations apportées par l'imagerie médicale et des différentes modalités biologiques et cliniques. Il s'appuie donc sur l'analyse des caractéristiques de l'image.

Une étude de la littérature de ces caractéristiques a montré que plusieurs paramètres pouvaient les faire varier. Premièrement, différentes définitions de ces caractéristiques existent dans la littérature alors qu'il serait plus judicieux d'harmoniser les pratiques. C'est pourquoi dans notre travail, nous allons utiliser les expressions mathématiques proposées par Orlhac et al. [Orlhac et al. 2015].

Beaucoup de méthodes de segmentation ont été proposées dans la littérature [Hatt et al. 2011a] [Onoma et al. 2012] Afin, de réduire le biais dû à la segmentation de la lésion, nous proposons d'utiliser la même méthode de segmentation automatique [Vauclin et al. 2009], simple à utiliser et reproductible, sur l'ensemble des données à notre disposition.

Nous comprenons tout à fait la problématique du calcul des caractéristiques de texture dans les lésions de faible volume énoncé par Brooks et al. [Brooks and Grigsby 2014]. Cependant, pour avoir un réel impact clinique, la radiomique doit pouvoir s'appliquer au plus grande nombre de patients. Par ailleurs, il est complexe dans notre problématique d'obtenir une base de données contenant un nombre de patients suffisant ayant une lésion de volume supérieur à 700 voxels. C'est pourquoi nous avons suivi les recommandations de Hatt et al. [Hatt et al. 2015] spécifiant de conserver les observations de faible volume tout en contrôlant les corrélations entre caractéristiques de texture et le MTV.

Il existe de nombreuses limites à l'utilisation des statistiques classiques utilisées actuellement [Chalkidou et al. 2015]. Plusieurs solutions ont été proposées afin d'éviter d'obtenir des conclusions erronées, telles que l'évaluation de la reproductibilité des caractéristiques, l'analyse des corrélations croisées, un taux d'événements adéquats (au moins 10-15 observations par caractéristiques testées) et l'utilisation d'une cohorte de validation externe. De plus, ces méthodes statistiques analysent principalement les caractéristiques individuellement. Ainsi, une caractéristique qui n'apparait pas comme pertinente dans une analyse univariée ne sera pas étudiée dans l'analyse multivariée, alors qu'elle pourrait apporter une information complémentaire à d'autres caractéristiques.

Pour pallier aux problèmes liés à ces statistiques, nous nous sommes tournés vers

les méthodes d'apprentissage automatique et de sélection des caractéristiques. Ces méthodes sont particulièrement adaptées à la sélection de caractéristiques parmi un grand nombre. Elles pourraient apporter un éclaircissement pour savoir quelles caractéristiques sont pertinentes, aussi bien individuellement qu'en termes de sous-ensemble. Ces méthodes sont présentées dans le chapitre suivant.

Chapitre 3

Les méthodes d'apprentissage

Sommaire

3.1 Introduction

La radiomique conduit à étudier plusieurs dizaines de caractéristiques. C'est pourquoi, les outils statistiques utilisés doivent être suffisamment puissants et robustes, comme évoqué dans le chapitre précédent.

Les méthodes basées sur l'apprentissage automatique peuvent être définies comme des méthodes statistiques capables de tirer des conclusions à partir de données d'apprentissage, ce qui permet d'automatiser et d'améliorer le processus de prédiction [Parmar et al. 2015]. Les méthodes d'apprentissage automatique possèdent la capacité de gérer de nombreuses caractéristiques et de capter leurs relations non-linéaires. Elles conduisent alors à une meilleure puissance discriminante lors de l'analyse de plusieurs dizaines de caractéristiques en comparaison aux statistiques classiques [El Naqa et al. 2009].

De plus, comme tout champs d'exploration de données à haut débit, la radiomique est soumise au "fléau de la dimension" [Bellman and Bellman 1961]. Ce phénomène apparait lorsque la dimension de l'espace des caractéristiques (c'est-à-dire le nombre de caractéristiques) augmente et que les données inclues deviennent distantes et dispersées. Dans ces situations, les méthodes statistiques classiques auront tendance à donner des résultats faussés et biaisés. Afin de corriger cette problématique, il peut être intéressant d'utiliser des stratégies de sélection de caractéristiques dans le but de réduire la dimension de l'espace des caractéristiques en retirant celles non-pertinentes.

Dans ce chapitre, sont tout d'abord présentés les différents principes des méthodes d'apprentissage automatique qui sont l'apprentissage supervisée et non-supervisée. Puis, les différentes méthodes de sélection des caractéristiques sont présentées (filtrante, enveloppante et intégrée). Enfin, les travaux de la littérature sur l'utilisation de ces méthodes en imagerie médicale sont présentés.

3.2 Apprentissage automatique

3.2.1 Principe

L'apprentissage automatique est une science du domaine de l'intelligence artificielle qui a pour objectif le développement de la capacité d'apprentissage des machines (d'où l'appellation anglaise "machine learning"). Ainsi, à l'aide d'un algorithme, une machine pourra modéliser une règle de décision concernant une problématique spécifique au sein d'une population, à partir d'un échantillon de cette population, appelé base de données d'apprentissage. Le modèle résultant de l'apprentissage est nommé classifieur et est utilisé dans la labellisation (étiquetage) de nouvelles observations jusque-là inconnues. Les labels des observations peuvent être quantitatifs ou qualitatives.

Les algorithmes d'apprentissage automatique sont utilisés dans de nombreux domaines tels que la vision par ordinateur, la recherche d'information ou la bio-informatique [Mohri et al. 2012]. Ces systèmes sont utilisés afin de remplir des tâches difficilement réalisables par des moyens algorithmiques plus classiques, lorsque la difficulté réside dans la complexité à décrire l'ensemble des comportements possibles des différentes entrées. On parle alors d'explosion combinatoire. On confie donc à des programmes le soin d'élaborer un modèle permettant d'assimiler cette complexité et de l'employer ensuite de manière optimale.

Le cadre mathématique de l'apprentissage automatique est le suivant (voir Figure 3.1). Au sein d'une population $\mathscr{P}$ étudiée, est tiré un échantillon représentatif D_{app}. Le but des méthodes d'apprentissage automatique est d'estimer la loi, répondant une problématique d'une population $\mathscr{P}$, à partir de D_{app}. Ainsi, lorsqu'une nouvelle entrée x est apportée au modèle généré, ce dernier doit être capable de prédire le label y_{pred} devant être le plus proche possible du vrai label y.

Il existe principalement 2 types d'apprentissage automatique :

— L'algorithme d'apprentissage supervisé apprend un modèle à partir de données d'apprentissage labellisées au préalable par un expert (ou oracle). Ainsi, l'échantillon d'apprentissage est composé de 2 parties : X_{app} et Y_{app}. X_{app} est une matrice

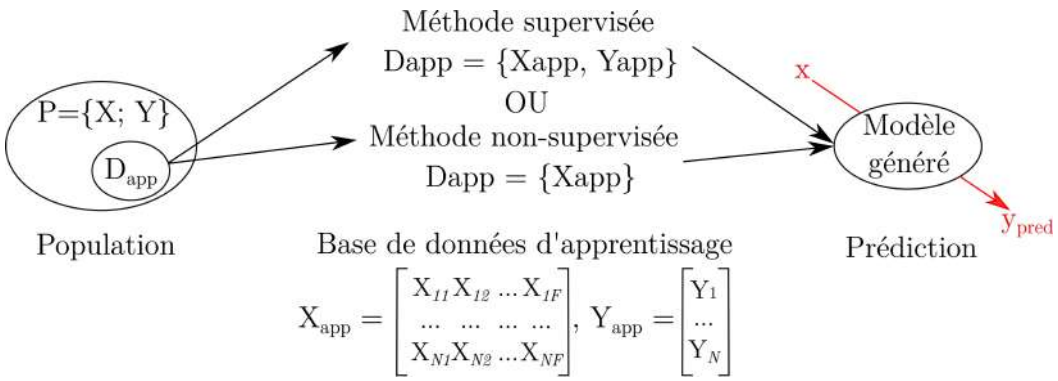


FIGURE 3.1 – Principe de base de l'apprentissage automatique. Après création d'une base de données, un modèle est généré par apprentissage automatique à partir de l'échantillon d'apprentissage D_{app}. Ce modèle permet de prédire un label y_{pred} d'une observation étudiée x.

comprenant N lignes de dimension F, représentant les caractéristiques des observations (par exemple une ligne par patient et une colonne par caractéristique) et Y_{app} est un vecteur à N valeurs représentant les labels de chaque observation. Lors de l'apprentissage, X_{app} est considérée comme une matrice d'entrée par l'algorithme d'apprentissage automatique et Y_{app} un vecteur de sortie. L'algorithme doit donc détecter le lien d'entrée-sortie qui existe entre X_{app} et Y_{app}.

— L'algorithme d'apprentissage non-supervisé dispose d'exemples, mais ne connait ni les labels, ni le nombre de classes et leur nature. Il doit découvrir par lui-même la structure plus ou moins cachée des données. L'avantage de cette méthode est qu'elle ne requiert la présence d'aucun expert pour la définition initiale de label. L'échantillon d'apprentissage D_{app} ne contient que la matrice des caractéristiques des observations X_{app}, utilisée comme entrée pour l'algorithme.

D'autres méthodes d'apprentissage existe comme l'apprentissage semi-supervisé, par renforcement et par transfert, elles ne seront pas développées dans cette thèse.

Comme l'algorithme génère un modèle à partir des données empiriques, la qualité de la base de données d'apprentissage est déterminante pour l'obtention de performances. La qualité du travail dépendra de facteurs liés à la base de données comme :

— Le nombre d'observations qui doit être suffisamment important, au moins plusieurs dizaines, pour répondre à la complexité du problème,

— La répartition des classes à prédire qui doit être la plus équilibrée possible, sous peine de fausser l'apprentissage,

— Le nombre et la qualité des caractéristiques décrivant ces observations (ratio entre nombre d'observations et de caractéristiques, résistance aux données manquantes),
— Le bruit généré par la présence de valeurs douteuses ou naturellement non-conformes à la distribution générale des observations.

De plus, dans les cas où la taille de la base de données d'apprentissage est faible et que le nombre de caractéristiques est important, il est possible d'être confronté à un phénomène de sur-apprentissage ("overfitting"). Par expérience en médecine, un ratio de 10 observations par caractéristique testée est préféré. Un modèle généré dans le cas contraire a de la peine à généraliser des données et perd ses pouvoirs de prédiction sur de nouvelles observations. Dans cette situation, l'utilisateur à le choix entre ajuster la complexité du modèle ou réduire le nombre de caractéristiques étudiées.

Dans les sous-sections suivantes, nous allons détailler davantage le principe des méthodes d'apprentissage supervisées et non-supervisées et nous allons expliquer le fonctionnement des algorithmes s'y rapportant.

3.2.2 Les méthodes supervisées

L'approche par apprentissage supervisé consiste à élaborer un modèle de décision à partir des données d'apprentissage labellisées ($D_{app} = \{X_{app}; Y_{app}\}$) en faisant l'hypothèse d'une distribution particulière des classes (approche générative) ou en traçant une frontière de décision séparant les classes en présence (approche discriminante).

Un modèle génératif apprend la distribution de probabilité jointe $p(x, y)$ modélisant la façon dont les données ont été générées afin de catégoriser un signal. Il pose la question : sur la base des hypothèses de génération, quelle catégorie est la plus susceptible de générer ces données? Un modèle discriminant apprend la distribution de probabilité conditionnelle $p(y|x)$. Il ne s'intéresse pas à la façon dont les données ont été générées, mais il les classe simplement. Il existe un lien entre ces 2 modèles en utilisant le théorème Bayes (Équation 3.1).

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{3.1}$$

Deux étapes successives sont utilisées par les méthodes d'apprentissage supervisées. Lors de la première phase (hors ligne, dites d'apprentissage) un modèle est élaboré à partir des données labellisées. L'ensemble d'apprentissage est composé de F caractéristiques recueillies, accompagnées des labels correspondants aux classes auxquelles ils appartiennent. La seconde phase (en ligne, dites de test) consiste à prédire le label d'une nouvelle donnée à partir du modèle établi lors de l'étape précédente. Certaines méthodes préfèrent associer une donnée, non pas à une classe unique, mais plutôt à une probabilité d'appartenance à chacune des classes prédéterminées. On parle alors d'apprentissage supervisé probabiliste.

Il existe 2 types de problèmes à résoudre avec l'apprentissage supervisé dépendant du type des variables à prédire :

— Si les variables à prédire sont continues, on se place dans le cas d'un problème de régression [Hosmer and Lemeshow 2000] [Specht 1991]. Ainsi le modèle généré peut estimer une valeur numérique en sortie en fonction des données d'entrée, comme l'estimation du poids d'un patient.

— Si les variables à prédire sont discrètes, il s'agit d'un problème de classification [Boughattas et al. 2014] [Ypsilantis et al. 2015]. On cherche alors à séparer les observations en différentes classes, par exemple les 2 classes patients répondeurs ou non-répondeurs à un traitement.

Parmi les études avec des variables à prédire continues, nous avons les méthodes suivantes.

La régression linéaire est un modèle simple et efficace pour l'apprentissage dans le cas linéaire. Le but est de rechercher, à partir d'un échantillon d'apprentissage $D_{app} = \{X_{app}; Y_{app}\}$, la droite permettant d'expliquer le comportement de Y_{app} comme étant une fonction affine de X_{app}. On peut l'exprimer mathématiquement par la notation suivante :

$$Y_{app} = X_{app}\beta + \epsilon \tag{3.2}$$

Où β représente le vecteur des paramètres du modèle et ϵ l'erreur.

Il existe plusieurs méthodes cherchant à optimiser β. On peut citer la méthode des

moindres carrés, le modèle par maximum de vraisemblance ou par inférence bayésienne ou la méthode par descente de gradient [Box and Tia 1973]. Ces méthodes convergent vers la solution optimale minimisant un critère d'évaluation. Cette méthode présente des performances limitées dû au fait qu'elle est restreinte aux cas linéaires.

Parmi les études avec des variables à prédire discrètes, nous avons les méthodes suivantes.

La régression logistique permet de rechercher la relation entre une caractéristique dichotomique (dans laquelle il n'y a que deux résultats possibles) et une ou plusieurs caractéristiques indépendantes. Contrairement à la régression linéaire, cette méthode n'est pas limitée aux cas linéaires. La régression logistique génère les coefficients d'une formule pour prédire une transformation "logit" de la probabilité de présence de la caractéristique d'intérêt telle que :

$$logit(p) = ln\frac{p}{1-p} \tag{3.3}$$

Où p est la probabilité de présence de la caractéristique d'intérêt binaire Y_{app}. La transformation "logit" est définie comme :

$$logit(p) = b_0 + X_{app}b \tag{3.4}$$

Où b_0 et b représentent les paramètres du modèle.

Ainsi, plutôt que de choisir des paramètres qui minimisent une erreur (comme dans la régression ordinaire), la régression logistique privilégie des paramètres qui maximisent la probabilité d'observer les valeurs de l'échantillon.

L'approche bayésienne est basée sur une approche probabiliste employant la règle de Bayes (voir Équation 3.1). Dans un cas à 2 classes y_1 et y_2, comme le dénominateur de la formule de Bayes ne dépend pas de y, nous ne nous intéressons qu'au numérateur. Les probabilités $P(y)$ de chaque classe ainsi que les distributions $P(x|y)$ doivent être préalablement estimées à partir d'un échantillon d'apprentissage. Le vecteur de caractéristiques x est assigné à la classe y_i si :

$$P(y_i|x) > P(y_j|x)\ \forall i,\ i \neq j \tag{3.5}$$

L'algorithme des *k-Plus Proches Voisins ou "k-Nearest Neighbors"* (kNN) [Blum and Langley 1997] est un exemple d'algorithme de classification par apprentissage supervisé. Pour estimer la sortie y_{pred} associée à une nouvelle observation x, la méthode des kNN consiste à prendre en compte de façon identique les k échantillons d'apprentissage les plus proches de cette observation (voir Figure 3.2). Cette approche est donc basée sur une estimation de distance. Pour un problème de classification, l'algorithme attribue à l'observation testée le label majoritaire des k individus de l'ensemble d'apprentissage les plus proches. En cas de régression, la valeur de l'observation x est définie comme la moyenne des k plus proches voisins. Si la valeur de k est faible, on obtient un classifieur dit de bonne résolution, très proche des données avec un biais faible. Si en revanche k est élevé, on obtient un classifieur qui peut lisser la frontière de décision et être plus robuste vis-à-vis du bruit.

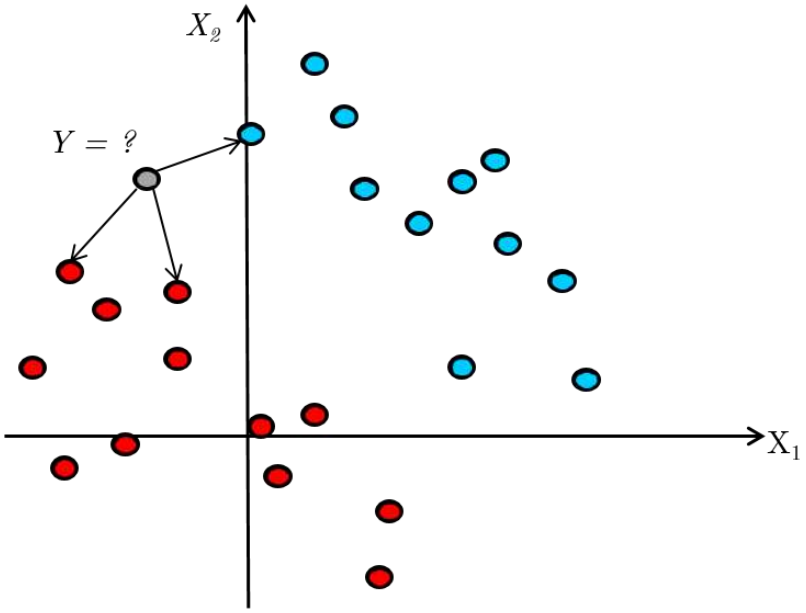


FIGURE 3.2 – Illustration d'un problème de classification kNN binaire en 2 dimensions. Le label de l'observation étudiée (en gris) est déterminé en fonction des labels des k plus proches voisins, ici $k = 3$. Le label Y est défini comme étant rouge.

Un autre exemple de méthode d'apprentissage automatique supervisée est la méthode *Séparateur à Vaste Marge ou "Support Vector Machine"* (SVM) [Cortes and Vapnik 1995]. C'est l'un des algorithmes les plus cités dans la littérature en raison de ses perfor-

mances et de sa généricité [Shawe-Taylor and Cristianini 2004] [Rakotomamonjy 2003].

Le but du SVM est de modéliser un hyperplan capable de séparer les données d'apprentissage en fonction de leur classe. Comme il existe une infinité d'hyperplan séparateur dans l'espace, le SVM cherche à maximiser une marge entre les données d'apprentissage (voir Figure 3.3). La marge est définie comme étant la distance entre les vecteurs de support qui représentent ces données d'apprentissage. La recherche de cet hyperplan est formulée comme un problème d'optimisation quadratique.

Afin de résoudre des problèmes non-linéaires, le SVM peut utiliser différentes fonctions noyaux afin de redistribuer l'espace. L'idée est de transformer l'espace de représentation des données d'entrée en un espace de plus grande dimension (espace de redescription), où il est probable qu'un séparateur linéaire obtienne de bonnes performances. Il existe plusieurs formes de noyaux, dont les plus utilisés sont les noyaux linéaires, polynomiales et gaussiens.

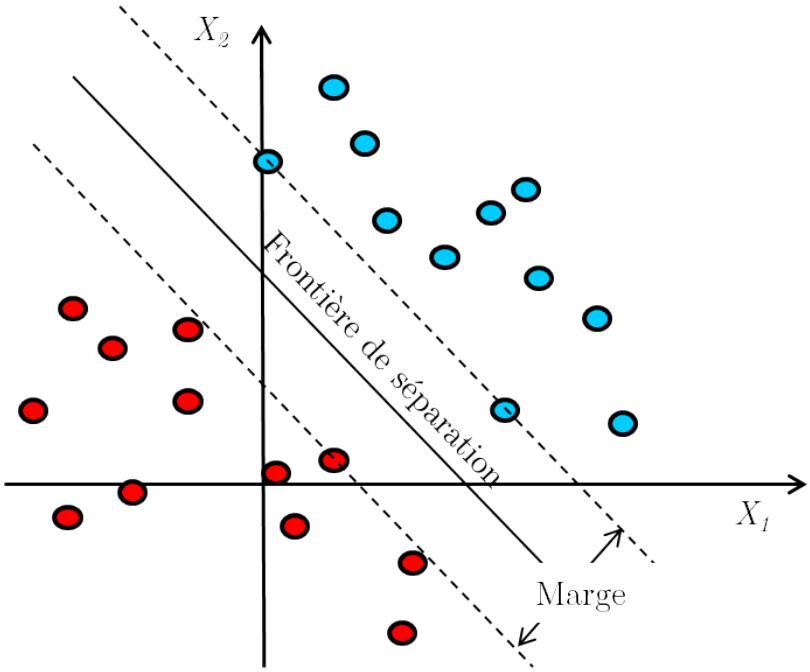


FIGURE 3.3 – Illustration d'un cas de classification SVM en 2 dimensions. La marge est établie entre les 2 vecteurs de supports représentant les échantillons d'apprentissage.

L'algorithme des *Réseaux de neurones artificiels ou "Artificial Neural Network"* (ANN) [McCulloch and Pitts 1943] est une méthode d'apprentissage supervisée (voir Figure 3.4) qui s'inspire du fonctionnement biologique des neurones. L'ANN est en fait l'association d'une multitude d'éléments de traitement (neurones artificiels) en réseau qui travaillent

à l'unisson pour résoudre des problèmes spécifiques. Les neurones sont organisés en couches successives où les résultats d'une couche servent d'entrées aux calculs de la suivante. Lors de l'apprentissage d'un modèle d'ANN, des poids reliant les neurones sont définis. Ainsi, le résultat du modèle généré dépend des poids appris et des caractéristiques présentées aux neurones.

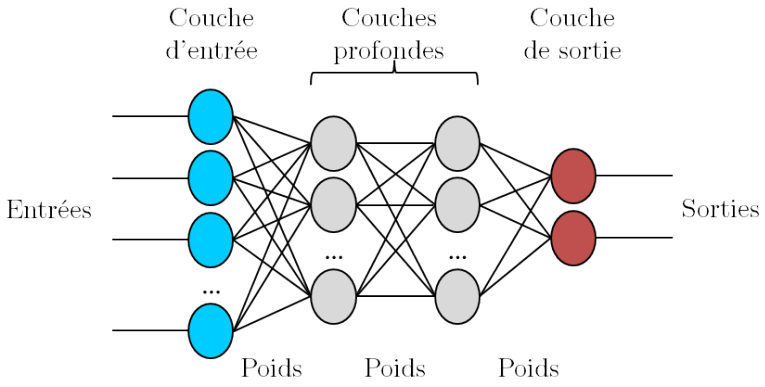


FIGURE 3.4 – Schéma de l'architecture d'un modèle de réseau de neurones artificiels.

Un problème lors de l'utilisation des ANN en traitement de l'image est le fait que chaque voxel est connecté à l'intégralité de ceux de la couche suivante ("fully-connected"). Cette construction engendre un grand nombre de poids à définir au sein du modèle pouvant engendrer un sur-apprentissage. Pour réduire cet impact, il existe des méthodes comme le "dropout" qui retire aléatoirement des connections entre neurones, allégeant ainsi le modèle [Srivastava et al. 2014]. Néanmoins, il peut être difficile de détecter les caractéristiques pertinentes d'une image en s'intéressant à celle-ci de manière globale.

Afin de répondre à cette problématique, ont été proposés les *Réseaux de neurones convolutifs* (CNN) [Fukushima 1980] [LeCun et al. 1998] (voir Figure 3.5). Pour chaque couche du CNN une série de convolutions est réalisée entre la sortie de la couche précédente et des filtres spécifiés par l'utilisateur. Ces convolutions mènent à des cartes de caractéristiques ("features map") représentant un nouveau niveau de caractéristiques locales qui serviront à leur tour comme entrées pour la couche suivante.

L'avantage de cette approche est que les neurones s'intéressent à des caractéristiques locales qui peuvent être plus pertinentes dans la résolution de certains problèmes, notamment lorsque le nombre de voxels de l'image étudiée est important.

Les CNN font partie de la famille des méthodes d'apprentissage profond (ou "deep

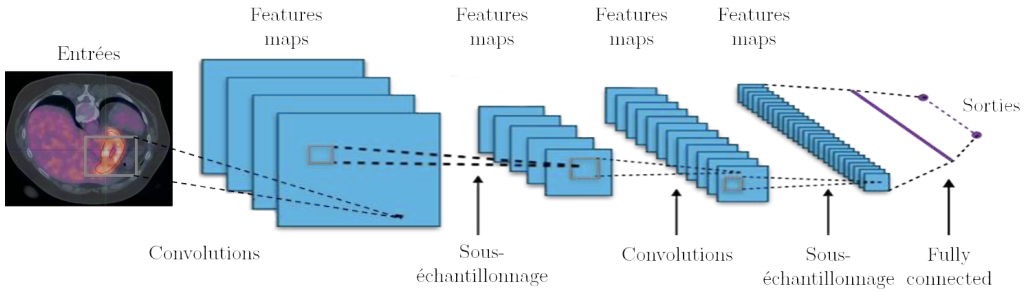


FIGURE 3.5 – Schéma de l'architecture d'un modèle de réseau de neurones convolutifs.

learning"). Ces méthodes basent leur processus d'apprentissage sur une abstraction successive des données grâce à une succession de couches cachées. Ces approches se sont développées ces dernières années grâce à la multiplication de la puissance de calcul des ordinateurs. A chaque couche, le modèle approfondit sa compréhension avec des concepts de plus en plus précis. Ce type de méthode nécessite une base d'apprentissage importante, avec plusieurs centaines d'observations, ce qui peut être difficile à obtenir dans différents domaines, notamment en imagerie médicale.

Enfin, on peut citer les méthodes issues des arbres de décisions comme les *Forêts Aléatoires ou "Random Forest"* (RF). La méthode des arbres de décision [Breiman et al. 1984] construit des classifieurs par arbre aussi bien en régression qu'en classification. Ainsi, un arbre est construit en divisant progressivement une population en 2 populations filles et ce dans l'optique d'optimiser l'homogénéité des populations en fonction de leur label. Plus tard, les méthodes des forêts aléatoires ont été proposées [Breiman 2001] qui regroupent une multitude d'arbres de décisions indépendants. Ces derniers sont construits à partir de la même base d'apprentissage à l'aide de différents processus d'aléas. Le fait de combiner plusieurs arbres de décision permet d'atténuer l'influence des données bruitées lors de la phase d'apprentissage. Les méthodes des arbres de décision et des forêts aléatoires sont présentés plus longuement dans le chapitre suivant car elles sont à la base de nos développements.

3.2.3 Les méthodes non-supervisées

Le but des méthodes à apprentissage non-supervisées est de rechercher les particularités de la distribution des observations étudiées. Pour cela, le système dispose en entrée

d'une base de données non-labellisées à analyser ($D_{app} = \{X_{app}\}$). De plus, le nombre de classes et leur nature ne sont pas toujours connus. L'algorithme doit alors découvrir par lui-même la structure plus ou moins cachée des données pour les classer en groupes homogènes ("clusters") selon leurs caractéristiques. Ainsi, les éléments d'un groupe homogène présentent des similarités entre eux et des différences vis-à-vis des autres "clusters".

La similarité est calculée selon un critère d'éloignement ou d'homogénéité entre individus. C'est ensuite à l'opérateur d'associer ou de déduire du sens à chaque groupe et motif ("pattern") d'apparition de groupe. L'avantage de cette méthode est qu'elle ne requiert pas la présence d'un expert. Il existe plusieurs familles de méthode de regroupement des données en "clusters" comme les méthodes basées sur une distance ou probabilistes.

Les méthodes basées sur une distance sont parmi les premières méthodes de "clustering" à avoir été proposées. Ces méthodes se basent sur la notion de distance entre les observations de la base de données. Ainsi, si 2 observations sont proches suivant cette distance, ils doivent appartenir au même "cluster". Il existe plusieurs types de distances utilisées. Les méthodes basées sur les prototypes définissent pour chaque "cluster" un repère associé à ce "cluster". Ce repère est appelé centroïde s'il est calculé comme la moyenne des éléments du "cluster" et médoïde s'il s'agit d'un objet particulier du "cluster". Les algorithmes k-Means [Macqueen 1967] et *"Fuzzy-C-Means"* (FCM) [Dunn 1974] sont les algorithmes les plus connus de cette famille. Ces méthodes permettent de trouver des formes de "clusters" convexes et sont très utilisées notamment à cause de leur coût algorithmique faible.

Le but de la méthode des k-Means est de diviser la base de données étudiées en k groupes, de façon à minimiser une fonction basée sur la distance entre un "point moyen" et les points associés à ce "cluster".

On peut représenter l'algorithme k-Means en 4 étapes :

— Choisir k points initiaux formant k "clusters"

— (Ré)affecter chaque point o au "cluster" C_i de centre μ_i tel que la distance entre les 2 soit minimale

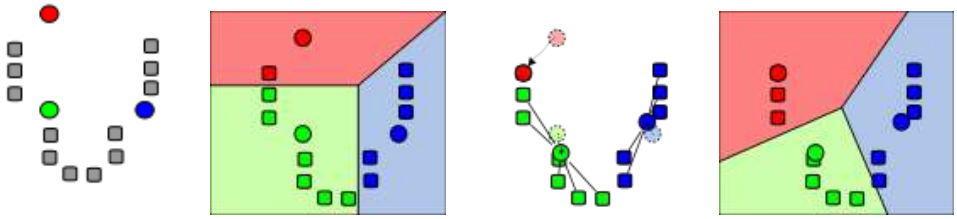
$$\arg\min_{C} \sum_{i=1}^{k} \sum_{o_j \in C_i} ||o_j - \mu_i||^2 \tag{3.6}$$

où μ_i est la moyenne des points au sein d'un "cluster" C_i, et C l'ensemble des "clusters".

— Recalculer μ_i de chaque "cluster" (le barycentre)

— Retour à l'étape 2 si on vient de faire une affectation

Il en résulte une séparation de l'espace en un diagramme de Voronoï composé de plusieurs cellules. Le fonctionnement de l'algorithme des k-Means est illustré Figure 3.6.



(a) (b) (c) (d)

FIGURE 3.6 – Principe de l'algorithme des k-means. (a) k points moyens initiaux sont créés aléatoirement (ici $k = 3$). (b) Les "clusters" sont créés en fonction du point moyen le plus proche. (c) Les positions des k points sont réévaluées en fonction des points appartenant aux "clusters". (d) Les étapes (b) et (c) sont répétées jusqu'à convergence vers un diagramme de Voronoï.

L'algorithme FCM est une généralisation de l'algorithme des k-means qui introduit la notion d'ensembles flous dans la définition des classes. Il s'agit d'une forme de "clustering" où chaque observation peut appartenir à plusieurs "clusters". Chaque "cluster" est caractérisé par son centroïde. FCM utilise un critère de minimisation des distances intra-classes et de maximisation des distances inter-classes en donnant un certain degré d'appartenance ($\in [0, 1]$) à chaque classe pour chaque observation. Ainsi, si une observation se trouve près du centre d'un "cluster", elle aura alors un degré d'appartenance élevé, alors qu'une observation plus éloignée présentera un degré plus faible. Au final, le FCM permet d'obtenir une partition floue. Le "cluster" auquel est associé une observation est celui dont le degré d'appartenance sera le plus élevé.

L'algorithme FCM se déroule selon les étapes suivantes :

— Initialisation des degrés d'appartenance de manière arbitraire à chaque point

— Calcul des centroïdes des k "clusters"

— Réajustement des degrés d'appartenance suivant les centroïdes obtenus à l'étape précédente

— Calcul du critère de minimisation et retour à l'étape 2 si non-convergence du critère l'algorithme

L'inconvénient de ces méthodes est qu'il est nécessaire de fixer au préalable le nombre de "clusters". Un choix inapproprié de k peut entraîner de mauvais résultats. Pour s'affranchir de ce problème, il existe des méthodes pour déterminer le nombre de "clusters" dans l'ensemble de données. Une autre limite de ces méthodes est l'utilisation de la distance euclidienne comme mesure de dispersion des "clusters". Les "clusters" doivent être de taille similaire, de sorte que l'affectation au centre du "cluster" le plus proche soit l'affectation correcte. Si par exemple, un modèle à 3 classes est présenté avec 2 "clusters" de petites tailles et un plus important, le modèle aura tendance à regrouper les 2 petits et à séparer le grand.

Les méthodes probabilistes supposent que les données suivent une certaine loi de probabilité. L'objectif est d'estimer les paramètres de cette loi et de définir un modèle de mélange de lois pour représenter les différents "clusters". Ces méthodes font l'hypothèse qu'à chaque "cluster" est associée une loi de probabilité qui permet de déterminer la probabilité d'appartenance de chaque object à un "cluster". On pourra notamment citer l'algorithme *"Expectation Maximization"* (EM) [Dempster et al. 1977].

3.2.4 Avantages et inconvénients des méthodes d'apprentissage

Chaque méthode d'apprentissage automatique possède des avantages et des inconvénients. C'est pour cela qu'il est nécessaire de bien sélectionner la méthode en fonction de la problématique. Que ce soit au niveau de la problématique posée (régression et classification, 2 classes ou plus), du type de données d'apprentissage dont l'utilisateur dispose (labellisées ou non, en grande quantité, équilibre entre les classes), le nombre de caractéristiques étudiées et le type de relation entre elles (linéaire ou non), une méthode sera privilégiée. De même, certaines méthodes mettent en avant la précision des résultats face à la durée d'apprentissage.

La méthode SVM est un bon choix pour les ensembles de caractéristiques de grande taille. Elle détecte bien la linéarité entre les caractéristiques et prédiction à l'aide des noyaux et présente une durée d'apprentissage relativement courte. Ainsi, les SVM sont en

mesure de séparer les classes plus rapidement et avec moins de sur-apprentissage que la plupart des autres algorithmes, tout en utilisant une petite quantité de mémoire. Cependant, la difficulté de traiter simultanément des données de différents types (continues, discrètes ...) reste un point faible de cette approche. Le SVM est initialement limité à la prédiction de 2 classes, cependant plusieurs auteurs se sont intéressés à la combinaison de multiples SVM pour augmenter ce nombre [Hsu and Lin 2002].

La méthode de classification par les ANN permet de gérer la prédiction de plusieurs classes à la fois et est extrêmement précise. Ces performances élevées ont toutefois un prix. L'apprentissage peut prendre beaucoup de temps (plusieurs minutes ou heures), en particulier pour les grands jeux de données avec un grand nombre de caractéristiques. De plus, cette méthode dépend d'un grand nombre de paramètres, plus que la plupart des autres méthodes, ce qui signifie que le temps de sélection des paramètres allonge grandement la complexité et le risque de sur-apprentissage. Néanmoins, des méthodes de régularisation existent afin de répondre à ce problème, c'est le cas par exemple du "dropout". Les CNN sont moins régis par cette question. Enfin, ces méthodes opèrent directement sur les images en extrayant des caractéristiques automatiquement. Il peut donc être difficile d'interpréter humainement des résultats obtenus par un réseau de neurones, ce qui peut provoquer une certaine réticence de son utilisation en clinique.

Les RF sont robustes au faible nombre de données d'apprentissage de part leur multiplication artificielle d'échantillons. Certains processus permettent également d'éviter le sur-apprentissage en limitant le nombre de divisions et le nombre minimum d'observations finales par feuille. Le modèle des forêts aléatoires construit est généralement très précis car chaque arbre compense l'erreur des précédents et permet la gestion de plusieurs classes à la fois. Cependant, plus le nombre d'arbres dans la forêt est élevé, plus cette méthode a tendance à utiliser de la mémoire. Cette méthode est également facilement modulable par l'utilisateur pour la détection des caractéristiques d'importances lors de l'apprentissage. Pour ces raisons, nous avons choisi d'étudier cette méthode pour résoudre nos problèmes.

3.3 Évaluation des méthodes d'apprentissage automatique

Une étape de test est nécessaire pour l'évaluation de la qualité du modèle généré par apprentissage automatique (voir Figure 3.7). Cette étape consiste à prédire les labels Y_{pred} d'une base de données de caractéristiques X_{test}, dont on connait la vérité terrain Y_{test}. La difficulté vient du fait qu'un modèle doit être testé par une autre base de données que celle ayant servie à sa construction, ou l'erreur sera sous-estimée. Des protocoles ont alors été proposés afin de résoudre cette problématique.

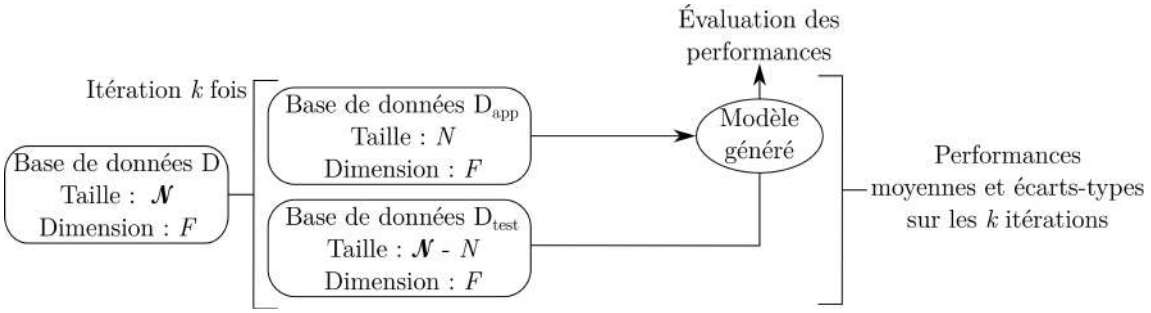


FIGURE 3.7 – Illustration du processus de test des méthodes d'apprentissage automatique.

Si l'on dispose initialement d'une base de données D contenant $\mathcal{N}$ observations et F caractéristiques, il est possible d'évaluer le modèle à l'aide d'approches de validation croisée. Il s'agit de méthodes basées sur des techniques d'échantillonnage. Ainsi, la base de données initiale D est scindée en 2, menant à une base d'apprentissage D_{app} à partir de laquelle est généré le modèle et une base de test D_{test} à partir de laquelle sont évaluées les performances. Ce processus est répété k fois pour au final générer une moyenne et un écart-type des performances sur l'ensemble des itérations.

Il existe plusieurs méthodes de validation croisée variant par leur nombre k d'itérations du processus et la séparation entre la base d'apprentissage et la base de test.

— La méthode *"Leave-p-out Cross-Validation"* (LpOCV) [Burman 1989], est une méthode de validation croisée exhaustive, c'est-à-dire, qu'elle permet d'apprendre et de tester toutes les façons possibles de diviser l'échantillon original en une base d'apprentissage et de test. Cette méthode consiste à utiliser p observations pour la création de l'ensemble de test et les observations restantes pour la base d'apprentissage. Ainsi, il existe C_p^N combinaisons possibles qui correspond aux nombres k

d'itérations réalisées.

