Analyse quantitative des données de routine clinique pour le pronostic précoce en oncologie

Cynthia Perier

université de BORDEAUX

THÈSE

pour l'obtention du grade de

Docteur de l'Université de Bordeaux

École Doctorale de Mathématiques et d'Informatique
Spécialité : Mathématiques appliquées et calcul scientifique

présentée et soutenue par

Cynthia Périer

le 14 novembre 2019

Analyse quantitative des données de routine clinique pour le pronostic précoce en oncologie

sous la direction de

Baudouin Denis de Senneville
Olivier Saut

Jury

Hermine Biermé	Professeure	Examinatrice
Irène Buvat	Directrice de recherche	Rapportrice
François Cornelis	Professeur	Rapporteur
Baudouin Denis de Senneville	Chargé de recherche	Directeur
Christèle Etchegaray	Chargée de recherche	Invitée
Jérôme Saracco	Professeur	Président
Olivier Saut	Directeur de recherche	Directeur

Résumé

L'évolution de la texture ou de la forme d'une tumeur à l'imagerie médicale reflète les modifications internes dues à la progression d'une lésion tumorale. Dans ces travaux nous avons souhaité étudier l'apport des caractéristiques delta-radiomiques pour prédire l'évolution de la maladie. Nous cherchons à fournir un pipeline complet de la reconstruction des lésions à la prédiction, en utilisant seulement les données de routine clinique.

Tout d'abord, nous avons étudié un sous ensemble de marqueurs radiomiques calculés sur IRM, en cherchant à établir quelles conditions sont nécessaires pour assurer leur robustesse. Des jeux de données artificiels et cliniques nous permettent d'évaluer l'impact de la reconstruction 3D des zones d'intérêt et celui du traitement de l'image. Une première analyse d'un cas clinique met en évidence des descripteurs de texture statistiquement associés à la survie sans évènement de patients atteints d'un carcinome du canal anal dès le diagnostic.

Dans un second temps, nous avons développé des modèles d'apprentissage statistique. Une seconde étude clinique révèle qu'une signature radiomique IRM en T2 à trois paramètres apprise par un modèle de forêts aléatoires donne des résultats prometteurs pour prédire la réponse histologique des sarcomes des tissus mous à la chimiothérapie néoadjuvante. Le pipeline d'apprentissage est ensuite testé sur un jeu de données de taille moyenne sans images, dans le but cette fois de prédire la rechute métastatique à court terme de patientes atteinte d'un cancer du sein. La classification des patientes est ensuite comparée à la prédiction du temps de rechute fournie par un modèle mécanistique de l'évolution des lésions.

Enfin nous discutons de l'apport des techniques plus avancées de l'apprentissage statistique pour étendre l'automatisation de notre chaine de traitement (segmentation automatique des tumeurs, analyse quantitative de l'œdème péri-tumoral).

Mots-clé : oncologie, analyse de textures, IRM, apprentissage statistique, radiomique, traitement d'image

Abstract

Tumor shape and texture evolution may highlight internal modifications resulting from the progression of cancer. In this work, we want to study the contribution of delta-radiomics features to cancer-evolution prediction. Our goal is to provide a complete pipeline from the 3D reconstruction of the volume of interest to the prediction of its evolution, using routinely acquired data only.

To this end, we first analyse a subset of MRI-extracted radiomics biomarquers in order to determine conditions that ensure their robustness. Then, we determine the prerequisites of features reliability and explore the impact of both reconstruction and image processing (rescaling, grey-level normalization). A first clinical study emphasizes some statistically-relevant MRI radiomics features associated with event-free survival in anal carcinoma.

We then develop machine-learning models to improve our results. Radiomics and machine learning approaches were then combined in a study on high grade soft tissu sarcoma (STS). Combining radiomics and machine-learning approaches in a study on high-grade soft tissue sarcoma, we find out that a T2-MRI delta-radiomic signature with only three features is enough to construct a classifier able to predict the STS histological response to neoadjuvant chemotherapy. Our ML pipeline is then trained and tested on a middle-size clinical dataset in order to predict early metastatic relapse of patients with breast cancer. This classification model is then compared to the relapsing time predicted by the mechanistic model.

Finally we discuss the contribution of deep-learning techniques to extend our pipeline with tumor automatic segmentation or edema detection.

Keywords : oncology, texture analysis, MRI, machine learning, radiomics, image processing

Cette thèse a été préparée dans le cadre de l'équipe-projet Inria Bordeaux Sud-Ouest MONC et de l'équipe Calcul scientifique et Modélisation de l'IMB.

Elle a été financée par le Laboratoire d'Excellence TRAIL ANR-10-LABX-57.

Remerciements

J'aimerais adresser en premier lieu des remerciements chaleureux à mes deux directeurs de thèse, *Olivier Saut* et *Baudouin Denis de Senneville*. Merci de m'avoir acceptée dans l'équipe, puis fait confiance pour entamer des travaux de thèse en son sein malgré mon parcours inhabituel. Vous avez tous deux enduré avec patience mes insécurités, moues dubitatives et poussées de pessimisme, que vous avez contrecarrées avec tranquillité, pragmatisme, confiance et beaucoup d'encouragements ! J'ai apprécié nos discussions et la complémentarité de vos enseignements et apports scientifiques. Merci entre autre de m'avoir répété que le diable est dans les détails et de m'avoir livré le secret magique pour se décharger d'une demande pénible formulée par un tiers [1].

Je souhaite à présent remercier *Irène Buvat* et *François Cornelis*, pour avoir accepté de rapporter mon manuscrit, pour le temps qu'ils lui ont consacré et l'intérêt qu'ils ont porté à mes travaux. Je suis en particulier reconnaissante envers Irène qui, après sa relecture particulièrement attentive, a mis en évidence avec indulgence les impairs de ce manuscrit et m'a guidée pour les améliorer.

Je remercie également *Hermine Biermé* et *Jérôme Saracco* pour leur participation à mon jury de thèse. Merci à *Christèle Etchegaray* d'avoir répondu favorablement à l'invitation, j'en suis ravie !

Merci à *Emeline Ribot* et *Nicolas Meunier*, membres de mon comité de thèse, pour leur suivi, leurs conseils et leurs encouragements.

Une grande partie de mes travaux dépend de l'association avec divers partenaires de travail.

Merci à *Claudia Pouypoudat* d'avoir été à l'origine des premières problématiques cliniques sur les tumeurs du pancréas. Ces travaux ont jeté les bases des recherches de cette thèse.

Merci à *Amandine Crombé* pour la riche collaboration qu'elle a initié sur les sarcomes et pour toutes les opportunités qu'elle apporte à l'équipe. J'en profite pour saluer *Michelle Kind* et la remercier d'avoir permis cette coentreprise.

Merci enfin à *Sébastien Benzekry* et à *Chiara Nicolò* de m'avoir intégrée à leur projet qui a été pour moi une belle opportunité pour améliorer mes compétences.

Je tiens à exprimer ma reconnaissance envers *Marie Beurton-Aimar*, pour avoir relu un chapitre de ce mémoire, pour m'avoir invitée à rejoindre son groupe de travail, pour le soutien depuis les débuts en master et pour sa gentillesse légendaire.

1. Par soucis de confidentialité, je ne révèlerai pas publiquement lequel d'entre vous a conseillé quoi. Je ne peux malheureusement pas non plus ébruiter ledit secret sous peine de le rendre caduc.

Mille mercis sont également à transmettre à *Annabelle Collin* qui m'a offert l'opportunité de donner des cours et m'a fournit un coup de pouce décisif pour la suite directe de ma carrière (la bonne fée de mes finances, quand on y pense !)

Je souhaite aussi adresser un vif remerciement à notre assistante d'équipe *Sylvie Embolla* ainsi qu'au personnel administratif de l'IMB car leur appui comme leur gentillesse facilitent incroyablement la vie des doctorant.e.s.

Mon'phis[2] dans son ensemble est une chouette équipe bicéphale qui m'a fourni conseils, coups de main et soutien pendant cinq ans. Je tiens en particulier à en remercier les membres non-permanents (mais pas que) pour leur contribution à créer un quotidien léger et agréable, quel que que soit l'étage de l'IMB. Cinq années, ça donne un peu trop de noms à lister pour ne pas en oublier[3]. Alors je préfère envoyer une bise aux visiteurs impromptus des bureaux, aux retardataires de la pause déjeuner comme à celles et ceux qui "font le tour", aux comploteurs transalpins des couloirs, aux évadé.e.s de prisons et à celles et ceux qui ont souvent prolongé un peu (beaucoup) la pause à la moindre occasion.

Mes co-bureaux en revanche n'échapperont pas à la dédicace nominative. Je salue les derniers en date, *Charles* puis *Pedro*, et leur souhaite bonne chance pour leur fin et leur début de thèse, respectivement. À *Olivier* et *Sergio*, pour les quatre ans de cohabitation avec chacun, je souhaite transmettre le plaisir que j'ai eu à les côtoyer au quotidien. ありがとう et *molte grazie* !

Je n'oublie pas un merci pour l'amitié et le soutien manifesté par mes acolytes d'école, d'études, de cours de danse ou d'ailleurs, coucou à vous si vous lisez ces lignes, ça me fait tellement plaisir !

J'ai été choyée et bien entourée par une famille (et une belle famille !) enthousiaste et affectueuse. Je vous embrasse, je suis fière et touchée de vous savoir présent.e.s et de pouvoir vous montrer mes travaux. Merci à mes parents de m'avoir encouragée dans mes études et soutenue de toutes les façons possibles. À eux et à ma petite sœur *Pauline*, j'adresse un tendre remerciement supplémentaire pour leur support inconditionnel.

Mes derniers mots s'adressent à *Boris*, mon collègue, ami et compagnon[4], dont le nom aurait pu figurer à plusieurs reprises déjà dans ces pages : aux côtés de celles et ceux qui m'ont aidée, conseillée et motivée, qui ont corrigé mon manuscrit ou qui m'ont fait répéter ma soutenance. Je te suis infiniment reconnaissante et je te dois ma progression. Désolée de t'avoir mené la vie dure ces derniers mois, je te remercie pour ta patience exceptionnelle et ton soutien précieux.

2. Orthographe non contractuelle.
3. Et pourtant, pendant la soutenance, ce passage est votre préféré, je vous connais.
4. Je me permets de me ré-approprier ta formule.

À *M. Dieudonné Zélé*,
qui m'a fait promettre de faire une thèse et l'a consigné dans sa dédicace du 10/10/10 pour que je ne l'oublie pas.

À *Alice*,
pour qui j'ai promis, sans preuve officielle, que je ferai quelque chose.

Table des matières

Introduction : contexte et motivations

Sommaire

En avril 2019, le troisième Plan Cancer en France entrait dans sa dernière année de mise en œuvre. Destiné à réduire les inégalités des patients face à la maladie, il se traduit par une liste d'objectifs et d'engagements, parmi lesquels l'accompagnement des projets d'innovation technologique. Deux mois plus tôt, le Ministère de l'Économie publiait un rapport intitulé "*Intelligence artificielle, état de l'art et perspectives pour la France*", établissant entre autre un bilan des opportunités générées par l'IA pour le secteur de la Santé. La personnalisation des traitements et la prise en compte de leur impact sur la qualité de vie des patients y sont notamment désignés comme "usages à développer en priorité".

C'est dans ce contexte que, dans le cadre de l'équipe Inria Monc et de l'initiative d'excellence TRAIL pour l'innovation en imagerie médicale, nous présentons le résultat de travaux de thèse visant à améliorer la description des tumeurs cancéreuses par l'analyse automatique

de l'image médicale ainsi que la prédiction précoce du pronostic des patients grâce aux possibilités offertes par l'IA.

1 Contexte biologique et clinique, de la maladie au

1.1 Le cancer et ses mécanismes de développement

Le cancer est une maladie décrite dès l'Antiquité par Hippocrate et dont des traces ont pu être trouvées sur des fragments de squelettes datant de la préhistoire. Les données actuelles du Centre international de Recherche sur le Cancer (CIRC) signalent quant à elles 9,6 millions de décès dus au cancer en 2018 dans le monde, pour 18,1 millions de nouveaux cas. Les cancers du poumon, du sein chez la femme et du colon-rectum constituent à eux seuls un tiers des incidences et figurent parmi les cinq premières causes de mortalité liées au cancer. Le rapport donne également une estimation de 43,8 millions de personnes survivant au cancer cinq ans après le diagnostic, mais ce chiffre varie selon la localisation des lésions.

Caractérisés par une remarquable complexité et longtemps réputés incurables, de nombreux cancers sont maintenant traités et guéris grâce aux progrès les plus récents de la médecine. Une compréhension plus fine de leurs mécanismes de développement a initié le développement de protocoles de traitements adaptés aux pathologies et aux patients.

Les mécanismes du cancer

L'équilibre d'un organisme vivant sain dépend de la balance entre la production de nouvelles cellules et leur destruction. Les cellules se divisent par mitose pour répliquer leur matériel génétique et leur mort est programmée par apoptose lorsqu'elles deviennent non fonctionnelles ou trop vieilles. Le cancer est le résultat d'une série de dysfonctionnements cellulaires issus de mutations génétiques impactant ce cycle cellulaire. Ces troubles permettent à certaines cellules de se multiplier de façon anarchique, conduisant à leur prolifération incontrôlée locale, puis à distance.

Une cellule cancéreuse est caractérisée par huit types de mutations qui dérèglent le cycle cellulaire et échappent aux différentes barrières de contrôle mises en place par l'organisme [**Pezzella2019**] (voir Fig. 1).

Dérèglement des signaux de croissance et de division. Leur émission devient chronique, les cellules grandissent et la mitose se produit n'importe quand et n'importe où.

Résistance aux inhibiteurs de croissance. Émises par les tissus environnants et par la cellule elle-même pour interrompre son cycle au moment opportun, ces molécules ne sont pas captées voire sont directement éliminées par les cellules cancéreuses.

Résistance à l'apoptose. Ce programme est normalement la principale barrière contre les cellules déviantes en déclenchant leur suicide par la destruction des organites [5] et l'ingestion

5. Compartiments différenciés remplissant une fonction précise au sein de la cellule.

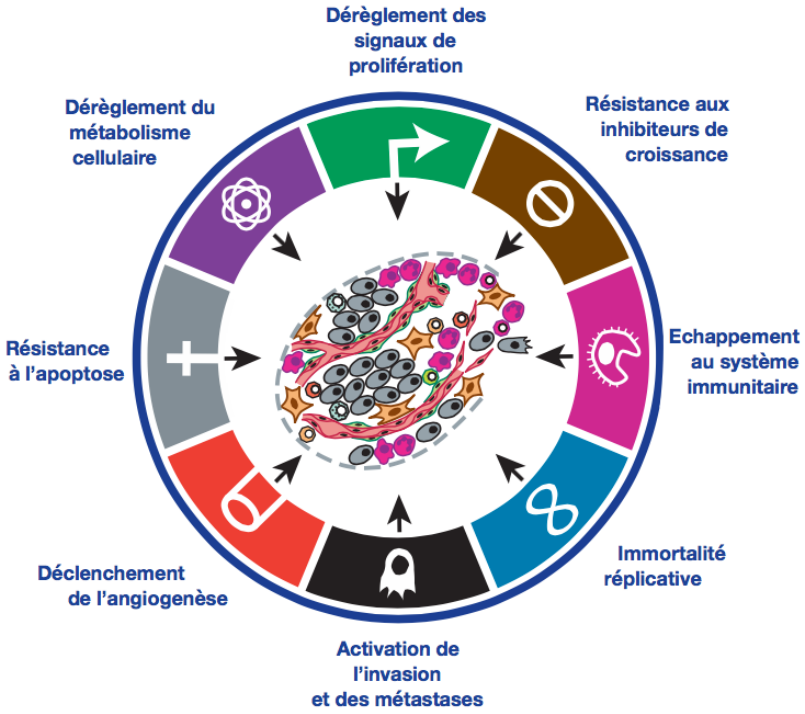


FIGURE 1 – Cycle cellulaire et mutations pouvant conduire à une cellule cancéreuse. *source : d'après "The hallmarks of Cancer", Oxford Textbook of Cancer Biology* [**Pezzella2019**]

des restes par les cellules macrophages [6].