Comme le nombre d'itération peut être extrêmement élevé avec cette méthode, d'autres approches ont été dérivées afin de réduire le temps de calcul nécessaire.

— La méthode *"Leave-One-Out Cross-Validation"* (LOOCV) est également une méthode de validation croisée exhaustive. Cette méthode est un cas particulier de la méthode LpOCV lorsque $p = 1$. Ainsi, l'apprentissage est réalisé sur $\mathcal{N} - 1$ observations et le test est réalisé sur l'observation restante. Ce processus est réalisé $\mathcal{N}$ fois ($k = \mathcal{N}$) pour que chaque observation de D serve une fois en tant que test D_{test}.

— La méthode *"k-Fold Cross-Validation"* (kFCV) est une méthode de validation croisée non-exhaustive, c'est-à-dire, que l'ensemble des combinaisons n'est testé que partiellement. Ici, la base de données D est divisée en k parties de même taille. L'une des parties sert de base de test D_{test} et les $k-1$ autres de base d'apprentissage D_{app}. Ce processus est réalisé k fois pour que chaque observation de D serve une fois de base de test D_{test}. Si $k = \mathcal{N}$, on retombe sur le cas précédent du LOOCV.

— La méthode des *Permutations Aléatoires* (PA) est une méthode de validation non-exhaustive [Dubitzky et al. 2007]. À chaque itération, un pourcentage des observations fixé par l'utilisateur est tiré aléatoirement sans remise parmi les $\mathcal{N}$ disponibles pour former la base de test alors que les observations restantes sont attribuées à la base d'apprentissage. Cette méthode ne fait pas partie de la famille des validations croisée car une observation peut être utilisée dans plusieurs bases de test. Une variante de cette approche propose de générer les échantillons de test et d'apprentissage en conservant le ratio des labels de la population initiale (ou la réponse moyenne en cas de régression). Ceci est particulièrement utile en classification si la répartition des classes est déséquilibrée.

D'autres méthodes peuvent encore être citées comme l'approche "Holdout" ou encore la méthode des boostraps. Cette dernière sera développée dans le chapitre suivant.

Chaque méthode possède ses propres avantages et inconvénients. Lors de l'utilisation des méthodes de validation croisée, une partie des données disponibles est conservée pour créer la base de test. Une base d'apprentissage plus petite peut dégrader la qualité du

modèle généré. La méthode de LOOCV est celle qui utilise le moins d'observations de test alors que pour la méthode kFCV, ce nombre varie en fonction de k. Si k augmente, la taille de l'échantillon d'apprentissage augmente. Les performances sont donc proportionnelles à k. Cependant, si la base de données de test n'est pas suffisamment importante, même avec un bon modèle, il est possible d'avoir de fausses prédictions. Ainsi, plus k augmente, plus le nombre d'observations de test diminue ce qui augmente la variance. La méthode LOOCV est celle présentant le moins de perte de données, cependant comme k est élevé, cette méthode coûte cher en temps de calcul.

L'avantage de la méthode des PA est que la proportion de la division des bases d'apprentissage/test ne dépend pas du nombre k d'itérations, ce qui laisse la liberté à l'utilisateur de fixer ses propres paramètres. Ainsi, si k augmente, on retrouve la méthode LpOCV testant l'intégralité possible des combinaisons. L'inconvénient de cette méthode est que certaines observations peuvent ne jamais être sélectionnées dans le sous-échantillon de test, tandis que d'autres peuvent être sélectionnées plus d'une fois. En d'autres termes, les sous-ensembles de validation peuvent se chevaucher.

En résumé, la méthode LOOCV apporte d'excellentes performances, mais prend plus de temps à réaliser. En cas d'importante base de données, il est possible de se replier sur des méthodes moins gourmandes en calculs comme la méthode kFCV ou la méthode des PA mais réduisant les performances. La méthode des PA est utilisée lorsque la base de données est encore trop faible pour utiliser la méthode kFCV.

À chaque itération, un modèle est généré à partir de D_{app}. Il est alors possible de comparer les labels estimés Y_{pred} et la vérité terrain Y_{test}. L'évaluation se fait à l'aide d'indices d'évaluation tels que l'erreur de classification, l'AUC, la *Sensibilité* (Se) et la *Spécificité* (Sp). La sensibilité et la spécificité sont obtenues à partir de la matrice de confusion qui compare les labels estimés et connus.

TABLEAU 3.1 – Matrice de confusion.

	Événement présent (1)	**Événement absent (0)**
Test Positif	Vrais Positifs (VP)	Faux Positifs (FP)
Test Négatif	Faux Négatifs (FN)	Vrais Négatifs (VN)

Les vrais positifs correspondent au nombre d'estimations correctes des labels positifs

(1), vrais négatifs aux estimations correctes des labels négatifs (0), faux négatifs aux estimations erronées des labels positifs et faux positifs aux estimations erronées des labels négatifs. Ainsi, la sensibilité et la spécificité sont calculées selon les Équations 3.7 et 3.8.

$$\mathrm{Se} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{3.7}$$

$$\mathrm{Sp} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}} \tag{3.8}$$

À partir de ces indices, il est également possible d'obtenir les courbes ROC, mesurant la performance d'un classificateur (voir sous-section 2.2.9 - Analyses statistiques, page 53). Il est possible de calculer l'aire sous cette courbe afin d'obtenir une AUC représentant la performance de classification.

A la fin du processus d'évaluation, k modèles sont générés et évalués. Les performances finales du processus sont donc obtenues en moyennant les k performances des différents modèles (moyenne μ et écart-type σ).

3.4 Sélection de caractéristiques

3.4.1 Principe

L'augmentation du nombre de caractéristiques qui modélisent un problème introduit des difficultés à plusieurs niveaux comme la complexité ou l'augmentation du temps de calcul. Pour améliorer les performances des analyses, il peut être intéressant de réduire la dimensionnalité initiale F_i d'une base de données, notamment dans le cas où la taille de la base d'apprentissage n'est pas très grande. La sélection de caractéristiques (en anglais "feature selection") est une méthode de réduction de la dimension consistant à trouver une représentation des données initiales dans un espace plus réduit. Cette réduction est réalisée par sélection des caractéristiques les plus pertinentes du phénomène étudié.

Il existe globalement 3 types de méthodes de sélection de caractéristiques [Chandrashekar and Sahin 2014] : la méthode filtrante ("filter"), la méthode enveloppante ("wrapper") et la méthode intégrée ("embedded").

3.4.2 Les méthodes filtrantes

La méthode filtrante a été l'une des premières méthodes utilisées en sélection de caractéristiques. Elle évalue la pertinence d'une caractéristique selon des mesures qui reposent sur les propriétés des données d'apprentissage, comme la corrélation entre une caractéristique et la prédiction. Cette méthode est davantage considérée comme une étape de prétraitement (filtrage) avant la phase d'apprentissage (voir Figure 3.8). Un avantage est le fait que l'évaluation se fait généralement indépendamment du classificateur [John et al. 1994]. Il est néanmoins nécessaire de spécifier une valeur de seuil ou un nombre fixe de caractéristiques pour obtenir le sous-ensemble final.

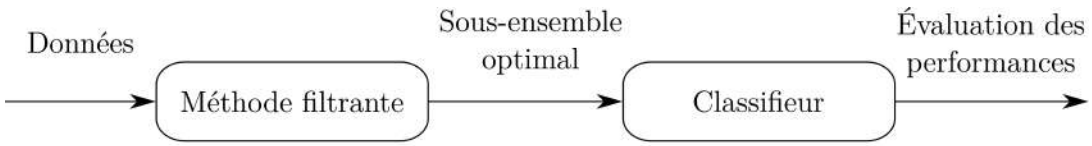


FIGURE 3.8 – Principe de la méthode filtrante de sélection de caractéristiques. Chaque caractéristique issue d'un ensemble de données est classée par la méthode filtrante selon un critère d'évaluation. Les meilleures caractéristiques sont ensuite sélectionnées en fonction d'un seuil défini au préalable par l'utilisateur.

Dans l'article de Guyon et al. [Guyon and Elisseef 2003] sont présentés plusieurs critères d'évaluation retrouvés dans la littérature, tels que le critère de corrélation, le critère de Fisher ou le "Signal-to-Noise Ratio".

La méthode *"RELevance In Estimating Features"* (RELIEF) [Kira and Rendell 1992] [Gilad-Bachrach et al. 2004]) est considérée comme l'une des méthodes filtrantes les plus efficaces. L'algorithme prend en entrée N observations, comprenant les caractéristiques et les labels, et va estimer un vecteur de poids pour chaque caractéristique. Puis, une observation est sélectionnée aléatoirement et comparée avec 2 plus proches voisins : un premier de la même classe et un de classe différente. En fonction de ces 3 observations, le vecteur de poids va être mis à jour. Ce processus est répété m fois tel que m est un paramètre prédéfini par l'utilisateur.

Plus récemment, il a été proposé la méthode *"Feature Assessment by Sliding Thresholds"* (FAST) [Chen and Wasikowski 2008], basée sur la valeur de l'AUC des courbes

ROC de chaque caractéristique, en faisant glisser les valeurs de seuil dans un espace unidirectionnel. Guyon et al. [Guyon and Elisseef 2003] ont fait remarquer que les caractéristiques non-pertinentes peuvent être associées à d'autres caractéristiques afin d'obtenir une combinaison plus utile que les caractéristiques séparément. Les auteurs ont alors étudié la pertinence des sous-ensembles de caractéristiques, en opposition aux méthodes de sélection basées sur un pouvoir discriminant individuel. La méthode de sélection *"Kernel Class Separability"* (KCS) classe les sous-ensembles de caractéristiques en fonction de la séparabilité des classes [Wang 2008] [Zhang et al. 2011]. Elle est décrite comme étant robuste aux échantillons de petites tailles et à la présence de bruit.

Le principal avantage des méthodes de filtrage est leur efficacité calculatoire et leur robustesse face au sur-apprentissage [Chandrashekar and Sahin 2014]. Malheureusement, ces méthodes ne tiennent pas compte des interactions entre caractéristiques et tendent à sélectionner des caractéristiques comportant de l'information redondante plutôt que complémentaire [Guyon and Elisseef 2003]. De plus, ces méthodes sont indépendantes des méthodes de classification qui suivent la sélection [Kohavi and John 1997].

3.4.3 Les méthodes enveloppantes

Les méthodes enveloppantes, à la différence des méthodes filtrantes, intègrent un classifieur au cœur de leur processus (voir Figure 3.9). Le but est de prendre en compte les performances de classification au cours de la sélection en utilisant un algorithme d'apprentissage [Kohavi and John 1997]. La phase d'apprentissage est divisée en 2 parties : une partie apprentissage et une de validation pour tester le sous-ensemble des caractéristiques sélectionnées. Il a été montré une amélioration des performance des méthodes enveloppantes par rapport à certaines méthodes filtrantes [Li and Guo 2008] [Huang et al. 2008].

Dans la littérature, la méthode *"Sequential Forward Selection"* (SFS) [Whitney 1971] [Theodoridis and Koutroumbas 2010] et la méthode *"Sequential Forward Floating Selection"* (SFFS) [Pudil et al. 1994] [Theodoridis and Koutroumbas 2010] sont 2 algorithmes représentatifs des méthodes enveloppantes de sélection des caractéristiques. SFS est un algorithme de recherche assez simple. Le but est de sélectionner le meilleur sous-ensemble

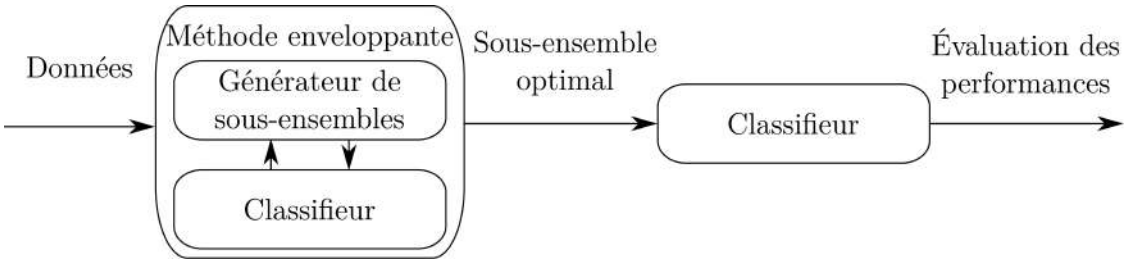


FIGURE 3.9 – Principe de la méthode enveloppante de sélection de caractéristiques. Chaque sous-ensemble de caractéristiques généré est injecté dans un classifieur. Le sous-ensemble optimal est celui présentant les meilleures performances de classification.

de caractéristiques en ajoutant les caractéristiques une à une [Tabesh et al. 2007]. De manière itérative, la meilleure caractéristique est ajoutée à un ensemble initialement vide selon un critère prédéfini. Le risque avec cette méthode est qu'une caractéristique puisse être piégée dans une solution non-optimale. Pour résoudre ce problème, la méthode SFFS effectue des mesures d'exclusion après chaque étape d'inclusion [Theodoridis and Koutroumbas 2010]. Malgré tout, elle traite toujours les caractéristiques individuellement. Plus récemment, Mi et al. [Mi et al. 2015] ont proposé une méthode nommée *"Hierarchical Forward Selection"* (HFS). Cette méthode recherche un sous-ensemble de caractéristiques et l'évalue à l'aide d'un classifieur SVM. Cette dernière a montré de bons résultats appliqués au domaine médical à l'aide de connaissances *a priori* lors de l'étude prédictive d'une cohorte de patients atteints d'un cancer du poumon. Malgré tout, le risque d'éliminer des caractéristiques apportant des informations complémentaires est encore présent.

La complexité de l'algorithme d'apprentissage rend les méthodes enveloppantes très coûteuses en temps de calcul. Pour éviter ou diminuer le sur-apprentissage, le mécanisme de validation croisée est fréquemment utilisé. De plus, l'évaluation des caractéristiques se fait par un seul classifieur lors de la sélection. Chaque classifieur a ses spécificités. C'est pourquoi, le sous-ensemble sélectionné dépend toujours du classifieur utilisé.

3.4.4 Les méthodes intégrées

A la différence des méthodes précédentes, les méthodes intégrées incorporent une étape de sélection de caractéristiques lors du processus d'apprentissage (voir Figure 3.10). Contrairement aux méthodes enveloppantes, l'utilisation de bases de validation n'est pas nécessaire pour tester le sous-ensemble de caractéristiques sélectionnées. Ainsi, les mé-

thodes intégrées peuvent se servir de tous les exemples d'apprentissage pour établir un modèle. Cela constitue un avantage qui peut améliorer les résultats. Un autre avantage de ces méthodes est leur plus grande rapidité par rapport aux autres méthodes parce qu'elles évitent que le classificateur recommence de zéro pour chaque sous-ensemble étudié.

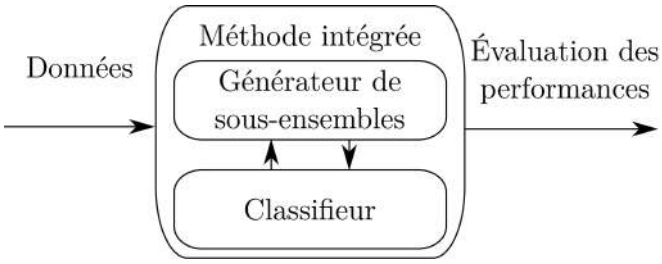


FIGURE 3.10 – Principe de la méthode intégrée de sélection de caractéristiques. Chaque sous-ensemble de caractéristiques généré à partir des données initiales est injecté dans un classifieur. Le sous-ensemble optimal est sélectionné en fonction des performances de classification.

La méthode *"Classification And Regression Tree"* (CART) possède un mécanisme intégré pour effectuer la sélection des caractéristiques [Breiman et al. 1984]. Pour diviser un nœud en deux, une caractéristique choisie par certaines règles est utilisée pour différencier les observations. Les RF, introduites dans [Breiman 2001], combinent un certain nombre d'arbres de décision construits sur différentes parties d'un même ensemble d'apprentissage. Chaque nœud d'un arbre est divisé en 2 en utilisant un sous-ensemble aléatoire de caractéristiques, dans le but de réduire le problème de sur-apprentissage d'un arbre. Guyon et al. [Guyon et al. 2002] ont proposé une méthode utilisant le SVM et basée sur une élimination récursive des caractéristiques *"Recursive Feature Elimination"* (RFE), nommée SVM-RFE. A partir de l'ensemble initial, la méthode SVM-RFE élimine progressivement la caractéristique la moins prometteuse, dont le retrait minimise la perte d'information. Cette étape est réalisée de manière itérative jusqu'à obtenir un sous-ensemble d'une taille définie au préalable.

3.5 Apprentissage et sélection des caractéristiques en imagerie médicale TEP en oncologie

3.5.1 Introduction

On constate ces dernières années une augmentation du nombre d'articles concernant l'utilisation de l'apprentissage automatique en médecine, et notamment en imagerie médicale et en radiomique. Plusieurs raisons expliquent que ces méthodes soient bien adaptées à cette problématique.

Tout d'abord, les raisonnements de diagnostic et de prise en charge médicale du patient sont similaires aux méthodes d'apprentissage automatique. En effet, les décisions médicales s'apparentent aux arbres de décisions. Les différentes explorations du patient fournissent de nombreuses informations qui sont ensuite analysées par le médecin à partir de leurs connaissances basées sur un apprentissage (formation universitaire, expérience professionnelle, guides des bonnes pratiques, etc). Cependant, plus le nombre d'informations augmente, plus l'analyse est complexe à réaliser, même pour un expert. De plus, comme ces caractéristiques sont issues du vivant, il existe une importante variabilité d'un individu à l'autre et la frontière séparant le sain et le pathologique n'est pas toujours très nette. Les approches d'apprentissage automatique permettent de répondre à ces problématiques.

De plus, les informations disponibles en médecine sont *a priori* très importantes, sauf en cas de pathologies rares. En effet, chaque patient possède un dossier médical tenu rigoureusement à jour lors de chaque examen (biologiques, cliniques, d'imagerie). Ainsi, une grande quantité d'informations est disponible pour être analysée par un algorithme d'apprentissage automatique. Par ailleurs, et ce de façon rétrospective, l'ensemble des données d'apprentissage peut être labellisé, par exemple à visée pronostique par la connaissance de la survie ou du décès du patient, permettant l'utilisation des méthodes d'apprentissage supervisées. Il existe tout de même quelques problèmes engendrés par la gestion de ces bases de données. Pour être en accord avec les droits du patient, il est nécessaire que ce dernier ait consenti à l'utilisation de ces données personnelles de manière

libre et éclairée. En outre, l'acquisition de données multicentriques peut créer une hétérogénéité des caractéristiques causée par la variation des protocoles des centres de soins. Néanmoins, dans le domaine médical, les caractéristiques pronostiques ou prédictives se doivent d'être robustes pour intégrer les bonnes pratiques médicales.

Dans le but d'étudier de grandes bases de données, les chercheurs utilisent majoritairement des bases rétrospectives. Elles permettent d'obtenir un nombre de patient plus important que les études prospectives et sont moins couteuses. En revanche, les données prospectives sont de bien meilleure qualité pour répondre à la question clinique posée. Lorsqu'il n'est pas possible de réaliser un essai clinique prospectif, il est nécessaire de respecter une certaine homogénéité des patients sélectionnés dans la base de données crée rétrospectivement (pathologie, traitement, acquisition des images, etc).

L'apprentissage automatique en imagerie médicale a mené à l'apparition d'outils d'*Aide aux Diagnostics ou "Computed Aided Diagnosis"* (CAD). Ces outils sont, par exemple, utilisés dans le dépistage du cancer du sein. Il a été montré que les techniques de CAD pouvaient mieux classer les anomalies détectées en lésions bénignes ou malignes qu'un radiologue non-spécialisé en mammographie [Roehrig and Castellino 1999]. L'apprentissage automatique en imagerie médicale est également utilisé pour la segmentation automatisées, notamment en cancérologie. En effet, il est nécessaire de réaliser une segmentation précise des structures anatomiques ou fonctionnelles à partir d'images médicales [Kerhet et al. 2009] [Kerhet et al. 2010]. On peut citer par exemple le cas de la radiothérapie, mais également l'établissement des critères PERCIST en cancérologie comme nous l'avons vu précédemment (voir 1.3.2.1 - Apports cliniques, page 24). Le concept de radiomique est également un domaine où il est intéressant d'utiliser les algorithmes d'apprentissage automatique en raison du nombre élevé de caractéristiques multimodales prises en compte.

Dans cette section est réalisé un état de l'art de l'utilisation de l'apprentissage automatique et de la sélection de caractéristiques dans le domaine de la radiomique. Cependant, au vu du nombre important de publications présentes dans la littérature, nous nous sommes focalisés sur l'utilisation de ces outils lors de l'étude radiomique des images TEP au FDG en oncologie afin de mettre en évidence la diversité des algorithmes utilisés.

3.5.2 Utilisation de la régression logistique

L'association de plusieurs types de caractéristiques radiomiques (caractéristiques du 1er ordre et de forme) a été étudiée dans l'article d'El Naqa et al. [El Naqa et al. 2009] pour la prédiction de la réponse au traitement. Après une analyse combinant ces caractéristiques par régression logistique, les auteurs ont proposé un modèle prédictif applicable aux cancers ORL combinant le V_{90} et "l'extent" menant à un coefficient de corrélation de Spearman de 0,87 avec la réponse au traitement et une AUC de 1. Par régression logistique, ces mêmes auteurs ont réalisé une combinaison de V_{10-90} et de l'énergie (matrice GLCM) pour la prédiction de la réponse au traitement chez des patientes atteintes d'un cancer du col de l'utérus.

Beukinga et al. [Beukinga et al. 2017] ont proposé une méthode pour améliorer la prédiction de la réponse au traitement des patients atteints d'un cancer de l'œsophage traités par RCT en se basant sur les caractéristiques (1er ordre et textures) extraites de l'imagerie TEP au FDG. Une base de données de 97 patients a été étudiée. Leur méthode consiste en une étape de sélection des caractéristiques par l'approche *"Least Absolute Shrinkage and Selection Operator"* (LASSO) et une classification par une régression logistique. La méthode LASSO a sélectionné les caractéristiques : histologie, stade T, LRLGE de l'image TEP et le RPr (matrice GLSZM) de l'image TDM. La régression logistique à partir de ces caractéristiques a montré, après validation, une AUC de 0,740 par rapport à 0,540 pour le modèle par régression logistique utilisant uniquement le SUV_{max}.

3.5.3 Utilisation des fonctions de croyance

Lian et al. [Lian et al. 2016] ont proposé l'évaluation de la réponse au traitement des patients sur 3 bases de données : une cohorte de 25 patients atteints d'un cancer des poumons non à petites cellules traités par radiothérapie et RCT, 36 patients atteints d'un cancer de l'œsophage de type carcinome épidermoïde traités par RCT et une cohorte de 45 patients avec un lymphome de type B. Dans chaque cohorte les patients ont bénéficié d'un examen d'imagerie TEP au FDG, d'où ont pu être extrait un grand nombre de caractéristiques images (1er ordre, 2ème ordre et temporelle). Les auteurs ont utilisé leur propre

méthode de sélection des caractéristiques (*i*EFS) basée sur la théorie de Dempster-Shafer [Shafer 1976]. Leurs résultats montrent une précision de classification de 94 % pour le cancer des poumons, de 83 % pour le cancer de l'œsophage et de 92 % pour le lymphome.

3.5.4 Utilisation des SVM

En 2010, Jayasurya et al. [Jayasurya et al. 2010] ont comparé 2 méthodes d'apprentissage automatique (approche bayésienne et SVM) afin de prédire la survie à 2 ans d'une base de données de patients atteints d'un cancer du poumon. De plus, une étape de sélection des caractéristiques est réalisée à l'aide d'une méthode filtrante : EM. Les modèles ont été entrainés à partir de 322 observations et testés sur 3 bases indépendantes de 35, 47 et 33 patients. Trois caractéristiques ont été détectées comme pronostiques de la survie à deux ans : la taille du volume tumoral, le stade OMS et le nombre de ganglions lymphatiques atteints sur un TEP. Le modèle issu de l'approche bayésienne basé sur ces caractéristiques, a présenté une AUC de 0,77, 0,72 et 0,7 respectivement pour les 3 bases. Le modèle SVM, basé sur les mêmes caractéristiques, a présenté des performances globales plus faible (AUC = 0,71, 0,68 et 0,69). Cependant, en écartant les patients ayant des données manquantes, les performances entre les 2 méthodes étaient similaires.

Afin d'améliorer l'évaluation diagnostique des ganglions lymphatiques médiastinaux dans le cas de patients atteints d'un cancer du poumon, Gao et al. [Gao et al. 2015] ont développé une nouvelle approche basée sur les SVM. Cette méthodologie est comparée à l'analyse visuelle de référence. Les auteurs ont étudié l'utilisation du SVM sur une cohorte de 132 patients atteints d'un cancer des poumons. Ces patients ont réalisé un examen TDM/TEP au FDG d'où ont pu être extraites 534 caractéristiques (512 de la TDM seule et 22 de la TEP au FDG). Ces caractéristiques sont des caractéristiques du 1er ordre (SUV_{max}, SUV_{moy}, diamètre maximal) ainsi que des caractéristiques de texture extraites des matrices de cooccurrence. A partir de cette base de données, 3 classifieurs SVM ont pu être construits : le premier est basé sur les caractéristiques TDM (SVM_{TDM}), un autre à partir des caractéristiques TEP (SVM_{TEP}) et un dernier à partir des caractéristiques combinées ($SVM_{TDM/TEP}$). Un noyau gaussien (RBF) est utilisé pour l'algorithme SVM et plusieurs paramètres (c et σ) ont été étudiés. Chaque classifieur est évalué par mesure de l'AUC des

courbes ROC. SVM_{TEP} présente une valeur d'AUC de 0,689, SVM_{TDM} possède une valeur d'AUC de 0,579 et $SVM_{TDM/TEP}$ une valeur d'AUC de 0,685. Le SVM_{TEP} est celui présentant les meilleurs résultats. En comparaison, le SUV_{max} et le diamètre tumoral maximal ont une valeur d'AUC de 0,652 et 0,684, respectivement.

Déjà abordé précédemment, Mi et al. [Mi et al. 2015] ont proposé une méthode de sélection des caractéristiques basée sur l'utilisation des SVM, intitulée HFS. Les auteurs ont étudié 79 caractéristiques cliniques et de l'imagerie TEP chez 25 patients atteints d'un cancer du poumon et traités par RCT. Leurs résultats ont montré que la combinaison du SUV_{max} avec deux autres caractéristiques issues de la matrice GLSZM était celle apportant les meilleurs résultats prédictifs. Ainsi, après classification SVM les performances allaient jusqu'à 100 % de bonne classification.

Mu et al. [Mu et al. 2015] ont classé 42 patientes atteints de cancer du col de l'utérus en 2 groupes (stade précoce et avancé) à l'aide de l'imagerieTEP. Grâce à la classification automatique par SVM et la caractéristique RPr, détectée comme étant la plus discriminante, le modèle a présenté une précision de 88 %.

3.5.5 Utilisation des réseaux de neurones

En 2014, Toney et al. [Toney and Vesselle 2014] ont étudié les performances obtenues par des classifieurs de type ANN construit à partir de caractéristiques cliniques et extraites de l'imagerie TEP pour déterminer le stade N de la maladie. Pour cela une base de données de 133 patients présentant un cancer pulmonaire non à petites cellules a été étudiée. Une précision de 99 % a été obtenue à l'aide du classifieur permettant de séparer les ganglions lymphatiques inflammatoires malins et bénins malgré leurs ressemblances.

Ypsilantis et al. [Ypsilantis et al. 2015] ont étudié l'utilisation de plusieurs méthodes d'apprentissage (ANN, SVM, RF et "Gradient Boosting"), dans le but de prédire la réponse au traitement par chimiothérapie néoadjuvante chez 107 patients présentant un cancer de l'œsophage. Pour cela, 103 caractéristiques radiomiques ont été extraites des examens TEP au FDG : 18 caractéristiques du 1^{er} ordre et 85 caractéristiques de texture (GLCM, GLRLM, GLSZM, GLDM et fractale). Leurs résultats montrent que l'ANN est la méthode d'apprentissage automatique la plus performante pour la prédiction de la réponse au trai-

tement avec une précision de bonne classification de 73,4 % ± 5 %, suivi du "Gradient Boosting" et des RF avec 66,8 % ± 6 % et 65,7 % ± 6 %, respectivement. Le SVM est la méthode la moins efficace avec 60,5 % ± 8 %. Néanmoins, ces méthodes permettent l'obtention de meilleures résultats que ceux obtenus par seuillage simple du SUV_{max} avec une précision de 41,0 % ± 5 %.

Wang et al. [Wang et al. 2017] ont comparé les résultats obtenus par une méthode d'apprentissage profond (le CNN) et 4 méthodes d'apprentissage automatique « classiques » (ANN, SVM, RF et AdaBoost) dans le but de classifier 1397 ganglions lymphatiques de 168 patients atteints d'un cancer du poumons non à petites cellules. Pour cela des caractéristiques TEP/TDM ont été étudiées. Pour chaque approche classique, différentes caractéristiques d'entrée ont été comparées pour sélectionner les sous-ensembles optimaux spécifiques à chacun d'entre eux. Les CNN ont présenté un taux de bonne classification de 86 % et une AUC de 0,910. Il n'y a pas de différence significative entre les résultats des CNN et les méthodes classiques, où les plus mauvaises performances sont obtenues par l'approche ANN (taux de bonne classification de 81 %) (voir Figure 3.11). Enfin, les 5 méthodes d'apprentissage automatique ont présenté des sensibilités plus élevées, mais des spécificités inférieures à celles des médecins.

3.5.6 Utilisation des forêts aléatoires

Dans l'article de Wang et al. [Wang et al. 2017], comparant différentes méthodes de classification associées à différents sous-ensembles de caractéristiques, la méthode des RF présente de bonnes performances avec un taux de bonne classification 85 %, une sensibilité de 82 % et une spécificité de 89 % (respectivement, 82, 73 et 90 % pour les experts) (voir Figure 3.11).

On trouve dans la littérature d'autres articles faisant ressortir les bonnes performances du RF lors de comparaison de méthodes. Ainsi, dans [Fernández-Delgado et al. 2014], les auteurs ont étudié 179 classifieurs provenant de 17 familles différentes (analyse discriminante, approche bayésienne, ANN, RF, SVM, ...). Ces classifieurs ont été testés sur 121 bases de données comprenant entre autres des données médicales comme des analyses de la thyroïde, de la maladie de Parkinson ou des images PET ayant chacune des particu-

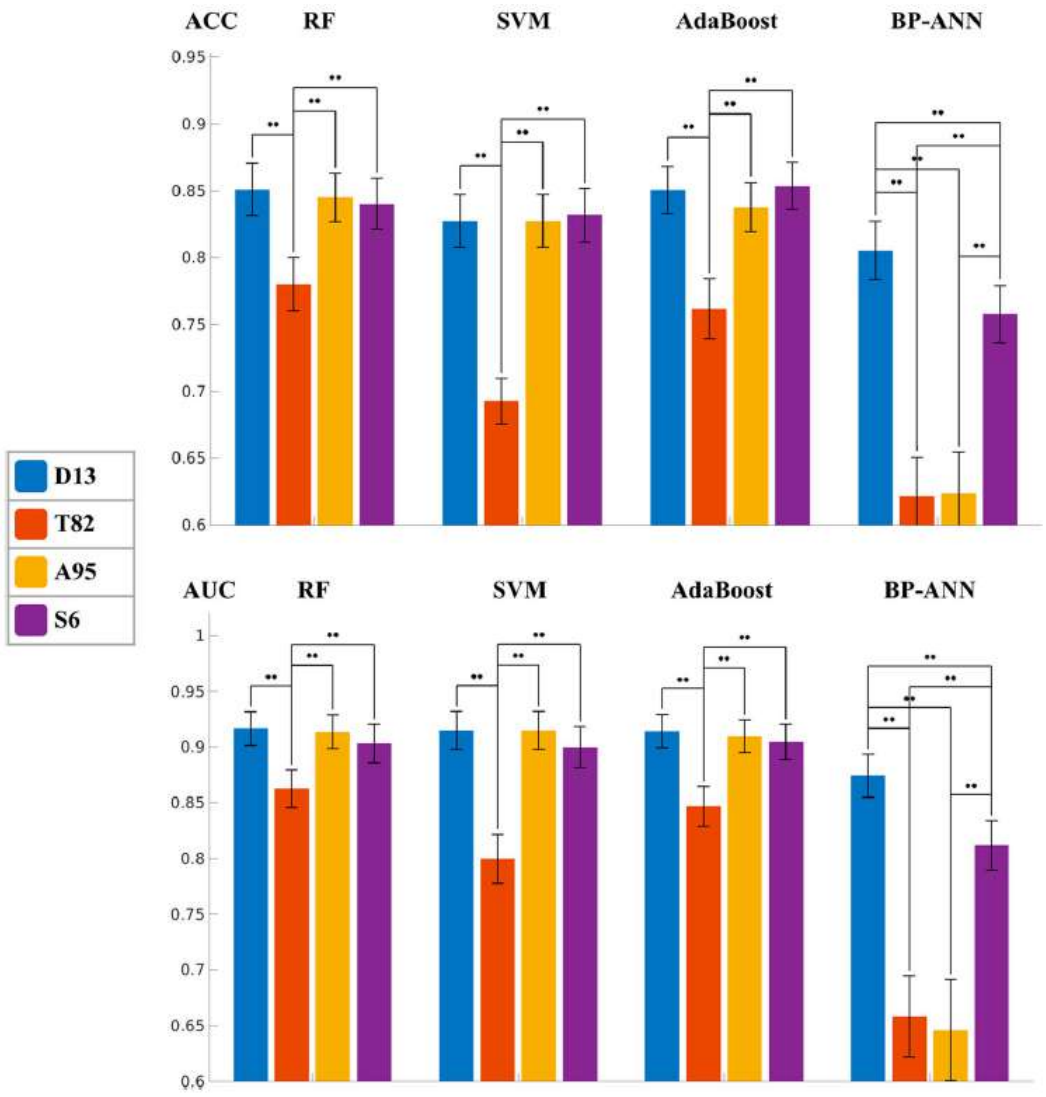


FIGURE 3.11 – Comparaison des résultats obtenus avec 4 méthodes de classifieurs (ANN, SVM, RF et AdaBoost) et 4 ensembles de caractéristiques différents (D13, T82, A95 et S6), basés sur les AUC et les taux de bonne classification moyens après validation croisée kFCV ($k = 10$). Les barres d'erreur indiquent un intervalle de confiance de 95 %. La valeur de p entre les différents ensembles de caractéristiques est tracée en tant que pont et étoiles, où deux étoiles signifie $p < 0,05$ après corrections de Bonferroni et FDR ("False Discovery Rate"), et une étoile signifie $p < 0,05$ uniquement après correction de FDR [Wang et al. 2017].

larités différentes. Leurs résultats ont montré que l'algorithme des RF était la méthode la plus performante. En effet, elle atteint 90 % de taux de bonne classification dans plus de 84 % des cas, atteignant jusqu'à 94 % de bonne classification.

Parmar et al. [Parmar et al. 2015] ont étudié la survie globale à 2 ans de 464 patients atteints d'un cancer du poumon. Les auteurs ont utilisé 12 méthodes d'apprentissage automatique et 14 méthodes filtrantes de sélection des caractéristiques. Ainsi, plusieurs dizaines de combinaisons ont pu être testées pour leur performance de prédiction et leur stabilité face à la perturbation des données. Pour cela, 440 caractéristiques radiomiques ont été extraites des examens TDM. Ces caractéristiques sont divisées en 4 groupes : les caractéristiques du 1^{er} ordre, de forme, de texture (GLCM, GLRLM), ainsi que des caractéristiques d'ondelettes. Concernant l'étude de la stabilité, les arbres de décision et le bootsting sont les 2 méthodes présentant une bonne résistance aux perturbations des données, alors que la méthode bayésienne présente une très faible stabilité. Concernant l'étude des performances des classifieurs, leurs résultats montrent que la méthode par forêts aléatoires est celle donnant les meilleures performances avec une AUC moyenne de 0,66 ± 0,03, alors que la méthode par arbres de décision donne les plus mauvais résultats (AUC = 0,54 ± 0,04). Concernant les méthodes de sélection des caractéristiques, la méthode par test de Wilcoxon décrite précédemment (voir sous-section 2.2.9 - Analyses statistiques, page 53), a montré les meilleures performances prédictives (AUC moyenne de 0,65 ± 0,02), alors que les méthodes Chi^2 et "Conditional Infomax Feature Extraction" donnent les moins bons résultats (AUC moyenne de 0,60 ± 0,03, et 0,60 ± 0,04, respectivement). D'après Parmar et al. [Parmar et al. 2015] le modèle des RF est donc la méthode de classification à privilégier en radiomique dans le cadre d'étude de la survie. De plus, cette méthode est présentée comme étant stable avec un ratio entre écart-type et moyenne de l'AUC de 3,52.

Giorgetti et al. [Giorgetti et al. 2016] ont étudié l'utilisation des RF afin de détecter les caractéristiques pronostiques parmi celles extraites de plusieurs examens TEP au FDG d'une base de données de 45 patients atteints d'un cancer de l'œsophage. Leur méthode a obtenu un taux d'erreur de 36 % et a identifié la variation du TLG (à 40 % du SUV_{max}) comme le facteur pronostique le plus important (importance relative de 100 %). De plus,

le stade T (17 %), N (5 %) et M (5 %), le stade de la maladie, l'histologie du cancer (11 %), le TLG (5 %) à la fin du traitement et la variation de TLG (17 % -5 %) ont également été associés à la survie du patient. La sélection de ces caractéristiques cliniques et d'images confirment les conclusions déjà observées dans la littérature et présenté précédemment (voir Tableaux dans la sous-section 1.2.4 - Apport clinique du SUV_{max}, page 16).

En 2016, Roman-Jimenez et al. [Roman-jimenez et al. 2016] ont proposé une méthode afin de prédire la récidive d'une lésion cancéreuse chez des patientes atteintes d'un cancer du col de l'utérus. Pour cela, 1026 caractéristiques extraites des examens TEP au FDG pré- et per-traitement ont été étudiées dans une base de données de 53 patientes. Deux méthodes ont été étudiées : une régression logistique univariée et les RF. Avec la régression logistique, 36 caractéristiques étaient prédictives de la récidive de la maladie à 3 ans ($p < 0,01$) avec une AUC allant de 0,720 à 0,830. Avec RF, le taux d'erreur de classification obtenu en utilisant la totalité des caractéristiques extraites a été de 26,4 %, avec une AUC = 0,720. Cependant, après élimination récurrente des caractéristiques les moins importantes, menant à un sous-ensemble de seulement 9 caractéristiques, le taux d'erreur du classificateur RF était de 13,2 %, avec une AUC = 0,900. Les résultats suggèrent que les examens TEP pré- et per- traitement fournissent des informations significatives pour prédire la récidive tumorale. Le classificateur RF est capable de gérer un très grand nombre de caractéristiques extraites et permet la combinaison des caractéristiques les plus pronostiques pour améliorer la prédiction.

3.6 Conclusion

Dans ce chapitre, nous avons présenté les différentes méthodes d'apprentissage automatique et de sélection des caractéristiques, ainsi que leurs utilisations dans le domaine médical. Ces méthodes présentent l'avantage d'être robuste à la présence d'importantes quantités de caractéristiques étudiées, ainsi qu'à leur relation non-linéaire avec la sortie à prédire (i.e. répondeur ou non-répondeur au traitement). Elles sont donc particulièrement adaptées aux problématiques de la radiomique car elles permettent une meilleure puissance discriminante lors de l'analyse de plusieurs dizaines de caractéristiques face

aux statistiques classiques. Cela dit, ces méthodes nécessitent l'utilisation de bases de données importantes ce qui est une limite dans notre domaine.

Les résultats encourageants de la littérature présentés dans ce chapitre nous ont poussés à rechercher la meilleure démarche de sélection de caractéristiques basée sur une méthode d'apprentissage automatique afin de réaliser des études prédictives de la réponse au traitement et pronostiques en cancérologie. Notre problématique concerne la classification supervisée en 2 classes. Il convient donc de privilégier une méthode d'apprentissage automatique supervisée robuste au faible nombre de données d'apprentissage et présentant un déséquilibre potentiel entre les 2 classes étudiées. En effet, l'existence d'une classe minoritaire par rapport à l'autre pourrait fausser l'apprentissage d'un modèle.

D'après leurs propriétés et les résultats rencontrés dans la littérature [Parmar et al. 2015], les forêts aléatoires présentent les qualités requises pour répondre à notre problématique. En effet, cet algorithme d'apprentissage automatique est supervisé et extrêmement modulable pour être ajusté à un problème particulier. Cette méthode est également robuste [Parmar et al. 2015] et ne nécessite pas une importante base de données d'apprentissage. De plus, l'algorithme des RF est capable de gérer des données multimodales binaires, discrètes ou continues. Enfin, la consommation de mémoire importante lors de l'apprentissage n'est pas un facteur limitant en médecine, une fois le modèle généré celui-ci peut être utilisé pour plusieurs nouvelles observations. Pour toutes ces raisons, nous nous sommes orientés vers l'utilisation des RF pour la mise en place d'une méthode de sélection de caractéristiques optimales pour des études prédictives de la réponse au traitement et pronostiques en imagerie TEP au FDG. Notre méthode est présentée dans le chapitre suivant.

Chapitre 4

Méthodes proposées de sélection des caractéristiques radiomiques

Sommaire

4.1 Introduction

Comme présenté dans le chapitre précédent, les forêts aléatoires présentent des résultats intéressants en radiomique dans la littérature [Parmar et al. 2015]. C'est pourquoi nous nous sommes orientés vers cette approche pour la mise en place d'une méthode de sélection de caractéristiques. Les caractéristiques sélectionnées permettront de générer 2 modèles. Un premier modèle pronostique tente de prédire la survie des patients et un deuxième, de prédire la réponse au traitement (voir Figure 3.1 et 3.7).

De plus, nous avons décidé d'interpréter la pertinence des caractéristiques en termes de groupes et non individuellement. En effet, nous considérons qu'une caractéristique unique n'est pas suffisamment puissante pour permettre la prédiction d'un événement aussi complexe qu'est l'évolution tumorale. Au contraire, le regroupement de plusieurs caractéristiques multi-modales complémentaires pourrait améliorer la prédiction. Cette approche a un fonctionnement contraire des statistiques univariées largement rencontrées dans la littérature médicale tel que le test de Mann-Whitney pour les études prédictives et l'étude des courbes de survie de Kaplan-Meier pour les études pronostiques.

Bien que l'algorithme des RF puisse être utilisé sans méthode de sélection des caractéristiques, les résultats de la littérature, la taille des données limitée et le risque de sur-apprentissage, ainsi que nos premières expérimentations nous ont poussées vers l'ajout d'une étape de sélection. En effet, comme abordé aux chapitres précédents, il est possible d'extraire plusieurs dizaines de caractéristiques pour chaque patient. Cependant, il est compliqué d'obtenir des bases de données suffisamment importantes pour étudier toutes ces caractéristiques. La génération d'un modèle par apprentissage automatique dans le cas où le nombre de caractéristiques étudiées est inférieur au nombre d'observations présente un risque de sur-apprentissage. De plus, comme présenté dans la littérature [Orlhac et al. 2014], certaines caractéristiques radiomiques sont corrélées entre elles, ce qui augmente le nombre de caractéristiques étudiées ce qui pourrait perturber la classification.

De ce fait, nous proposons une méthode composée de deux sélections successives (Figure 4.1). Dans un premier temps, nous proposons l'utilisation d'une méthode de sé-

lection filtrante (voir Section 3.4.2 - Méthode filtrante, page 78) à partir de l'étude des corrélations entre caractéristiques. Dans un second temps, nous proposons l'utilisation d'une méthode de sélection enveloppante (voir Section 3.4.3 - Méthode enveloppante, page 80) basée sur l'algorithme des forêts aléatoires. Pour cette dernière, nous proposons 2 stratégies possibles. La première est basée sur l'estimation du coefficient d'importance des caractéristiques que l'on peut calculer avec l'algorithme des RF. Nous l'avons appelé *"Forest's Coefficient Importance"* (FIC). La seconde est basée sur l'utilisation d'un algorithme génétique [Holland 1992] également intégrant l'algorithme des forêts aléatoires. Nous l'avons appelé *"Genetic Algortihm based on Random Forest"* (GARF).

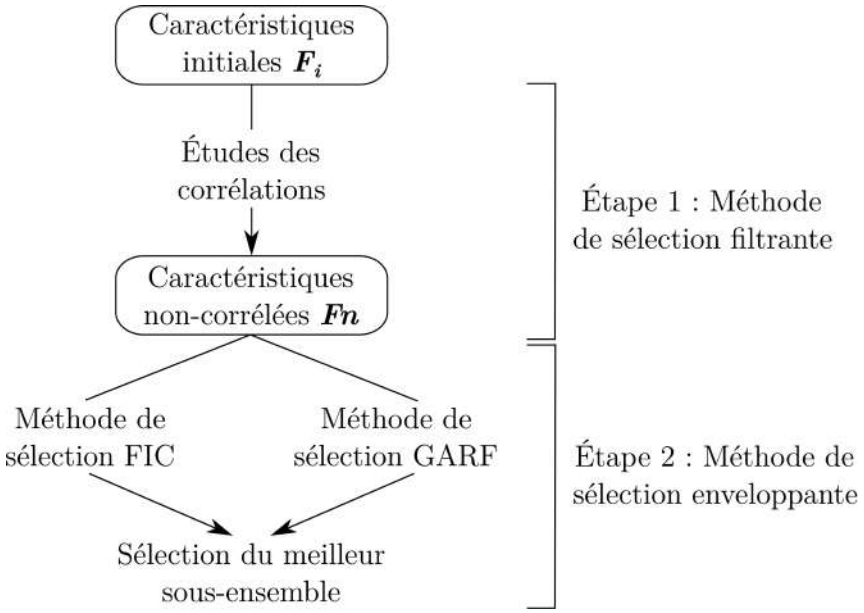


FIGURE 4.1 – Principe des différentes étapes de sélection des caractéristiques de nos méthodes FIC ("Forest's Importance Coefficient") et GARF ("Genetic Algorithm based on Random Forest").

L'algorithme des forêts aléatoires est une méthode d'apprentissage automatique supervisée s'appuyant sur un regroupement d'arbres de décision. Il se place au cœur de notre méthode de sélection des caractéristiques. C'est pourquoi nous allons commencer par la présentation du principe des arbres de décision et des forêts aléatoires. Puis, nous développerons la procédure de notre première étape de sélection filtrante basée sur l'analyse des corrélations. Enfin, nous aborderons les différentes approches envisagées de sélection enveloppantes afin de sélectionner des sous-ensembles de caractéristiques.

4.2 Les arbres de décision

La méthode CART [Breiman et al. 1984] construit des classifieurs par arbres aussi bien en régression qu'en classification. Le principe général est le partitionnement récursif d'une population de façon dyadique afin d'obtenir des sous-partitions optimales pour la prédiction ou la classification.

Pour chaque nœud i d'un arbre, une caractéristique X_j est tirée aléatoirement parmi les F caractéristiques disponibles pour séparer le nœud père en 2 nœuds fils. Un seuil d_i est défini pour scinder la population (Pop) d'un nœud en 2. La répartition des observations o peut être représentée mathématiquement par l'Équation 4.1 suivante :

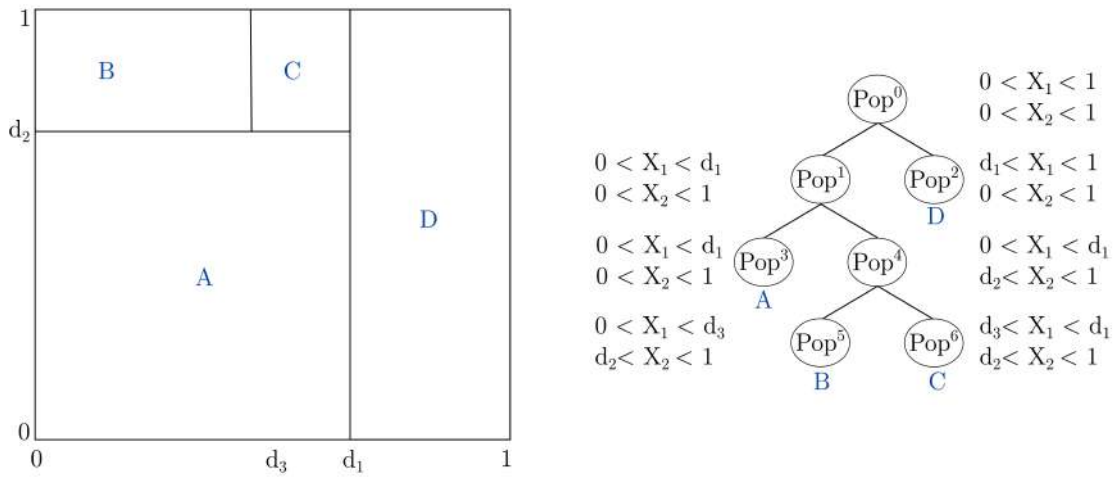
$$\forall o \in \text{Pop} \begin{cases} \text{nœud gauche si } X_j(o) \leq d_i \\ \text{nœud droit sinon} \end{cases} \tag{4.1}$$

Cela signifie que toutes les observations présentant une valeur de la caractéristique X_j inférieure ou égale à d_i iront vers le nœud fils de gauche, et toutes celles avec une valeur X_j supérieure à d_i iront vers le nœud fils de droite. A chaque étape du partitionnement, une partie de l'espace est divisée en deux sous-parties.

Un arbre binaire (ou arbre de décision) est alors naturellement généré au cours de la partition. La Figure 4.2 illustre la correspondance entre une partition dyadique et un arbre binaire. Chacun des nœuds d'un arbre est associé à un sous-ensemble des éléments de la population. Par exemple, la racine de l'arbre est associée à la population initiale, ses deux nœuds fils sont associés aux deux sous-populations obtenues par la première division et ainsi de suite.

Le but est ainsi d'obtenir des nœuds terminaux homogènes. Ainsi, la règle de la division varie en fonction du problème posé :

— En régression, on cherche à minimiser la variance var des nœuds fils. Si un nœud père a été scindé par la caractéristiques X_j en 2 nœuds fils. La variance du nœud fils i est définie comme l'écart à la moyenne ($\overline{X_j}$) de la valeur de X_j des n observations, selon l'Équation 4.2 suivante :



(a) (b)

FIGURE 4.2 – (a) Exemple de la partition dyadique du carré unité et (b) son arbre CART associé d'après [Genuer 2010].

$$var^i = \sum_n (X_j - \overline{X_j})^2 \tag{4.2}$$

— En classification, où le but est la séparation de la population initiale en L classes, on cherche à minimiser la dispersion statistique des nœuds fils, par exemple, par l'indice de Gini [Gini 1921]. Soit $\widehat{p_c^i}$ la proportion d'observations de la classe c sur le nœud i, alors l'indice de Gini du nœud i (G^i) est défini par l'Équation 4.3 suivante :

$$G^i = \sum_{c=1}^{L} \widehat{p_c^i}(1 - \widehat{p_c^i}) \tag{4.3}$$

Rechercher l'indice de Gini minimal revient à chercher l'homogénéité maximale des nœuds obtenus.

L'arbre est construit ainsi progressivement de la racine aux nœuds terminaux, aussi appelés feuilles (voir Figure 4.3). Il est possible de stopper la création de l'arbre développé par l'ajout d'une règle d'arrêt. Une règle classique consiste à ne pas découper les nœuds respectant déjà une certaine homogénéité ou ne contenant qu'un certain nombre d'observations.

On peut caractériser un arbre en fonction de 2 paramètres [Hastie et al. 2009] :

— Son biais qui représente la différence entre la prédiction et la réalité ;

— Sa variance qui représente la différence de prédiction entre 2 itérations.

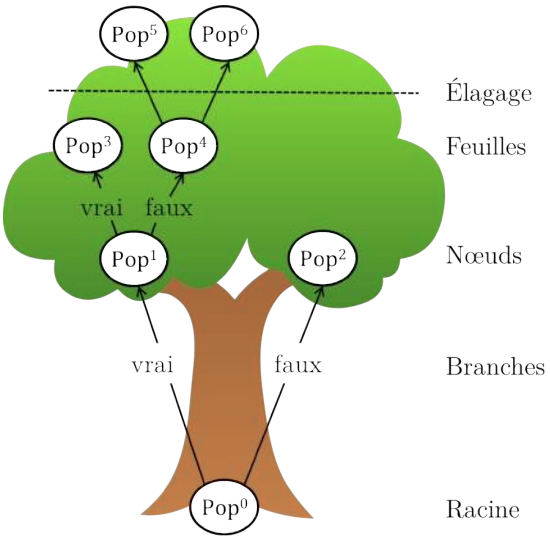


FIGURE 4.3 – Illustration d'un arbre de décision avec ou sans élagage.

Ainsi, plus un arbre est complexe, plus il présente un faible biais et une variance élevée. Par exemple, un arbre constitué uniquement de la racine (peu complexe) présente une faible variance mais un biais élevé, alors que l'arbre entier (complexe) possède une très grande variance et un biais faible. Le but est de trouver un compromis minimisant le biais et la variance.

Il est possible de réaliser une étape d'élagage de l'arbre développé, c'est-à-dire de couper l'arbre entier après un certain nombre de nœuds. Cette étape permet de trouver un compromis entre la variance et le biais.

Au final, à chaque feuille de l'arbre est associée une partition de la population, ainsi qu'un vecteur des caractéristiques représentatif de cette partition. Plusieurs paramètres entre en jeu dans la construction d'un arbre. On retrouve l'utilisation ou non de l'élagage, le choix de la règle de division et de celle d'arrêt.

4.3 Les forêts aléatoires

4.3.1 Le principe de l'apprentissage

Les forêts aléatoires ont également été introduites par Breiman [Breiman 2001]. Cette méthode est basée sur la technique CART et fait partie des méthodes d'ensemble de classifieurs. À partir d'une base de données d'apprentissage (voir Figure 3.7), la construction

d'un modèle basé sur les RF est la suivante (voir Figure 4.4). Soit une base de données d'apprentissage D_{app} représentée par une matrice de taille $(N \times F)$ et les labels correspondant, une multitude T d'arbres de décisions sont construits indépendamment à partir de cette base.

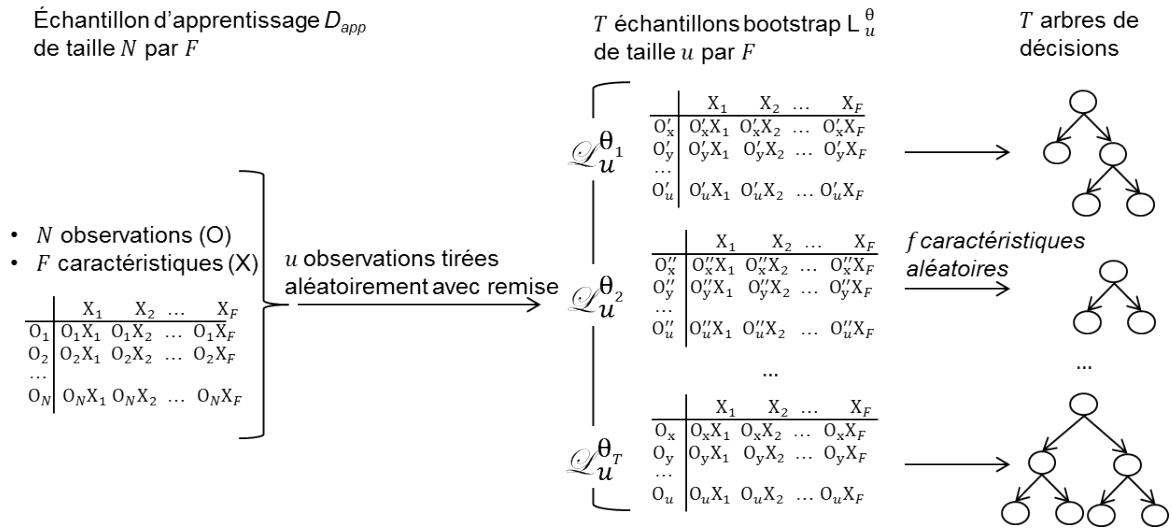
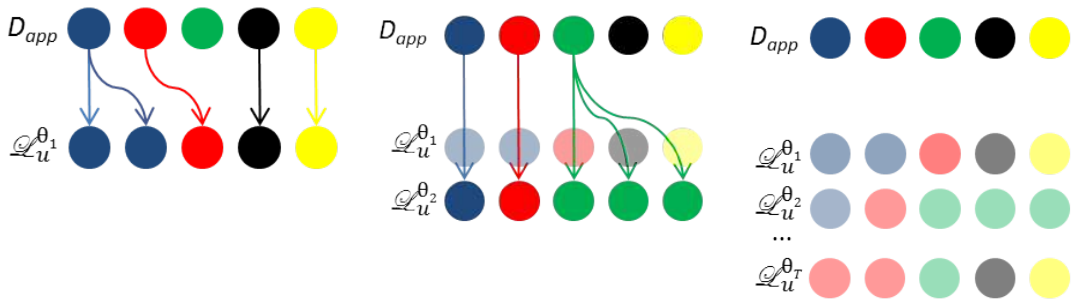


FIGURE 4.4 – Principe de l'algorithme des forêts aléatoires. T échantillons bootstraps contenant u lignes de taille F sont créés par tirage aléatoire sans remise à partir d'un échantillon d'apprentissage D_{app} de taille $N \times F$. Chaque bootstrap est utilisé pour la construction d'un arbre de décision.

Le principe des RF est, tout d'abord, de générer plusieurs échantillons bootstraps $\mathscr{L}_u^{\Theta_1}, \ldots, \mathscr{L}_u^{\Theta_T}$ à partir de D_{app} (voir Figure 4.5). La technique du bootstrap permet de multiplier artificiellement le nombre d'échantillons d'apprentissage à partir d'une population. Un échantillon bootstrap $\mathscr{L}_u^{\Theta\ i}$ est obtenu en tirant aléatoirement u observations $(u \leq N)$ avec remise dans la base de données d'apprentissage D_{app}, chaque observation ayant une probabilité $1/N$ d'être tirée. Il est possible qu'une observation apparaisse plusieurs fois dans l'échantillon bootstrap. Θ est une variable représentant le tirage aléatoire. Chacun des T échantillons bootstraps créés sert comme ensemble d'apprentissage pour un arbre de décision de la forêt.

Pour chaque nœud de l'arbre, f caractéristiques sont tirées aléatoirement sans remise parmi F et de façon équiprobable. Pour chaque nœud d'un arbre, le but est de trouver la caractéristique et son seuil d (voir Équation 4.1) minimisant l'impureté dans les nœuds fils. Cette impureté est mesurée par un critère de partitionnement, tel que le critère de Gini, mais d'autres critères existent, tel que le gain d'information. Ainsi, les f caractéristiques disponibles sont associées à des seuils. Le but est de détecter le couple caractéris-



(a) (b) (c)

FIGURE 4.5 – Illustration de la méthode de création d'échantillons bootstraps. (a) et (b) Une base de données d'apprentissage D_{app} de N observations sert de base à la création (c) des T échantillons bootstraps $\mathscr{L}_u^{\Theta_T}$ obtenus en tirant aléatoirement u observations avec remise (ici $u = N$).

tique/seuil optimisant la dichotomie du nœud étudié. Il existe des méthodes rendant la fixation du seuil aléatoire [Geurts et al. 2006].

Cette étape est répétée pour tous les nœuds jusqu'à ce que les observations soient correctement séparées en accord avec la vérité terrain (labels Y). Le choix de f est un paramètre fixé au début de la construction de la forêt. Il est identique pour tous les arbres. Une taille de f importante réduit l'aléa entre les arbres, car la probabilité qu'une caractéristique soit présente dans le sous-ensemble étudié augmente. Il n'est donc pas intéressant d'utiliser une valeur de f trop importante, car cela réduit la variabilité des arbres au sein de la forêt. Cependant, il est nécessaire d'augmenter cette valeur dans le cas de l'étude d'une base de données présentant d'avantage de caractéristiques non-pertinentes pour le problème étudié [Genuer 2010].

Il existe donc 2 sources d'aléas dans les RF : l'aléa dû à la création des bootstraps et l'aléa du choix des f caractéristiques pour découper chaque nœud d'un arbre.

4.3.2 Le principe de la classification

La collection d'arbres obtenus est agrégée afin d'aboutir à un modèle de classifieur RF. Dans un cas de classification, chaque arbre t va chercher à prédire le label (i.e. patient non-répondeur - label 0 ou patient répondeur - label 1) d'une nouvelle observation o, noté $\widehat{Y_t}(o)$ (voir Figure 4.6). Le label estimé final prédit par le classifieur RF, noté $\widehat{Y}(o)$, est obtenu par vote majoritaire sur l'ensemble des arbres T. En régression, la prédiction finale est obtenue en moyennant les valeurs obtenues par les arbres.

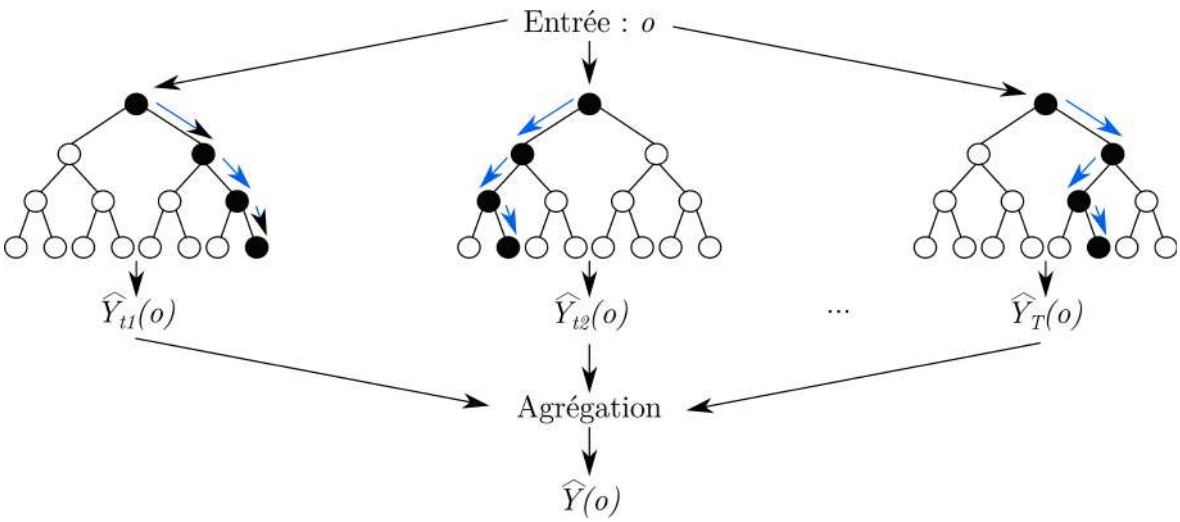


FIGURE 4.6 – Principe de l'algorithme des forêts aléatoires et d'agrégation des arbres de décision. Le chemin parcourus par l'observation o testée est représenté en noir.

4.3.3 Les paramètres de l'algorithme

Il existe plusieurs paramètres affectant la construction des RF. On trouve le nombre d'arbres T d'une forêt correspondant également au nombre de bootstraps, la taille f de la liste des caractéristiques tirées aléatoirement pour chaque nœuds, ainsi que le nombre d'observations u dans les échantillons bootstraps. D'autre paramètres plus spécifiques existent, comme le nombre minimal d'échantillons par feuille, la profondeur maximale des arbres de décision, etc.

Il n'existe pas de règle générale pour le choix du nombre d'arbres T de décision dans une forêt, ni pour le choix de la valeur de f. C'est pourquoi, des expérimentations doivent être réalisées afin de rechercher les paramètres optimaux en fonction de notre problématique. Breiman [Breiman 2001] indique néanmoins une préférence à l'utilisation d'un nombre f égale à la racine carré du nombre total de caractéristiques étudiées F. De plus, l'arbre construit doit être complètement développé et non-élagué [Breiman 2001]. Concernant le nombre d'observations dans les échantillons bootstraps (u), d'après Breiman [Breiman 2001], fixer u égale à N (le nombre total d'observations) permet d'obtenir de bons résultats. Plusieurs de ces paramètres devront être testés lors d'une étape de validation afin d'utiliser la méthode la plus adéquat.

4.3.4 Les indicateurs de performances

Il est possible d'extraire différents indicateurs de performances des RF. Ces indicateurs sont liés aux propriétés des forêts aléatoires.

4.3.4.1 L'erreur Out-Of-Bag

L'avantage de la procédure *"Out-Of-Bag"* (OOB) est qu'elle ne nécessite pas de découper l'échantillon. Elle utilise le fait que les arbres sont construits sur des échantillons bootstrap et que, par conséquent, ils n'utilisent pas toutes les observations de l'échantillon d'apprentissage.

Soit, i une observation $\{X(i), Y(i)\}$ de l'échantillon d'apprentissage, on désigne $\mathcal{T}_i$ l'ensemble des arbres de la forêt qui ne contiennent pas cette observation i dans leur échantillon bootstrap. On note alors $\widehat{Y_{OOB}}(i)$, le label estimé de i par l'agrégation de ces $\mathcal{T}_i$ arbres. En répétant, ce processus sur les N observations, on obtient un vecteur des labels estimés par approche OOB, noté $\widehat{Y_{OOB}}$. On comparant les labels connus Y des N observations avec les $\widehat{Y_{OOB}}$ prédit, il est possible de calculer une erreur de classification OOB (OOB_{err}).

$$\text{OOB}_{\text{err}} = \frac{1}{N}\sum_{i}\delta(\widehat{Y_{OOB}(i)}, Y(i)) \tag{4.4}$$

En classification, δ est égal à 1 si les labels étudiés sont différents et en régression, δ est une fonction d'erreur quadratique [Breiman 2001].

L'erreur OOB permet facilement et rapidement d'estimer la précision du classifieur. De cette manière, les performances du classifieur peuvent être évaluées à partir de l'échantillon d'apprentissage sans découpage supplémentaire. Cependant, il est important de préciser que pour chaque OOB_{err}, seul un sous-ensemble d'arbre est utilisé, et non la forêt dans son intégralité. Cette erreur estime donc l'erreur de généralisation d'une forêt à partir des prédictions des agrégations d'arbres de la forêt. En conséquence, l'erreur OOB a pour inconvénient d'être souvent considérée comme optimiste.

4.3.4.2 Les courbes ROC

Grâce aux bootstraps, il est également possible de calculer pour chaque observation i le pourcentage d'arbres prédisant correctement son label parmi les $\mathcal{T}_i$. Ces pourcentages sont recueillis dans un vecteur Φ de dimension N. Si le pourcentage d'arbres prédisant correctement le label est élevé, alors cette prédiction est considérée comme robuste. Ainsi, plus la moyenne du vecteur Φ est élevée, plus le classifieur RF est robuste.

Il est possible d'utiliser Φ dans la création de courbes ROC. Pour cela, les valeurs de Φ sont triées par ordre décroissant, puis, à l'aide d'un seuil glissant de la valeur la plus haute à la plus faible, une sensibilité et une spécificité sont calculées. La courbe ROC représente l'ensemble des couples sensibilité/spécificité obtenus. L'aire sous cette courbe ROC lié aux RF est appelée AUC_{RF}. Cette technique permet, à l'aide des RF, d'obtenir une AUC à partir d'un sous-ensemble de caractéristiques, ce qui est impossible à réaliser directement.

4.3.4.3 Le coefficient d'importance

Enfin, grâce aux RF, un coefficient d'importance des caractéristiques utilisées dans la construction de la forêt peut être calculé [Breiman 2001].

On pose OOB_t, l'échantillon OOB regroupant l'ensemble des observations absentes du processus de création de l'arbre t. On désigne maintenant $\widetilde{\text{OOB}}_t^j$, l'échantillon OOB_t dans lequel on a perturbé aléatoirement les valeurs de la $j^{\text{ème}}$ variable, c'est-à-dire que les valeurs de cette caractéristique ont été inter-changées aléatoirement entre les observations. Les erreurs de prédiction OOB obtenues à l'aide de OOB_t et de $\widetilde{\text{OOB}}_t^j$ sont calculées.

Le coefficient d'importance C_i d'une caractéristique j est alors défini comme étant la moyenne des différences d'erreur de prédiction entre l'échantillon normal et le perturbé sur l'ensemble des T arbres de la forêt (Équation 4.5).

$$C_i(j) = \frac{1}{T}\sum_t (\widetilde{\text{OOB}_{\text{err}}}_t^j - \text{OOB}_{\text{err}\,t}) \tag{4.5}$$

Plus cette moyenne est grande, plus la caractéristique est considérée comme étant

contributive dans le modèle de classification. A l'inverse, si les permutations n'ont quasiment aucun effet sur l'erreur, la variable est considérée comme une variable présentant peu d'importance. Enfin, au contraire des 2 autres indicateurs de performances, le coefficient d'importance est centré sur la caractéristique et non sur les observations.

Grâce à ces différents indicateurs de performances des RF, différentes approches de sélection de caractéristiques ont été testées. Ces méthodes sont décrites dans les sections suivantes.

4.4 Première étape de sélection des caractéristiques non-corrélées

Comme abordé précédemment, la faible taille des bases de données souvent rencontrée en imagerie médicale crée un déséquilibre entre nombre de caractéristiques étudiées (plusieurs dizaines). La génération d'un modèle par apprentissage automatique dans le cas où le nombre de caractéristiques étudiées est inférieur au nombre d'observations présente un risque de sur-apprentissage. Malgré le fait que l'algorithme des RF puisse être utilisé sans sélection des caractéristiques, nous avons privilégié une approche réduisant le nombre de caractéristiques étudiées en supprimant celles qui sont corrélées.

La première étape de la méthode de sélection que nous proposons porte sur l'utilisation d'une sélection filtrante des caractéristiques. Contrairement aux autres méthodes de sélection de caractéristiques, l'approche filtrante est indépendante du classifieur utilisé dans l'étape suivante. Cela permet d'évaluer, sans biais, l'apport de cette étape. D'après Parmar et al. [Parmar et al. 2015], l'utilisation d'une approche filtrante est plus efficace en temps de calcul que les autres méthodes de sélection de caractéristiques. Les auteurs ont réalisé l'étude de 12 méthodes filtrantes de sélection de caractéristiques sur une base de données de 440 caractéristiques radiomiques extraites des examens TDM de patients atteints d'un cancer du poumon. Leurs résultats ont montré que la méthode de sélection de Wilcoxon donne les meilleurs résultats avec la majorité des classificateurs (voir Section 82 - Apprentissage et sélection des caractéristiques en imagerie médicale TEP, page 82). Cette méthode est une méthode statistique non-paramétrique, basée sur les rangs

pour comparer la médiane de la population des deux classes. Cette méthode univariée ne prend pas en compte la redondance des caractéristiques sélectionnées au cours de leur classement.

Plusieurs auteurs ont montré qu'il existait de nombreuses corrélations entre les caractéristiques (voir Sous-section 2.2.5 - Corrélations entre les caractéristiques, page 48). La réduction du nombre de caractéristiques permet, notamment, d'éviter le phénomène de sur-apprentissage causé par une information répétitive apportée par des caractéristiques corrélées. Ainsi, Orlhac et al. [Orlhac et al. 2014] se sont intéressés aux corrélations de 3 bases de données : 72 lésions issues d'un cancer colorectal, 24 d'un cancer des poumons et 54 d'un cancer du sein. A partir des images TEP, 41 caractéristiques dont 31 de texture ont été extraites. Les auteurs ont calculé le coefficient de corrélation de Pearson (r) et ont considéré qu'une corrélation était pertinente si $|r| \geq 0.8$ avec une p-value inférieure à 5 % (voir Section 2.2.5 - Corrélation entre les caractéristiques, page 48).

Une fois l'ensemble des corrélations calculées, Orlhac et al. ont associé les caractéristiques répondant à ces critères au sein de groupes de corrélation. Ainsi, 11 groupes de corrélation ont pu être créés (Tableau 4.1), dont 2 avec des caractéristiques célibataires ("busyness" et LZLGE). Aucune caractéristique du 1er ordre, excepté le TLG, n'est corrélé avec le MTV (groupe 7), alors que le groupe 8 contient 6 caractéristiques du 1er ordre fortement corrélées avec le SUV_{max}. Ces caractéristiques ont été obtenues à partir d'images ré-échantillonnées de manière relative et non absolue. Comme nous l'avons vu précédemment (voir Sous-section 2.2.5 - Corrélations entre les caractéristiques, page 48), avec un ré-échantillonnage absolu, les corrélations entre caractéristiques ont plutôt tendances à s'organiser autour du SUV_{max} [Orlhac et al. 2015].

Dans notre méthodologie, nous nous sommes inspirés de celle d'Orlhac et al. [Orlhac et al. 2014] car l'étude de nombreuses caractéristiques potentiellement corrélées peut réduire la qualité de la sélection de caractéristiques. Cependant, contrairement à ces auteurs qui étudient les corrélations de Pearson, nous avons privilégié l'utilisation de la méthode des corrélation des rangs de Spearman [Spearman 1904]. En effet, cette approche, calculant la dépendance statistique non-paramétrique entre 2 caractéristiques (ρ), est privilégiée pour son efficacité à détecter les corrélations non-linéaires entre caractéris-

TABLEAU 4.1 – Groupes des caractéristiques corrélées d'après [Orlhac et al. 2014].

Groupes	Caractéristiques
1	Homogénéité (GLCM) - Corrélation (GLCM) - Contraste (GLCM) - Dissimilarité (GLCM) - Contraste (GLDM)
2	Énergie (GLCM) - Entropie (GLCM) - "Coarseness"
3	SRE - RPr - SZE - ZP
4	LRE - LZE - LZHGE
5	LGRE - SRLGE - LRLGE - LGZE - SZLGE
6	HGRE - SRHGE - LRHGE - HGZE - SZHGE
7	MTV - TLG - GLNUr - RLNU - GLNUz - ZLNU
8	SUV_{max} - SUV_{moy} - SUV_{peak} - SD - Entropie - Énergie
9	"Skewness" - "Kurtosis"
10	"Busyness"
11	LZLGE

tiques. Ainsi, F_i caractéristiques initiales sont analysées 2 à 2 menant à $(F_i \times (F_i - 1))/2$ analyses de corrélation.

Soit une liste X composée de N observations de dimension F_i ; X_1 et X_2, 2 caractéristiques de cette liste. Les observations sont classées par ordre croissant en fonction des valeurs de chaque caractéristique. Leur rang est relevé, correspondant à la position qu'elles occupent une fois le classement effectué. Le rang selon la caractéristique X_1 est noté $rg(X_1)$ et celui selon X_2 est noté $rg(X_2)$. On définit ensuite d_{rg} comme étant les différences de rang pour chaque observation des 2 caractéristiques étudiées, soit $d_{rg} = rg(X_1) - rg(X_2)$. Le coefficient de corrélation des rangs de Spearman ρ est alors défini par l'équation suivante (Équation 4.6) :

$$\rho = 1 - \sum_{i=1}^{N} d_{rg}(i)^2 \times \frac{6}{N(N^2-1)} \tag{4.6}$$

Les caractéristiques inter-corrélées répondant aux critères ($|\rho| \geq 0,8$ et p-value ≤ 5 %) sont associées au sein d'un groupe de corrélation, comme proposé par Orlhac et al. [Orlhac et al. 2014]. Les informations apportées par les caractéristiques d'un groupe sont donc redondantes et chaque groupe apporte une information différente. Afin de réduire le nombre de caractéristiques étudiées, une seule caractéristique est retenue par groupe.

Comme critère de sélection, nous nous sommes basés sur la robustesse des caractéristiques vis-à-vis des différentes causes de variation dans la littérature (voir Sous-section

2.2 - Étude des caractéristiques, page 42). Ainsi, la caractéristique présentée comme la plus robuste a été sélectionnée comme représentante de son groupe.

D'après Yan et al. [Yan et al. 2015], certaines caractéristiques sont extrêmement robustes aux paramètres de reconstruction de l'image (variation ≤ 5 %) : le SUV_{moy}, le SUV_{peak}, l'entropie (1^{er} ordre), la différence d'entropie, la différence inverse normalisée et le moment différentiel inverse (matrice GLCM) et le LGZE (matrice GLSZM). D'autres ont présentés une robustesse un peu plus faible mais néanmoins convenable (variation ≤ 10 %) : le SUV_{max}, l'énergie, le kurtosis, le COV (1^{er} ordre), le somme moyenne, la somme des entropies, l'homogénéité, l'entropie (matrice GLCM), le LZE, HGZE et le LZLGE (GLSZM). Ces résultats concordent avec ceux rencontrés précédemment dans la littérature [Tixier et al. 2012] et [Hatt et al. 2013], où des caractéristiques comme l'entropie (matrice GLCM), l'homogénéité (matrice GLCM) ou le ZP (matrice GLSZM) apparaissaient robustes.

Cette deuxième étape mène à une sélection de F_{nc} caractéristiques non-corrélées (Figure 4.7).

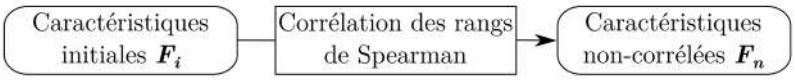


FIGURE 4.7 – Réduction du nombre de caractéristiques par analyse des corrélations de Spearman.

4.5 Deuxième étape de sélection

Après la première étape de sélection, 2 méthodes de sélection enveloppante sont proposées pour une sélection plus fine (voir Figure 4.1).

4.5.1 Première approche de sélection basée sur les coefficients d'importance des RF_{err}

La première approche de recherche d'un sous-ensemble de caractéristiques optimales proposée, nommées FIC est composée de 2 étapes (voir Figure 4.8) basées principalement

sur les coefficients d'importance et l'erreur OOB des RF.