Immortalité réplicative. La longévité cellulaire est déterminée par les télomères, séquences hautement répétitives non codantes situées à l'extrémité des chromosomes, qui perdent quelques nucléotides à chaque réplication de la cellule. Lorsque la taille des télomères passe sous un certain seuil, la cellule saine stoppe son cycle. Mais l'érosion de l'extrémité des chromosomes d'une cellule cancéreuse se poursuit : soit jusqu'à leur instabilité totale, soit à l'infini, en prolongeant régulièrement les télomères par recombinaisons.

Angiogenèse. Arrivée à un certain stade, la tumeur nécessite un apport supplémentaire en glucose et en oxygène ainsi qu'un moyen d'évacuation des déchets pour éviter la nécrose de ses cellules. Elle envoie alors des signaux induisant la création de nouveaux vaisseaux sanguins pour se raccorder au système vasculaire de l'organisme.

Métastases. Certaines cellules cancéreuses de haut grade perdent leurs capacités d' adhérence avec leurs voisines et deviennent invasives. Elles migrent dans les tissus adjacents, puis au travers du système vasculaire mis en place à l'étape précédente pour se disséminer dans d'autres organes. Si elles s'adaptent et survivent dans leur nouvel environnement, elles finissent pas se regrouper et générer des métastases macroscopiques. La régularisation du

6. Ce dernier point constitue la principale différence avec la nécrose (cellules mourant par défaut d'oxygène ou d'énergie et dont les débris provoquent des réactions inflammatoires).

processus de colonisation est complexe et implique des mécanismes intra-cellulaires et des cellules de support.

Dérèglement du métabolisme. En altérant leur métabolisme énergétique, les cellules s'adaptent à l'environnement. Elles favorisent notamment le système glycolytique, qui privilégie la consommation du glucose ou de ses dérivés (y compris le lactate, censé être toxique pour les cellules).

Échappement au système immunitaire. Puisque qu'un système immunitaire fonctionnel développe une certaine tolérance à ses propres antigènes, une tumeur peut éviter la surveillance et les attaques en n'exprimant que des antigènes reconnus comme 'normaux'.

Différents types de cancers à différents stades d'évolution de la maladie présentent différentes combinaisons de ces mutations. La recherche s'est donc attachée à diversifier les traitements de façon à optimiser le soin en ciblant le maximum de mécanismes défectueux simultanément.

1.2 Les traitements

La section qui suit ne vise pas à proposer un catalogue exhaustif de tous les traitements existants : il s'agit d'un sujet extrêmement vaste dépassant largement le cadre de ces travaux. Toutefois, on cherchera à mettre en évidence la diversité d'action des techniques présentées, ainsi que leurs limitations respectives.

Détruire la tumeur

Par suppression globale. La façon la plus intuitive de soigner une tumeur solide est sa suppression pure et simple par chirurgie. L'exérèse complète R0 (tumeur entière et marges de sécurité) est radicale et souvent indispensable au traitement de certaines pathologies, mais elle n'est pas toujours possible en raison de son caractère intrinsèquement invasif. Les lésions ne doivent pas être trop nombreuses, leurs marges doivent être bien définies et non envahissantes et leur taille raisonnable pour permettre l'opération sans risquer la perte d'une fonction corporelle. Le patient lui-même n'est pas toujours en état de subir une opération. Un traitement complémentaire est souvent exigé pour améliorer les conditions et prévenir tout risque de rechute locale due à l'oubli ou à l'impossibilité de traiter certaines marges et lésions invisibles à l'œil nu ou distantes.

Par blocage de la division cellulaire. La radiothérapie est l'utilisation locale de radiations ionisantes pour provoquer des lésions dans l'ADN des cellules cancéreuses afin d'empêcher leur multiplication cellulaire et à terme, de les détruire. Les rayons provoquent des dégâts sur toutes les cellules qu'ils atteignent, ce qui explique les nombreux effets secondaires subis par les patients lorsque des tissus sains ont été touchés.

La chimiothérapie agit également sur le mécanisme de multiplication. La majorité des substances chimiothérapeutiques sont *cytotoxiques* : elles interrompent la mitose voire provoquent l'apoptose. Elles agissent sur l'ensemble des cellules à division rapide : cellules cancé-

reuses mais aussi cellules sanguines ou cellules responsables de la croissance des cheveux, etc. Les effets indésirables induits sont nombreux (anémies, hémorragies, infections ...) Certaines molécules d'hormonothérapie, comme le trastuzumab dans les cancers du sein, vont plutôt bloquer les récepteurs liés à la prolifération cellulaire.

La radiothérapie est un traitement dit *locorégional* (qui cible une zone locale) alors que la chimiothérapie et l'hormonothérapie sont des traitements *systémiques* (qui ciblent le corps entier).

Cependant, comme les cellules du centre des tumeurs de stade avancé ne se divisent plus (par manque d'espace, d'oxygène et de nutriments), elles sont insensibles aux méthodes de blocage de la division cellulaire et échappent au traitement.

Par destruction des cellules tumorales. Souvent utilisée pour le traitement délicat des métastases vertébrales ou du cancer du rein, la cryothérapie congèle la tumeur en diffusant localement un gaz très froid qui s'introduit dans les cellules. Sous l'effet du réchauffement subséquent, leur membrane finit par exploser. Les débris sont ensuite nettoyés par le système immunitaire.

À l'inverse, le traitement par radiofréquence vise à détruire les petites tumeurs par des sondes diffusant localement un courant qui provoque une chaleur supérieure à 60 °C, soit la température de dénaturation des protéines. La radio-ablation est couramment utilisée pour les petites tumeurs difficiles d'accès ou sur des parents non opérables (souvent pour des cancers du foie ou du poumon).

L'électroporation utilise également un courant électrique, cette fois appliqué de façon à stimuler l'ouverture des pores d'une cellule afin de la perméabiliser. Il s'agit soit d'enclencher un processus irréversible immédiatement mortel, soit de faciliter l'entrée de molécules médicamenteuses (et donc de diminuer les doses) pour agir sur le métabolisme de la cellule.

Agir sur les voies métaboliques

Des molécules anticancéreuses ont été développées pour cibler des caractéristiques plus spécifiques du cancer. Dans les cancers du côlon, le cetuximab rétablit la communication et débloque le signal induisant l'apoptose en se fixant sur un facteur de croissance de l'EGFR. Les thérapies anti-angiogéniques (bevacizumab, sunitinib) empêchent la vascularisation sanguine de la tumeur en ciblant la voie VEGF dans certains cas avancés de cancer du poumon, côlon, rein, sein ...

Rétablir le système immunitaire

Certaines molécules comme le pembrolizumab ou le nivolumab sont capables de réactiver le système immunitaire pour qu'il cible à nouveau les cellules cancéreuses. C'est le principe de l'immunothérapie.