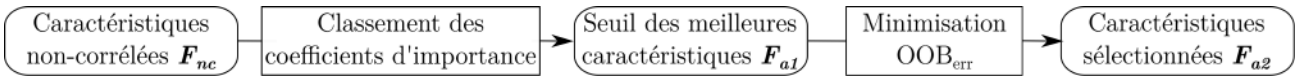


FIGURE 4.8 – Processus de l'approche FIC de sélection de caractéristiques basée sur les coefficients d'importance mesurés par RF.

Après l'analyse des corrélations et la réduction de F_i à F_{nc} caractéristiques, le coefficient d'importance C_i de chacune des caractéristiques restantes est calculé à partir d'un premier classifieur RF. Un classement est ensuite réalisé en fonction de C_i. Les meilleures caractéristiques de ce classement sont conservées en fonction d'un pourcentage S du coefficient d'importance maximal C_{imax} (Équation 4.7).

$$\text{Sélection de la caractéristique } j \begin{cases} \text{oui, si } C_i(j) \geq S \times C_{imax} \\ \text{non, sinon} \end{cases} \tag{4.7}$$

Cette étape mène à une première sélection de F_{a1} caractéristiques. Puis, de nouveaux classifieurs RF sont générés à partir de l'ensemble des combinaisons de caractéristiques possibles. Les performances des classifieurs servent à évaluer les différentes combinaisons de caractéristiques. Au final, le sous-ensemble de F_{a2} caractéristiques minimisant l'OOB_{err} est retenu.

Le nombre de combinaisons étudiées lors de cette deuxième étape augmente exponentiellement en fonction du nombre de caractéristiques étudiées, par exemple il existe 31 combinaisons possibles pour 5 caractéristiques étudiées, 1023 pour 10 et plus d'un million pour 20. C'est pour cette raison qu'une 1ère étape basée sur les coefficients d'importance limitant le nombre de caractéristiques étudiées est nécessaire pour réduire le temps de calcul de cette approche.

Cette méthode part du principe que les caractéristiques pertinentes sont celles possédant un coefficient d'importance élevée. Cependant, il est possible qu'une combinaison de caractéristiques, dont certaines présentent un faible coefficient, soit plus pertinente que la combinaison de caractéristiques avec un fort coefficient, car elles apportent des informations complémentaires. L'utilisation des coefficients d'importance peut donc être limitée. C'est pourquoi nous avons proposé une deuxième approche de sélection.

4.5.2 Deuxième approche de sélection basée sur l'association de l'algorithme génétique et les RF

La deuxième approche, nommée GARF, recherche le sous-ensemble optimal de caractéristiques de manière itérative. Comme le nombre de combinaisons étudiées est important, il est nécessaire d'être aidé par un outil capable de réduire le temps de calcul, c'est pourquoi nous nous sommes tourné vers l'algorithme génétique (*"Genetic Algorithm"* (GA)) [Holland 1992]. Ainsi, à partir des F_{nc} caractéristiques non-corrélées, une succession de combinaison de caractéristiques, appelé chromosome, est testée à l'aide de GA permettant la construction de plusieurs modèles RF convergeant vers le modèle optimal (voir Figure 4.9).

FIGURE 4.9 – Processus de l'approche GARF de sélection de caractéristiques basée sur l'algorithme génétique et les RF.

L'algorithme génétique est une méthode récursive, similaire au processus de sélection naturelle, basée sur des générations successives de population (voir Figure 4.10). GA converge vers une solution optimale à l'aide d'une fonction d'évaluation ("fitness function").

À partir d'une base de données d'apprentissage X_{app} de taille (N, F_{nc}), une population initiale est créée composée de $nPop$ chromosomes. Un chromosome est un vecteur de taille F_{nc} représentant les caractéristiques étudiées comme des chaînes binaires de 0 et 1 (1 représentent les caractéristiques étudiées). Chaque chromosome est évalué à l'aide d'une fonction défini par l'utilisateur. Lorsque l'ensemble des chromosomes d'une population est évalué, le meilleur est celui minimisant la fonction d'évaluation. Le ou les meilleures sont sélectionnés, puis subissent des modifications menant à une nouvelle population. Cette nouvelle population représente une nouvelle génération de chromosomes. Après un nombre $nGen$ de générations, les chromosomes convergent vers la

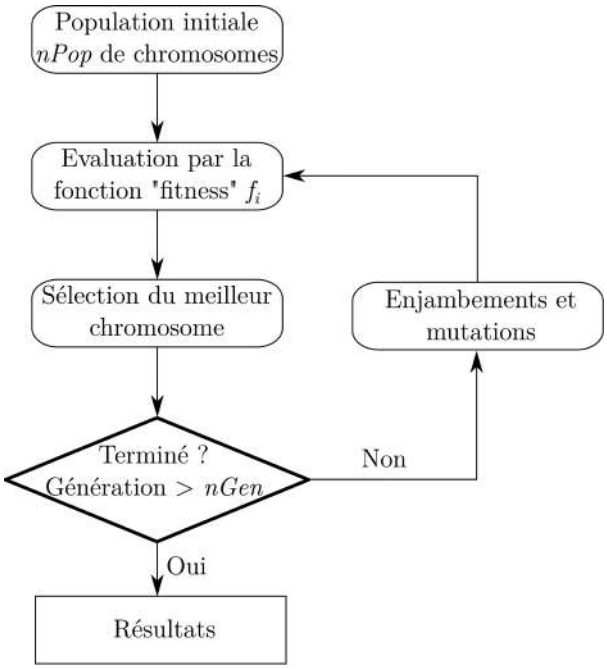


FIGURE 4.10 – Principe de l'algorithme génétique GA.

meilleure solution représentant le sous-ensemble optimal de caractéristiques.

Nous avons proposé une définition du sous-ensemble optimal de caractéristiques selon les 3 critères suivants :

— Le taux minimal d'erreur OOB de classification (OOB_{err}),

— La mesure minimale de l'AUC_{RF} calculée à partir du score Φ obtenu par le classifieur RF (voir page 102),

— Le nombre minimal de caractéristiques traduit par une contrainte parcimonieuse P définie comme le ratio entre le nombre de caractéristiques sélectionnées et le nombre total de caractéristiques étudiées à cette étape.

Le but est de chercher le chromosome optimal comme étant celui qui, avec un minimum de caractéristiques, présente la plus faible erreur de classification OOB et l'AUC_{RF} la plus élevée. On peut remarquer que P, AUC_{RF} et OOB_{err} varient tout les 3 entre [0, 1].

De ce fait, nous proposons la fonction d'évaluation f_b suivante :

$$f_b = \frac{P + \alpha(1 - \text{AUC}_{\text{RF}}) + \beta \text{OOB}_{\text{err}}}{\alpha + \beta + 1} \tag{4.8}$$

où f_b présente une relation linéaire avec les 3 paramètres P, AUC_{RF} et OOB_{err}, où α et β sont des coefficients de pondération ($\in [1,10]$) respectivement appliqués à l'AUC_{RF} et

à l'OOB_{err}. Ils sont utilisés afin d'augmenter l'importance des indices de performance (AUC_{RF} et OOB_{err}) face à la taille du sous-ensemble de caractéristiques. Un poids de 1 est attribué à P car c'est la plus faible valeur que peuvent avoir α et β. Ainsi, au fur et à mesure que α et β augmentent, l'importance du nombre de caractéristique diminue. Enfin, on retrouve la somme des poids au dénominateur de f_b afin de normaliser la fonction.

Le contenu de la population initiale est généré aléatoirement, même si des *a priori* peuvent y être incorporé. La sélection des chromosomes entre 2 générations se fait en fonction d'un score correspondant à l'adaptation du chromosome au problème. Il existe plusieurs techniques de sélection, dont voici les principales [Michalewicz 1996] :

— Méthode de sélection par rang qui choisit toujours les individus possédant les meilleurs scores d'adaptation. Le hasard n'entre pas dans ce mode de sélection. Ainsi, la sélection appliquée consiste à conserver les k meilleurs individus parmi n d'après la fonction d'évaluation suivant une probabilité qui dépend du rang (et pas de la fonction d'évaluation).

— Méthode de sélection uniforme qui choisit les individus aléatoirement, uniformément et sans intervention de la valeur d'adaptation. Chaque individu a la même probabilité d'être sélectionné.

— Méthode de sélection par tournoi qui affronte des paires de chromosomes aléatoirement, puis sélectionne celui ayant le meilleur score d'adaptation.

— Méthode de sélection proportionnelle à l'adaptation (roulette) où la probabilité de tirer au sort un individu est proportionnelle à son adaptation au problème d'après la fonction d'évaluation.

Les modifications réalisées sont similaires à celles présentes naturellement dans le génome. Ainsi des mécanismes d'enjambement ("crossover") et de mutation sont réalisés (voir Figure 4.11) [Michalewicz 1996].

— Les enjambements correspondent au fait que 2 chromosomes échangent une partie de leurs informations. Ainsi, la section d'information *"a"* exprimée initialement sur le chromosome *"A"* est échangée avec la section *"b"* exprimée par le chromosome *"B"*. La probabilité de croisement est nommée p_c.

— Les mutations correspondent à un changement aléatoire au sein d'un chromo-

some. Ainsi, si un chromosome ne sélectionne pas une caractéristique lors d'une génération, il est possible que ce chromosome muté la prenne en compte. Le taux de mutation lors des changements de population est nommé p_m. La mutation sert à éviter une convergence prématurée de l'algorithme vers un extremum local.

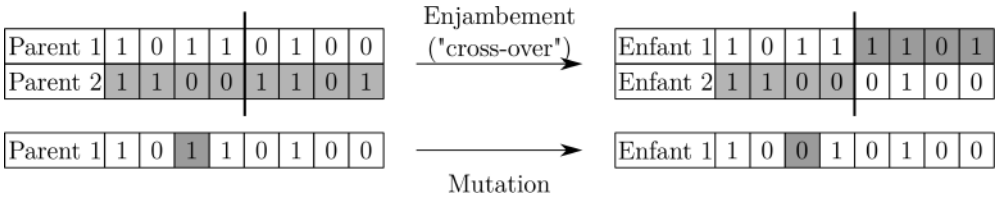


FIGURE 4.11 – Principe des modifications de l'algorithme génétique : enjambement et mutation.

Enfin, il existe des critères d'arrêt pour stopper le processus avant l'étude de toutes les générations. Ce critère peut être par exemple une variation trop faible de la valeur de la fonction d'évaluation entre les générations ou un temps de calcul limité.

Il existe plusieurs paramètres pouvant faire varier l'efficacité de l'algorithme génétique. On retrouve par exemple le nombre de génération $nGen$, la taille $nPop$ de la population, le type de sélection utilisé, les paramètres affectant les étapes d'enjambements et de mutations (p_c et p_m) et enfin, les critères d'arrêt. À notre connaissance, il n'existe pas dans la littérature de recommandations permettant d'adapter ces paramètres à notre problématique. C'est pourquoi, il nous sera nécessaire d'expérimenter plusieurs de ces paramètres pour rechercher la valeur optimisant nos résultats.

Au terme de cette étape, une sélection d'un sous-ensemble de F_b caractéristiques prédictives ou pronostiques est réalisée.

4.6 Conclusion

Dans ce chapitre, nous avons présenté 2 étapes de sélection des caractéristiques. La première étape consiste à étudier les corrélations de Spearman entre les caractéristiques, puis de limiter leur nombre aux caractéristiques non-corrélées.

La deuxième étape est la sélection du sous-ensemble de caractéristiques prédictif ou pronostique parmi ces caractéristiques non-corrélées. Deux approches ont été proposées. La première (FIC) utilise principalement la capacité intrinsèque du RF à calculer un

coefficient d'importance pour chaque caractéristique afin de générer un classement. Seul un certain pourcentage des meilleures caractéristiques de ce classement est conservé. Les possibles combinaisons de ces caractéristiques sont évaluées par un modèle RF, d'où est extrait une erreur de classification. Le sous-ensemble optimal de f_a caractéristiques est celui minimisant cette erreur. La deuxième (GARF) approche utilise directement les caractéristiques non-corrélées au sein d'un algorithme génétique, où un modèle RF est généré afin d'en extraire l'OOB_{err} et l'AUC_{RF}. Le sous-ensemble de f_b caractéristiques est celui qui trouve un juste milieu entre maximisisation de l'AUC_{RF} et minimisation du nombre de caractéristique et de l'OOB_{err}. Pour connaitre laquelle des méthodes de sélection présentée est la meilleure, elles ont été testées sur 2 bases de données.

Chapitre 5

Évaluation et résultats

Sommaire

5.1 Introduction

Dans le chapitre précédent, nous avons présenté deux méthodologies de sélection des caractéristiques basées sur l'utilisation des RF (FIC et GARF). À présent, nous allons étudier les performances de ces deux méthodes à l'aide de deux bases de données de patients atteints de différentes lésions cancéreuses : une cohorte de patients atteints d'un cancer de l'œsophage et une autre de patients atteints d'un cancer du poumon.

En premier lieu, nous présenterons les matériels et méthodes utilisés tout au long de ce chapitre en commençant par les différentes bases de données. De même, nous présenterons les caractéristiques utilisées pour chaque base, ainsi que la manière dont elles ont été extraites.

Plusieurs expérimentations ont été réalisées à partir de ces 2 bases. Tout d'abord, une classification sans méthode de sélection de caractéristiques a été réalisée. Puis, plusieurs méthodes de sélection de caractéristiques ont ensuite été étudiées en commençant par les 2 méthodes proposées dans le chapitre précédent FIC et GARF (voir Chapitre 4 - Méthodes proposées de sélection des caractéristiques radiomiques, page 93). Comme plusieurs paramètres sont utilisés au sein de ces 2 méthodes (seuil du coefficient de corrélation de Spearman, type de ré-échantillonnage, hyper-paramètres du RF, seuil du coefficient d'importance pour la méthode FIC et hyper-paramètres de l'algorithme génétique pour la méthode GARF), afin d'obtenir les meilleurs résultats, nous avons optimisé ces paramètres à partir d'une série d'expérimentations sur la base de données des patients atteints d'un cancer de l'œsophage. Les paramètres sélectionnés pour cette cohorte de patients seront utilisés pour la cohorte de patients atteints d'un cancer du poumon.

Une fois les paramètres optimaux définis, les performances de FIC et de GARF ont pu être comparées à celles obtenues par d'autres méthodes de sélection de caractéristiques couramment rencontrées dans la littérature du traitement de l'information : *"Sequential Forward Floating Selection"* (SFFS), HFS, RFE et LASSO (voir Section 3.4 - Sélection des caractéristiques). Enfin, des méthodes de statistiques largement utilisées dans la littérature médicale ont été étudiées : un test U de Mann-Whitney pour l'étude prédictive et une analyse univariée de Kaplan-Meier pour l'étude pronostique.

5.2 Matériel et méthodes générales

5.2.1 Présentation des bases de données

Dans cette section sont présentées les 2 bases de données utilisées dans l'évaluation de nos méthodes de sélection de caractéristiques.

5.2.1.1 Présentation de la cohorte du cancer de l'œsophage

Soixante-cinq patients atteints d'un cancer de l'œsophage localement avancé ont été inclus dans cette cohorte obtenue à partir des travaux précédents de notre équipe [Lemarignier et al. 2014]. Cette base a été obtenue dans le cadre de l'essai clinique RTEP3 étudiant l'intérêt prédictif de la TEP au FDG réalisée chez des patients traités par radiochimiothérapie pour un cancer de l'œsophage. Des détails cliniques concernant les patients sont regroupés dans le Tableau 5.1.

À partir des données démographiques, cliniques et biologiques, 16 caractéristiques issues du dossier médical des patients ont été obtenues et intégrées dans cette étude (voir Tableau 5.2). Des détails sur les classifications Stade TNM et OMS sont rajoutés en annexes (voir les Sous-sections A.1.1 et A.1.2). Les patients ont tous été traités entre 2005 et 2012 par RCT selon le schéma de Herskovic [Herskovic et al. 1992]. Ce traitement comprend une radiothérapie ininterrompue délivré par une technique à deux champs de fraction de 2 Gy par jour, à raison de cinq séances par semaine, pour un total de 50 Gy au volume cible tumoral, ainsi qu'une chimiothérapie au platine ou au 5-fluoro-uracile.

Les patients ont bénéficié d'un bilan initial comprenant *a minima* une endoscopie à ultrasons, un examen TDM avec injection de produit de contraste et éventuellement une biopsie de l'œsophage. Les patients ont également bénéficié de plusieurs examens d'imagerie TDM/TEP au FDG, dont un premier a été réalisé en pré-traitement lors de l'étape de stadification et de localisation de la tumeur. Les patients ont bénéficié d'autres examens TDM/TEP au FDG une fois le traitement terminé pour leur suivi systématique (à 1 mois et 3 ans post-traitement) ou en cas de suspicion de récidive.

L'ensemble des examens TDM/TEP a été pratiqué sur le même appareil Biograph®

TABLEAU 5.1 – Liste des caractéristiques cliniques issues du dossier médical des patients atteints d'un cancer de l'œsophage.

Caractéristiques	**Nombre de patients**
Démographique	
Age des patients (années)	
Médiane (gamme)	63 (46-85)
Moyenne (écart-type)	63,3 ± 9,9
Sexe des patients	
Homme	54 (83 %)
Femme	11 (17 %)
Clinique	
Localisation tumorale	
Tiers supérieur	18 (27 %)
Tiers moyen	31 (47 %)
Tiers inférieur	22 (33 %)
Histologie	
Adénocarcinome (ADC)	8 (12 %)
Carcinome épidermoïde (SCC)	57 (88 %)
Stade TNM	
II	17 (26 %)
III	39 (60 %)
IV	9 (14 %)
Stade OMS	
0	31 (48 %)
1	30 (46 %)
2	4 (6 %)
Albuminémie (g/L)	
Médiane (gamme)	39 (24-50)
Moyenne (écart-type)	38,8 ± 5,1
Résultats	
Survie à 3 ans (OS)	
Vivant	25 (38 %)
Décédé	40 (62 %)
Réponse à 1 mois	
Réponse Complète (RC)	41 (63 %)
Réponse Non-Complète (RNC)	24 (37 %)
Durée du suivi (mois)	
Médiane (gamme)	23 (6-79)
Moyenne (écart-type)	27,6 ± 18,0

TABLEAU 5.2 – Liste des 16 caractéristiques cliniques étudiées pour le cancer de l'œsophage.

Type des caractéristiques	**Caractéristiques**
Clinique	Age, Sexe, Poids actuel (kg) du patient, Poids de forme (kg), Perte de poids du patient (%), Albuminémie (g/l), *"Nutritional Risk Index"* (NRI), Malnutrition*, Localisation tumorale, Histologie, Stade T, N, M, stade OMS Longueur tumorale par endoscope (cm)

*absent si NRI > 97,5, moyen si 83,5 ≤ NRI ≤ 97,5 et sévère si NRI < 83.5

Sensation 16 High-Rez (Siemens Medical Solutions, Knoxville, Tennessee, États-Unis) suivant un protocole stricte. Après un jeûne d'au moins six heures et un repos d'au moins vingt minutes, une activité totale de 5 MBq/kg de FDG a été injectée au patient. Soixante minutes plus tard (± 10 minutes), six à huit positions de lit par patient ont été acquises en utilisant un protocole corps entier, à raison de 3 minutes par position. Les images TEP ont été reconstruites en utilisant les méthodes *"Fourier Rebinning"* (FORE) et *"Attenuation-Weighted Ordered Subset Expectation Maximisation"* (AW-OSEM). Les images ont été corrigées des coïncidences fortuites, de la diffusion, ainsi que de l'atténuation. Enfin, les images TEP au FDG ont été lissées avec un filtre gaussien de largeur à mi-hauteur de 5 mm. La taille des voxels de l'image reconstruite est de $4,06 \times 4,06 \times 2$ mm^3.

L'évaluation de la réponse au traitement a été réalisée un mois après la fin du traitement par RCT. Elle a consisté en un examen clinique, des biopsies de l'œsophage, ainsi que d'un examen d'imagerie par TDM/TEP au FDG. Les patients ont été classés comme présentant une réponse clinique complète (RC, 41 patients) au traitement si aucune tumeur résiduelle n'était détectée à l'endoscopie (biopsies négatives) et si aucune maladie locorégionale ou à distance n'a été identifiée sur l'examen TEP. Les patients ont été classés comme RNC (24 patients) si la présence d'une tumeur résiduelle ou d'une maladie locorégionale ou à distance a été détectée ou si le décès était survenu. Le suivi moyen de la population étudiée est de 27,6 ± 18,0 mois. L'OS utilisé pour l'étude pronostique a été estimée trois ans après la fin de la RCT. En fin du suivi, 25 patients étaient vivants et 40 morts (Tableau 5.1). Après la RCT, 14 patients ont bénéficié d'une chirurgie (4 patients en stade II, 8 en stade III et 2 en stade IV). Elle n'a néanmoins pas été utilisée dans notre

étude.

Quarante-cinq caractéristiques ont été extraites des images TEP (voir Tableau 5.3), selon la procédure suivante. En premier lieu, les lésions ont été segmentées sur les images TEP à l'aide d'un algorithme par seuillage adaptatif [Vauclin et al. 2009] menant à l'obtention du MTV. À partir de ce volume, dix-neuf caractéristiques du premier ordre ont été extraites. Ces caractéristiques sont présentées dans le Chapitre 1 (voir Section 1.3.2.1 - Caractéristiques statistiques du 1er ordre, page 24).

Afin d'extraire les caractéristiques d'ordres supérieurs un ré-échantillonnage préalable des niveaux de gris a été réalisée (voir Section 2.2.4 - Influence du ré-échantillonnage de l'image, page 45). Deux types de ré-échantillonnage ont été étudiés : un relatif des SUV du volume [Orlhac et al. 2014] (voir Équation 2.2, page 45) et un ré-échantillonnage absolu [Orlhac et al. 2015] (voir Équation 2.4, page 46). Pour le ré-échantillonnage relatif, chaque lésion a été ré-échantillonnée pour obtenir 64 niveaux de gris ($D = 64$) alors que pour le ré-échantillonnage absolu, un pas de ré-échantillonnage de 0,5 a été utilisé pour prendre en compte l'intégralité de la gamme des SUV de la cohorte ($B = 0,5$). À partir de ces MTV ré-échantillonnés, 3 matrices de texture ont été extraites : la matrice GLCM, la matrice GLDM et la matrice GLSZM. Concernant la matrice GLCM, les 13 matrices issues des 13 directions spatiales 3D, ont été calculées, puis moyennées en une matrice moyenne unique. De cette dernière, ont pu être extraits 10 caractéristiques de texture. De plus, 5 caractéristiques ont été extraites de la matrice GLDM et 11 pour la matrice GLSZM menant à deux jeux de 26 caractéristiques de texture (relatif et absolu). Les expressions mathématiques des caractéristiques sont présentées en Annexes Tableaux A.3, A.5 et A.6. L'association des 16 caractéristiques cliniques et de celles extraites de l'image TEP a conduit à l'obtention d'un nombre de caractéristiques initiales $F_i = 61$.

5.2.1.2 Cohorte du cancer du poumon

La deuxième base de données étudiée comprend 25 patients présentant un cancer du poumon non à petites cellules. Cette base vient de l'essai clinique RTEP2 étudiant l'intérêt prédictif de la TEP au FDG réalisée en cours de radiothérapie ou RCT pour des patients atteints d'un cancer du poumon [Calais et al. 2015]. Les caractéristiques ont été direc-

TABLEAU 5.3 – Liste des 45 caractéristiques TEP étudiées pour le cancer de l'œsophage.

Type des caractéristiques	**Caractéristiques**
1er ordre	SUV_{max}, SUV_{moy}, SUV_{peak}, Somme des SUV (SUV_{sum}) MTV, TLG, Écart-type (SD), COV, Sphéricité, Skewness, Kurtosis, Énergie, Entropie, SUV_{10}, SUV_{90}, SUV_{10}-SUV_{90}, V_{10}, V_{90}, V_{10}-V_{90}
Textures	GLCM : Variance, Énergie, Entropie, Corrélation, Dissimilarité, Contraste, Homogénéité, Moment Différentiel Inverse (IDM), Cluster Shade, Cluster Tendency GLSZM : Short Zone Emphasis (SZE), Long Zone Emphasis (LZE), Low Gray level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Short Zone Low Gray-level Emphasis (SZLGE), Long Zone Low Gray-level Emphasis (LZLGE), Short Zone High Gray-level Emphasis (SZHGE), Long Zone High Gray-level Emphasis (LZHGE), Zone Percentage (ZP), Gray Level Non Uniformity (GLNUz), Zone Length Non Uniformity (ZLNU) GLDM : Coarseness, Contrast, Busyness, Complexity, Strength

tement récupérées au sein d'un tableau créé lors de travaux précédents réalisés au sein de notre laboratoire [Mi et al. 2015]. Comme nous n'avons pas eu accès aux dossiers médicaux et aux images, il n'a pas été possible de rajouter les caractéristiques cliniques et radiomiques autres que celles déjà existantes. Ainsi, 13 caractéristiques correspondant à des données démographiques, cliniques et biologiques ont été étudiées (voir Tableaux 5.4 et 5.5). Les patients ont tous bénéficié d'un examen TDM/TEP au FDG pré-traitement sur le Biograph® Sensation 16 High-Rez (Siemens Medical Solutions, Knoxville, Tennessee, États-Unis) pour la stadification initiale, puis 1 mois après la fin du traitement pour l'évaluation de la réponse.

Le protocole de traitement et d'acquisition des examens TDM/TEP au FDG est identique à celui présenté pour la cohorte de patients atteints d'un cancer de l'œsophage. Les critères de séparation entre les patients RC et RNC sont les mêmes que pour la cohorte précédente, menant à 6 patients classés RC et 19 comme RNC (voir Tableau 5.4). Seize caractéristiques de texture ont été extraites des images TEP (voir Tableau 5.5), selon la même procédure que pour la cohorte précédente, mais seule les caractéristiques de la matrice GLSZM ont été étudiées permettant la comparaison de nos résultats avec des

TABLEAU 5.4 – Liste des caractéristiques cliniques issues du dossier médical des patients atteints d'un cancer du poumon.

Caractéristiques	**Nombre de patients**
Démographique	
Sexe des patients	
Homme	30 (83 %)
Femme	6 (17 %)
Clinique	
Localisation tumorale	
Tiers supérieur	10 (28 %)
Tiers moyen	20 (56 %)
Tiers inférieur	10 (28 %)
Dysphagie	
0	4 (11 %)
1	11 (31 %)
2	17 (47 %)
3	3 (8 %)
Stade Stade TNM	
II	7 (19 %)
III	21 (58 %)
ND	7 (20 %)
OMS	
0	19 (53 %)
1	17 (47 %)
Résultats	
Réponse à 1 mois	
Réponse Complète (RC)	13 (64 %)
Réponse Non-Complète (RNC)	22 (36 %)

TABLEAU 5.5 – Liste des 29 caractéristiques cliniques et radiomiques étudiées pour le cancer pulmonaire.

Type des caractéristiques	**Caractéristiques**
Clinique	Age, Sexe, Poids actuel (kg) du patient, Poids de forme (kg), Perte de poids du patient (%), Localisation tumorale, Histologie, Stade T, N, M, stade OMS, Longueur tumorale par endoscope (cm)
1er ordre	SUV_{max}, SUV_{moy}, SUV_{peak} MTV, TLG
Textures	*GLSZM* : Short Zone Emphasis (SZE), Long Zone Emphasis (LZE), Low Gray level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Short Zone Low Gray-level Emphasis (SZLGE), Long Zone Low Gray-level Emphasis (LZLGE), Short Zone High Gray-level Emphasis (SZHGE), Long Zone High Gray-level Emphasis (LZHGE), Zone Percentage (ZP), Gray Level Non Uniformity (GLNUz), Zone Length Non Uniformity (ZLNU)

travaux antérieurs réalisés au sein de notre équipe [Mi et al. 2015].

5.2.2 Protocole des expérimentations

Pour la classification sans sélection de caractéristiques, l'intégralité des caractéristiques (F_i) a été utilisée afin de générer un classifieur RF. Pour cela, nous avons utilisé les paramètres par défaut suivants : un nombre d'arbre $T = 500$ et un nombre f de caractéristiques étudiées pour chaque nœud d'un arbre égale à $\sqrt{F}$, comme proposé par [Breiman 2001].

Le principe général du protocole expérimental basé sur les algorithmes de classification est donnée Figure 5.1. À partir de la base de donnée initiale D de taille N par F_i, l'algorithme de sélection de caractéristiques est itéré k fois. A chaque itération, la base D est divisée en une base de données d'apprentissage D_{app} et une base de test D_{test}. La base D_{app} est utilisée dans le processus de génération du classifieur permettant la sélection des caractéristiques, alors que la base D_{test} est conservée pour l'évaluation des performances. Comme les observations de test ne sont pas utilisées lors du processus d'apprentissage, cela a permis de minimiser le sur-apprentissage. Une fois un sous-ensemble de caracté-

ristiques sélectionné, la dimension des base D_{app} et D_{test} est réduite au nombre de ces caractéristiques (D'_{app} et D'_{test}). La base D'_{app} sert alors à la génération d'un nouveau classifieur qui est évalué grâce à D'_{test}.

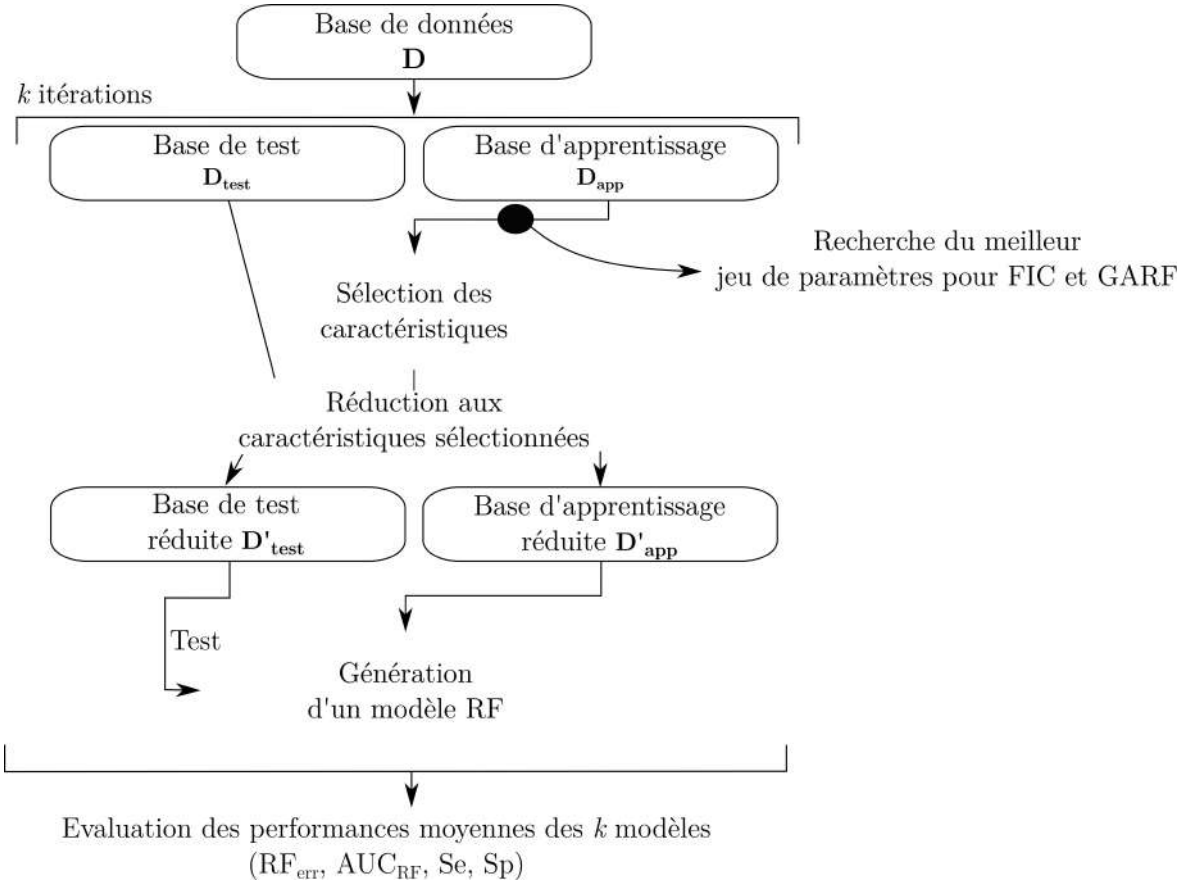


FIGURE 5.1 – Schéma du protocole de sélection des caractéristiques d'optimisation des paramètres de FIC et GARF incluant une méthode de recherche des paramètres optimaux.

L'évaluation des méthodes de sélection de caractéristiques a été réalisée de la façon suivante :

— Pour la base de données du cancer de l'œsophage, la méthode de validation des permutations aléatoires a été utilisée (voir Section 3.3 - Évaluation des méthodes d'apprentissage automatique, page 74) en utilisant 10 itérations et 2 tiers des observations pour l'apprentissage ($k = 10$).

— Pour la base de données du cancer du poumon, la méthode de validation croisée "Leave One Out" a été privilégiée en raison du faible nombre d'observations dans cette base. Ainsi, chaque observation a été utilisée individuellement comme base de test ($k = 25$).

Les indices de performances étudiés ont été l'AUC_{RF}, le RF_{err}, la sensibilité (Se) et la

spécificité (Sp). Le RF_{err} est présenté en termes de pourcentage (%), ainsi qu'en nombre de patients mal classés (pat). Ces indicateurs ont été moyennés sur les k itérations. L'existence d'une différence significative entre les différentes expériences a été étudiée à l'aide d'un test statistique de "Wilcoxon signed rank test" [Demšar 2006], associé à une correction de Benjamini-Hochberg [Hochberg and Benjamin 1990].

Avant la sélection des caractéristiques définitives, une étape de validation a été réalisée (voir Figure 5.1) afin d'optimiser les paramètres utilisés dans les méthodes FIC et GARF. Comme D'_{test} est conservée pour l'évaluation finale du classifieur, la recherche des paramètres est uniquement réalisée à partir de la base D'_{app}. Cette dernière a été séparée en 2 nouvelles bases : une nouvelle base d'apprentissage D_{app2} et une base de validation D_{val} servant à l'évaluation des modèles générés.

Ainsi, la base de donnée initiale D est divisée en 2, k fois, donnant k bases d'apprentissage D_{app} différentes. Chacune de ces bases est de nouveau divisée en 2, donnant une nouvelle base d'apprentissage D_{app2} et une base de validation D_{val}. Ces étapes sont répétées plusieurs fois afin de tester différents jeux de paramètres de FIC et GARF générant à chaque fois un modèle de RF. Ces modèles ont été évalués de la manière décrite précédemment et le test statistique de Wilcoxon a permis de mettre en évidence la présence de différences significatives entre les jeux de paramètres.

Pour chaque expérience, un tableau récapitulatif des matériels et méthodes utilisés est présenté.

5.2.3 Outils informatiques

Les images TEP reconstruites ont été importées avec leurs contours de segmentation à l'aide du logiciel OncoPlanet (Dosisoft, Cachan, France). Ces images ont ensuite été traitées à l'aide de plusieurs outils mathématiques que nous avons développé sur Matlab (version R2016a, The MathWorks, Inc., Natick, Massachusetts, USA). Nous avons ainsi mis au point des outils de calcul des SUV, de ré-échantillonnage relatif et absolu et d'extraction des différentes caractéristiques (1^{er} ordre, texture et forme).

Les expériences ont été réalisées à l'aide du logiciel Matlab. Concernant les différentes étapes de nos processus de sélection des caractéristiques, les fonctions de base procurées

par Matlab ont été utilisées pour l'analyse des corrélations de Spearman, la classification par RF ou SVM et l'utilisation de l'algorithme génétique.

Concernant les autres méthodes de sélection, la méthode de sélection de caractéristiques SFFS a été développé par Theodoridis et al. [Theodoridis and Koutroumbas 2010] et la méthode HFS a été développé au sein de notre équipe par Mi et al. [Mi et al. 2015]. Ces mêmes auteurs ont également retranscrit en Matlab la méthode RFE. Matlab présente une version de la méthode LASSO ainsi que des statistiques classiques. Seules les courbes de survie de Kaplan-Meier et les courbes ROC ont été générées à l'aide du logiciel MedCalc (version 12.7, MedCalc Software, Ostend, Belgium) pour une question d'esthétisme.

L'intégralité des expériences a été réalisée sur un ordinateur avec un processeur Intel Core i5-2500 (3,30 GHz) et 8 Go de mémoire RAM.

5.3 Étude de la base de données du cancer de l'œsophage

Dans cette section sont présentés les différents résultats des classifications appliquées à la base de données du cancer de l'œsophage. La présentation commence par les résultats de classification sans sélection de caractéristiques, puis avec différentes méthodes de sélection en développant les étapes de recherche de paramètres optimaux pour la méthode FIC et GARF.

5.3.1 Classification sans méthode de sélection de caractéristiques

5.3.1.1 Matériel et méthode

Dans le Tableau 5.8 sont présentés les détails méthodologiques concernant la classification par RF sans méthode de sélection de caractéristiques pour la cohorte de patients atteints d'un cancer de l'œsophage.

En s'appuyant sur la littérature [Breiman 2001], le nombre T d'arbres utilisés dans la forêts a d'abord été fixé à 500. De même, le nombre f de caractéristiques étudiées par nœud d'un arbre a été fixé à la racine carré du nombre de caractéristiques présent dans la base d'apprentissage (avec un nombre de caractéristiques F).

TABLEAU 5.6 – Paramètres utilisés pour les classifications par RF sans méthode de sélection de caractéristiques pour la cohorte de patients atteints d'un cancer de l'œsophage.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app} (43 observations)
Base de test	D_{test} (22 observations)
Paramètres :	
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$

5.3.1.2 Résultats et discussion

Les résultats des classifications RF sans sélection de caractéristiques pour les études prédictives et pronostiques des patients atteints d'un cancer de l'œsophage sont présentés dans le Tableau 5.7.

TABLEAU 5.7 – Résultats des classifications par RF sans méthode de sélection de caractéristiques pour les études prédictives et pronostiques des patients atteints d'un cancer de l'œsophage. La moyenne et l'écart-type des indices de performance sont présentés.

Étude	**RF_{err} (%)**	**RF_{err} (pat)**	**AUC_{RF}**	**Se (%)**	**Sp (%)**
Étude prédictive	25,5 ± 6,5	5,6 ± 1,4	0,798 ± 0,084	74 ± 10	88 ± 12
Étude pronostique	34,3 ± 6,3	7,5 ± 1,4	0,677 ± 0,097	78 ± 10	65 ± 22

On note que les résultats de l'étude prédictive sont meilleurs que ceux obtenus pour l'étude pronostique. De plus, on observe des écarts-types importants sur l'ensemble des indices de performances. Si l'on s'intéresse d'avantage à l'erreur de classification en termes d'individu (RF_{err} (pat)), on note un écart-type de 1,4 pour les études prédictives et pronostiques. Cette haute valeur peut s'expliquer par le faible nombre de patients utilisés dans la base de test (22 observations).

Ces résultats préliminaires nous ont incité à utiliser une stratégie de sélection de caractéristiques pour améliorer les performances de l'algorithme RF.

5.3.2 Étude de l'influence des paramètres des méthodes de sélection FIC et GARF

Dans le Tableau 5.8 sont donnés les paramètres par défaut que nous avons utilisé pour notre première évaluation de FIC et GARF. Nous avons ensuite étudié les paramètres des 2 algorithmes afin de les optimiser et d'évaluer leur influence sur les méthodes de sélection.

TABLEAU 5.8 – Valeurs des paramètres par défaut des méthodes de sélection de caractéristiques (FIC et GARF).

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\lvert\rho\rvert$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	50
Taille des populations du GA $nPop$	30
Poids de l'AUC_{RF} α	8
Poids de l'OOB_{err} β	8

Lors de nos premières expériences, nous avons considéré qu'une valeur absolue de ρ de l'analyse des corrélations de Spearman supérieure ou égale à 0,8 associée à une valeur de p inférieure à 5 % était significative, comme proposé dans [Orlhac et al. 2014]. Les valeurs des paramètres du RF (T et f) sont les mêmes que pour l'expérimentation précédente, c'est-à-dire T=500 et $f = \sqrt{F}$. En revanche, les autres paramètres par défaut assignés aux méthodes FIC et de GARF ont été définis de manière arbitraire.

5.3.2.1 Étude du seuil du coefficient de Spearman

5.3.2.1.1 Matériel et méthode

La première étape de notre méthode de sélection des caractéristiques est une sélection filtrante basée sur l'étude des corrélations de rang de Spearman entre les caractéristiques, comme décrit précédemment (Section 4.4 - Première étape de sélection des caractéristiques non-corrélées, page 104). Le seuil du coefficient de Spearman (ρ) indique à partir de quelle valeur une corrélation est considérée comme étant significative.

Nous avons étudié l'impact de ce seuil sur les résultats des classifications avec FIC et GARF en faisant varier ce seuil de 0,7 à 0,9 par pas de 0,1. Dans le Tableau 5.9 sont résumés les paramètres utilisés lors de ces expérimentations.

TABLEAU 5.9 – Valeurs des paramètres utilisés pour l'étude de l'influence du seuil du coefficient de Spearman. En gras sont représentés les différents seuils étudiés.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\|\rho\|$	**0,7, 0,8 et 0,9**
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	50
Taille des populations du GA $nPop$	30
Poids de l'AUC_{RF} α	8
Poids de l'OOB_{err} β	8

5.3.2.1.2 Résultats

Les résultats de ces études sont présentés dans le Tableau 5.10. Ainsi, pour les seuils de $|\rho|$ de 0,7, 0,8 et 0,9, nous obtenons un nombre F_{nc} de caractéristiques non-corrélées de 23, 28 et 36, respectivement. On note que lorsque l'on augmente la valeur de $|\rho|$ certains groupes se trouvent divisés. Par exemple, si l'on observe le groupe 10 obtenu avec un seuil de 0,9, on note qu'il était présent dans le groupe 8 avec un seuil de 0,8. Ainsi, les 3 caractéristiques du groupe 10 (LZE, LZHGE et LZLGE de la matrice GLSZM) sont fortement corrélées entre elles, mais elles le sont moins ($< 0,9$) avec les autres membres du groupe 8 et forment, de ce fait, leur propre groupe.

TABLEAU 5.10 – Corrélations des caractéristiques absolues pour un seuil de $|\rho|$ de 0,7, à 0,9. En gras sont représentées les caractéristiques sélectionnées F_{nc} pour l'étape suivante (F_{nc} = 23, 28 et 36, respectivement). En rouge, sont présentées les caractéristiques identiques pour les 3 cas étudiés. Les caractéristiques non-corrélées sont regroupées en tant que caractéristiques indépendantes (Indpt).

ρ	Grp	Caractéristiques
0,7	1	**Poids du patient** - Poids de forme
	2	***"Nutritional Risk Index"*** **(NRI)** - Albumibémie - Malnutrition
	3	**V_{10-90}** - V_{90}
	4	**ZLNU** - Cluster Shade (GLCM) - SZE
	5	**Énergie** - Entropie - Kurtosis - Skewness
	6	**MTV** - TLG - SUV_{sum} - Corrélation (GLCM) - "Coarseness" (GLDM) - "Busyness" (GLDM) - GLNUz
	7	**SUV_{max}** - SUV_{10} - Variance (GLCM) - HGZE - CT (GLCM) - SUV_{moy} - SUV_{peak} - SZHGE - SD - "Complexity" (GLDM) - SUV_{10-90} - LGZE - Entropie (GLCM) - Contraste (GLCM) - Dissimilarité (GLCM) - ZP - "Strength" (GLDM) - SUV_{90}
	8	**Homogénéité (GLCM)** - IDM (GLCM) - "Contrast" (GLDM) - Énergie (GLCM) - LZE - LZHGE - LZLGE
	Indpt	**11 caractéristiques cliniques - V_{10} - COV - SZLGE - Sphéricité**
0,8	4	**ZLNU** - Cluster Shade (GLCM)
	5	**Énergie** - Entropie
	6	**MTV** - TLG - SUV_{sum} - Corrélation (GLCM)
	7	**SUV_{max}** - SUV_{10} - Variance (GLCM) - HGZE - CT (GLCM) - SUV_{moy} - SUV_{peak} - SZHGE - SD - "Complexity" (GLDM) - SUV_{10-90} - LGZE
	8	**Homogénéité (GLCM)** - IDM (GLCM) - "Contrast" (GLDM) - Énergie (GLCM) - LZE - LZHGE - LZLGE - Dissimilarité (GLCM) - Contraste (GLCM) - ZP - Entropie (GLCM) - "Strength" (GLDM)
	9	**"Busyness" (GLDM)** - "Coarseness" (GLDM) - Sphéricité
	Indpt	**11 caractéristiques cliniques - V_{10} - COV - SZLGE - SUV_{90} - Kurtosis - Skewness - SZE - GLNUz**
0,9	4	**ZLNU** - Cluster Shade (GLCM)
	5	**Énergie** - Entropie
	6	**MTV** - TLG - SUV_{sum}
	7	**SUV_{max}** - SUV_{10} - Variance (GLCM) - HGZE - CT (GLCM) - SUV_{moy} - SUV_{peak} - SZHGE - SD
	8	**Homogénéité (GLCM)** - IDM (GLCM) - "Contrast" (GLDM) - Dissimilarité (GLCM) - Contraste (GLCM) - ZP - Entropie (GLCM)
	9	**"Busyness" (GLDM)** - "Coarseness" (GLDM) - Sphéricité
	10	**LZE** - LZHGE - LZLGE
	Indpt	**11 caractéristiques cliniques - V_{10} - COV - SZLGE - Skewness - Kurtosis - SUV_{90} - SUV_{10-90} - Énergie (GLCM) - Corrélation (GLCM) - SZE - LGZE - GLNUz - "Complexity" (GLDM) - "Strength" (GLDM)**

Sur la Figure 5.2 sont présentés les résultats des classifications réalisées à partir de ces F_{nc} caractéristiques et de nos méthodes de sélection (FIC et GARF).

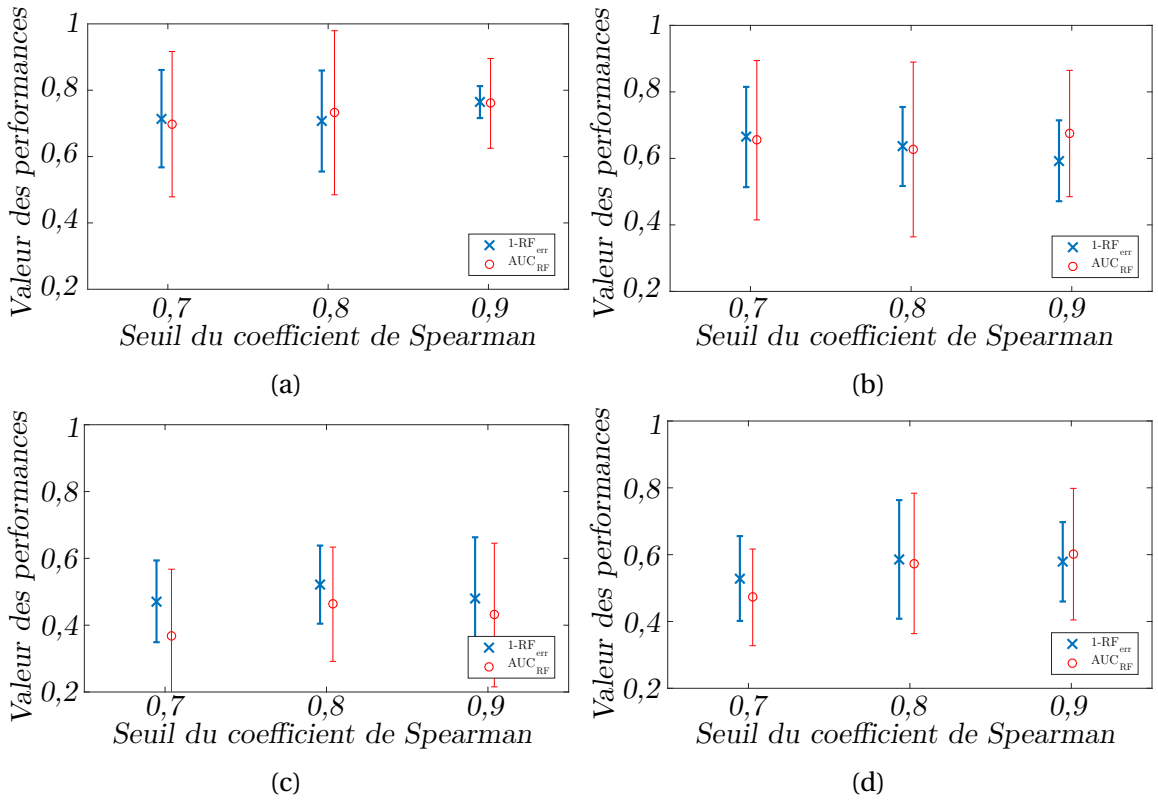


FIGURE 5.2 – Résultats des classifications RF en fonction des différents seuils de Spearman ρ utilisés pour l'étude prédictive avec (a) FIC et (b) GARF et pour l'étude pronostique avec (c) FIC et (d) GARF. Les tests de Wilcoxon ne révèlent aucune différence significative ($p > 0,05$).

Les résultats des classifications RF réalisées entre les 3 différents groupes sont voisins. Les tests de Wilcoxon réalisés pour chaque cas ne révèlent aucune différence significative ($p > 0,05$).

Dans [Orlhac et al. 2014], une valeur de $|r| = 0,8$ a été utilisée. Le but est, avec cette valeur intermédiaire, de détecter les caractéristiques suffisamment corrélées tout en évitant d'éliminer celles pouvant apporter une information complémentaire. Cette valeur a été conservée pour la suite de nos expériences.

Les 2 premiers groupes concernent uniquement les caractéristiques cliniques :

— Le groupe 1 est composé des caractéristiques cliniques de poids du patient. **Le poids du patient** le jour de l'examen a été privilégié car cette mesure est précise et datée au contraire du poids de forme du patient qui est une estimation variable entre les patients.

— Le groupe 2 est composé de caractéristiques dérivés de l'albuminémie, il est donc normal qu'elles soient toutes les 3 corrélées. Nous avons favorisé **le NRI** car cette caractéristique a déjà montré sa pertinence dans l'évaluation du cancer de l'œsophage [Di Fiore et al. 2014].

Pour les groupes 3 à 9, nous avons privilégié les caractéristiques radiomiques les plus robustes face aux différents paramètres de reconstruction des images d'après la littérature [Yan et al. 2015] (voir Sous-section 4.4, page 104).

— Le groupe 3 est représenté par le $\mathbf{V_{10\text{-}90}}$, il n'existe pas d'étude comparant sa robustesse face au V_{90}, donc ce choix est potentiellement interchangeable.

— Le groupe 4 est représenté par le ZLNU car il est plus robuste que le "Cluster Shade" (GLCM).

— Le groupe 5 est représenté par **l'énergie du 1er ordre**. Cependant, d'après [Yan et al. 2015], sa robustesse est similaire à celle de l'entropie, le choix est donc interchangeable.

— Le groupe 6 et 7 sont représentés respectivement par **le MTV et le $\mathbf{SUV_{max}}$** car ce sont les caractéristiques de référence lors de l'étude de robustesse en radiomique.

— Le groupe 8 regroupe un nombre important de caractéristiques des matrices GLCM et GLSZM. Plusieurs d'entre elles présentent de bonnes robustesses dans la littérature, comme l'homogénéité, l'IDM, l'énergie, l'entropie de la matrice GLCM. Nous avons sélectionné **l'homogénéité (GLCM)** car elle est également présenté comme robuste à la méthode de segmentation selon [Tixier et al. 2012].

— Le groupe 9 est représenté par **la "Busyness" de la matrice GLDM**, même si sa robustesse est plutôt faible. Dans [Yan et al. 2015], la "Coarseness" ne montre pas une meilleure robustesse que la "Busyness". À notre connaissance, il n'existe pas d'évaluation de la robustesse de la sphéricité.

5.3.2.2 Étude du type de ré-échantillonnage utilisé

5.3.2.2.1 Matériel et méthode

Nous avons ensuite étudié l'impact du type de ré-échantillonnage utilisé pour les caractéristiques de texture (relatif et absolu) sur nos méthodes de sélection FIC et GARF.

Dans le Tableau 5.11 sont présentés les paramètres utilisés lors de ces expérimentations.

TABLEAU 5.11 – Valeurs des paramètres utilisés pour l'étude de l'influence du type de ré-échantillonnage. En gras sont représentés les différents ré-échantillonnages étudiés.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\|\rho\|$	0,8
Type de ré-échantillonnage	**Absolu et relatif**
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	50
Taille des populations du GA $nPop$	30
Poids de l'AUC_{RF} α	8
Poids de l'OOB_{err} β	8

5.3.2.2.2 Résultats

Dans le Tableau 5.12 sont données les groupes de corrélation obtenus avec $|\rho| > 0,8$ et l'utilisation d'un ré-échantillonnage relatif .

Les résultats des classifications RF après sélection des caractéristiques par les méthodes FIC et GARF avec les différents ré-échantillonnage sont présentées Figure 5.3 pour les études prédictives et pronostiques.

En comparant les groupes de corrélations obtenus après un ré-échantillonnage relatif et absolu (Tableaux 5.10 et 5.12 pour un seuil de 0,8), les corrélations ont tendance à s'organiser autour du SUV_{max} lorsque les images sont ré-échantillonnées de manière absolue (Équation 2.3 et 2.4), alors qu'avec un ré-échantillonnage de type relatif, les corrélations sont plutôt autour du MTV. Ces remarques sont en accord avec les résultats d'Orlhac et al. [Orlhac et al. 2015] dans le cancer du poumon.

TABLEAU 5.12 – Corrélations des caractéristiques obtenues par un ré-échantillonnage relatif pour un seuil de $|\rho|$ de 0,8. En gras sont représentées les caractéristiques sélectionnées F_{nc} pour l'étape suivante ($F_{nc} = 32$). Les caractéristiques non-corrélées sont regroupées en tant que caractéristiques indépendantes (Indpt).

Groupe	Caractéristiques
1	**Poids du patient** - Poids de forme
2	**NRI** - Albumibémie - Malnutrition
3	**V_{10-90}** - V_{90} - Variance (GLCM)
4	**MTV** - TLG - $_{sum}$SUV - Énergie - Entropie - Énergie (GLCM) - Entropie (GLCM) - "Strength" (GLDM) - ZLNU (GLSZM) - GLNUz (GLSZM)
5	**SUV_{max}** - SUV_{10} - SUV_{10}-SUV_{90} - SUV_{moy} - SUV_{peak} - SD
6	**Homogénéité (GLCM)** - Dissimilarité (GLCM) - Contraste (GLCM) - IDM (GLDM)
7	**Kurtosis** - Skewness - LZHGE (GLSZM) - HGZE (GLSZM)
8	**ZP (GLSZM)** - LZE (GLSZM) - SZE (GLSZM)
9	**LGZE (GLSZM)** - SZLGE (GLSZM)
10	**"Coarseness" (GLDM)** - Sphericité
Indpt	**11 caractéristiques cliniques** - **V_{10}** - **SUV_{90}** - **COV** - **Cluster shade (GLCM)** - **Cluster tendency (GLCM)** - **Correlation (GLCM)** - **Contrast (GLDM)** - **Busyness (GLDM)** - **Complexity (GLDM)** - **SZHGE (GLSZM)** - **LZLGE (GLSZM)**

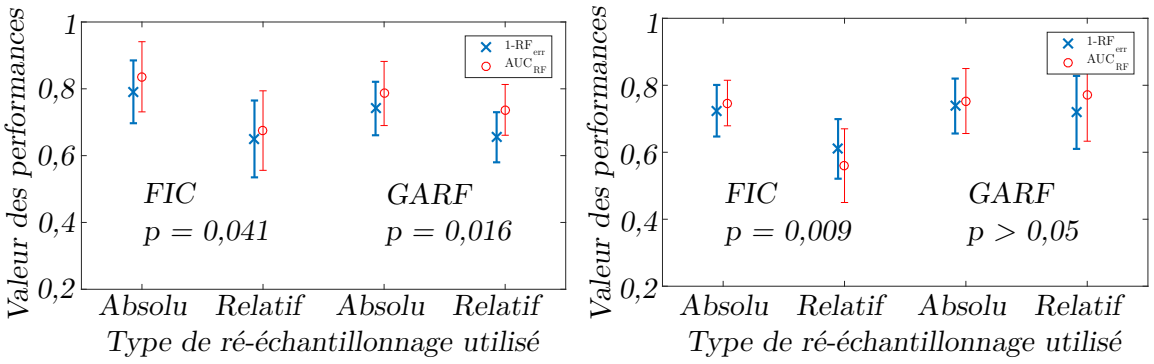


FIGURE 5.3 – Résultats de la classification par RF (AUC_{RF} et 1-RF_{err}) en fonction des différents types de ré-échantillonnage utilisés pour (a) l'étude prédictive et (b) l'étude pronostique.

Les tests de "Wilcoxon signed rank" montrent une différence significative en faveur de l'utilisation du ré-échantillonnage absolu (au risque $\alpha = 5$ %) à l'exception des résultats de l'étude pronostique obtenus après utilisation de la méthode GARF (p-value >5 %).

En s'appuyant sur ces résultats et ceux de la littérature [Orlhac et al. 2015], la méthode de ré-échantillonnage absolu a été retenue.

5.3.2.3 Étude de l'influence des paramètres du RF

5.3.2.3.1 Matériel et méthode

Nous nous sommes ensuite intéressés à 2 paramètres d'importance du RF : le nombre d'arbres T utilisé dans la construction des forêts et le nombre f de caractéristiques tirées pour chaque nœud d'un arbre. Ainsi, nous avons fait varier les valeurs de T entre 50 et 500 par pas de 50. En plus de $\sqrt{F}$ proposé par Breiman et al. [Breiman 2001], 2 autres méthodes de calcul ont été testées pour f avec des valeurs supérieures ou égales à la moitié de F ($F/2$ et $3F/4$). Prendre une valeur de f supérieure à $F/2$ a montré de bons résultats dans la littérature lorsque de nombreuses caractéristiques non-pertinentes sont présentes dans la base de données d'apprentissages [Genuer 2010]. Dans le Tableau 5.13 sont présentés les paramètres utilisés lors de ces expérimentations.

5.3.2.3.2 Résultats

Les résultats des classifications RF après sélection des caractéristiques par les méthodes FIC et GARF avec les différents paramètres du RF (T et f) sont présentés Figure 5.4 pour l'étude prédictive et Figure 5.5 pour l'étude pronostique.

D'après ces résultats, aucune tendance n'est observée ni pour l'étude prédictive (Figure 5.4), ni pour l'étude pronostique (Figure 5.5). Les résultats des classifications RF réalisées entre les 30 différents groupes sont similaires. Les tests de "Wilcoxon signed rank" réalisés pour chaque cas ne révèlent aucune différence significative (au risque $\alpha = 5$ %).

Pour la suite de nos expériences, une valeur de $T = 500$ a été retenue car elle constitue un compromis entre temps de calcul et performances de classification [Yu 2003]. Concernant la valeur du nombre des caractéristiques étudiées par nœud, nous avons décidé de conserver une valeur de $f = \sqrt{F}$ en s'appuyant sur la littérature [Breiman 2001].

TABLEAU 5.13 – Valeurs des paramètres utilisées pour l'étude de l'influence des paramètres du RF. En gras sont représentés les paramètres étudiés.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\lvert\rho\rvert$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	**100, 200, 300, 400 et 500**
Type de f	$\sqrt{F}$, $F/2$ et $3F/4$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	50
Taille des populations du GA $nPop$	30
Poids de l'AUC_{RF} α	8
Poids de l'OOB_{err} β	8

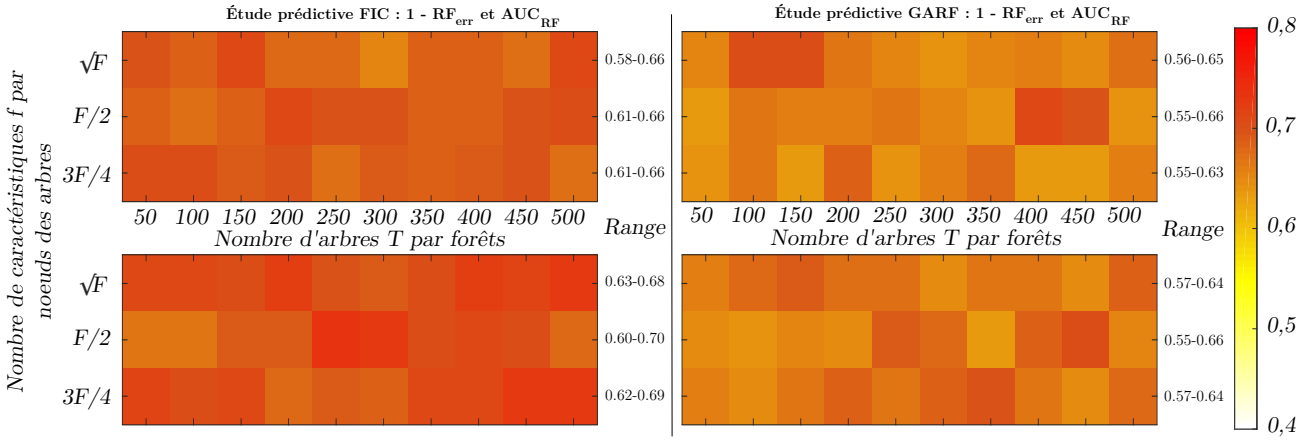


FIGURE 5.4 – Résultats de la classification par RF (AUC_{RF} et 1-RF_{err}) en fonction des différents nombres d'arbres T en abscisse et de la méthode de calcul de f en ordonnée pour nos méthodes de sélection FIC et GARF concernant l'étude prédictive. La première ligne correspond à l'inverse de l'erreur de classification (1-RF_{err}) et la deuxième ligne à l'AUC_{RF}.

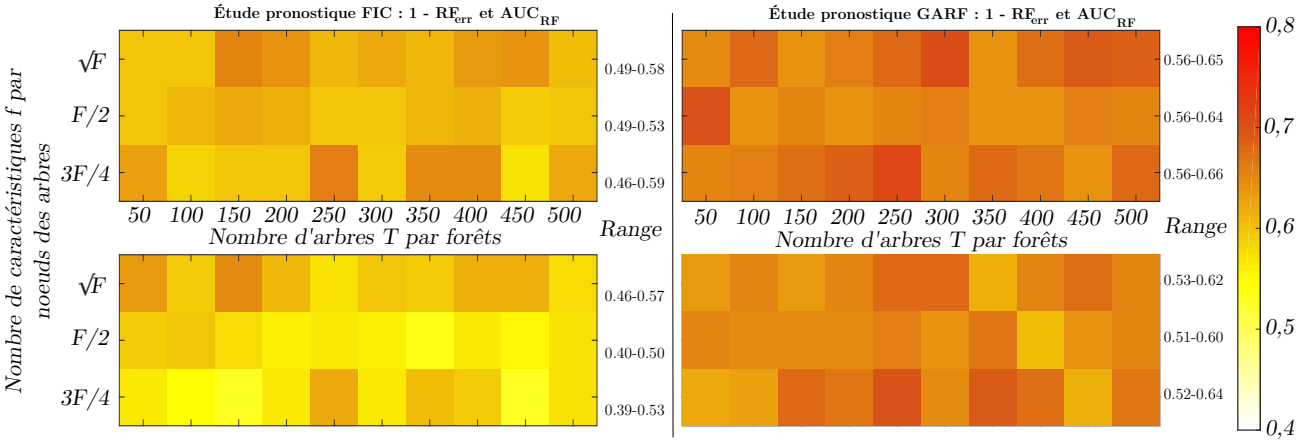


FIGURE 5.5 – Résultats de la classification par RF (AUC_{RF} et 1-RF_{err}) en fonction des différents nombres d'arbres T en abscisse et de la méthode de calcul de f en ordonnée pour nos méthodes de sélection FIC et GARF concernant l'étude pronostique. La première ligne correspond à l'inverse de l'erreur de classification (1-RF_{err}) et la deuxième ligne à l'AUC_{RF}.

5.3.2.4 Étude de l'influence des paramètres de la méthode FIC : seuil S des coefficients d'importance

5.3.2.4.1 Matériel et méthode

Lors de l'utilisation de la méthode FIC, il est nécessaire de paramétrer le seuil S de sélection des caractéristiques qui sont conservées en fonction de leur coefficient d'importance. Plusieurs seuils de coefficient d'importance ont été testés, entre 10 % et 50 % du coefficient le plus élevé par pas de 10 %. Le Tableau 5.14 résume les paramètres utilisés pour cette étude.

5.3.2.4.2 Résultats

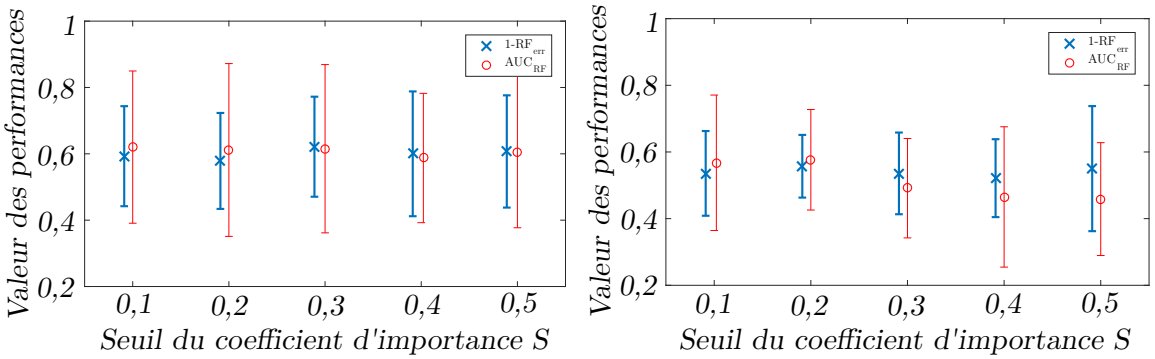
Les résultats des classifications RF après sélection des caractéristiques par la méthode FIC avec les différentes valeurs du seuil S des coefficients d'importance sont présentés Figure 5.6 pour l'étude prédictive et pour l'étude pronostique.

D'après ces résultats, aucune tendance n'est observée pour les études prédictives et pronostiques (Figure 5.6). Les résultats des classifications RF réalisées entre les 5 différents groupes sont similaires. Les tests de "Wilcoxon signed rank" réalisés pour chaque cas ne révèlent aucune différence significative (au risque $\alpha = 5$ %).

Une augmentation du nombre de caractéristiques sélectionnées par le calcul des coefficients d'importance (C_i) n'est pas synonyme d'une amélioration des performances,

TABLEAU 5.14 – Valeurs des paramètres utilisées pour l'étude de l'influence des paramètres du seuil S du coefficient d'importance. En gras sont représentés les paramètres étudiés.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\lvert\rho\rvert$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	**10, 20, 30, 40 et 50** %



(a) (b)

FIGURE 5.6 – Résultats de la classification par RF après sélection FIC en fonction des différents seuils S des coefficients d'importance étudiés pour (a) l'étude prédictive et (b) pour l'étude pronostique.

mais a pour conséquence une augmentation considérable du temps de calcul. De ce fait, nous avons décidé de conserver une valeur de S de 50 %.

5.3.2.5 Étude de l'influence des paramètres de la méthode GARF : paramètres du GA

5.3.2.5.1 Matériel et méthode

Comme la méthode GARF intègre l'utilisation d'un algorithme génétique, il est nécessaire d'étudier l'influence du nombre de génération $nGen$, la taille des populations $nPop$ par génération et les valeurs de poids α et β au sein de la fonction d'évaluation (Équation 4.8, page 110). Le Tableau 5.15 résume les paramètres utilisés pour cette étude.

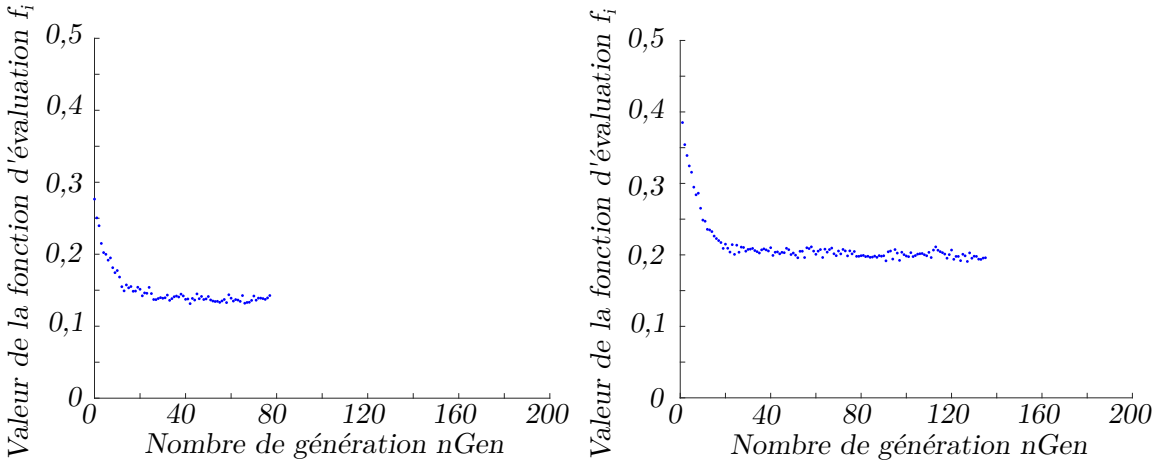
TABLEAU 5.15 – Valeurs des paramètres utilisées pour l'étude de l'influence des paramètres du seuil S du coefficient d'importance. En gras sont représentés les paramètres étudiés.

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app2} (29 observations)
Base de test	D_{val} (14 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\lvert\rho\rvert$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
GARF :	
Nombre de génération du GA $nGen$	**1 à 200**
Taille des populations du GA $nPop$	**20, 40, 60, 80 et 100**
Poids de l'AUC_{RF} α	**2, 4, 6, 8, 10**
Poids de l'OOB_{err} β	**2, 4, 6, 8, 10**

5.3.2.5.2 Résultats

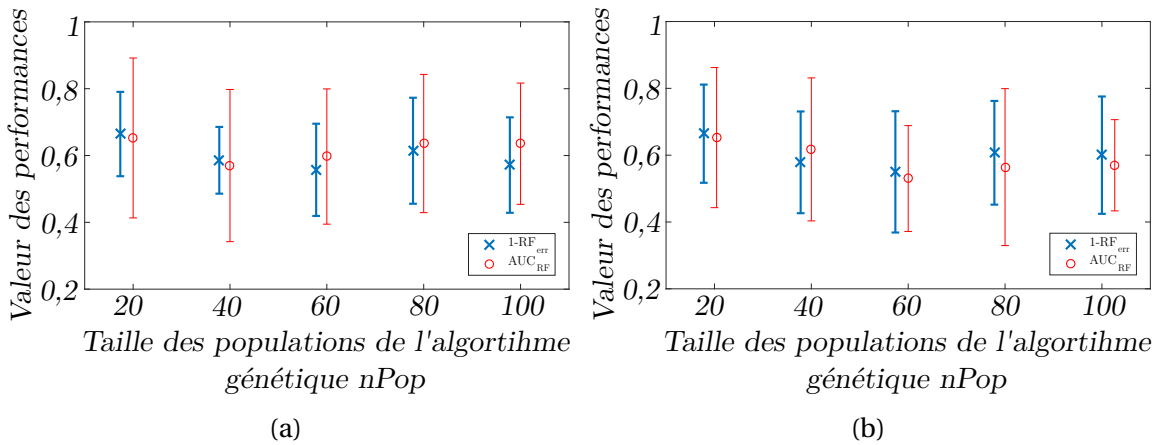
Les résultats concernant la recherche des valeurs optimales pour $nGen$ et $nPop$ sont présentés Figures 5.7 et 5.8, respectivement.

Les résultats concernant la recherche des valeurs optimales de α et β sont présentés Figure 5.9.



(a) (b)

FIGURE 5.7 – Courbes de la valeur moyenne de la fonction d'évaluation en fonction du nombre de génération de l'algorithme génétique (a) pour l'étude prédictive et (b) pour l'étude pronostique.



(a) (b)

FIGURE 5.8 – Résultats de l'erreur de classification du RF et de l'AUC (AUC_{RF} et 1-RF_{err}) en fonction de la taille de la population dans l'algorithme génétique (a) pour l'étude prédictive et (b) pour l'étude pronostique.

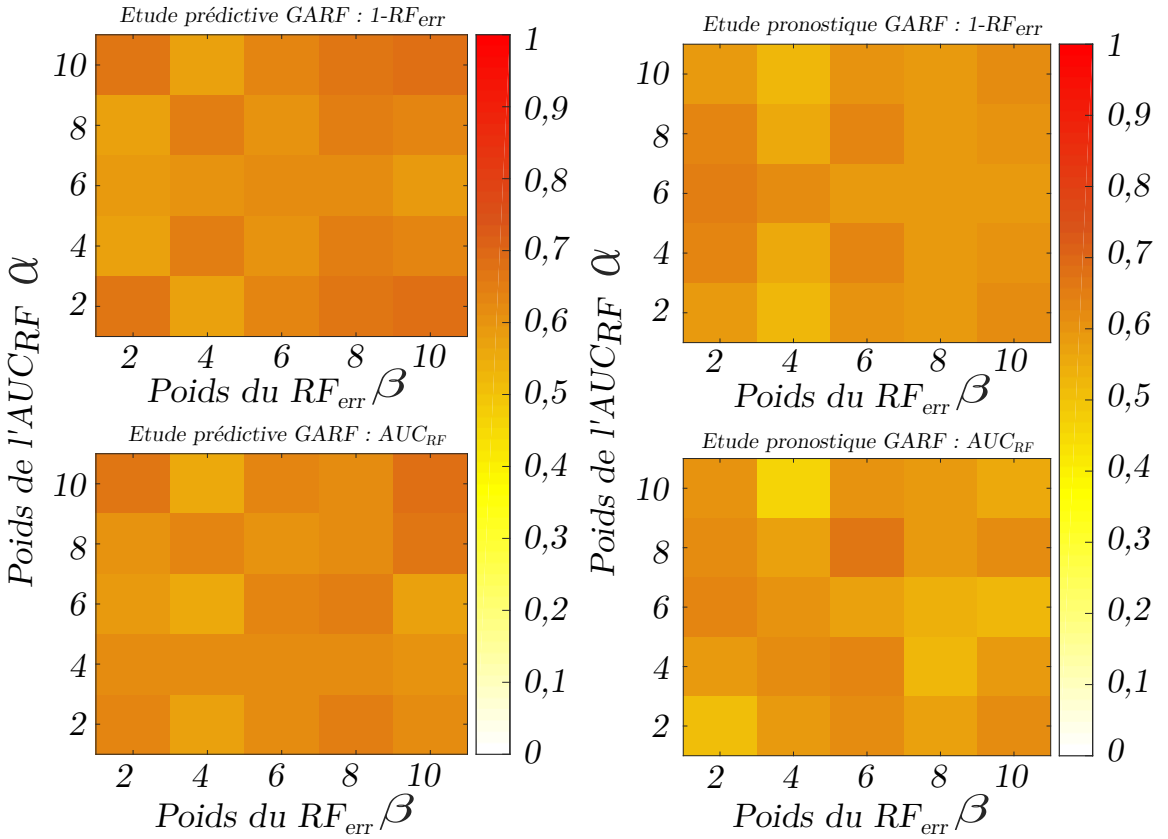


FIGURE 5.9 – Résultats de la classification par RF en fonction des différentes valeurs d'α et de β pour l'algorithme génétique de la méthode GARF (a) pour l'étude prédictive et (b) pour l'étude pronostique. La première ligne correspond à l'inverse de l'erreur de classification (1-RF_{err}) et la deuxième ligne à l'AUC_{RF}. Les tests de "Wilcoxon signed rank" ne révèlent aucune différence significative ($p > 0,05$).

Concernant l'étude du nombre de génération $nGen$ (voir Figure 5.7), on note que la fonction d'évaluation converge vers une valeur de plateau, aussi bien pour l'étude prédictive que pour l'étude pronostique. Ce plateau est atteint autour de 20 générations pour l'étude prédictive et 40 générations pour l'étude pronostique. De plus, l'algorithme a stoppé son processus de recherche avant d'atteindre les 200 générations car les variations de performances entre 2 générations successives sont très faibles. Une valeur de $nGen = 40$ a été sélectionnée car cette valeur permet d'obtenir les mêmes résultats que pour 200 générations mais dans un temps plus court, aussi bien pour l'étude prédictive que pronostique.

Concernant la taille optimale des populations d'une génération $nPop$ (voir Figure 5.8), des tests de "Wilcoxon signed rank" ont été réalisés pour l'étude des 5 cas. Aucune différence significative n'a été révélée par ce test (au risque $\alpha = 5$ %). Nous avons donc décidé de conserver une taille de population $nPop = 20$ aussi bien pour l'étude prédictive que pour l'étude pronostique. En effet, l'augmentation de $nPop$ n'a pas montré d'amélioration significative des performances, au détriment du temps de calcul.

Concernant la recherche des valeurs optimales de α et β pour la fonction d'évaluation du GA (voir Figure 5.9), les tests de "Wilcoxon signed rank" réalisés pour l'étude des 25 cas, n'ont montré aucune différence significative au risque $\alpha = 5$ %. Nous avons alors fixé les valeurs d'α et de β égalent à 10. Ainsi, la sélection GARF privilégiera 10 fois plus l'amélioration de ces indices de performance face au nombre de caractéristiques F dans le sous-ensemble sélectionné.