1.3 Le suivi de la maladie

La détection et la caractérisation du type de cancer dont souffre le patient, ou diagnostic, permettent de choisir et de mettre en place un protocole de traitement (Fig. 2). Ce protocole est lui même éventuellement adapté en cours de route. L'objectif est :

— de traiter de façon adaptée et personnalisée pour optimiser l'efficacité, stopper la progression et limiter les rechutes,
— de maintenir le confort de vie des patients en limitant les traitements lourds, les toxicités et autres effets secondaires.

Les décisions sont prises à mesure de l'estimation de l'impact du traitement sur les tissus cancéreux et sur la santé du patient en général : c'est **l'évaluation de la réponse au traitement** ou *réponse thérapeutique*.

Cette réponse elle-même n'est cependant qu'une donnée intermédiaire, aussi appelée donnée de substitution. La finalité clinique réelle est de fournir des indications sur le **pronostic** du patient, c'est à dire d'estimer deux paramètres :

— la survie sans évènement : à partir du début du traitement, c'est le temps écoulé avant que le patient ne présente une rechute locale (= de la tumeur primaire) et/ou à distance (apparition de métastases) ou tout autre complication que le traitement est censé éliminer.
— la survie globale : à partir de la date du diagnostic ou de celle du début du traitement, c'est la durée de vie du patient.

On a donc besoin d'un diagnostic juste, d'un bilan d'extension [7] et d'un suivi de l'évolution des lésions sous traitement pour obtenir des informations permettant de prédire la survie et d'adapter le protocole.

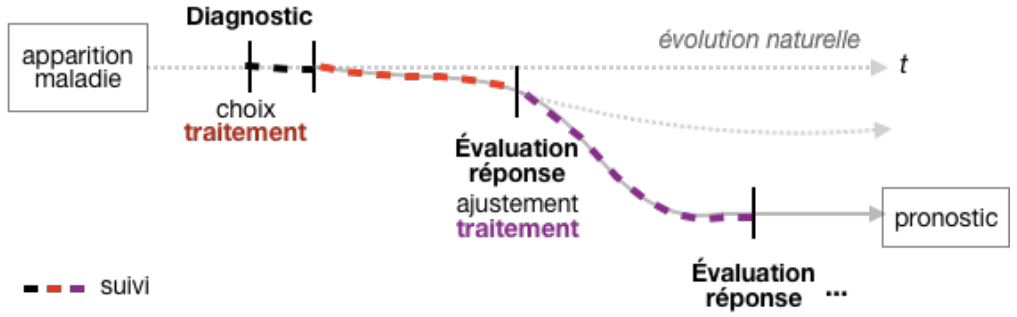


FIGURE 2 – Le parcours de soin d'un patient est ajusté à mesure de sa réponse au traitement.

Ces informations peuvent provenir d'outils d'évaluation qualitative, comme par exemple l'anamnèse ou l'examen physique. L'analyse quantitative passe par le suivi objectif de marqueurs biologiques : des caractéristiques mesurées numériquement qui donnent des indicateurs chiffrés sur les processus biologiques normaux, pathologiques ou issus d'une réponse au traitement [**Hodgson2009**].

Dans les deux cas, les principales sources d'information clinique sont les mesures in vivo des caractéristiques physiologiques (activité cardiaque, tension, température), les mesures in

7. Ensemble des examens visant à estimer l'étendue d'un cancer et la présence éventuelle de métastases distantes.

vitro (sang, urine, etc.), avec notamment l'analyse histopathologique des tissus prélevés par biopsie, l'imagerie médicale, ou encore les analyses génomiques [**Sawyers2008**].

L'imagerie médicale est un cas particulier d'outil d'évaluation in vivo jouant un rôle essentiel dans toutes les étapes du traitement du cancer [**Levy2011**] car l'intégralité des lésions peuvent être apréhendées de façon non-invasive et répétée. L'importance de l'imagerie dans la prise de décision thérapeutique n'a fait qu'augmenter ces dernières années et son évolution vers la quantification de l'information apportée par les images a permis d'en améliorer la précision et la reproductibilité [**Prescott2012**].

2 L'imagerie médicale pour le diagnostic, le suivi et la prise de décisions

Nous donnerons ici quelques concepts clés des principes et fonctionnement de l'imagerie médicale et une évaluation de l'information qui peut en être extraite.

Toute image peut être représentée par une matrice dont les cellules contiennent une valeur d'intensité de signal associée à un point dans l'espace (x,y,z). Ce sont les pixels en 2D, ou voxels en 3D. Les valeurs les plus hautes correspondent généralement à des pixels clairs et les valeurs basses à des pixels foncés. Ces éléments sont ordonnés sur une grille de façon régulière [8]. Le voxel s'avère particulièrement adapté à la représentation de volumes à partir de plusieurs images de coupes, ce qui est le cas des images médicales.

Plus précisément, une image médicale est caractérisée en particulier par :

- sa taille globale, c'est à dire ses **dimensions** (nombre de pixels/voxels) dans les deux/trois axes de l'espace x, y (et z),
- la taille de ses voxels ou **pas d'espace** (en anglais *spacing*), lui aussi en x, y (et z),
- son **origine** dans l'espace physique, c'est à dire ses coordonnées dans l'espace réel, indexé à 0,
- son **orientation** dans l'espace physique : représentée par une matrice, ce sont les directions de chacun des axes x, y (et z).

2.1 L'IRM

L'IRM (Imagerie par résonance magnétique) est une technique d'imagerie non-invasive qui fournit des images anatomiques 3D de haut niveau de résolution spatiale et spectrale sans induire de radiations nocives pour les patients. Son fonctionnement est basé sur l'excitation puis la détection des changements de direction de l'axe de rotation du proton présent dans chaque atome d'hydrogène du corps. L'IRM utilise des aimants puissants destinés à créer un champ magnétique suffisant pour forcer les protons à s'aligner dans un *état d'équilibre*. Le spin des protons est ensuite dévié par une onde radio perturbatrice excitatrice, puis se réaligne avec le champ magnétique initial lorsque le courant cesse. Il devient alors possible de mesurer l'énergie dégagée et le temps pris par les protons pour se réorienter. Puisque ces

8. On ne traitera pas ici des grilles non ordonnées.

deux propriétés magnétiques dépendent de l'environnement et du type de molécules présentes dans le tissu, leurs valeurs sont enregistrées et discrétisées dans l'espace de façon à obtenir une image exploitable.

Séquences

En IRM le signal mesuré d'un tissu dépend des paramètres instrumentaux et physico-chimiques, des mouvements (macro ou micro) du patient et de la séquence utilisée. La séquence est une image au contraste donné, qui dépend de l'ajustement du temps de répétition (TR, intervalle entre deux excitations) et du temps d'écho (TE, intervalle entre l'excitation et la survenue du signal IRM). Le calcium (donc l'os) et l'air apparaissent noirs quelle que soit la séquence, mais le signal des autres tissus dépendent de la pondération choisie.

La pondération T1 reflète le temps mis par les protons pour revenir dans l'axe de l'aimant (= le temps de relaxation longitudinal). L'hypersignal [9] dans une image T1 est caractéristique de la présence de graisses, de lésions hémorragiques, de mélanine, etc., mais aussi de la présence d'artefacts métalliques. L'eau apparaît hypointense.

Des agents chélatés de contraste comme le gadolinium peuvent être administrés par intraveineuse avant ou pendant l'examen pour augmenter le temps de relaxation des protons. Ils produisent des images plus contrastées en T1 qu'on désignera par la suite séquences T1-gado.

Dans une image pondérée en T2, le signal dépend du temps mis par les protons pour revenir à leur état initial (= le temps de relaxation transverse). Cette fois, les molécules d'eau sont hyperintenses et les graisses plus sombres.