5.3.2.6 Discussion

On observe globalement sur l'ensemble des expérimentations que les résultats obtenus lors de l'étude pronostique sont moins bons que ceux de l'étude prédictive. Le fait que le GA nécessite 2 fois plus de générations pour l'étude pronostique que pour l'étude prédictive (Figure 5.7) illustre cette situation.

Au final, nous avons choisi d'utiliser des paramètres identiques entre les études prédictives et pronostiques, afin d'uniformiser le processus. Ces paramètres optimaux ont été défini comme étant ceux optimisant les performances de classification. Cependant,

pour beaucoup de nos expérimentations le test de "Wilcoxon signed rank" n'a pas pu mettre en évidence de différence significative entre les cas étudiés. On précise également que l'exploration des paramètres a été réalisée en profondeur, elle est donc non-exhaustive. De cette façon, il est possible que nous soyons passés à côté d'une combinaison de paramètres optimale. Cependant, pour des raisons de temps de calcul, une approche systématique n'a pas pu être employée.

Le manque de significativité pourrait s'expliquer par le manque d'effectif dans notre base de données d'expérimentations (43 patients). L'utilisation de 2 tiers de ces données comme données d'apprentissage ne permet pas de générer un modèle suffisamment représentatif. Ce manque de données pourrait être comblé par l'augmentation du nombre d'itération k. Cependant, à la vue de notre base de données trop petite, il est difficile d'augmenter ce nombre k. En cas de non-significativité de l'expérience, notre choix a été dicté comme étant un compromis entre les données de la littérature et le temps de calcul nécessaire à la sélection.

Malgré le manque de données, l'utilisation du ré-échantillonnage absolu montre un apport significatif aux classifications face au ré-échantillonnage relatif. Cela révèle l'importance de cette étape dans le processus d'étude des caractéristiques d'images.

5.3.3 Classifications par méthode sélection FIC et GARF

5.3.3.1 Matériel et méthode

L'ensemble des paramètres optimales des méthodes FIC et GARF ont été définis et regroupés au sein du Tableau 5.16. Une fois les paramètres des méthodes de sélection optimisés, des classifications réalisées à partir des sous-ensembles de caractéristiques sélectionnées par ces méthodes peuvent être effectuées et leurs performances comparées à ceux obtenus sans sélection.

Pour l'étude pronostique, un test de Kaplan-Meier a été utilisé pour estimer la distribution de survie une fois les meilleurs sous-ensembles déterminés. Elle permet d'obtenir une survie médiane, un pourcentage de décès dans chaque groupe et un HR.

Comme les processus de sélection se déroulent en 2 étapes, sélection filtrante ba-

TABLEAU 5.16 – Valeurs des paramètres optimaux des méthodes de sélection de caractéristiques (FIC et GARF).

Légende	**Détails**
Lésion	Cancer de l'œsophage
Méthode de validation	Permutations aléatoires : $c = 2/3$ et $k = 10$
Base d'apprentissage	D_{app} (43 observations)
Base de test	D_{test} (22 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\|\rho\|$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	40
Taille des populations du GA $nPop$	20
Poids de l'AUC α	10
Poids de l'OOBerr β	10

sée sur l'analyse des corrélations de Spearman, suivie d'une deuxième sélection enveloppante inclue dans FIC ou GARF (voir Figure 4.1), nous avons également étudié l'influence de chacune d'elles.

Les classifications RF ont été réalisées après sélection des caractéristiques en utilisant les paramètres optimaux. Une première série d'expériences a été réalisée en utilisant les algorithmes FIC et GARF sans l'analyse des corrélations de Spearman préalable, c'est-à-dire à partir des $F_i = 61$ caractéristiques initiales. Une deuxième série d'expériences a été réalisée où seule l'analyse des corrélations de Spearman a été effectuée avant la classification, à partir des $F_{nc} = 28$ caractéristiques non-corrélées.

5.3.3.2 Résultats

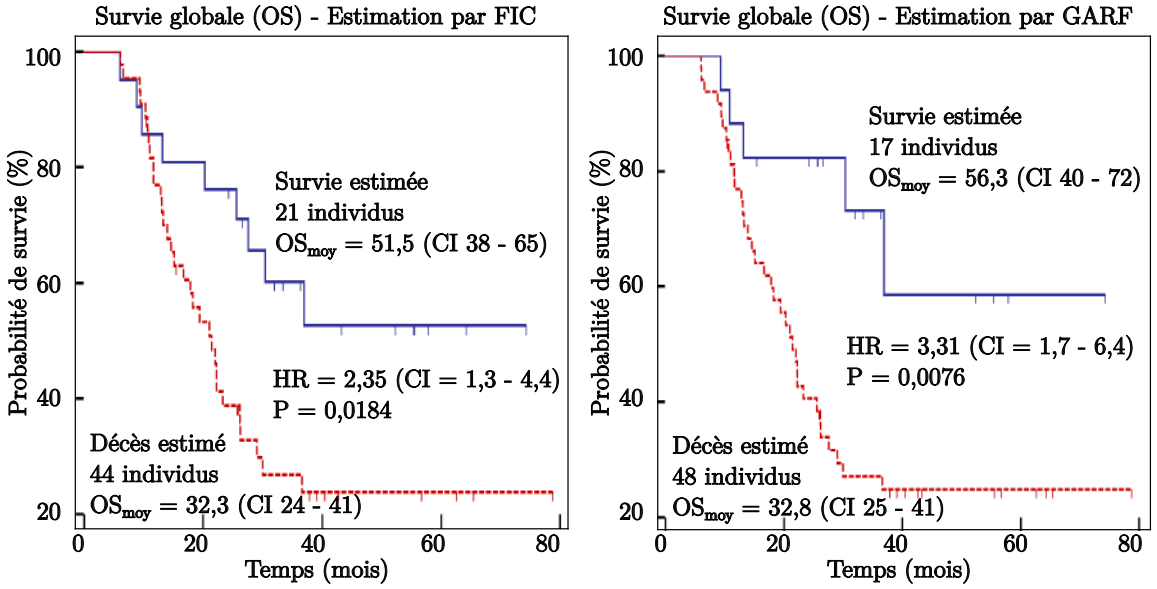
Concernant l'étude prédictive, la méthode FIC associée à l'analyse de corrélation de Spearman a sélectionné un sous-ensemble de 2 caractéristiques : le MTV et l'homogénéité (GLCM) alors que la méthode GARF également associée à l'analyse de corrélation de Spearman en a sélectionné 9 : le sexe, le poids usuel, la perte de poids, la localisation, le stade T, le MTV, le SUV_{max}, la "skewness" et l'homogénéité (GLCM). Les performances de classification entre ces 2 méthodes sont non-significativement différentes ($p > 5$ %).

Concernant l'étude pronostique, la méthode FIC associée à l'analyse de corrélation de Spearman a sélectionné un sous-ensemble de 3 caractéristiques : le stade OMS, le NRI et le MTV et la méthode GARF associée à l'analyse de corrélation de Spearman en a sélectionné 8 : l'histologie, la localisation, le stade OMS, le NRI, le SUV_{max}, la COV, le SUV_{90} et le V_{10}. Les courbes de survie de Kaplan-Meier, sont présentées Figure 5.10. Comme pour l'étude prédictive, les performances de classification entre ces 2 méthodes sont non-significativement différentes.

Les résultats des classifications RF obtenus sans sélection de caractéristiques, avec la sélection filtrante de Spearman seule, ou associées à FIC et à GARF sont présentés dans le Tableau 5.17. On retrouve également dans ce tableau, les résultats concernant les classifications réalisées après la sélection par la méthode enveloppante FIC et GARF uniquement.

TABLEAU 5.17 – Résultats des classifications RF réalisées sans et avec sélection de caractéristiques par les procédures de sélection proposées avec ou sans étape 1 d'analyse des corrélations de Spearman (Sp), ou étape 2 de sélection (FIC et GARF) pour les études prédictives et pronostiques. La moyenne et l'écart-type de chaque indice de performance sont présentés. F_{sel} est la taille des sous-ensembles de caractéristiques sélectionnées par les méthodes. En gras sont présentés les meilleurs résultats.

Étude	Méthode	F_{sel}	RF_{err} (%)	RF_{err} (pat)	AUC_{RF}	Se (%)	Sp (%)
Prédictive	Sans sélection	/	25,5 ± 6,5	5,6 ± 1,4	0,798 ± 0,084	74 ± 10	88 ± 12
	Sp seul	28	28,2 ± 4,7	6,2 ± 1,0	0,788 ± 0,074	76 ± 7	85 ± 11
	FIC seul	4	30,5 ± 7,7	6,7 ± 1,7	0,745 ± 0,092	62 ± 16	90 ± 14
	Sp + FIC	2	**21,4 ± 10,5**	**4,7± 2,3**	0,826 ± 0,104	**81 ± 8**	91 ± 12
	GARF seul	21	33,6 ± 7,2	7,4 ± 1,6	0,742 ± 0,080	62 ± 15	90 ± 13
	Sp + GARF	9	25,5 ± 7,2	5,6 ± 1,6	**0,842 ± 0,099**	77 ± 14	**92 ± 11**
Pronostique	Sans sélection	/	34,3 ± 6,3	7,5 ± 1,4	0,677 ± 0,097	78 ± 10	65 ± 22
	Sp seul	28	31,4 ± 9,8	6,9 ± 2,2	0,698 ± 0,085	80 ± 15	68 ± 12
	FIC seul	3	38,1 ± 8,1	8,4 ± 1,8	0,654 ± 0,083	68 ± 17	71 ± 14
	Sp + FIC	3	27,7 ± 4,5	6,1 ± 1,0	**0,822 ± 0,059**	**79 ± 9**	**95 ± 6**
	GARF seul	20	34,8 ± 5,0	7,6 ± 1,1	0,653 ± 0,111	62 ± 21	79 ± 29
	Sp + GARF	8	**26,7 ± 7,5**	**5,6 ± 1,6**	0,773 ± 0,101	75 ± 17	80 ± 16



(a) (b)

FIGURE 5.10 – Courbes de survie de Kaplan-Meier réalisées à partir des labels estimés obtenus après (a) la méthode de sélection FIC et (b) la méthode GARF.

5.3.3.3 Discussion

Tout d'abord, on note que les 2 méthodologies de sélection proposées (FIC et GARF associées à l'analyse de corrélation de Spearman) sont les méthodes apportant les meilleurs résultats de sélection, que ce soit pour l'étude prédictive ou l'étude pronostique.

Concernant l'étude prédictive, la méthode FIC précédée de l'analyse des corrélations de Spearman (Sp + FIC) est celle présentant le meilleur taux de bonnes classifications alors que la méthode GARF précédée de l'analyse des corrélations de Spearman (Sp + GARF) présente la meilleure AUC_{RF}. Cependant, il n'existe pas de différence significative entre ces 2 méthodes.

Les observations sont similaires pour l'étude pronostique, où nos méthodes de sélection des caractéristiques apportent respectivement les meilleurs RF_{err} et AUC_{RF}. Le test de "Wilcoxon signed rank" ne révèle pas de différence significative des performances entre ces 2 méthodes, mais montre un apport de ces méthodologies en comparaison avec la classification sans sélection de caractéristiques ($p = 0{,}03$). Concernant les courbes de survie de Kaplan-Meier (Figure 5.10), on note que le HR obtenu par la méthode GARF est supérieur à celui obtenu par FIC (3,31 pour GARF contre 2,35 pour FIC). Cela signifie que la séparation des patients lors de l'analyse pronostique est de meilleure qualité avec le sous-ensemble de caractéristiques sélectionnées par la méthode GARF.

L'obtention de meilleures valeurs de RF_{err} par la méthode FIC que GARF peut s'expliquer par le fait que FIC cherche uniquement à minimiser cet indice de performance alors que la méthode GARF incorpore plus de contraintes en cherchant également à maximiser l'AUC_{RF}. Cela explique également le fait que la méthode FIC sélectionne des sous-ensembles de caractéristiques plus petits que la méthode GARF (2 caractéristiques contre 9 pour l'étude prédictive et 3 caractéristiques contre 8 pour l'étude pronostique). Comme la méthode GARF doit à la fois minimiser l'erreur de classification et maximiser l'AUC_{RF}, plus de caractéristiques sont requises.

L'utilisation du GA lors de l'utilisation de GARF ne permet pas d'obtenir de meilleurs performances qu'avec l'utilisation de FIC. En effet, le GA permet notamment de converger vers la solution optimale et donc d'éviter d'avoir à tester l'ensemble des combinai-

sons possibles et donc de limiter le temps de calcul. A l'opposé, la méthode FIC analyse l'ensemble des combinaisons de caractéristiques possibles après analyse des coefficients d'importance, ce qui peut être extrêmement chronophage en fonction du nombre de caractéristiques.

Globalement, le MTV (groupe 6) semble avoir un rôle prédictif important car il est présent dans les sous-ensembles sélectionnés par nos 2 méthodes. On le retrouve également dans le sous-ensemble de caractéristiques pronostiques sélectionnées par la méthode FIC. Ces observations confirment l'intérêt particulier de cette caractéristique dans le suivi de la réponse au traitement, ainsi que pour la survie du patient [Van De Wiele et al. 2013]. La conclusion concernant l'apport du SUV_{max} (groupe 7) est plus mitigée. En effet, alors que cette caractéristique est considérée dans la littérature comme une caractéristique incontournable [Van De Wiele et al. 2013], elle n'est pas sélectionnée par la méthode FIC, mais 2 fois par la méthode GARF (étude prédictive et pronostique). De nouveau, le fait que la méthode GARF cherche à minimiser une AUC pourrait être responsable de la sélection du SUV_{max}.

On observe une sélection importante de caractéristiques cliniques par nos méthodes, excepté pour la méthode FIC lors de l'étude prédictive (GARF : 5/9 pour l'étude prédictive et 4/8 pour l'étude pronostique, FIC : 2/3 pour l'étude pronostique). Le fait que les meilleures sous-ensembles de caractéristiques sont des combinaisons des caractéristiques cliniques et de l'image TEP (1^{er} ordre et indices de texture) montre que ces 2 sources d'informations sont complémentaires. Les caractéristiques de texture semblent présenter un intérêt majoritairement sur la réponse au traitement des patients. En effet, l'homogénéité (GLCM, groupe 8) est la seule caractéristique de texture sélectionnée par nos méthodes de sélection, et ce, pour l'étude prédictive.

Concernant les résultats de classification à partir des $F_{nc} = 28$ caractéristiques non-corrélées (Tableau 5.17), on note une augmentation de l'erreur de classification et une réduction de l'AUC par rapport aux classifications réalisées sans sélection de caractéristiques $F_i = 61$ et avec le processus de sélection complet. Ces différences ne sont pas significatives à l'exception de l'AUC de l'étude pronostique ($p = 0{,}002$ pour FIC et $p = 0{,}016$ pour GARF). Le fait de ne retenir que les 28 caractéristiques non-corrélées ne signifie pas

nécessairement une amélioration des performances de classification de l'algorithme des RF. Ceci s'explique par le fait que la 1ère étape de sélection des caractéristiques est basée sur un test de rang. Dans ce cadre, 2 caractéristiques fortement corrélées au sens du test de Spearman sont quand même susceptibles d'apporter des informations complémentaires pour les RF.

Bien que les résultats ne soient pas significatifs, on constate également que l'utilisation d'une méthode de sélection telle que FIC seul et GARF seul, n'améliore pas non plus les performances pronostiques et prédictives. De ce fait, c'est l'ensemble des 2 étapes de sélection (FIC et GARF associés à l'analyse des corrélations de Spearman) qui améliore les performances de classification, et ce, de façon significative.

5.3.4 Comparaisons à d'autres méthodes de sélection

Nous avons comparé nos résultats avec d'autres méthodes proposées dans la littérature, aussi bien des statistiques classiques que d'autres méthodes de sélection de caractéristiques. Ces comparaisons ont été réalisées à partir des F_{nc} caractéristiques non-corrélées.

5.3.4.1 Analyse statistique classique

Matériel et méthode

Nous avons étudié les résultats obtenus par un test U de Mann-Whitney (étude prédictive), ainsi qu'à une analyse univariée de Kaplan-Meier (étude pronostique), toutes les 2 largement utilisées dans la littérature médicale.

Pour les 2 études, une analyse des courbes ROC a été réalisée. Cette étude a permis la définition de la valeur de seuil la plus discriminante permettant la différenciation de 2 groupes de patients (RC contre RNC par exemple). Une sensiblité, une spécificité et une AUC ont également pu être obtenues grâce à cette analyse.

Concernant l'étude prédictive, les relations entre les caractéristiques et la réponse au traitement à un mois ont été étudiées à l'aide du test U de Mann-Whitney. Une valeur de p inférieure à 5 % a été considérée comme statistiquement significative.

Pour évaluer la valeur pronostique des caractéristiques, un test de Kaplan-Meier a été utilisé pour estimer la distribution de survie. L'OS a été calculée à partir de la date du diagnostic initial jusqu'à la date du décès ou jusqu'à la fin du suivi. L'association entre l'OS et chaque caractéristique a été effectuée après un processus de dichotomisation à l'aide de la valeur de seuil définie au préalable par l'analyse des courbes ROC. La valeur pronostique de chaque caractéristique en termes de survie globale a été évaluée à l'aide du test log-rank.

Pour éviter de fausses conclusions, des corrections statistiques appropriées pour les erreurs de type I ont été effectuées selon [Chalkidou et al. 2015]. Une correction de Benjamini-Hochberg [Hochberg and Benjamin 1990] pour des tests d'hypothèses multiples a été appliquée pour chaque valeur de p calculée dans les études prédictives et pronostiques. De plus, pour l'étude pronostique, une correction des valeurs de p obtenues à partir des seuils optimaux a été effectuée en utilisant la formule d'Altman [Altman et al. 1994].

Une fois les caractéristiques prédictives (ou pronostiques) obtenues, plusieurs d'entre elles ont été combinées en une caractéristique unique à l'aide d'une *Régression Logistique* (RL). Le test U de Mann-Whitney et l'analyse de survie de Kaplan-Meier ont été réitérés afin d'étudier si la combinaison de ces caractéristiques permet d'améliorer les résultats.

Résultats et discussion

Les résultats de l'étude prédictive obtenus par statistiques classiques (test U de Mann-Whitney) sont présentés dans le Tableau 5.18.

Cinq caractéristiques ont montré un pouvoir prédictif significatif. Les caractéristiques "busyness" (GLDM, grp 9) et MTV (grp 6) sont celles présentant les meilleures performances avec une AUC respective de 0,810 et de 0,802, accompagnée d'une valeur de p inférieure à 0,0001. On peut constater la spécificité de 100 % du MTV. Si l'on compare nos résultats avec ceux obtenus par le test de Mann-Whitney U, on note que uniquement ces 2 meilleures caractéristiques montrent des performances proches de celles obtenues par les sous-ensembles sélectionnées par FIC et GARF (AUC moyenne de 0,826 et de 0,842, respectivement). Concernant les combinaisons par régression logistique, le meilleur ré-

TABLEAU 5.18 – Résultats de l'étude prédictive utilisant nos méthodes de sélection (Sp + FIC et Sp + GARF) ainsi que le test U de Mann-Whitney ($p < 0.05$) des patients atteints d'un cancer de l'œsophage. Les courbes ROC ont été créées pour obtenir une sensibilité (Se), une spécificité (Sp), une AUC et les valeurs de seuil

Caractéristiques	**Se (%)**	**Sp (%)**	**AUC**	**Seuil**	***p*-value**
Sp + FIC					
Sous-ensemble de caractéristiques :	81±8	91±12	0,826±0,104	-	-
- MTV (grp 6) et Homogénéité (GLCM, grp 8)					
Sp + GARF					
Sous-ensemble de caractéristiques :	77±14	92±11	0,842±0,099	-	-
- Sexe, Poids de forme (grp 1),					
- Perte de Poids, Localisation,					
- Stade T, MTV (grp 6), SUV_{max} (grp 7),					
- "Skewness" et Homogénéité (GLCM, grp 8)					
Test de Mann-Whitney U					
"Busyness" (GLDM)	66	88	0,810	$8{,}9e^{11}$	< 0,0001
MTV	51	100	0,802	9,330	0,0001
Perte de Poids	61	83	0,737	6,2	0,0015
Énergie	54	88	0,723	0,026	0,0030
GLNUz	76	75	0,718	10,55	0,0037
Combinaison des 5 par RL	71	67	0,680	-	0,0164

sultat est obtenu en combinant les 5 caractéristiques prédictives, cependant, ce processus ne permet pas d'améliorer les résultats. Cette méthode n'est pas assez efficace pour combiner des caractéristiques si différentes pour en sortir une information pertinente.

Concernant l'étude pronostique, aucune caractéristique n'a été détectée comme étant significativement pronostique en utilisant l'analyse de survie de Kaplan-Meier (Tableau 5.19).

TABLEAU 5.19 – Résultats de l'étude pronostique des patients atteints d'un cancer de l'œsophage utilisant les méthodes de sélection FIC et GARF ainsi que l'analyse univariée de Kaplan-Meier ($p < 0,05$).

Caractéristiques	**Se (%)**	**Sp (%)**	**AUC**
FIC			
Sous-ensemble de caractéristiques :	79 ± 9	95 ± 6	0,822 ± 0,059
- Stade OMS, NRI (grp 2) et MTV (grp 6)			
GARF			
Sous-ensemble de caractéristiques :	75±17	80±16	0,773±0,101
- Histologie, Localisation, Stade OMS			
- NRI (grp 2), SUV_{max} (grp 7), COV, SUV_{90} et V_{10}			
Analyse univariée de Kaplan-Meier			
Aucune	-	-	-

Ce dernier résultat est très en faveur des méthodes d'apprentissage automatique par RF par rapport aux statistiques classiques.

5.3.4.2 Analyse d'autres méthodes de sélection de caractéristiques

Matériel et méthode

Nos stratégies de sélection des caractéristiques RF ont été comparées, pour l'étude prédictive et pronostique, à d'autres méthodes de sélection de caractéristiques couramment rencontrées dans la littérature : SFFS, HFS, RFE et LASSO (voir Section 3.4 - Sélection des caractéristiques, page 78). Concernant la méthode HFS, basée sur le SVM, un noyau gaussien a été utilisé avec comme paramètres σ et c égalent à 1.

Comme nos méthodes de sélection ont été optimisées avec l'algorithme RF, nous avons également étudié l'influence du classifieur utilisé une fois la sélection réalisée. À cet effet, nous avons effectué une classification par SVM (noyau gaussien avec $\sigma = 1$ et

$c = 1$). Cette classification a mené à une erreur de classification notée SVM_{err}.

Résultats et discussion

Les résultats des classifications RF réalisées après sélection de caractéristiques par les différentes approches sont présentés Tableau 5.20 pour les études prédictives et pronostiques.

TABLEAU 5.20 – Résultats des études prédictives et pronostiques utilisant plusieurs méthodes de sélection de caractéristiques (FIC, GARF, SFFS, HFS, RFE et LASSO). Les moyennes et les écarts-types des indices de performances obtenus par validation sont indiqués. F_{sel} est la taille des sous-ensembles de caractéristiques sélectionnées par les méthodes. En gras sont présentés les meilleurs résultats.

Étude	Méthode	F_{sel}	RF_{err} (%)	SVM_{err} (%)	AUC	Se (%)	Sp (%)
Prédictive	**FIC**	**2**	**21 ± 11**	**35 ± 6**	**0,826 ± 0,104**	**81 ± 8**	**91 ± 12**
	GARF	9	26 ± 7	29 ± 6	0,842 ± 0,099	77 ± 14	92 ± 11
	SFFS	3	41 ± 6	35 ± 7	0,569 ± 0,087	53 ± 30	75 ± 31
	HFS	4	29 ± 12	41 ± 6	0,814 ± 0,093	77 ± 15	86 ± 15
	RFE	3	30 ± 9	38 ± 7	0,783 ± 0,113	67 ± 14	93 ± 11
	LASSO	12	29 ± 3	42 ± 7	0,779 ± 0,049	69 ± 10	86 ± 15
Pronostique	FIC	3	28 ± 5	33 ± 5	0,822 ± 0,059	79 ± 9	95 ± 6
	GARF	**8**	**27 ± 8**	**35 ± 4**	**0,773 ± 0,101**	**75 ± 17**	**80 ± 16**
	SFFS	3	42 ± 12	35 ± 7	0,547 ± 0,119	46 ± 29	79 ± 18
	HFS	3	48 ± 9	40 ± 9	0,526 ± 0,124	46 ± 29	79 ± 18
	RFE	3	36 ± 10	35 ± 6	0,621 ± 0,078	70 ± 13	68 ± 20
	LASSO	10	28 ± 7	29 ± 9	0,750 ± 0,064	88 ± 25	66 ± 28

Concernant la comparaison de nos méthodes à celles rencontrées dans la littérature pour l'étude prédictive, aucune des méthodes comparées ne présentent d'aussi bons résultats que nos méthodes FIC et GARF. Cependant, le test de "Wilcoxon signed rank" ne révèle aucune différence significative, à l'exception de la méthode SFFS qui apparait nettement comme moins performante avec une augmentation de l'erreur de classification RF de 20 % par rapport à FIC ($p = 0{,}0004$) et de 16 % par rapport à la méthode GARF ($p = 0{,}0002$).

Concernant la comparaison de nos méthodes à celles rencontrées dans la littérature pour l'étude pronostique (voir Tableau 5.20), aucune des méthodes comparées ne présentent d'aussi bons résultats que FIC et GARF. Le test de "Wilcoxon signed rank" ré-

vèle que les méthodes FIC, GARF, ainsi que LASSO permettent l'obtention significative de meilleurs résultats par rapport aux 3 autres ($p \leq 0{,}039$).

Concernant l'utilisation du SVM comme classifieur, on observe une tendance à une augmentation de l'erreur de classification pour les études prédictives et pronostiques sauf lors de l'utilisation du SFFS. La méthode HFS, qui pourtant est liée au SVM dans son processus de sélection, ne montre pas d'amélioration dans ses performances. Enfin, les résultats concernant nos méthodes de sélection sont détériorés. Pour nos méthodes FIC et GARF, ce résultat pourrait s'expliquer par le fait que nos méthodes ont été optimisées pour l'algorithme RF et que, par conséquent, utiliser un autre classifieur, comme le SVM, lors de l'évaluation réduit les performances finales.

Néanmoins, nos méthodes donnent tout de même les meilleurs performances de classification, à l'exception de l'étude pronostique après la méthode LASSO qui donne des performances très stables ($RF_{err} = 28 \pm 7$, $SVM_{err} = 29 \pm 9$). Cependant, il n'existe pas de différence significative entre les performances obtenues par cette méthode et celles obtenues par nos méthodes ($p > 5$ %).

Ces résultats sont à nuancer car le classifieur SVM n'a pas bénéficié de la même optimisation de ses paramètres (σ et c) que les RF pour l'étude de la base de données du cancer de l'œsophage. Nous avons privilégié l'utilisation des paramètres définis lors de travaux antérieurs réalisés au sein de notre laboratoire sur la base de données du cancer pulmonaire [Mi et al. 2015].

5.4 Étude de la base de données du cancer du poumon

5.4.1 Matériel et méthode

En utilisant la même procédure que précédemment, nous avons étudié l'apport de nos méthodes de sélection des caractéristiques sur une autre base de données de patients atteints d'un cancer du poumon. Les paramètres sélectionnés pour cette étude sont les mêmes que ceux définis comme optimaux pour le cancer de l'œsophage et sont résumés Tableau 5.21.

TABLEAU 5.21 – Valeurs des paramètres optimaux des méthodes de sélection de caractéristiques (FIC et GARF).

Légende	**Détails**
Lésion	Cancer du poumon
Méthode de validation	"Leave One Out"
Base d'apprentissage	D_{app} (17 observations)
Base de test	D_{test} (8 observations)
Paramètres :	
Étude des corrélations de Spearman :	
Seuil de $\|\rho\|$	0,8
Type de ré-échantillonnage	Absolu
RF :	
Nombre d'arbres T	500
Valeur de f	$\sqrt{F}$
FIC :	
Seuil du coefficient d'importance S	50 %
GARF :	
Nombre de génération du GA $nGen$	40
Taille des populations du GA $nPop$	20
Poids de l'AUC α	10
Poids de l'OOB_{err} β	10

5.4.2 Résultats et discussion

Les résultats de l'analyse des corrélations de rang de Spearman sont présentés Tableau 5.22.

TABLEAU 5.22 – Corrélations des caractéristiques des patients atteints d'un cancer du poumon pour un seuil de $|\rho|$ de 0,8. En gras sont représentées les caractéristiques sélectionnées F_{nc} pour l'étape suivante ($F_{nc} = 17$).

ρ	Grp	Caractéristiques
0,8	1	**SUV_{max}** - SUV_{peak}
	2	**MTV** - TLG
	3	**LGZE (GLSZM)** - SZLGE (GLSZM)
	4	**LZHGE (GLSZM)** - LZE (GLSZM) - GLNUz (GLSZM) - ZP (GLSZM) - LZLGE (GLSZM)

Parmi les 24 caractéristiques initiales, 4 groupes de caractéristiques sont créés. On remarque que les groupes 3 et 4 sont constitués uniquement de caractéristiques de la matrice GLSZM. Ainsi, 7 caractéristiques sont éliminées pour mener à un nombre $F_{nc} = 17$ caractéristiques non-corrélées utilisées pour l'étape suivante.

Concernant l'étude des caractéristiques à l'aide du test U de Mann-Whitney, on note

qu'aucune caractéristique n'a été détectée comme significativement prédictive alors que nos 2 méthodes de sélection de caractéristiques proposent des solutions classifiant convenablement les observations.

Les résultats des classifications de l'étude prédictive sans et avec différentes méthodes de sélection de caractéristiques (FIC, GARF, SFFS, HFS, RFE et LASSO) pour les patients atteints d'un cancer du poumon sont présentées Tableau 5.23.

TABLEAU 5.23 – Résultats des études prédictives sans et avec différentes méthodes de sélection de caractéristiques (FIC, GARF, SFFS, HFS, RFE et LASSO). Les moyennes et les écarts-types des indices de performance obtenus par validation sont indiqués. F_{sel} est la taille des sous-ensembles de caractéristiques sélectionnées par les méthodes.

Étude	Méthode	F_{sel}	RF_{err} (%)	RF_{err} (pat)	AUC	Se (%)	Sp (%)
Prédictive	Sans sélection	/	26,3 ± 7,1	2,1 ± 0,6	0,450 ± 0,148	55 ± 35	80 ± 26
	FIC	2	12,5 ± 8,3	1,0 ± 0,7	0,871 ± 0,149	83 ± 21	95 ± 16
	GARF	4	20,0 ± 12,1	1,6 ± 1,0	0,804 ± 0,178	83 ± 21	90 ± 21
	SFFS	3	30,0 ± 8,7	2,4 ± 0,7	0,504 ± 0,117	52 ± 34	75 ± 26
	HFS	2	40,0 ± 17,5	3,2 ± 1,4	0,713 ± 0,122	48 ± 23	100 ± 0
	RFE	3	21,3 ± 6,0	1,7 ± 0,5	0,688 ± 0,173	83 ± 21	75 ± 26
	LASSO	12	30,0 ± 16,9	2,4 ± 1,4	0,575 ± 0,121	68 ± 25	75 ± 26

Concernant l'étude prédictive, l'amélioration de l'erreur de classification est de 13,8 % pour la méthode FIC et de 6,3 % pour la méthode GARF. Les conclusions sont les mêmes concernant l'AUC. La méthode FIC est la méthode présentant les meilleures performances prédictives avec une faible erreur de classification (12,5 %) et une AUC élevée (0,871). La méthode HFS est celle présentant les plus faibles résultats avec 40,0 % d'erreur de classification. D'après le "Wilcoxon signed rank test", les méthodes FIC, GARF et RFE ne sont pas significativement différentes entre elles (au risque $\alpha = 5$ %) et présentent des performances significativement meilleures que les autres méthodes de sélection SFFS, HFS et LASSO.

Au terme de l'analyse de cette cohorte de patients, on note que les méthodes de sélection proposées (FIC et GARF) apportent véritablement une amélioration lors de la classification par RF, notamment la méthode FIC. De plus, on note que l'utilisation des paramètres optimaux définis sur la base de données de l'œsophage convient également à la base du cancer du poumon.

5.5 Discussion générale

Dans ce chapitre, nous avons choisis d'étudier 2 bases de données (œsophage et poumon). Nous avons privilégié l'homogénéité des protocoles de traitement et d'acquisition des images TEP afin de réduire les biais engendrés par ces derniers. Ce choix nous a mené à l'étude de bases de tailles modestes (65 patients pour l'œsophage et 25 pour le poumon).

La taille des bases de données ne nous permet pas d'étudier l'intégralité des caractéristiques utilisées dans la littérature, comme d'autres caractéristiques de forme ou les fractales. C'est pour cette raison que nous nous sommes limités à l'étude des caractéristiques de l'image TEP, sans prendre en compte celles extraites des images TDM. Concernant la base de données du cancer du poumon, seules 5 caractéristiques du premier ordre et les textures issus de la matrice GLSZM étaient à notre disposition. Cela réduit la portée des conclusions pour cette base de données. C'est aussi pour réduire le nombre de caractéristiques qu'une première élimination des caractéristiques corrélées a été réalisée à l'aide de l'étude de Spearman.

Il existe un déséquilibre entre les classes aussi bien pour l'étude prédictive que pour l'étude pronostique (voir Tableaux 5.4 et 5.2). Ce déséquilibre engendre potentiellement un biais lors de l'analyse des résultats. En effet, il est possible d'obtenir une faible erreur de classification mais qu'en réalité le modèle privilégie la classe majoritaire et n'arrive pas à bien identifier les patients de la classe minoritaire. S'intéresser en parallèle à d'autres indices de performance comme l'AUC permet de contrôler ce genre de phénomène car il prend en compte la répartition des classifications.

La méthode d'évaluation a été ajustée à la taille des populations observées. Ainsi, la méthode des k-fold n'a pas été utilisée en raison du faible nombre d'observations dans les bases de données. Nous avons ainsi privilégié les permutations aléatoires pour l'œsophage, offrant un bon compromis entre résultats et temps de calcul, et une validation croisée de type "leave-one-out" pour le cancer du poumon, car le nombre de patients dans cette base est faible. De plus, ce manque de patients peut engendrer une difficulté à révéler des différences significatives lors de nos expérimentations dû à un faible nombre d'itérations ($k = 10$).

A propos des paramètres par défaut utilisés lors des premières classifications (voir Tableau 5.8), nous nous sommes basés sur la littérature lorsque cela a été posssible. C'est le cas pour les paramètres de l'étude des corrélations et du RF. En revanche, d'autres paramètres par défaut assignés aux méthodes FIC et de GARF ont été définis de manière arbitraire. Pour ces derniers, nous les avons définis en fonction d'un compromis entre performances et temps de calculs. Malgré le manque de données significatives, l'utilisation du ré-échantillonnage absolu montre un apport significatif aux classifications face au ré-échantillonnage relatif, révélant l'importance de cette étape dans l'étude des caractéristiques d'images.

Concernant nos résultats, on note une amélioration des performances de classification lorsque les méthodes FIC et GARF sont combinées à une analyse des corrélations de Spearman. De cette manière, la sélection de caractéristiques permet d'améliorer les résultats des classification réalisées sans sélection des caractéristiques ou avec d'autres méthodes rencontrées dans la littérature. Nos résultats sont particulièrement révélateurs lors de l'étude pronostique du cancer de l'œsophage. Les résultats obtenus par les autres méthodes sont médiocres, en particulier l'analyse de Kaplan-Meier qui ne détecte aucune caractéristique significative. En revanche, nos 2 méthodes aboutissent à des résultats satisfaisants. Seule la méthode LASSO aboutit à des performances équivalentes à nos méthodes lors des études prédictives et pronostiques du cancer de l'œsophage. Cependant, ses performances sont réduites lors de l'étude du cancer du poumon.

5.6 Conclusion

Au terme de ce chapitre, une tendance de l'amélioration des performances de classification a été observée lorsque nos propositions de méthodes de sélection des caractéristiques sont utilisées. Ces tendances ont été observées aussi bien en prédictif qu'en pronostique, pour les différentes pathologies étudiées (cancer de l'œsophage et du poumon) par rapport aux statistiques classiques généralement utilisées en médecine.

Nous avons dû au préalable rechercher les paramètres de ces méthodes permettant d'optimiser les résultats. Plusieurs séries d'expérimentations réalisées sur la cohorte de

patients atteints d'un cancer de l'œsophage ont permis de définir un ensemble de paramètres optimaux (voir Tableau 5.16). Ces derniers ont été utilisés pour les études des 2 bases de données. Les bons résultats obtenus sur la deuxième base nous font penser que ces paramètres ne sont pas spécifiques à la base de données de l'œsophage. Nous avons également montré l'importance des 2 étapes de sélection successives dans nos méthodes (analyse des corrélations de Spearman suivie d'une sélection de type enveloppante).

Enfin, nous avons comparé nos méthodes de sélection face à d'autres méthodes largement utilisées dans la littérature du traitement de l'image (SFFS, HFS, RFE et LASSO) ou médicale (test de Mann-Whitney ou Kaplan-Meier). Les performances de nos méthodes sont significativement meilleures concernant l'étude pronostique pour la cohorte des patients atteints d'un cancer de l'œsophage. Cependant, les améliorations ne sont pas significatives pour l'étude prédictive de cette cohorte malgré l'existence d'une tendance. Enfin, l'apport prédictif de nos méthodes appliquées à la cohorte de patient atteints d'un cancer pulmonaire est significatif.

Conclusions et perspectives

Synthèse

Dans cette thèse, nous avons répondu à la problématique liée à la définition de sous-ensembles de caractéristiques présentant une valeur pronostique ou prédictive de la réponse au traitement en cancérologie. Cette problématique se place dans le concept de radiomique cherchant à mettre en place une médecine personnalisée basée sur un grand nombre de caractéristiques multimodales. L'objectif prometteur est d'anticiper, à partir d'un examen TEP au FDG initial, la survie du patient ou la réponse à un traitement.

Notre travail de thèse a débuté par la création d'une base de données de patients atteints du cancer de l'œsophage ayant bénéficiés d'un examen TEP au FDG et le développement d'outils de calcul des caractéristiques radiomiques. À partir des dossiers médicaux et des examens d'imagerie, nous avons extrait une liste de caractéristiques cliniques et radiomiques en nous basant sur celles qui sont utilisées dans la littérature. Toujours d'après la littérature, nous avons réalisé un état de l'art de l'étude des caractéristiques extraites des images TEP pour le cancer de l'œsophage et du poumon. Nous avons également listé les différents paramètres pouvant modifier leur valeur et leur robustesse pour établir la réponse à un traitement ou leur valeur pronostique. À la fin de notre chapitre 2, nous avons recensé un grand nombre de problèmes liés à l'utilisation de la radiomique (acquisition des images, méthode de segmentation et de ré-échantillonnage, etc). De plus, les méthodes de statistiques classiques utilisées dans la littérature médicale ne paraissent pas adaptées à cette problématique en raison du grand nombre de caractéristiques étudiées.

C'est pour pallier à ces problèmes, que nous nous sommes tournés vers les méthodes

d'apprentissage automatique avec la sélection de caractéristiques. Les méthodes d'apprentissage automatique sont particulièrement adaptées à la sélection de caractéristiques parmi un grand nombre et apportent un éclaircissement pour savoir quelles caractéristiques sont pertinentes, aussi bien en terme individuelle que de sous-ensemble. Nous avons fait le choix d'utiliser l'algorithme des RF car cette méthode est robuste au faible nombre de données d'apprentissage grâce à la multiplication artificielle des échantillons et aux processus aléatoires entrant en jeu dans la construction des arbres. De plus, l'algorithme des RF est capable de gérer des données multimodales binaires, discrètes ou continues.

Nous avons ainsi développé nos propres algorithmes de sélection des caractéristiques radiomiques ayant une valeur pronostique ou prédictive de la réponse au traitement en cancérologie basées les forêts aléatoires. Nos méthodes, intitulées FIC [Desbordes et al. 2016] et GARF [Desbordes et al. 2017], portent sur 2 étapes successives de sélection. La première est une sélection filtrante basée sur l'analyse des corrélations de rang de Spearman permettant d'éliminer les caractéristiques corrélées. La deuxième étape diffère selon nos méthodes mais reste basée sur l'utilisation des RF et aboutit à la création de sous-ensembles de caractéristiques réduisant l'erreur de classification. Ainsi, la méthode FIC est axée sur l'étude des caractéristiques les plus importantes dans la minimisation de l'erreur de classification au sein du RF, alors que la méthode GARF utilise une fonction d'évaluation incluse au sein d'un algorithme génétique pour minimiser l'erreur de classification, maximiser une AUC tout en respectant une contrainte parcimonieuse.

Les paramètres utilisés au cœur de nos méthodes ont d'abord été optimisés. Malgré le manque de données significatives, probablement dû au faible nombre de patients dans notre base de données, l'utilisation du ré-échantillonnage absolu montre un apport significatif des classifications face au ré-échantillonnage relatif, révélant l'importance de cette étape dans l'étude des caractéristiques images. Après optimisation de leurs paramètres, ces méthodes ont permis la sélection de sous-ensembles de caractéristiques prédictives et pronostiques sur 2 bases de données (cancer de l'œsophage et du poumon).

Nous avons tout d'abord montré que les meilleurs performances ont été obtenues lors de la combinaison des 2 sélections successives. Finalement, nous avons comparé nos mé-

thodes de sélection face à d'autres méthodes largement utilisées dans la littérature du traitement de l'image (SFFS, HFS, RFE et LASSO) ou médicale (test de Mann-Whitney ou Kaplan-Meier). Nos algorithmes sont les méthodes apportant les meilleures performances de classification parmi l'ensemble des méthodes étudiées.

Ainsi, pour l'étude prédictive de la base de données du cancer de l'œsophage, l'erreur de classification RF pour la méthode FIC associée à l'analyse de Spearman a été de 21,4 % ± 10,5 avec une AUC de 0,826 ± 0,104 et pour la méthode GARF associée à l'analyse de Spearman de 25,5 % ± 7,2 avec une AUC de 0,842 ± 0,099. Ces résultats ont montré une tendance à l'amélioration des performances sans sélection de caractéristiques (25,5 % ± 6,5 avec une AUC de 0,798 ± 0,084) et également face aux méthode concurrentes. Cependant, cette tendance est non-significative selon le test "Wilcoxon signed rank" ($p > 5$ %). Les observations sont du même ordre concernant l'étude prédictive de la base de données du cancer du poumon, où nos méthodes donnent les meilleures performances.

Au contraire, l'amélioration des résultats est significative pour l'étude pronostique de la base de données du cancer de l'œsophage ($p \leq 0{,}039$). Ainsi, l'erreur de classification RF pour la méthode FIC associée à l'analyse de Spearman a été de 27,7 % ± 4,5 avec une AUC de 0,822 ± 0,059 et pour la méthode GARF associée à l'analyse de Spearman de 26,7 % ± 7,5 avec une AUC de 0,773 ± 0,101.

Concernant les caractéristiques, le volume tumoral métabolique présente un rôle prédictif et pronostique important car il est présent dans les sous-ensembles sélectionnés par nos 2 méthodes, excepté l'étude pronostique avec la méthode GARF. Ces observations confirment l'intérêt particulier de cette caractéristique trouvé dans la littérature concernant le suivi de la réponse au traitement, ainsi que la survie du patient [Van De Wiele et al. 2013]. La conclusion concernant l'apport du SUV_{max} est plus mitigée. Cette dernière est uniquement sélectionnée par la méthode GARF lors des études prédictives et pronostiques, alors qu'elle est considérée dans la littérature comme étant une caractéristique incontournable [Van De Wiele et al. 2013].

Concernant les caractéristiques de texture, elles semblent présenter un intérêt pour la prédiction de la réponse au traitement des patients. En effet, nos 2 méthodes ont sélectionné l'homogénéité (*"Gray Level Cooccurrence Matrix"*) lors de l'étude prédictive. Enfin,

le fait que les meilleures sous-ensembles de caractéristiques sont des combinaisons des caractéristiques cliniques et de l'image TEP (1er ordre et indices de texture) montre que ces 2 sources d'informations sont complémentaires.

Perspectives

Nous avons proposé plusieurs algorithmes de sélection des caractéristiques radiomiques ayant une valeur pronostique ou prédictive de la réponse au traitement en cancérologie basées sur une méthode d'apprentissage automatique des forêts aléatoires. Nos résultats montrent que l'utilisation des méthodes d'apprentissage automatique donnent de bonnes performances en comparaison aux statistiques classiques. L'utilisation de ces méthodes est donc encourageante et nous pousse à continuer dans cette voie. Cependant, l'étude de la littérature montre qu'une certaine rigueur est requise dans la modélisation et l'évaluation de ces méthodes [Chalkidou et al. 2015].

Afin de respecter ces contraintes et d'obtenir des résultats significatifs, il est nécessaire d'avoir à disposition une importante base de données de patients homogènes afin de limiter les biais pouvant apparaitre lors de l'acquisition des images. Cependant, cela réduit le nombre de patients étudiés au sein d'une cohorte. Il pourrait être intéressant de réaliser une étude multicentrique afin de regrouper un nombre important de patients pour augmenter la taille des bases de données. La mise en place d'un protocole d'étude regroupant la méthodologie à suivre pour ces différents centres est nécessaire (paramètres de reconstruction des images, méthodes de segmentation, type de ré-échantillonnage utilisé, etc). Ce type d'étude soulève cependant la problématique de la robustesse des caractéristiques radiomiques vis-à-vis de l'imageur utilisé.

De plus, notre étude s'est concentrée sur l'utilisation des caractéristiques cliniques associées à celles extraites de l'imagerie TEP. Cependant, nous avons vu dans la littérature que de nombreuses autres caractéristiques pouvaient être intégrées dans les études radiomiques. On peut citer par exemple les données protéomiques, génomiques ainsi que des caractéristiques pouvant être extraites d'autres modalités d'imagerie comme le TDM et l'IRM. Il est possible d'étudier l'intégralité de ces caractéristiques afin d'en étudier les

différents apports en cancérologie mais cela soulève le problème de la multiplication des biomarqueurs.

Dans notre étude, 2 bases de données de patients présentant des cancers différents ont été étudiées (œsophage et poumon). Il serait intéressant d'étudier d'autres types de lésions et d'étudier les caractéristiques qui ressortent comme étant prédictives ou pronostiques dans chaque cas. Cela permettrait de mettre en avant une ou des caractéristiques déterminantes dans le fonctionnement globale du cancer. Ainsi, il pourrait être intéressant de confronter les experts à ces différents sous-ensembles de caractéristiques créés afin qu'ils puissent interpréter ces résultats et relier les caractéristiques à des événements biologiques.

Concernant les caractéristiques extraites des images, l'étude de caractéristiques temporelles, mesurant leur évolution au cours du temps, pourrait permettre l'amélioration des performances lors des études pronostiques et prédictives. Il peut peut être difficile d'obtenir des performances suffisamment fiables à partir d'un examen TEP initial pour que les classifications soient utilisables en routine clinique, c'est pourquoi l'étude de plusieurs examens pourrait aider. Cependant, la multiplication des caractéristiques pose à nouveau le problème de la taille des bases de données.

Ces dernières années, nous avons été témoin de l'explosion de l'intérêt des approches de "deep learning" en imagerie médicale. Ces approches présentent une amélioration croissante de leurs résultats et semblent très prometteurs. De plus, elles permettent de s'affranchir de l'étape d'extraction des caractéristiques en privilégiant des caractéristiques plus "profondes" ou abstraites.

Bibliographie personnelle sur le sujet

Articles

Toledano M, **Desbordes P**, Banjar A, Gardin G, Vera P, Ruminy P, Jardin F, Tilly H, Becker S. Combination of baseline FDG-PET/CT metabolic tumor volume and gene-expression profile have a robust predictive value in patients with Diffuse Large B-Cell Lymphoma. *Clinical Cancer Research*, 2017. Soumis.

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Predictive Value of Initial FDG PET Features for Treatment Response and Survival in Esophageal Cancer Patients Treated with Chemo-Radiation Therapy Using a Random Forest Classifier. *Plos One*, 2017. doi : 10.1371/journal.pone.0173208.

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Comput Med Imag Grap*, 2016. doi : 10.1016/j.compmedimag.2016.12.002.

Desbordes P, Petitjean C, Ruan S. Segmentation of lymphoma tumor in PET images using cellular automata : A preliminary study. *IRBM*. 2016 ; 37 : 3—10.

Desbordes P, Petitjean C, Ruan S. 3D Automated Lymphoma Segmentation in PET Images Based on Cellular Automata. *4th International Conference on IPTA 2014*. doi : 10.1109/IPTA.2014.7

Présentations à des conférences internationales

Desbordes P, Modzelewski R, Ruan S, Vera P, Gardin I. Predictive Value of Initial FDG-PET Features for Treatment Response and Survival in Esophageal Cancer Patients Treated with Chemoradiotherapy using a Random Forest Classifier. SNMMI 2017 Annual Meeting (Denver, USA, Juin 2017).

Gardin I, Fdhila MT, **Desbordes P**, Lebtahi R, Dieudonne A. Predictive value of dosimetry indices for treatment response in Liver Cancer Patients Treated with Yttrium 90 Microspheres using a Random Forest Algorithm. SNMMI 2017 Annual Meeting (Denver, USA, Juin 2017).

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Prognostic and Predictive Values of Initial 18FDG PET Features Using Random Forest Classifier : Application to Patients after Chemo-radiotherapy for Oesophageal Cancer. 28th Annual EANM Congress (Hamburg, Germany, Oct 2015).

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Prognostic and Predictive Feature Selection for Oesophageal Cancer Using Random Forest Classifier. 18th MICCAI Conference (Munich, Germany, Oct 2015).

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Détermination des valeurs prédictives et pronostiques des caractéristiques TEP initiales au 18FDG basée sur la méthode du random forest chez les patients atteints d'un cancer de l'œsophage traités par radio-chimiothérapie. GDR-isis : "Statistical Learning Methods and Applications to health"(Paris, France, Dec 2015).

Présentations affichées à des conférences internationales

Desbordes P, Modzelewski R, Ruan S, Vauclin S, Pineau P, Vera P, Gardin I. Prognostic and Predictive Feature Selection for Oesophageal Cancer Using Random Forest Classifier. 18th MICCAI Conference (Munich, Germany, Oct 2015).

Gouel P, **Desbordes P**, Di Fiore F, Vera P, Modzelewski R. Predictive value of Intra Tumor Heterogeneity on Baseline 18F-FDG PET Images in Esophageal Cancer. SNM conference (Boston, USA, Aout 2015).

Annexe A

Description des caractéristiques

A.1 Caractéristiques cliniques

A.1.1 Classification TNM

La classification du Stade TNM est un système international permettant de classer les cancers selon leur extension anatomique. Les trois lettres symbolisent la propagation de la maladie cancéreuse sur le site de la tumeur primitive (T), dans les ganglions lymphatiques voisins (N) et à distance pour d'éventuelles métastases (M).

Le stade T est évalué à l'aide d'un examen clinique, de l'imagerie et/ou de l'exploration chirurgicale. Par exemple, pour le cas du cancer de l'œsophage, cette caractéristique correspond à la profondeur d'envahissement de la paroi œsophagienne par la tumeur primitive et son développement dans les tissus environnants :

— T0 : pas de signe de tumeur primitive

— T1 : la tumeur envahit la muqueuse ou la sous-muqueuse (T1a et T1b)

— T2 : la tumeur envahit la musculeuse

— T3 : la tumeur envahit l'adventice

— T4 : la tumeur envahit les structures adjacentes (autres organes, ...)

Le stade N est le nombre de ganglions lymphatiques entourant qui contiennent des cellules cancéreuses et leur emplacement :

— N0 : pas de signe d'atteinte des ganglions lymphatiques régionaux

— N1 : propagation dans 1 ou 2 ganglions lymphatiques voisins

— N2 : propagation dans 3 à 6 ganglions lymphatiques voisins

— N3 : propagation dans plus de 7 ganglions lymphatiques voisins

Le stade M (métastase) correspond à la propagation du cancer à d'autres parties du corps :

— M0 : absence de métastase à distance

— M1 : présence de métastase, le cancer s'est propagé à une autre partie du corps

Une fois ces caractéristiques définies elles sont combinées entre elles formant ainsi le Stade TNM (Tableau A.1).

TABLEAU A.1 – Regroupement des stades T (tumor), N (nodes) et M (metastatis) en stades TNM.

Stade TNM	**Stade T**	**Stade N**	**Stade M**
Stade 0	T *in situ*	N0	M0
Stade IA	T1	N0	M0
Stade IB	T2	N0	M0
Stade IIA	T3	N0	M0
Stade IIB	T1, T2	N1	M0
Stade IIIA	T4a	N0	M0
-	T3	N1	M0
-	T1, T2	N2	M0
Stade IIIB	T3	N2	M0
Stade IIIC	T4a	N1, N2	M0
-	T4b	tous N	M0
-	tous T	N3	M0
Stade IV	tous T	tous N	M1

A.1.2 Échelle de performance

L'échelle de performance est un indice qui permet d'évaluer en oncologie l'état de santé général du patient, ainsi que ses activités quotidiennes. Cette évaluation peut par exemple déterminer si l'état du patient permet l'administration d'une chimiothérapie et si les doses doivent être ajustées. Plusieurs systèmes d'évaluation sont disponibles, comme le score de Karnofsky [Karnofsky and Burchenal 1949] et celui de Zubrod [Oken et al. 1982] employés par l'OMS.

— Grade 0 : le patient est capable d'une activité identique à celle précédent la maladie, aucune restriction.

— Grade 1 : l'activité physique du patient est diminuée, mais il est toujours capable de mener un travail et de se déplacer

— Grade 2 : le patient est capable de se déplacer et prendre soin de lui, cependant il est incapable de travailler. De plus, il est alité plus de la moitié de son temps.

— Grade 3 : le patient est uniquement capable de réaliser quelques soins personnels et est alité plus de la moitié de son temps.

— Grade 4 : le patient est incapable de prendre soin de lui-même. Il est alité ou assis en permanence.

A.1.3 Nutritional Risk Index

Le NRI est un indice nutritionnel simple représentant la dénutrition d'un patient dépendant de l'albuminémie et de son poids (Équation A.1).

$$NRI = \text{albumine plasmatique (g/l)} \times 1,519 + 41,7\frac{\text{poids actuel (kg)}}{\text{poids de forme (kg)}} \tag{A.1}$$

L'albumine est une protéine essentielle pour le maintien de la pression osmotique indispensable à la bonne répartition des liquides entre les vaisseaux sanguins et les tissus. L'intervalle normal de concentration en albumine dans le sang est de 34 à 46 g.L^{-1}.

A partir du NRI, les patients peuvent être répartis en 3 classes :

— NRI > 97,5 : absence de dénutrition ;

— 83,5 < NRI ≤ 97,5 : dénutrition moyenne ;

— NRI ≤ 83,5 : dénutrition sévère.

A.2 Caractéristiques de texture

Dans la Figure A.1 est illustré en 2D la notion de couple distance/direction nécessaire à la construction de certaines matrices de texture. En 3D, la direction peut être représentée par un vecteur tridimensionnelle.

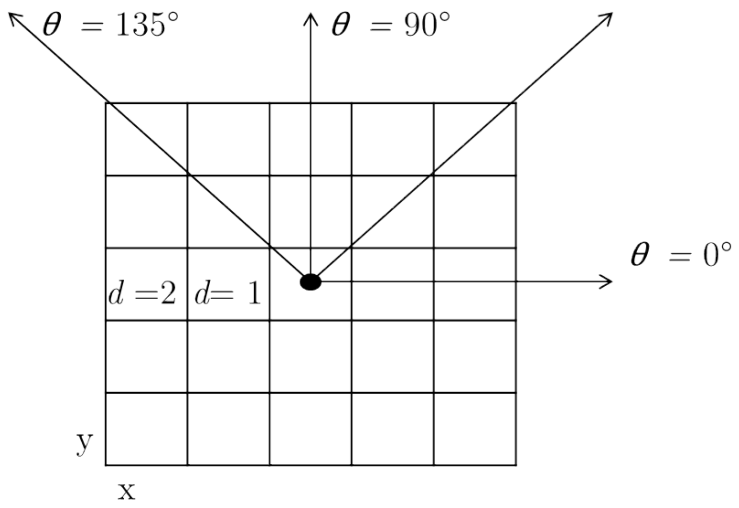


FIGURE A.1 – Illustration 2D de la notion de couple distance/direction dans la construction des matrices de texture.

Il existe 13 vecteurs indépendants dans l'espace 3D (Tableau A.2) qui sont, en général, utilisées avec une distance de 1.

TABLEAU A.2 – Liste des 13 directions 3D indépendantes.

1	0	0
0	1	0
1	1	0
1	-1	0
0	0	1
0	1	1
0	-1	1
-1	0	1
1	0	1
1	1	1
1	-1	1
1	-1	-1
1	1	-1

A.2.1 Matrice de cooccurrence (GLCM)

La construction des GLCM se fait de la manière suivante. Un élément de la matrice de cooccurrence $C(i,j)$ d'une image I de taille $\omega = X \times Y \times Z$ est défini comme la probabilité d'avoir 2 voxels d'intensité i et j, séparés par un vecteur (d,θ) correspondant à un

déplacement $(\Delta x, \Delta y, \Delta z)$:

$$C_{d,\theta}(i,j) = \frac{1}{\omega} \sum_{x,y,z}^{X,Y,Z} \begin{cases} 1, \text{ si } I(x,y,z) = i \text{ et } I(x+\Delta x, y+\Delta y, z+\Delta z) = j \\ 0 \text{ sinon} \end{cases} \tag{A.2}$$

Où l'image I peut être considérée comme une fonction de trois variables x, y et z avec $x \in [1,X]$, $y \in [1,Y]$ et $z \in [1,Y]$. $I(x,y,z)$ peut prendre des valeurs discrètes telles que $i = 0,1,\ldots,G-1$, où G est le nombre de niveaux de gris dans l'image. La construction d'une matrice est illustré dans la Figure A.2.

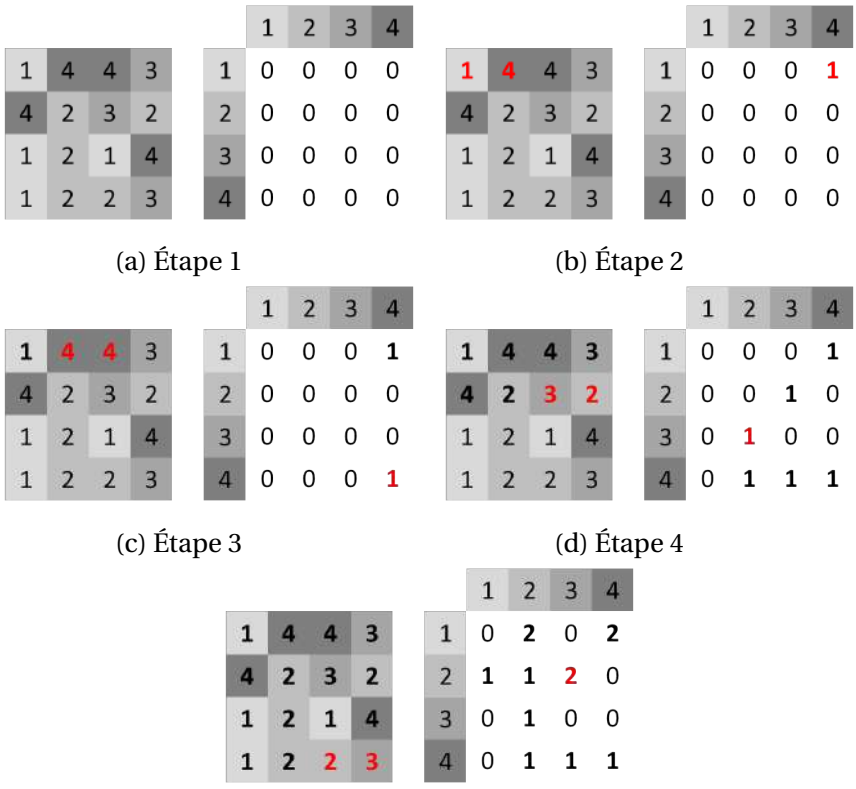


FIGURE A.2 – Illustration de la construction par étape d'une matrice de cooccurrence en fonction d'un couple $\theta = 0$ et $d = 1$.

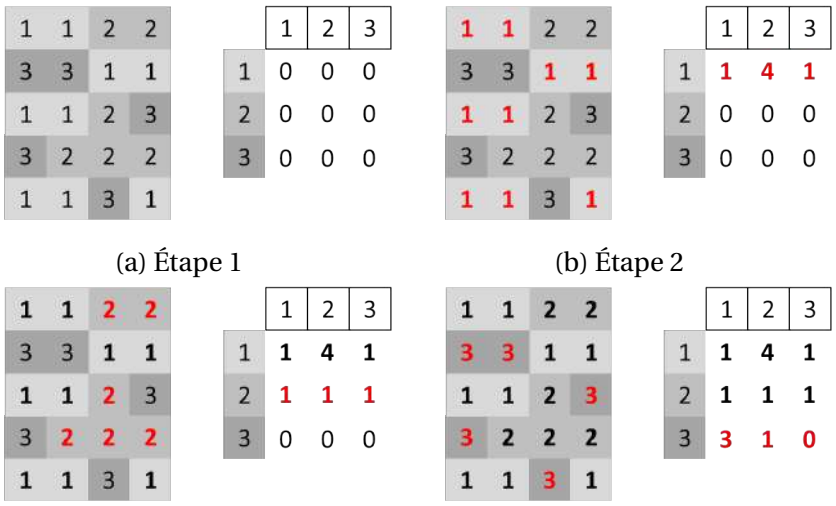
À partir des matrices de cooccurrence, il est possible d'extraire plusieurs caractéristiques de texture, aussi appelées paramètres d'Haralick (Tableau A.3).

TABLEAU A.3 – Caractéristiques issues de la matrice de cooccurrence C. G est la valeur d'intensité rééchantillonnée maximale, et (μ_i, σ_i) ou (μ_j, σ_j) sont la moyenne et la variance en colonne ou en ligne de la matrice C [Orlhac et al. 2014] supp data.

Caractéristiques	**Formules**
Variance	$\sum_{i,j}^{G} C(i,j)[(i-\mu_x)^2 + (j-\mu_y)^2]$
Énergie	$\sum_{i,j}^{G} C(i,j)^2$
Entropie	$-\sum_{i,j}^{G} C(i,j)\log(C(i,j))$
Corrélation	$\sum_{i,j}^{G} C_{ij} \times \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y}$
Dissimilarité	$\sum_{i,j}^{G} C(i,j)\lvert i-j\rvert$
Contraste	$\sum_{i,j}^{G} C(i,j)(i-j)^2$
Homogénéité	$\sum_{i,j}^{G} \frac{C(i,j)}{1+\lvert i-j\rvert}$
Moment Différentiel Inverse (IDM)	$\sum_{i,j}^{G} \frac{C(i,j)}{1+(i-j)^2}$
Cluster Shade	$\sum_{i,j}^{G} C(i,j) \times (i+j-\mu_x-\mu_y)^3$
Cluster Tendency	$\sum_{i,j}^{G} C(i,j) \times (i+j-\mu_x-\mu_y)^4$

A.2.2 Matrice des longueurs de plages homogènes (GLRLM)

La construction des GLRLM est illustré dans la Figure A.3.



(a) Étape 1 (b) Étape 2

(c) Étape 3 (d) Étape 4

FIGURE A.3 – Illustration de la construction par étape d'une matrice de longueur de plage homogène en fonction d'un couple $\theta = 0$ et $d = 1$.

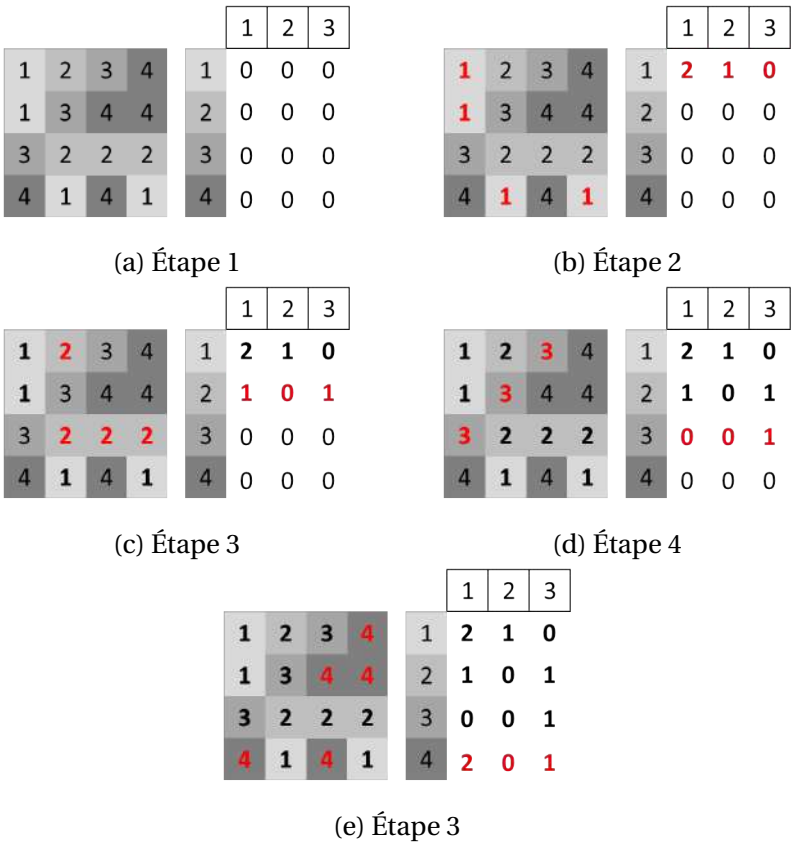
Le premier élément de la matrice correspond au nombre d'itération qu'un voxel d'intensité 1 possède un voisin de la même intensité dans la direction 0 degré. Les caractéristiques présentées dans le tableau A.4 peuvent être extraites des matrices GLRLM.

TABLEAU A.4 – Caractéristiques régionales issues de la matrice de longueurs de plages homogènes R. N_r est le nombre total de zones homogènes, G est la valeur d'intensité rééchantillonée maximale, N la taille maximales des zones et $Z(i,j)$ un élément de la matrice ayant pour intensité i et pour taille j [Orlhac et al. 2014] supp data.

Caractéristiques	**Formules**
SRE pour Small Run Emphasis	$\frac{1}{N_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{R(i,j)}{j^2}$
LrE pour Large Run Emphasis	$\frac{1}{N_r}\sum_{i=1}^{M}\sum_{j=1}^{N}R(i,j)\times j^2$
LGRE pour Low Intensity Run Emphasis	$\frac{1}{N_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{Z(i,j)}{i^2}$
HGRE pour High Intensity Run Emphasis	$\frac{1}{N_r}\sum_{i=1}^{G}\sum_{j=1}^{N}R(i,j)\times i^2$
SRLGE (SRE combiné avec LGRE)	$\frac{1}{N_r}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{R(i,j)}{i^2\times j^2}$
SRHGE (SRE combiné avec HGRE)	$\frac{1}{N_r}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{R(i,j)\times i^2}{j^2}$
LRLGE (LRE combiné avec LGRE)	$\frac{1}{N_r}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{R(i,j)\times j^2}{i^2}$
LRHGE (LRE combiné avec HGRE)	$\frac{1}{N_r}\sum_{i=1}^{G}\sum_{j=1}^{N}R(i,j)\times i^2\times j^2$
GLNUr ou non uniformité des niveaux de gris	$\frac{1}{N_r}\sum_{i=1}^{G}\left[\sum_{j=1}^{N}R(i,j)\right]^2$
RLNU ou non uniformité des longueurs de plage	$\frac{1}{N_r}\sum_{j=1}^{N}\left[\sum_{i=1}^{G}R(i,j)\right]^2$
RP ou distribution des plages dans l'image	$N_r/\sum_{i=1}^{G}\sum_{j=1}^{N}(R(i,j)\times j)$

A.2.3 Matrice des tailles de zones homogènes (GLSZM)

La construction des GLSZM est illustré dans la Figure A.4.



(a) Étape 1

(b) Étape 2

(c) Étape 3

(d) Étape 4

(e) Étape 3

FIGURE A.4 – Illustration de la construction par étape d'une matrice des tailles de zones homogènes.

Le premier élément de la matrice correspond au nombre d'itération qu'un voxel d'intensité 1 possède une succession de voisins de la même intensité dans son voisinage. Les caractéristiques présentées dans le tableau A.5 peuvent être extraites des matrices GLSZM.

TABLEAU A.5 – Caractéristiques extraites de la matrice de tailles de zones homogènes Z. N_z est le nombre total de zones homogènes, G est la valeur d'intensité rééchantillonnée maximale, N la taille maximales des zones et $Z(i,j)$ un élément de la matrice ayant pour intensité i et pour taille j [Orlhac et al. 2014] supp data.

Caractéristiques	**Formules**
SZE ou Small Zone Emphasis	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{Z(i,j)}{j^2}$
LZE ou Large Zone Emphasis	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}Z(i,j)\times j^2$
LGZE ou Low Intensity Zone Emphasis	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{Z(i,j)}{i^2}$
HGZE ou High-Intensity Zone Emphasis	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}Z(i,j)\times i^2$
SZLGE (SZE combiné avec LGZE)	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{Z(i,j)}{i^2\times j^2}$
SZHGE (SZE combiné avec HGZE)	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{Z(i,j)\times i^2}{j^2}$
LZLGE (LZE combiné avec LGZE)	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}\frac{Z(i,j)\times j^2}{i^2}$
LZHGE (LZE combiné avec HGZE)	$\frac{1}{N_z}\sum_{i=1}^{G}\sum_{j=1}^{N}Z(i,j)\times i^2\times j^2$
GLNUz ou non uniformité des niveaux de gris	$\frac{1}{N_z}\sum_{i=1}^{G}\left[\sum_{j=1}^{N}Z(i,j)\right]^2$
ZLNU ou non uniformité des tailles de zone	$\frac{1}{N_z}\sum_{j=1}^{N}\left[\sum_{i=1}^{G}Z(i,j)\right]^2$
ZP ou distribution des zones dans l'image	$N_z/\sum_{i=1}^{G}\sum_{j=1}^{N}(Z(i,j)\times j)$

A.2.4 Matrice des différences de niveaux de gris (GLDM)

La construction des GLDM se fait de la manière suivante. L'élément $N(i,1)$ correspond à la probabilité d'apparition du niveau i et l'élément$(i,2)$ correspond à la somme de la différence en valeur absolue entre les voxels d'intensité i et la moyenne de leur voisinage respectif :

$$N(i,2) = \sum_{x,y,z}^{X,Y,Z} \begin{cases} 1, |\overline{M}(x,y,z) - i| \text{ si } I(x,y,z) = i \\ 0, \text{ sinon} \end{cases} \tag{A.3}$$

Où $\overline{M}(x,y,z)$ représente la moyenne de l'intensité des voisins du voxel de coordonnées (x,y,z).

Les caractéristiques présentées dans le Tableau A.6 peuvent être extraites des matrices GLDM.

TABLEAU A.6 – Caractéristiques issues de la matrice de différences de niveaux de gris N. E est le nombre de voxels dans le VOI et G est le nombre de niveaux de gris [Orlhac et al. 2014] supp data.

Caractéristiques	**Formules**
Coarseness	$1 \Big/ \sum_{i=1} (N(i,1) \times N(i,2))$
Contrast	$\frac{1}{E \times G(G-1)} \left[\sum_{i=1} \sum_{j=1} N(i,1) \times N(j,1) \times (i-j)^2 \right] \times \left[\sum_{i=1} N(i,2) \right]$
Busyness	$\sum_{i=1} N(i,1) \times N(i,2) \Big/ \sum_{i=1} \sum_{j=1} (i \times N(i,1) - j \times N(j,1))$
Complexity	$\sum_{i=1} \sum_{j=1} \frac{\lvert i-j \rvert}{E(N(i,1)+N(j,1))} \times (N(i,1) \times N(i,2) + N(j,1) \times N(j,2))$
Strength	$\sum_{i=1} \sum_{j=1} (N(i,1)+N(j,1))(i-j)^2 \Big/ \sum_{i=1} N(i,2)$

Bibliographie

Al-Taan O S, Eltweri A, Sharpe D, et al. Prognostic value of baseline FDG uptake on PET-CT in esophageal carcinoma. *World J Gastrointest Oncol*, 2014 ;6(5) :139–144. doi :10.4251/wjgo.v6.i5.139.

Altman D G, Lausen B, Sauerbrei W, et al. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *Semin Oncol*, 1994 ;86(11) :829–835.

Amadasun M and King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern*, 1989 ;19(5) :1264–1273. doi :10.1109/21.44046.

Bellman R and Bellman R E. *Adaptive Control Processes : A Guided Tour*. Princeton Legacy Library. Princeton University Press, 1961.

Ben-Haim S and Ell P. 18F-FDG PET and PET/CT in the Evaluation of Cancer Treatment Response. *J Nucl Med*, 2008 ;50(1) :88–99. doi :10.2967/jnumed.108.054205.

Berghmans T, Dusart M, Paesmans M, et al. Primary Tumor Standardized Uptake Value (SUVmax) Measured on Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) is of Prognostic Value for Survival in Non-small Cell Lung Cancer (NSCLC) : A Systematic Review and Meta-Analysis (MA) by the European Lu. *J Thorac Oncol*, 2008 ; 3(1) :6–12. doi :10.1097/JTO.0b013e31815e6d6b.

Beukinga R J, Hulshoff J B, van Dijk L V, et al. Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment ¹⁸ F-FDG PET/CT Imaging. *J Nucl Med*, 2017 ;58(5) :723–729. doi : 10.2967/jnumed.116.180299.

Blom R L G M, Steenbakkers I R, Lammering G, et al. PET/CT-based metabolic tumour volume for response prediction of neoadjuvant chemoradiotherapy in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging*, 2013 ;40(10) :1500–1506. doi : 10.1007/s00259-013-2468-x.

Blum A L and Langley P. Selection of relevant features and examples in machine learning. *Artif Intell*, 1997 ;97(1-2) :245–271. doi :10.1016/S0004-3702(97)00063-5.

Boellaard R. Standards for PET Image Acquisition and Quantitative Data Analysis. *J Nucl Med*, 2009 ;50(Suppl_1) :11S–20S. doi :10.2967/jnumed.108.057182.

Boellaard R, Delgado-Bolton R, Oyen W J G, et al. FDG PET/CT : EANM procedure guidelines for tumour imaging : version 2.0. *Eur J Nucl Med Mol Imaging*, 2015 ;42(2) :328–354. doi :10.1007/s00259-014-2961-x.

Boellaard R, Krak N C, Hoekstra O S, et al. Effects of Noise, Image Resolution, and ROI Definition on the Accuracy of Standard Uptake Values : A Simulation Study. *J Nucl Med*, 2004 ;45(9) :1519–1527.

Boellaard R, Oyen W J G, Hoekstra C J, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging*, 2008 ;35(12) :2320–2333. doi :10.1007/s00259-008-0874-2.

Boughattas N, Berar M, Hamrouni K, et al. Brain tumor segmentation from multiple MRI sequences using multiple kernel learning. In *IEEE-ICIP*. Paris, 2014 ; .

Bourgier C, Colinge J, Aillères N, et al. D{é}finition et applications cliniques des radiomics. *Cancer/Radiotherapie*, 2015 ;19(6-7) :532–537. doi :10.1016/j.canrad.2015.06.008.

Box G and Tia G. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

Breiman L. Random Forests. *Mach Learn*, 2001 ;45(1) :5–32. doi :10.1023/A:1010933404324.

Breiman L, Friedman J, Olshen R, et al. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Breki C M, Dimitrakopoulou-Strauss A, Hassel J, et al. Fractal and multifractal analysis of PET/CT images of metastatic melanoma before and after treatment with ipilimumab. *EJNMMI Res*, 2016 ;6(1) :61. doi :10.1186/s13550-016-0216-5.

Brooks F J and Grigsby P W. The Effect of Small Tumor Volumes on Studies of Intratumoral Heterogeneity of Tracer Uptake. *J Nucl Med*, 2014 ;55(1) :37–42. doi :10.2967/jnumed.112.116715.

Bryant A S and Cerfolio R J. The Maximum Standardized Uptake Values on Integrated FDG-PET/CT Is Useful in Differentiating Benign From Malignant Pulmonary Nodules. *Ann Thorac Surg*, 2006 ;82(3) :1016–1020. doi :10.1016/j.athoracsur.2006.03.095.

Burman P. No TitleA Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika,*, 1989 ;76(3) :503–514.

Buvat I. Les limites du SUV. *Med Nucl*, 2007 ;31(4 SPEC. ISS.) :165–172. doi :10.1016/j.mednuc.2007.03.003.

Buvat I, Orlhac F, and Soussan M. Tumor Texture Analysis in PET : Where Do We Stand ? *J Nucl Med*, 2015 ;56(11) :1642–1644. doi :10.2967/jnumed.115.163469.

Calais J, Thureau S, Dubray B, et al. Areas of High 18F-FDG Uptake on Preradiotherapy PET/CT Identify Preferential Sites of Local Relapse After Chemoradiotherapy for Non-Small Cell Lung Cancer. *J Nucl Med*, 2015 ;56(2) :196–203. doi :10.2967/jnumed.114.144253.

Chalkidou A, O'Doherty M J, and Marsden P K. False discovery rates in PET and CT studies with texture features : A systematic review. *PLoS One*, 2015 ;10(5) :1–18. doi :10.1371/journal.pone.0124165.

Chandrashekar G and Sahin F. A survey on feature selection methods. *Comput Electr Eng*, 2014 ;40(1) :16–28. doi :10.1016/j.compeleceng.2013.11.024.

Chang C, Liu W, and Zhang H. Image retrieval based on region shape similarity. *Photonics West 2001-Electronic Imaging*, 2001 ;4315 :31–38.

Chen H H W, Lee B F, Su W C, et al. The increment in standardized uptake value determined using dual-phase 18F-FDG PET is a promising prognostic factor in non-small-cell lung cancer. *Eur J Nucl Med Mol Imaging*, 2013 ;40(10) :1478–1485. doi : 10.1007/s00259-013-2452-5.