L'hypersignal des tissus graisseux en T1 peut masquer le réhaussement (hypersignal relatif) après injection de gadolinium et constituer une gêne. Le signal des graisses en T2 peut être difficile à discerner de celui de l'œdème.

Pour faciliter l'interprétation de ces séquences, il existe donc plusieurs techniques de suppression de ce signal :

— le *Fat-Sat* se base sur la saturation de la graisse et l'excitation sélective de l'eau, grâce aux différences de fréquence de résonance de l'hydrogène entre les deux molécules.

— le STIR (*short TI inversion recovery*) utilise l'inversion-récupération pour supprimer le signal d'un tissu en fonction de son temps de relaxation. Il est classiquement employé en T2.

Le radiologue, en analysant les différentes séquences, peut évaluer le caractère sain ou pathologique des tissus étudiés. En oncologie, les IRM sont utilisées pour la détection et la caractérisation des lésions cancéreuses et des éventuels signes de leur propagation ainsi que pour la planification de la chirurgie et/ou de la radiothérapie.

9. Signal d'un tissu supérieur au signal des tissus qui l'entourent. Se traduit par des pixels plus clairs.

2.2 Les autres modalités d'imagerie

Dans le cadre de ces travaux de thèse, la principale modalité d'imagerie recueillie dans nos jeux de données est l'IRM. Toutefois, d'autres modalités pourront être évoquées dans le présent document et font donc ici l'objet d'une courte définition.

CT-scan La tomodensitométrie (TDM ou CT-scan) est un examen qui permet d'obtenir des images du corps en coupes millimétriques au moyen d'un balayage du patient par rayons X. Leur absorption par les tissus est mesurée et convertie par traitement informatique. Les tissus se distinguent les uns des autres par leur densité, comptées en unités Hounsfield.

La radiographie et l'angiographie utilisent également des rayons X mais sont des modalités planaires. Elles fournissent des images de qualité moindre et sont surtout utilisées pour l'imagerie osseuse et l'imagerie vasculaire respectivement.

TEP-scan L'imagerie fonctionnelle permet de visualiser le fonctionnement métabolique des tissus. Grâce à l'injection en intraveineuse d'un traceur faiblement radioactif, la tomographie par émission de photons (TEP) met en évidence les organes où il se fixe. On choisit généralement un composé proche du glucose mais non dégradable comme le ^{18}F-FDG [10], assimilé puis bloqué en grande quantité par les cellules cancéreuses en pleine prolifération. L'atome radioactif, choisi pour sa demi-vie d'une à deux heures, s'y désintègre. Le système d'imagerie capte les photons émis à cette occasion et génère une cartographie représentative des zones de stockage par recoupement.

2.3 Mesure des biomarqueurs classiques à l'imagerie anatomique

Utilisée à la base comme un outil de contrôle qualitatif, l'imagerie est à présent considérée comme une source d'informations quantifiées sur l'évolution de la maladie. La référence d'évaluation des traitements se base sur la mesure de la taille des lésions (primaires et secondaires) à l'imagerie médicale.

Parmi les critères les plus communément utilisés figurent le critère RECIST 1.1 et le critère WHO (Fig. 3).

Le critère RECIST 1.1 (*Response Evaluation Criteria In Solid Tumor*) est la mesure sur IRM ou TDM du plus grand diamètre d'une lésion isolée ou de la somme des plus grands diamètres des cinq plus grandes lésions d'un patient. Les lésions cibles (primitives ou secondaires) sont listées quatre semaines avant traitement, puis suivies durant son application pour évaluer la réponse du patient. Le résultat appartient à l'une des quatre catégories suivantes : réponse complète (disparition totale des lésions), réponse partielle (-30 % de la somme des lésions), maladie stable, ou maladie progressive (+20% de la somme des lésions).

Le critère WHO [11] calcule la somme des produits des deux plus grandes dimensions perpendiculaires de l'ensemble des lésions détectées. La différence entre les mesures durant traitement et la mesure initiale est toujours classée selon les mêmes catégories, mais les seuils diffèrent (mesure diminuée de plus 50% pour la réponse partielle et augmentée de plus 25% pour la maladie progressive).

10. Fluorodesoxyglucose marqué au fluor 19.
11. *World Health Organization*

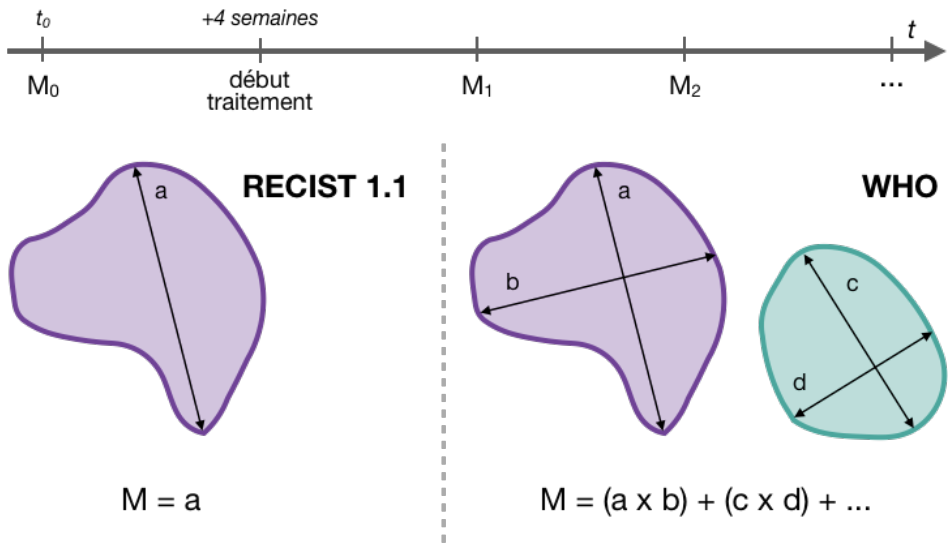


FIGURE 3 – Évaluation de la réponse au traitement suivant les critères RECIST 1.1 et WHO. M_1, M_2, etc. sont comparés à M_0 suivant le barème propre à chaque critère.

Les critères RECIST et WHO présentent l'avantage d'être faciles à mesurer et non invasifs. Toutefois, les études les plus récentes remettent en cause l'utilisation exclusive de ces critères dans la prise de décision clinique [**Burton2007**]. Basés uniquement sur la mesure d'une ou deux dimensions de la tumeur, ils ne prennent pas en compte l'intégralité de son information morphologique, ni l'information du signal de l'image ni même l'évolution de son aspect. Pourtant, le changement de taille n'est pas lié à la réponse de la lésion pour tous les types de tumeur. Il existe par exemple des cas de neuroblastomes ayant une réponse histologique au traitement très claire sans que leur taille n'ait réduit selon le critère RECIST [**McHugh2019**].

En outre, les lésions (pour le critère RECIST) et les plus grands axes sont choisis manuellement et donc sujets à imprécision : la variabilité intra et surtout inter-opérateurs est souvent significative [**Wilson2018** ; **Kuhl2019**] et amène à considérer les résultats avec précaution.

On peut également évoquer les critères PERCIST, mesures combinant l'information morphologique et métabolique utilisées pour l'évaluation de la réponse à partir d'un TEP-scan.

2.4 Un domaine prometteur : l'analyse radiomique

Les critères qualitatifs couramment utilisés par les radiologues et médecins nucléaires sont variés (forme et bords de la lésion, marges, densité, quantité de nécrose, etc.). Cependant, simplement basés sur l'évaluation visuelle du constraste, ils sont subjectifs et souffrent de variabilité intra et inter-observateurs [**Tixier2014**]. D'un autre côté, les critères quantitatifs de référence précédemment évoqués sont peu à peu remis en question car la description qu'ils proposent est incomplète [**Burton2007**].