Chen X w and Wasikowski M. FAST : a roc-based feature selection metric for small samples and imbalanced data classification problems. *Proceeding 14th ACM SIGKDD Int Conf Knowl Discov data Min - KDD 08*, 2008 ;pages 124–132. doi :10.1145/1401890.1401910.

Cook G J R, Yip C, Siddique M, et al. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non–Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy ? *J Nucl Med*, 2013 ;54(1) :19–26. doi :10.2967/jnumed.112.107375.

Cortes C and Vapnik V. Support-Vector Networks. *Mach Learn*, 1995 ;20(3) :273–297. doi : 10.1023/A:1022627411411.

Cox D R and Oakes D. *Analysis of Survival Data.* Chapman {&} Hall, New York, 1984.

Das B K and Das B K. Positron Emission Tomography- A Guide for Clinicians. Springer edition, 2015 ;.

de Langen A J, Vincent A, Velasquez L M, et al. Repeatability of 18F-FDG Uptake Measurements in Tumors : A Metaanalysis. *J Nucl Med*, 2012 ;53(5) :701–708. doi : 10.2967/jnumed.111.095299.

Dempster A P, Laird N M, and Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol*, 1977 ;39(1) :1–38. doi :http://dx.doi.org/10.2307/2984875.

Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J Mach Learn Res*, 2006 ;7 :1–30. doi :10.1016/j.jecp.2010.03.005.

Desbordes P, Ruan S, Modzelewski R, et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Comput Med Imaging Graph*, 2016 ;doi :10.1016/j.compmedimag.2016.12.002.

Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemoradiation therapy using a random forest classifier. *PLoS One*, 2017 ;12(3) :e0173208. doi :10.1371/journal.pone.0173208.

Di Fiore A, Lecleire S, Gangloff A, et al. Impact of nutritional parameter variations during definitive chemoradiotherapy in locally advanced oesophageal cancer. *Dig Liver Dis*, 2014 ;46(3) :270–275. doi :10.1016/j.dld.2013.10.016.

Diehn M, Nardini C, Wang D S, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc Natl Acad Sci*, 2008 ;105(13) :5213–5218. doi :10.1073/pnas.0801279105.

Dong X, Xing L, Wu P, et al. Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma. *Nucl Med Commun*, 2013 ; 34(1) :40–46. doi :10.1097/MNM.0b013e32835ae50c.

Doumou G, Siddique M, Tsoumpas C, et al. The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer. *Eur Radiol*, 2015;25(9) :2805–2812. doi :10.1007/s00330-015-3681-8.

Du Bois D and Du Bois E. A Formula To Estimate The Approximate Surface Area If Height And Weight be Known. *Arch Intern Med*, 1916;.

Dubitzky W, Granzow M, and Berrar D. *Fundamentals of data mining in genomics and proteomics*. Springer US, 2007.

Dunn J C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J Cybern*, 1974;4(1) :95–104. doi :10.1080/01969727408546059.

Eisenhauer E A, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours : Revised RECIST guideline (version 1.1). *Eur J Cancer*, 2009;45(2) :228–247. doi :10.1016/j.ejca.2008.10.026.

El Naqa I, Grigsby P W, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit*, 2009;42(6) :1162–1171. doi :10.1016/j.patcog.2008.08.011.

Erselcan T, Turgut B, Dogan D, et al. Lean body mass-based standardized uptake value, derived from a predictive equation, might be misleading in PET studies. *Eur J Nucl Med*, 2002;29(12) :1630–1638. doi :10.1007/s00259-002-0974-3.

Fave X, Cook M, Frederick A, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imaging Graph*, 2015;44 :54–61. doi :10.1016/j.compmedimag.2015.04.006.

Fernández-Delgado M, Cernadas E, Barro S, et al. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ? *J Mach Learn Res*, 2014;15 :3133–3181.

Fledelius J, Winther-Larsen A, Khalil A A, et al. (18)F-FDG-PET/CT for very early response evaluation predicts CT response in Erlotinib treated NSCLC patients - A comparison of assessment methods. *J Nucl Med*, 2017;page jnumed.117.193003. doi :10.2967/jnumed.117.193003.

Fukushima K. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*, 1980;36(4) :193–202. doi :10.1007/BF00344251.

Galavis P E, Hollensen C, Jallow N, et al. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*, 2010;49(7) :1012–1016. doi :10.3109/0284186X.2010.498437.

Galloway M M. Texture analysis using gray level run lengths. *Comput Graph Image Process*, 1975;4(2) :172–179. doi :http://dx.doi.org/10.1016/S0146-664X(75)80008-6.

Gao X, Chu C, Li Y, et al. The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. *Eur J Radiol*, 2015;84(2) :312–317. doi :10.1016/j.ejrad.2014.11.006.

Genuer R. *For{ê}ts al{é}atoires : aspects th{é}oriques, s{é}lection de variables et applications*. Ph.D. thesis, 2010.

Geurts P, Ernst D, and Wehenkel L. Extremely randomized trees. *Mach Learn*, 2006; 63(1) :3–42. doi :10.1007/s10994-006-6226-1.

Gilad-Bachrach R, Navot A, and Tishby N. Margin based feature selection - theory and algorithms. In *Proc. 21st Int. Confer- ence Mach. Learn.* 2004; .

Gini C. Measurement of Inequality of Incomes. *Econ J*, 1921;31(121) :124–126. doi :10.1017/CBO9781107415324.004.

Giorgetti A, Pallabazzer G, Ripoli A, et al. Prognostic Significance of 2-Deoxy-2-[18F]-Fluoro-D-Glucose PET/CT in Patients With Locally Advanced Esophageal Cancer Undergoing Neoadjuvant Chemoradiotherapy Before Surgery : A Nonparametric Approach. *Medicine (Baltimore)*, 2016;95(13) :e3151. doi :10.1097/MD.0000000000003151.

Guyon I and Elisseef A. An Introduction to Variable and Feature Selection Isabelle. *J ofMachine Learn Res*, 2003;3 :1157–1182. doi :10.1016/j.aca.2011.07.027.

Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines. *Mach Learn*, 2002;46(4) :389–422. doi :10.1109/5254.708428.

Haralick R M, Shanmugam K, and Dinstein I. Textural features for image classification. 1973. doi :10.1109/TSMC.1973.4309314.

Hastie T, Tibshirani R, and Friedman J. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2009.

Hatt M, Cheze Le Rest C, Albarghach N, et al. PET functional volume delineation : A robustness and repeatability study. *Eur J Nucl Med Mol Imaging*, 2011a;38(4) :663–672. doi :10.1007/s00259-010-1688-6.

Hatt M, Groheux D, Martineau A, et al. Comparison Between 18F-FDG PET Image-Derived Indices for Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer. *J Nucl Med*, 2013;54(3) :341–349. doi :10.2967/jnumed.112.108837.

Hatt M, Majdoub M, Vallières M, et al. 18F-FDG PET Uptake Characterization Through Texture Analysis : Investigatin the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi–Cancer Site Patient Cohort. *J Nucl Med*, 2015;56(1) :38–44. doi :10.2967/jnumed.114.144055.

Hatt M, Visvikis D, Albarghach N M, et al. Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. *Eur J Nucl Med Mol Imaging*, 2011b;38(7) :1191–1202. doi :10.1007/s00259-011-1755-7.

Hatt M, Visvikis D, Pradier O, et al. Baseline 18F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. *Eur J Nucl Med Mol Imaging*, 2011c;38(9) :1595–1606. doi :10.1007/s00259-011-1834-9.

Herlidou S, Rolland Y, Bansard J Y, et al. Comparison of automated and visual texture analysis in MRI : Characterization of normal and diseased skeletal muscle. *Magn Reson Imaging*, 1999;17(9) :1393–1397. doi :10.1016/S0730-725X(99)00066-1.

Herskovic A, Martz K, Al-Sarraf M, et al. Combined Chemotherapy and Radiotherapy Compared with Radiotherapy Alone in Patients With Cancer of the Esophagus. *N Engl J Med*, 1992 ;326 :1593–1598. doi :10.1056/NEJM199206113262403.

Hilsenbeck S G, Clark G M, and McGuire W L. Why do so many prognostic factors fail to pan out ? *Breast Cancer Res Treat*, 1992 ;22(3) :197–206. doi :10.1007/BF01840833.

Hochberg Y and Benjamin Y. More Powerful Procedures for Multiple Significance Testing. *Stat Med*, 1990 ;9 :811–818.

Hofheinz F, Lougovski A, Zöphel K, et al. Increased evidence for the prognostic value of primary tumor asphericity in pretherapeutic FDG PET for risk stratification in patients with head and neck cancer. *Eur J Nucl Med Mol Imaging*, 2014 ;42(3) :429–437. doi : 10.1007/s00259-014-2953-x.

Holland J H. *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA, USA, 1992.

Hosmer D and Lemeshow S. *Applied logistic regression*. New York, 2000.

Hsu C W and Lin C J. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Trans Neural Networks*, 2002 ;13(2) :415–425.

Huang C J, Yang D X, and Chuang Y T. Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Syst Appl*, 2008 ;34(4) :2870–2878. doi : 10.1016/j.eswa.2007.05.035.

Hyun S H, Choi J Y, Shim Y M, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol*, 2010 ;17(1) :115–122. doi :10.1245/s10434-009-0719-7.

Ioannidis J P A. Why most published research findings are false. *PLoS Med*, 2005 ;2(8) :696–701. doi :10.1371/journal.pmed.0020124.

Jackson A, O'Connor J P B, Parker G J M, et al. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. *Clin Cancer Res*, 2007 ;13(12) :3449–3459. doi :10.1158/1078-0432.CCR-07-0238.

Jayachandran P, Pai R K, Quon A, et al. Postchemoradiotherapy positron emission tomography predicts pathologic response and survival in patients with esophageal cancer. *Int J Radiat Oncol Biol Phys*, 2012 ;84(2) :471–477. doi :10.1016/j.ijrobp.2011.12.029.

Jayasurya K, Fung G, Yu S, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Med Phys*, 2010 ;37(4) :1401–1407. doi :10.1118/1.3352709.

John G H, Kohavi R, and Pfleger K. Irrelevant Features and the Subset Selection Problem. In *Proc. Elev. Int. al Conf.* 1994 ; pages 121–129. doi :10.1145/2663598.2629474.

Kalff V, Duong C, Drummond E G, et al. Findings on 18F-FDG PET scans after neoadjuvant chemoradiation provides prognostic stratification in patients with locally advanced rectal carcinoma subsequently treated by radical surgery. *J Nucl Med*, 2006 ; 47(1) :14–22. doi :10.1056/NEJM199704033361402.

Kaplan E L and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 1958 ;53(282) :457–481. doi :10.2307/2281868.

Karnofsky D and Burchenal J. The Clinical Evaluation of Chemotherapeutic Agents in Cancer. In MacLeod CM, editor, *Eval. Chemother. Agents*, page 196. 1949 ;.

Kato H, Fukuchi M, Miyazaki T, et al. Prediction of response to definitive chemoradiotherapy in esophageal cancer using positron emission tomography. *Anticancer Res*, 2007 ; 27(4 C) :2627–2633.

Kenney J M, Marinelli L D, and Woodard H Q. Tracer Studies with Radioactive Phosphorus in Malignant Neoplastic Disease1. *Radiology*, 1941 ;37(6) :683–690. doi :10.1148/37.6.683.

Kerhet A, Small C, Quon H, et al. Segmentation of Lung Tumours in Positron Emission Tomography Scans : a Machine Learning Approach. *Artif Intell Med*, 2009 ;2 :146–155.

Kerhet A, Small C, Quon H, et al. Application of machine learning methodology for PET-based definition of lung cancer. *Curr Oncol*, 2010 ;17(1) :41–47.

Kiers H A L and Smilde A K. A comparison of various methods for multivariate regression with highly collinear variables. *Stat Methods Appl*, 2007 ;16(2) :193–228. doi :10.1007/s10260-006-0025-5.

Kim K, Kim S J, Kim I J, et al. Prognostic value of volumetric parameters measured by F-18 FDG PET/CT in surgically resected non-small-cell lung cancer. *Nucl Med Commun*, 2012 ;33(6) :613–620. doi :10.1097/MNM.0b013e328351d4f5.

Kira K and Rendell L A. The feature selection problem : Traditional methods and a new algorithm. In *AAAI-92 Proc.* 1992 ; pages 129–134. doi :10.1016/S0031-3203(01)00046-2.

Kohavi R and John G H. Wrappers for feature subset selection. *Artif Intell*, 1997 ;97(1-2) :273–324. doi :10.1016/S0004-3702(97)00043-X.

Kruskal W H and Wallis W A. Use of Ranks in One-Criterion Variance Analysis. *Source J Am Stat Assoc*, 1952 ;47(260) :583–621. doi :10.1080/01621459.1952.10483441.

Lambin P, Rios-Velazquez E, Leijenaar R T H, et al. Radiomics : Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 2012 ; 48(4) :441–446. doi :10.1016/j.ejca.2011.11.036.

Larson S, Erdi Y, Akhurst T, et al. Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging The Visual Response Score and the Change in Total Lesion Glycolysis. *Clin Positron Imaging*, 1999 ; 2(3) :159–171.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*, 1998 ;86(11) :2278–2323. doi :10.1109/5.726791.

Lee P, Bazan J G, Lavori P W, et al. Metabolic tumor volume is an independent prognostic factor in patients treated definitively for nonsmall-cell lung cancer. *Clin Lung Cancer*, 2012 ;13(1) :52–58. doi :10.1016/j.cllc.2011.05.001.

Leijenaar R T H, Carvalho S, Velazquez E R, et al. Stability of FDG-PET Radiomics features : an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*, 2013 ; 52(7) :1391–1397. doi :10.3109/0284186X.2013.812798.

Leijenaar R T H, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics : the need for standardized methodology in tumor texture analysis. *Sci Rep*, 2015 ;5(1) :11075. doi :10.1038/srep11075.

Lemarignier C, Di Fiore F, Marre C, et al. Pretreatment metabolic tumour volume is predictive of disease-free survival and overall survival in patients with oesophageal squamous cell carcinoma. *Eur J Nucl Med Mol Imaging*, 2014 ;i(2014) :2008–2016. doi : 10.1007/s00259-014-2839-y.

Levine E a, Farmer M R, Clark P, et al. Predictive Value of 18-Fluoro-Deoxy-Glucose-Positron Emission Tomography (18F-FDG-PET) in the Identification of Responders to Chemoradiation Therapy for the Treatment of Locally Advanced Esophageal Cancer. *Ann Surg*, 2006 ;243(4) :472–478. doi :10.1097/01.sla.0000208430.07050.61.

Li Y and Guo L. TCM-KNN scheme for network anomaly detection using feature-based optimizations. In *Proc. 2008 ACM Symp. Appl. Comput. - SAC '08*. ACM Press, New York, New York, USA, 2008 ; page 2103. doi :10.1145/1363686.1364194.

Lian C, Ruan S, Denœux T, et al. Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction. *Med Image Anal*, 2016 ;32 :257–268. doi :10.1016/j.media.2016.05.007.

Lowe V J, Fletcher J W, Gobar L, et al. Prospective investigation of positron emission tomography in lung nodules. *J Clin Oncol*, 1998 ;16(3) :1075–1084.

Macqueen J. Some methods for classification and analysis of multivariate observations. *Proc Fifth Berkeley Symp Math Stat Probab*, 1967 ;1(233) :281–297. doi : citeulike-article-id:6083430.

Mandelbrot B B. A fractal's lacunarity, and how it can be tuned and measured. *Fractals Biol Med*, 1994 ;doi :10.1007/978-3-0348-8501-0_2.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 1966 ;50(3) :163–170.

Marusyk A, Almendro V, and Polyak K. Intra-tumour heterogeneity : a looking glass for cancer ? *Nat Rev Cancer*, 2012 ;12(5) :323–334. doi :10.1038/nrc3261.

McCulloch W S and Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, 1943 ;5(4) :115–133. doi :10.1007/BF02478259.

Meyer Y and Salinger D H. *Wavelets and Operators*. Cambridge University Press, Cambridge, 1992.

Mi H, Petitjean C, Dubray B, et al. Robust feature selection to predict tumor treatment outcome. *Artif Intell Med*, 2015 ;64(3) :195–204. doi :10.1016/j.artmed.2015.07.002.

Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. 1996. doi : 10.1007/978-3-662-03315-9.

Miller a B, Hoogstraten B, Staquet M, et al. Reporting results of cancer treatment. *Cancer*, 1981 ;47(1) :207–214. doi :10.1002/1097-0142(19810101)47:1<207::AID-CNCR2820470134>3.0.CO;2-6.

Miller T R, Pinkus E, Dehdashti F, et al. Improved prognostic value of 18F-FDG PET using a simple visual analysis of tumor characteristics in patients with cervical cancer. *J Nucl Med*, 2003 ;44(2) :192–197.

Miwa K, Inubushi M, Wagatsuma K, et al. FDG uptake heterogeneity evaluated by fractal analysis improves the differential diagnosis of pulmonary nodules. *Eur J Radiol*, 2014 ; 83(4) :715–719. doi :10.1016/j.ejrad.2013.12.020.

Mohri M, Rostamizadeh A, and Talwalkar A. *Foundations of Machine Learning*. The MIT Press, 2012.

Mu W, Chen Z, Liang Y, et al. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. *Phys Med Biol*, 2015 ; 60(13) :5123–5139. doi :10.1088/0031-9155/60/13/5123.

Nakajo M, Jinguji M, Nakabeppu Y, et al. Texture analysis of 18F-FDG PET/CT to predict tumour response and prognosis of patients with esophageal cancer treated by chemoradiotherapy. *Eur J Nucl Med Mol Imaging*, 2017 ;44(2) :206–214. doi :10.1007/s00259-016-3506-2.

Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of {<}sup{>}18{<}/sup{>}F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. *J Nucl Med*, 2005 ; 46(8) :1342–1348. doi :46/8/1342[pii].

Ng F, Kozarski R, Ganeshan B, et al. Assessment of tumor heterogeneity by CT texture analysis : Can the largest cross-sectional area be used as an alternative to whole tumor analysis ? *Eur J Radiol*, 2013 ;82(2) :342–348. doi :10.1016/j.ejrad.2012.10.023.

Oken M, Creech R, Tormey D, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. 1982. doi :10.1097/00000421-198212000-00014.

Onoma D P, Gardin I, Modzelewski R, et al. Segmentation des tumeurs en imagerie m{é}dicale TEP bas{é}e sur la marche al{é}atoire 3D. *RFIA 2012 (Reconnaissance des Formes Intell Artif*, 2012 ;pages 978–2–9539515–2–3.

Orlhac F, Nioche C, Soussan M, et al. Understanding Changes in Tumor Texture Indices in PET : A Comparison Between Visual Assessment and Index Values in Simulated and Patient Data. *J Nucl Med*, 2017 ;58(3) :387–392. doi :10.2967/jnumed.116.181859.

Orlhac F, Soussan M, Chouahnia K, et al. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS One*, 2015 ;10(12) :1–16. doi :10.1371/journal.pone.0145063.

Orlhac F, Soussan M, Maisonobe J A, et al. Tumor Texture Analysis in 18F-FDG PET : Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. *J Nucl Med*, 2014 ;55(3) :414–422. doi :10.2967/jnumed.113.129858.

Palie O, Michel P, Ménard J F, et al. The predictive value of treatment response using FDG PET performed on day 21 of chemoradiotherapy in patients with oesophageal squamous cell carcinoma. A prospective, multicentre study (RTEP3). *Eur J Nucl Med Mol Imaging*, 2013 ;40(9) :1345–1355. doi :10.1007/s00259-013-2450-7.

Parmar C, Grossmann P, Bussink J, et al. Machine Learning methods for Quantitative Radiomic Biomarkers (Supplement). *Sci Rep*, 2015 ;5(1) :13087. doi :10.1038/srep13087.

Petkovska I, Shah S K, McNitt-Gray M F, et al. Pulmonary nodule characterization : A comparison of conventional with quantitative and visual semi-quantitative analyses using contrast enhancement maps. *Eur J Radiol*, 2006 ;59(2) :244–252. doi :10.1016/j.ejrad.2006.03.005.

Pudil P, Novovieova J, and Kittler J. Pattern Recognition Letters. *Pattern Recognit Lett*, 1994 ;15(June 1993) :1119–1125. doi :10.1016/j.patrec.2010.09.010.

Pyka T, Bundschuh R A, Andratschke N, et al. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. *Radiat Oncol*, 2015 ;10(1) :100. doi :10.1186/s13014-015-0407-7.

Rakotomamonjy A. Variable Selection Using SVM-based Criteria. *J ofMachine Learn Res*, 2003 ;3 :1357–1370. doi :10.1162/153244303322753706.

Rizk N P, Tang L, Adusumilli P S, et al. Predictive value of initial PET-SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. *J Thorac Oncol*, 2009 ;4(7) :875–879. doi :10.1097/JTO.0b013e3181a8cebf.

Roehrig J and Castellino R A. The promise of computer aided detection in digital mammography. *Eur J Radiol*, 1999 ;31(1) :35–39. doi :10.1016/S0720-048X(99)00067-4.

Roman-jimenez G, Acosta O, Leseur J, et al. Random forests to predict tumor recurrence following cervical cancer therapy using pre- and per-treatment 18 F-FDG PET parameters. 2016 ;pages 2444–2447.

Segal E, Sirlin C B, Ooi C, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol*, 2007 ;25(6) :675–680. doi :10.1038/nbt1306.

Shafer G. *A Mathematical Theory of Evidence*. Princeton University Press., 1976.

Shawe-Taylor J and Cristianini N. *Kernel Methods for Pattern Analysis*. 2004.

Soussan M, Chouahnia K, Maisonobe J A, et al. Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. *Eur J Nucl Med Mol Imaging*, 2013 ;40(5) :668–676. doi : 10.1007/s00259-012-2321-7.

Spearman C. The Proof and Measurement of Association between Two Things. *Am J Psychol*, 1904 ;15(1) :72–101.

Specht D. A general regression neural network. *IEEE Trans Neural Networks*, 1991 ; 2(6) :568–576. doi :10.1109/72.97934.

Srivastava N, Hinton G, Krizhevsky A, et al. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*, 2014 ;15 :1929–1958. doi :10.1214/12-AOS1000.

Swisher S G, Erasmus J, Maish M, et al. 2-Fluoro-2-deoxy-d-glucose positron emission tomography imaging is predictive of pathologic response and survival after preoperative chemoradiation in patients with esophageal carcinoma. *Cancer*, 2004 ;101(8) :1776–1785. doi :10.1002/cncr.20585.

Tabesh A, Teverovskiy M, Pang H Y, et al. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Trans Med Imaging*, 2007 ;26(10) :1366–1378. doi :10.1109/TMI.2007.898536.

Takeshita T, Morita K, Tsutsui Y, et al. The influence of respiratory motion on the cumulative SUV-volume histogram and fractal analyses of intratumoral heterogeneity in PET/CT imaging. *Ann Nucl Med*, 2016 ;30(6) :393–399. doi :10.1007/s12149-016-1071-1.

Tan S, Kligerman S, Chen W, et al. Spatial-Temporal FDG-PET Features for Predicting Pathologic Response of Esophageal Cancer to Neoadjuvant Chemoradiotherapy. *Int J Radiat Oncol Biol Phys*, 2013a ;85(5) :1375–1382. doi :10.1016/j.ijrobp.2012.10.017.

Tan S, Zhang H, Zhang Y, et al. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in 18F-FDG uptake patterns. *Med Phys*, 2013b ;40(10) :101707. doi :10.1118/1.4820445.

Teyton P, Metges J P, Jestin-Le Tallec V, et al. Valeur pronostique de la TEP au FDG dans le bilan initial du cancer de l'oesophage. *Med Nucl*, 2008 ;32(6) :323–331. doi :10.1016/j.mednuc.2008.02.013.

Theodoridis S and Koutroumbas K. *Introduction to Pattern Recognition : A Matlab Approach*. 2010.

Therasse P, Arbuck S G, Eisenhauer E A, et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *Clin Trials*, 2005 ;12(3) :16–27.

Thibault G, Fertil B, Navarro C, et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. *Pattern Recognit Inf Process*, 2009 ;pages 140–145. doi :Artn1357002\rDoi10.1142/S0218001413570024.

Tixier F, Cheze-Le Rest C, Hatt M, et al. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. *J Nucl Med*, 2011 ;52(3) :369–378. doi :10.2967/jnumed.110.082404.

Tixier F, Hatt M, Cheze-Le Rest C, et al. Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET. *J Nucl Med*, 2012 ;53(5) :693–700. doi :10.2967/jnumed.111.099127.

Toney L K and Vesselle H J. Neural Networks for Nodal Staging of Non–Small Cell Lung Cancer with FDG PET and CT : Importance of Combining Uptake Values and Sizes of Nodes and Primary Tumor. *Radiology*, 2014 ;270(1) :91–98. doi :10.1148/radiol.13122427.

Van De Wiele C, Kruse V, Smeets P, et al. Predictive and prognostic value of metabolic tumour volume and total lesion glycolysis in solid tumours. *Eur J Nucl Med Mol Imaging*, 2013 ;40(2) :290–301. doi :10.1007/s00259-012-2280-z.

van Velden F H P, Kramer G M, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies : Impact of Reconstruction and Delineation. *Mol Imaging Biol*, 2016 ;18(5) :788–795. doi :10.1007/s11307-016-0940-2.

Vauclin S, Doyeux K, Hapdey S, et al. Development of a generic thresholding algorithm for the delineation of 18FDG-PET-positive tissue : application to the comparison of three thresholding models. *Phys Med Biol*, 2009 ;54(22) :6901–6916. doi :10.1088/0031-9155/54/22/010.

Vera P, Mezzani-Saillard S, Edet-Sanson A, et al. FDG PET during radiochemotherapy is predictive of outcome at 1 year in non-small-cell lung cancer patients : a prospective multicentre study (RTEP2). *Eur J Nucl Med Mol Imaging*, 2014 ;41 :1057–1065. doi :10.1007/s00259-014-2687-9.

Wahl R L, Jacene H, Kasamon Y, et al. From RECIST to PERCIST : Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med*, 2009 ;50 Suppl 1(5) :122S—-50S. doi :10.2967/jnumed.108.057307.

Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI Res*, 2017 ;7(1) :11. doi :10.1186/s13550-017-0260-9.

Wang L. Feature selection with kernel class separability. *IEEE Trans Pattern Anal Mach Intell*, 2008 ;30(9) :1534–1546. doi :10.1109/TPAMI.2007.70799.

Weber W A. Quantitative analysis of PET studies. *Radiother Oncol*, 2010 ;96(3) :308–310. doi :10.1016/j.radonc.2010.07.004.

Whitney A W. A Direct Method of Nonparametric Measurement Selection. *IEEE Trans Comput*, 1971 ;C-20(9) :1100–1103. doi :10.1109/T-C.1971.223410.

Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bull*, 1945 ;1(6) :80–83.

Willaime J M Y, Turkheimer F E, Kenny L M, et al. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. *Phys Med Biol*, 2013 ;58(2) :187–203. doi :10.1088/0031-9155/58/2/187.

Xu R, Kido S, Suga K, et al. Texture analysis on 18F-FDG PET/CT images to differentiate malignant and benign bone and soft-tissue lesions. *Ann Nucl Med*, 2014 ;28(9) :926–935. doi :10.1007/s12149-014-0895-9.

Yan H, Wang R, Zhao F, et al. Measurement of tumor volume by PET to evaluate prognosis in patients with advanced non-small cell lung cancer treated by non-surgical therapy (Retracted article. See vol. 53, pg. 592, 2012). *Acta radiol*, 2011 ;52(6) :646–650. doi :10.1258/ar.2011.100462.

Yan J, Chu-Shern J L, Loi H Y, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J Nucl Med*, 2015 ;56(11) :1667–1673. doi :10.2967/jnumed.115.156927.

Yip S S F and Aerts H J W L. Applications and limitations of radiomics. *Phys Med Biol*, 2016 ;61(13) :R150—-R166. doi :10.1088/0031-9155/61/13/R150.

Ypsilantis P P, Siddique M, Sohn H M, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. *PLoS One*, 2015 ; 10(9) :e0137036. doi :10.1371/journal.pone.0137036.

Yu C H. Resampling methods : concepts, applications, and justification. *Pract Assessment, Res Eval*, 2003 ;8(19) :1.

Zaidi H and El Naqa I. PET-guided delineation of radiation therapy treatment volumes : A survey of image segmentation techniques. *Eur J Nucl Med Mol Imaging*, 2010 ; 37(11) :2165–2187. doi :10.1007/s00259-010-1423-3.

Zhang H, Wroblewski K, Appelbaum D, et al. Independent prognostic value of whole-body metabolic tumor burden from FDG-PET in non-small cell lung cancer. *Int J Comput Assist Radiol Surg*, 2013 ;8(2) :181–191. doi :10.1007/s11548-012-0749-7.

Zhang N, Ruan S, Lebonvallet S, et al. Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. *Comput Vis Image Underst*, 2011 ;115(2) :256–269. doi :10.1016/j.cviu.2010.09.007.

Zhu W, Xing L, Yue J, et al. Prognostic significance of SUV on PET/CT in patients with localised oesophagogastric junction cancer receiving neoadjuvant chemotherapy/ chemoradiation : A systematic review and meta-analysis. *Br J Radiol*, 2012 ;85(1017) :694–701. doi :10.1259/bjr/29946900.

Liste des figures

Liste des tableaux

Glossaire

ADC *Adénocarcinome*. 18, 47, 118, 201

AJCC "American Joint Committee on Cancer". 40

ANN *Réseaux de neurones artificiels ou "Artificial Neural Network"*. 68, 69, 74, 87–89, 198

AUC *"Area Under ROC Curves"*. 30, 33, 77–79, 84–90, 103, 110, 111, 113, 124, 127–129, 134, 135, 137, 138, 140–142, 145, 147–152, 154–158, 162, 163, 198–200, 202

AW-OSEM *"Attenuation-Weighted Ordered Subset Expectation Maximisation"*. 119

BSA *"Body Surface Area"*. 15, 16

CAD *Aide aux Diagnostics ou "Computed Aided Diagnosis"*. 83

CART *"Classification And Regression Tree"*. 81, 96–98, 198

CNN *Réseaux de neurones convolutifs*. 69, 70, 74, 87

COV *"Coefficient Of Variation"*. 27, 50, 107, 121, 135, 146

CTV *"Volume Tumoral Clinique"*. 28, 197

EM *"Expectation Maximization"*. 73, 85

FAST *"Feature Assessment by Sliding Thresholds"*. 79

FCM *"Fuzzy-C-Means"*. 71, 72

FDG *2-[18]-Fluoro-2-désoxy-D-glucose*. 1, 2, 8, 9, 13–18, 20, 22–24, 27, 28, 31, 33, 34, 39, 41, 44, 51, 84–87, 90, 92, 117, 119–121, 161, 197, 201

FIC *"Forest's Coefficient Importance"*. 95, 107, 108, 112, 115, 116, 124–126, 128, 129, 132–139, 144–158, 162, 163, 198–203

FORE *"Fourier Rebinning"*. 119

GA *"Genetic Algorithm"*. 109, 110, 128, 129, 134, 137, 140, 143, 145, 148, 155, 199

GARF *"Genetic Algortihm based on Random Forest"*. 95, 109, 113, 115, 116, 124–126, 128, 129, 132–138, 140, 142–158, 162, 163, 198–203

GLCM *"Gray Level Cooccurrence Matrix"*. 30–33, 44, 46, 49–52, 84, 87, 89, 107, 120, 121, 131, 133, 135, 146, 149, 152, 174

GLDM *"Gray Level Difference Matrix"*. 31–33, 42, 49, 51, 87, 120, 121, 131, 133, 135, 151, 152, 181

GLNUr *"Gray Level Non-Uniformity"*. 32, 49, 106

GLNUz *"Gray Level Non-Uniformity"*. 32, 33, 106, 135, 152, 197

GLRLM *"Gray Level Run Length Matrix"*. 31, 32, 46, 48–51, 87, 89, 177

GLSZM *"Gray Level Size Zone Matrix"*. 22, 31, 32, 42, 46, 48, 50, 51, 85–87, 107, 120, 121, 123, 130, 133, 135, 156, 157, 179

HFS *"Hierarchical Forward Selection"*. 81, 86, 116, 126, 153, 154, 156, 157, 159, 163, 202, 203

HGRE *"High Gray-level Run Emphasis"*. 32, 49–51, 106

HGZE *"High Gray-level Zone Emphasis"*. 32, 46, 50, 52, 106, 107, 135

HIV *Histogramme Intensité-Volume*. 26, 28, 197

HR *"Hazard-Ratio"*. 54, 144, 148

IDM *Moment différentiel inverse*. 32, 121, 133

IRM *Imagerie par Résonance Magnétique*. 1, 7, 12, 13, 39, 41, 164

ITER *"Iterative-Vue Point"*. 49

KCS *"Kernel Class Separability"*. 79

kFCV *"k-Fold Cross-Validation"* . 76, 77, 88, 198

kNN *k-Plus Proches Voisins ou "k-Nearest Neighbors"*. 67, 198

LASSO *"Least Absolute Shrinkage and Selection Operator"*. 84, 85, 116, 126, 153, 154, 156, 157, 159, 163, 202, 203

LBM *"Lean Body Mass"*. 15, 16

LGRE *"Low Gray level Run Emphasis"*. 32, 49, 50, 106

LGZE *"Low Gray level Zone Emphasis"*. 32, 46, 50, 106, 107, 135

LOOCV *"Leave-One-Out Cross-Validation"*. 75–77

LpOCV *"Leave-p-out Cross-Validation"*. 75, 77

LRE *"Long Run Emphasis"*. 32, 46, 51, 106

LRHGE *"Long Run High Gray-level Emphasis"*. 32, 106

LRLGE *"Long Run Low Gray-level Emphasis"*. 32, 85, 106

LZE *"Long Zone Emphasis"*. 32, 106, 107, 135

LZHGE *"Long Zone High Gray-level Emphasis"*. 32, 106, 135

LZLGE *"Long Zone Low Gray-level Emphasis"*. 32, 105–107, 135

MIP *"Maximum Intensity Projection"*. 14, 197

MP *Maladie Progressive*. 8, 9

MS *Maladie Stable*. 8, 9, 28

MTV *"Metabolically Tumour Volume"*. 18, 24, 27, 34, 37, 47, 48, 50–53, 57, 105, 106, 120, 121, 123, 133, 135, 136, 146, 148, 151, 152, 156, 201

NGLDM *"Neighboring Gray Level Dependence Matrix"*. 49, 50

NRI *"Nutritional Risk Index"*. 119, 133, 135, 146, 173

OMS *Organisation Mondiale de la Santé*. 7, 85, 117–119, 122, 123, 146, 172

OOB *"Out-Of-Bag"*. 102, 103, 107, 108, 110, 111, 113, 128, 129, 134, 137, 140, 145, 155

ORL *Oto-rhino-laryngologie*. 17, 29, 30, 39, 84

OS *"Overall Survival" ou survie globale*. 7, 27, 30, 33, 118, 119, 150

OSEM *"Ordered Subset Expectation Maximisation"*. 49

PA *Permutations Aléatoires*. 76, 77

PERCIST *"PET Response Criteria in Solid Tumors"*. 9, 27, 84

PET *Positron Emission Tomography*. 10, 89

PFS *"Progression Free Disease" ou survie sans récidive*. 7, 27, 30, 33, 34

PSF *"Point-Spread Function"*. 50

RC *Réponse Complète*. 8, 9, 20, 27, 28, 118, 119, 121, 122, 150

RCT *Radio-ChimioThérapie*. 7, 20, 22, 27, 31, 33, 42, 84–86, 117, 119, 120, 201

RECIST *"Response Evaluation Criteria in Solid Tumors"*. 8, 23

RELIEF *"RELevance In Estimating Features"*. 79

RF *Forêts Aléatoires ou "Random Forest"*. 70, 74, 81, 87–91, 93–95, 99–104, 107–113, 116, 123–129, 132, 134–142, 144–149, 153–158, 162, 163, 198–202

RFE *"Recursive Feature Elimination"*. 82, 116, 126, 153, 154, 156, 157, 159, 163, 202, 203

RL *Régression Logistique*. 151, 152

RLNU *"Run Length Non-Uniformity"*. 32, 46, 106

RNC *Réponse Non-Complète*. 118, 119, 121, 122, 150

ROC *"Receiver Operating Characteristic"*. 20, 22, 27, 31, 33, 53, 54, 78, 79, 86, 103, 126, 150, 197

RP *Réponse Partielle*. 8, 9

RPr *"Run Percentage"*. 32, 40, 51, 85, 86, 106

SCC *Carcinome épidermoïde*. 18, 47, 118, 201

Se *Sensibilité*. 77, 124

SFFS *"Sequential Forward Floating Selection"*. 80, 81, 116, 126, 153, 154, 156, 157, 159, 163, 202, 203

SFS *"Sequential Forward Selection"*. 80

SNE *"Small Number Emphasis"*. 49, 50

Sp *Spécificité*. 77, 78, 124

SRE *"Short Run Emphasis"*. 32, 46, 51, 106

SRHGE *"Short Run High Gray-level Emphasis"*. 32, 106

SRLGE *"Short Run Low Gray-level Emphasis"*. 32, 106

Stade TNM "Tumor, Nodes, Metastasis Stage". 40, 117, 118, 122, 171, 172

SUV *"Standardized Uptake Value"*. 15, 16, 23–26, 41, 43, 45, 47, 49, 120, 121, 123, 125, 131, 133, 135, 136, 146, 149, 152, 156, 197

SUV_{BW} *"Standardized Uptake Value Body Weight"*. 15, 16

SVM *Séparateur à Vaste Marge ou "Support Vector Machine"*. 67, 68, 73, 81, 82, 85–89, 125, 153, 154, 198

SZE *"Short Zone Emphasis"*. 32, 51, 106, 135

SZHGE *"Short Zone High Gray-level Emphasis"*. 32, 106, 135

SZLGE *"Short Zone Low Gray-level Emphasis"*. 32, 106, 135

TDM *TomoDensitoMétrie*. 1, 7, 8, 10, 12, 13, 23, 26, 31, 33, 39–41, 56, 85–87, 89, 104, 117, 119, 121, 157, 164, 197

TEP *Tomographie par Émission de Positons*. 1, 2, 7–10, 12–16, 18, 20, 22, 24, 26–28, 30, 31, 33, 34, 39–42, 44–53, 56, 84–87, 90, 92, 104, 105, 117, 119–121, 125, 149, 157, 161, 164, 165, 197, 201

TL *"Tumor Length"*. 18, 24, 201

TLG *"Total Lesion Glycolysis"*. 18, 24, 90, 105, 106, 121, 123, 135, 156, 201

TOF *"Time Of Flight"*. 50

VOI *"Volume Of Interest"*. 24–26, 29, 50, 52

WHO *"World Health Organization"*. 7, 8

ZLNU *"Zone Length Non-Uniformity"*. 32, 33, 106, 133, 135, 197

ZP *"Zone Percentage"*. 32, 51, 106