D'autres caractéristiques quantitatives de l'image ont été reliées à de multiples phénomènes biologiques. Dans l'imagerie TEP, la prise de contraste au ^{18}F-FDG est liée au nombre de cellules cancéreuses viables et au processus de croissance de la tumeur. L'hétérogénéité intra-tumorale en général, qu'elle soit spatiale ou temporelle, est provoquée par les varia-

tions régionales du métabolisme comme la vasculature ou l'oxygénation et par l'expression génétique [YA16]. Elle est ainsi un signe avéré de malignité [**Fisher2013**]. Sa quantification peut ainsi permettre de différencier les types de tumeurs, les grades ou encore leur réponse [**Alic2014**].

Les caractéristiques de la forme d'une tumeur ont aussi été associées à son agressivité. [**Limkin2019**] ont récemment dressé une liste d'études ayant trouvé un intérêt significatif à une variable morphologique. Les lésions ayant des bords mal définis ont par exemple une plus grande capacité à envahir les structures adjacentes et sont associées à des tumeurs de stade avancé [**Cuccurullo2011**].

L'augmentation de la taille des jeux de données bio-médicaux et du nombre d'outils d'analyse a conduit au développement massif de méthodes d'extraction mathématique de ces caractéristiques descriptives difficiles à évaluer à l'œil nu. Ces mesures sont appelées *caractéristiques radiomiques* [Lam+12].

La radiomique a été étudiée dans de nombreuses applications diagnostiques en oncologie (détection de tissus pathologiques, classification du grade, etc.), largement détaillées par Yip et al dans leur revue de littérature [YA16]. Une étude de Parmar et al. [**Parmar2015**] a notamment révélé que les descripteurs associés à la survie du cancer du poumon ne sont pas les mêmes que ceux associés aux cancers des voies aérodigestives supérieures, montrant ainsi que certaines mesures sont spécifiques à une pathologie donnée.

3 Prédire l'évolution des cancers à court ou long terme

Une fois que des biomarqueurs radiomiques utiles à la description des tumeurs sont identifiés et susceptibles de constituer des facteurs pronostiques pertinents, des modèles combinant ces indicateurs peuvent être construits de façon à aller plus loin, en proposant des prédictions du devenir des lésions.

3.1 Méthodes de prédiction

Un modèle est une traduction des observations du réel dans un formulisme scientifique (mathématiques, algorithme) dans l'objectif d'exprimer des relations entre les variables. Une traduction inverse des relations trouvées par le modèle vers le réel permet d'obtenir des prédictions.

Nous décrivons ici les trois stratégies de modélisation évoquées dans nos travaux.

Méthodes statistiques classiques. Les méthodes statistiques descriptives, univariées et multivariées sont couramment utilisées dans la littérature médicale pour l'analyse de données. Elles permettent d'effectuer des comparaisons statistiques entre les variables, les cohortes et les patients et d'étudier la significativité de leurs différences (notamment par le calcul d'une *p-value*).

On réalise des analyses multivariées (qui prennent en compte plusieurs variables ainsi que les facteurs de confusions) à l'aide de modèles statistiques. Les plus fréquemment utilisés en

médecine sont les régressions (linéaires ou logistiques), les analyses de Kaplan-Meier et les modèles de Cox (voir chapitre 3).

Ce type d'analyse est essentiellement dédié à trouver des relations entre variables et à les expliquer, même s'il peut aussi être étendu à la prédiction [**Steyerberg2019**].

Modélisation mathématique. L'équipe INRIA Monc s'est spécialisée dans la construction et l'analyse de modèles mathématiques dédiés à l'oncologie. Une partie des recherches se concentre plus spécifiquement sur le développement de modèles de prédiction de l'évolution des tumeurs à partir de l'imagerie médicale.

L'idée est de prédire le temps de croissance d'une tumeur ou sa réponse à un traitement en utilisant des modèles d'équations aux dérivées partielles (EDP) non linéaires avec une série longitudinale d'examens IRM ou CT en entrée. L'objectif est de trouver un modèle EDP simple avec un nombre limité de paramètres indépendants, puis d'estimer les valeurs optimales de ces paramètres de façon à retranscire au mieux les observations. Ces paramètres sont ensuite utilisés pour réaliser des prédictions. La croissance des métastases du poumon et du foie, ainsi que celle des glioblastomes et des méningiomes a ainsi été étudiée.

Plus récemment, des membres de l'équipe ont tenté de calculer d'autres paramètres que le volume des lésions. Plusieurs descripteurs ont ainsi été utilisés dans les modèles : un critère d'hétérogénéité des métastases des GIST [**Lefebvre2015**], un estimateur de la quantité de cellules nécrosées et proliférantes dans les tumeurs rénales [**Peretti2017**], un indicateur de forme des gliomes pour prédire leur croissance et une nouvelle mesure d'intensité du signal en TEP pour prédire leur rechute [**Kritter2018_these**].

Intelligence artificielle. Les algorithmes d'intelligence artificielle, ou plus précisément d'apprentissage statistique (*machine learning*) et d'apprentissage profond (*deep learning*), ont fait de grands progrès dans le domaine de la perception (détection de motifs complexes) et de l'analyse quantitative de l'imagerie médicale. Les capacités prédictives des modèles d'apprentissage sont décuplées par l'augmentation des données de radiologie.

3.2 Étude préliminaire : prédiction de la survie sans progression des cancers du pancréas

Au cours de travaux antérieurs à cette thèse, nous avons mené une étude préliminaire mêlant radiomique et apprentissage pour la prédiction de la survie sans progression du cancer du pancréas, en collaboration avec l'Institut Bergonié. Les travaux initiés dans le cadre de cette thèse découlent directement des conclusions et des limitations que nous avons observées lors de la réalisation de cette étude préliminaire.

Contexte

Le cancer pancréatique est une pathologie rare dont l'incidence augmente en Europe et en fait la troisième cause de décès lié au cancer [**Siegel2014**]. De pronostic très défavorable, il a un taux de rechute à 80% et sa médiane de survie à 5 ans est estimée à 5% [**Carpelan2005** ; **Cartwright2008**]. Le seul traitement à ce jour est la résection chirurgicale R0 associée à la chimiothérapie adjuvante. En cas de lésion non opérable, notamment pour les cas nombreux

de diagnostic tardifs, la chimiothérapie ou radio-chimiothérapie peut augmenter les chances d'effectuer une chirurgie secondaire pour environ un tiers des patients. Malgré cela, si la chirurgie seule améliore le taux de survie, l'opération est lourde et risquée et l'hospitalisation qui suit est longue. L'utilisation de la radiochimiothérapie pour que la chirurgie soit tolérée est elle aussi lourde d'effets secondaires.

L'étude préliminaire initiée par Claudia Pouypoudat, radiothérapeute, vise à évaluer l'impact réel de la radiochimiothérapie sur le taux de résections secondaires des tumeurs primaires en **estimant si la survie sans progression et la survie globale des patients atteints d'un adénocarcinome pancréatique sont prédictibles avant la chirurgie [Pouypoudat2017]**.

Notre approche

Nous avons essayé de classer une cohorte de patients atteints d'un adénocarcinome du pancréas en fonction de 3 critères : survenue d'un évènement (présence *vs* absence de rechute), survie à court terme (avant le temps de rechute médian *vs* après ou non concerné), survie globale (décès lié au cancer *vs* non lié ou non concerné).

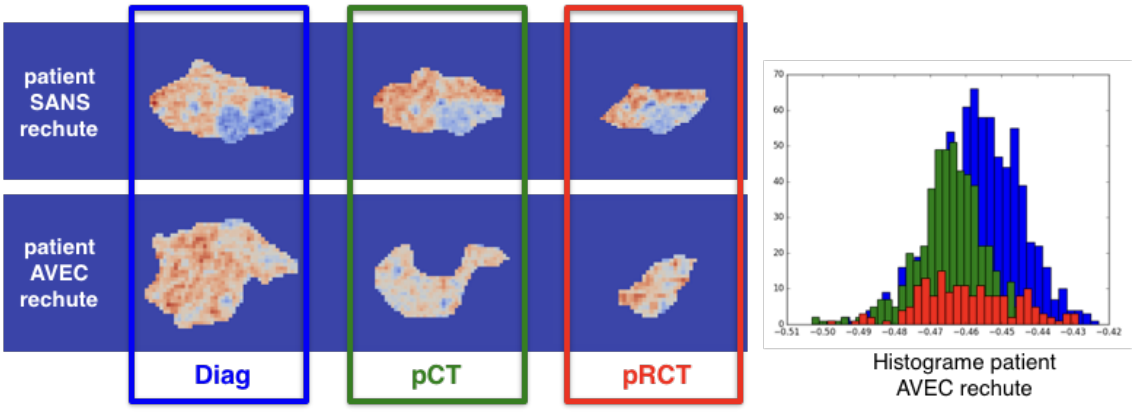


FIGURE 4 – Exemples : plus grandes ROI 2D sur CT de la tumeur du pancréas de deux patients au trois temps étudiés. À droite, l'histogramme des trois temps pour le patient qui rechute.

Nous avons constaté une évolution de l'intensité et de la texture des tumeurs du diagnostic (*Diag*) au traitement par chimiothérapie (*pCT*) et au traitement par radio-chimiothérapie (*pRCT*) (voir exemple Fig. 4). Nous avons donc décidé d'étudier l'évolution de l'hétérogénéité des lésions en passant par une approche delta-radiomique (voir section 1.5). Nous avons entraîné des modèles d'apprentissage statistique à partir de données cliniques et de descripteurs de texture des ROI 2D sur CT. Les détails d'implémentation sont donnés en annexe A.

Résultats préliminaires

Les principaux résultats de l'apprentissage statistique se résument comme suit :

• Les modifications de texture se sont avérées plus significatives entre le pCT et le pRCT (voir histogramme Fig. 4).

• Les modèles de prédiction de la survie sans évènement obtiennent les meilleurs résultats. Les faux négatifs sont peu nombreux.

• Les modèles de prédiction de la survie globale obtiennent globalement de mauvaises performances. Les modèles de prédiction de la survie à court ou long terme échouent également à classifier les patients selon leur temps médian de rechute.

• Les modèles construits avec les variables cliniques et les variables radiomiques obtiennent de meilleurs résultats que ceux qui n'utilisent qu'un seul type de variables.

Les pistes d'améliorations

Les résultats obtenus sont encourageants et suggèrent que l'impact de la RCT sur le devenir des patients est mesurable avec l'information de texture. Cette information pourrait permettre d'estimer les chances de rechute d'un patient.

Sans surprise le décès lié au cancer est plus difficile à prédire, car il dépend de facteurs non pris en compte, comme la tolérance générale du patient aux toxicités du traitement et à l'opération. La classification de la survie à court terme ne fonctionne pas non plus. Le problème peut venir du seuil (la médiane, environ 300 jours). La régression directe sur le temps de survie sans évènement n'a pas été testée, mais la taille de la cohorte est potentiellement trop faible pour qu'un apprentissage sur données continues soit fiable.

Plus généralement, la taille du jeu de données elle-même est un frein puisqu'elle limite la confiance dans la capacité des modèles à généraliser leur prédiction. Les résultats n'ont ainsi pas été validés sur un ensemble indépendant

Cette étude souffre en outre de plusieurs limitations techniques. Les premières viennent du contourage manuel. Les surfaces sont difficiles à délinéer et nous n'avons pas fait vérifier les contours par un second opérateur. Du fait des difficultés de délinéation, la texture a été analysée sur une seule coupe. La forme de la tumeur n'est pas prise en compte et une grande quantité d'informations a peut être été manquée. D'autre part, malgré une qualité parfois médiocre (artefacts, etc.), le pré-traitement des images est limité.

Du côté des modèles de classification, le principal frein est le déséquilibre de la distribution des classes dans le jeu de données, qui rend les performances plus difficiles à évaluer.

Ces limitations nous ont permis de tirer des enseignements pour la suite de nos recherches. Nous souhaitons ainsi utiliser des cohortes plus grandes de lésions où la texture est mieux discernable à l'image. Nous porterons une plus grande attention à ses artefacts et renforcerons les pré-traitements. L'étude de l'ensemble du volume des tumeurs est à privilégier, de façon à pouvoir intégrer l'information de forme. Nous souhaitons aussi mieux sélectionner les métriques d'analyse pour une vue fiable des performances des classifieurs.

4 Contributions cliniques, méthodologiques et logicielles

Nous cherchons à prédire l'échappement des lésions, leur rechute et le décès des patients pour les prévenir, mais aussi aider à limiter les traitements inutiles. Nous avons pour cela travaillé en collaboration avec l'Institut Bergonié, l'hôpital Saint André et l'hôpital Haut l'évêque à Bordeaux.

4.1 Problématiques cliniques et contributions

Chaque pathologie étudiée a ses enjeux propres et on se propose à présent de discuter des exemples de problématiques cliniques auxquelles nous avons été confrontés.

Décrire la survie sans progression des carcinomes du canal anal par l'analyse des caractéristiques de texture à l'IRM

L'objectif est d'évaluer les résultats fournis par une étude des descripteurs de texture des carcinomes du canal anal sans métastases sur l'imagerie IRM. Nous cherchons à savoir si leur évolution est liée à la survie sans progression des patients. Pour cela, nous disposons des examens IRM de 28 patients traités à l'hôpital Haut-L'évêque.

Classer la réponse au traitement des sarcomes de haut grade grâce à une analyse delta-radiomique IRM

Le traitement standard des patients atteints d'un sarcome des tissus mous (STS) de haut grade sont en cours de redéfinition, afin d'inclure la chimiothérapie néo-adjuvante (NAC) qui a récemment montré des effets positifs sur le pronostic des patients. Cependant, l'évaluation de la réponse au traitement durant les essais cliniques repose encore sur le critère RECIST 1.1.

L'objectif de cette étude est d'estimer l'apport d'une approche combinant variables radiologiques et delta-radiomiques dans la prédiction précoce de la réponse au traitement NAC de patients atteints d'un STS. Nous utilisons pour cela une base de données de l'Institut Bergonié comprenant 65 patients avec une IRM au diagnostic, une autre acquise après deux cycles de traitement ainsi qu'un ensemble de variables et mesures démographiques, cliniques et radiologiques.

Prédire la rechute métastatique précoce des cancers du sein

L'estimation de la probabilité de rechute métastatique du cancer du sein de stade précoce est essentielle à la prise de décision thérapeutique, notamment en ce qui concerne les traitements adjuvants. Les modèles de prédictions actuels se basent uniquement sur des statistiques classiques de type modèles de Cox. En utilisant un jeu de données fourni par l'Institut Bergonié (environ 600 patientes), nous pouvons évaluer le potentiel des algorithmes d'apprentissage pour la prédiction de la rechute métastatique à 5 ans et les comparer aux modèles mécanistiques développées par l'équipe Monc.

4.2 Besoins et contraintes méthodologiques

L'objectif à terme de ces trois problématiques cliniques est de prédire l'évolution du patient suite à son parcours de soin de façon à pouvoir le réorienter. Pour cela, nous devons construire des modèles simples basés sur des variables dont l'interprétation biologique et visuelle est pertinente.

Les motivations additionnelles sont d'ordre méthodologique et consistent à développer une méthode de traitement et d'analyse réutilisable et fiable basée sur les données rétrospectives de routine. Cela implique d'utiliser une quantité limitée de patients et d'examens constituant

des cohortes potentiellement déséquilibrées. Leur temps de suivi n'est pas toujours très long ni homogène ce qui peut limiter les études de survie.

D'autres contraintes techniques comme l'hétérogénéité des IRM, l'hyper-paramétrage des algorithmes ou la sélection des multiples variables descriptives sont également à adresser.

4.3 Développement logiciel

L'avancement de ces travaux de thèse a initié et/ou amélioré le développement de bibliothèques d'outils codées principalement en Python.

La principale est **papriK** [12] (*Python Analysis and PRocessing of Images toolKit*), que nous avons créée et employée pour la lecture, le traitement et l'analyse des images médicales. Cette bibliothèque a été utilisée dans la quasi totalité des travaux de la thèse, notamment pour les pré-traitements d'images et le calcul de caractéristiques radiomiques. Elle est majoritairement basée sur les bibliothèques *numpy* (manipulation de tableaux), *ITK* (lecture, traitement, analyse d'image [**Yoo2002**]) et *VTK* (idem).

À présent co-maintenue avec l'équipe Monc, la bibliothèque fait l'objet d'une intégration continue avec le service CI-INRIA et son code source est lui même disponible auprès d'Inria. Un conteneur *docker* rassemblant le code et ses dépendances peut être fourni à la demande par l'équipe.

La seconde bibliothèque, **maliK** [13] *(MAchine LearnIng toolKit)* rassemble les routines nécessaires à la création des chaînes de traitement de données et d'apprentissage statistique décrits dans les chapitres de la seconde partie de la thèse. Elle utilise notamment les modules Python *scipy* (statistiques), *pandas* (organisation de données hétérogènes en tableaux structurés), *seaborn* (visualisation, graphes), *scikit-learn* (apprentissage statistique [**Pedregosa2012**]) et *keras* (*deep-learning*). Elle est disponible sur *bitbucket* (cperier/maliK).

4.4 Organisation du manuscrit

La suite du présent document est constituée de deux grandes parties de trois chapitres chacune et d'une discussion générale finale.

Dans la première partie sont exposées nos recherches sur le traitement des images IRM et l'extraction de caractéristiques radiomiques pour décrire une zone d'intérêt.

Nous détaillons dans un premier chapitre la reconstruction des volumes d'intérêt et les indicateurs radiomiques que nous utilisons. Ces caractéristiques se répartissent en deux catégories : les mesures décrivant la forme et la morphologie 3D d'une lésion et celles décrivant les niveaux de gris des voxels. Nous explicitons les méthodes de calcul choisies et leurs limitations, avant d'étudier brièvement la reproductibilité des mesures de morphologie

12. À prononcer "paprica".
13. De même, à prononcer 'malika'.

suivant les caractéristiques du support de l'image. Nous observons que même pour un volume modeste, les erreurs de mesure dues à la discrétisation et à la reconstruction de l'image s'avèrent négligeables.

Le deuxième chapitre porte sur le traitement des images IRM dans l'objectif d'étudier les caractéristiques radiomiques de plusieurs examens et de plusieurs patients. L'objectif est d'extraire des descripteurs fiables et comparables d'un examen à l'autre. Nous décrivons d'abord les traitements servant à homogénéiser les défauts d'une image provoqués par le système d'imagerie et le processus d'acquisition. Ensuite, nous détaillons les traitements de normalisation des niveaux de gris et de la taille des voxels d'une image à l'autre, l'objectif étant d'harmoniser la signification des caractéristiques extraites. Nous étudions également leur reproductibilité en fonction des traitements d'image choisis. Ces tests nous permettent d'établir des recommandations pour limiter les erreurs de mesure dues à la transformation de l'image.

Le troisième chapitre présente les résultats de notre première problématique clinique. Nous avons extrait les caractéristiques radiomiques d'intensité et de texture sur l'IRM de diagnostic de 28 patients atteints d'un carcinome du canal anal. En raison de la taille modérée du jeu de données, seule une étude statistique classique a été réalisée. Nous montrons que deux mesures sont significativement associées à la survie sans progression. Nous expliquons que cette étude gagnerait à étudier l'évolution de ces marqueurs dans le temps ainsi que leur valeur prédictive.

La seconde partie du manuscrit est consacrée aux applications prédictives en oncologie.

Le chapitre 4 est conçu comme une introduction à l'apprentissage statistique et un guide pratique de l'implémentation de notre chaîne de traitement, de la donnée à la prédiction. Cette section n'est pas un catalogue exhaustif des techniques de *machine learning* : il se concentre sur la classification et tend à couvrir les notions et outils que nous utilisons dans les chapitres suivants. Le lecteur déjà familier avec ces mécanismes peut simplement s'attarder sur la justification de nos choix d'implémentation.

Le chapitre 5 est une application directe de cette méthodologie et présente les résultats de notre deuxième problématique clinique. Nous cherchons ici à prédire au plus tôt la réponse au traitement néoadjuvant des sarcomes des tissus mous. Grâce à deux examens IRM, nous avons calculé l'évolution des caractéristiques radiomiques que nous avons combinées aux variables démographiques et radiologiques de 65 patients. Un de nos classifieurs, entraîné et testé sur les 50 premiers patients, a obtenu des scores de prédiction élevés. La particularité de ce modèle est qu'il est construit avec une signature composée de seulement trois variables comprenant une caractéristique de forme, une caractéristique d'intensité et une caractéristique qualitative radiologique, l'évolution de la quantité d'œdème. Ce modèle est validé sur le second ensemble de 15 patients. Après analyse, les erreurs restantes sont toutes identifiées comme étant des cas très particuliers. Nous présentons enfin des premiers résultats prometteurs de détection automatique de l'œdème dans l'objectif de le quantifier automatiquement.

Enfin, le sixième chapitre couvre les résultats de la prédiction de la rechute à distance du cancer du sein d'une cohorte rétrospective de près de 600 patients. Cette étude n'utilise pas l'imagerie. Elle vise à comparer les qualités respectives de la classification par apprentissage statistique et des modèles mathématiques inspirés de la biologie dans la perspective de combiner les deux approches. Nous avons ainsi développé des classifieurs pour prédire la rechute à court ou long terme de la cohorte.

Première partie

Analyse quantitative de l'image IRM

Chapitre 1

Caractériser la morphologie et la texture d'une zone d'intérêt avec la radiomique

Sommaire

Analyse quantitative des régions d'intérêt

Le terme **radiomique** (*radiomics*) apparaît dans la littérature en 2010 [Gil+10]. Il désigne l'extraction numérique et l'analyse de biomarqueurs quantitatifs à partir d'images médicales [Tra+18]. Ces attributs sont généralement invisibles à l'œil humain, même expert [YA16]. L'objectif est de caractériser les traits phénotypiques significatifs d'une tumeur, souvent dans un but prédictif.

Dans ce chapitre sont décrits les biomarqueurs [1] utilisés dans les études mentionnées en introduction. Ils peuvent être regroupés en deux catégories principales : les indicateurs sur la morphologie d'une région d'intérêt et ceux sur l'intensité des pixels et les motifs de texture qu'ils forment.

On sous-entend ici que la région ou le volume d'intérêt est connu et qu'il existe donc un masque applicable sur l'image de base qui restreint la zone d'extraction des indicateurs.

1.1 Reconstruction des volumes d'intérêt

Les examens produits par les différents systèmes d'imagerie sont stockés dans des fichiers aux formats adaptés, propriétaires ou non. Nous utilisons des fichiers de norme DICOM [Par95] ("*Digital imaging and communications in medicine*") qui est le standard de gestion numérique des données et méta-données d'imagerie (stockage, transfert), généralisé en routine clinique. Les fichiers contenant les contours des régions d'intérêt cliniques sont également pris en charge.

1.1.1 Identification des volumes d'intérêt

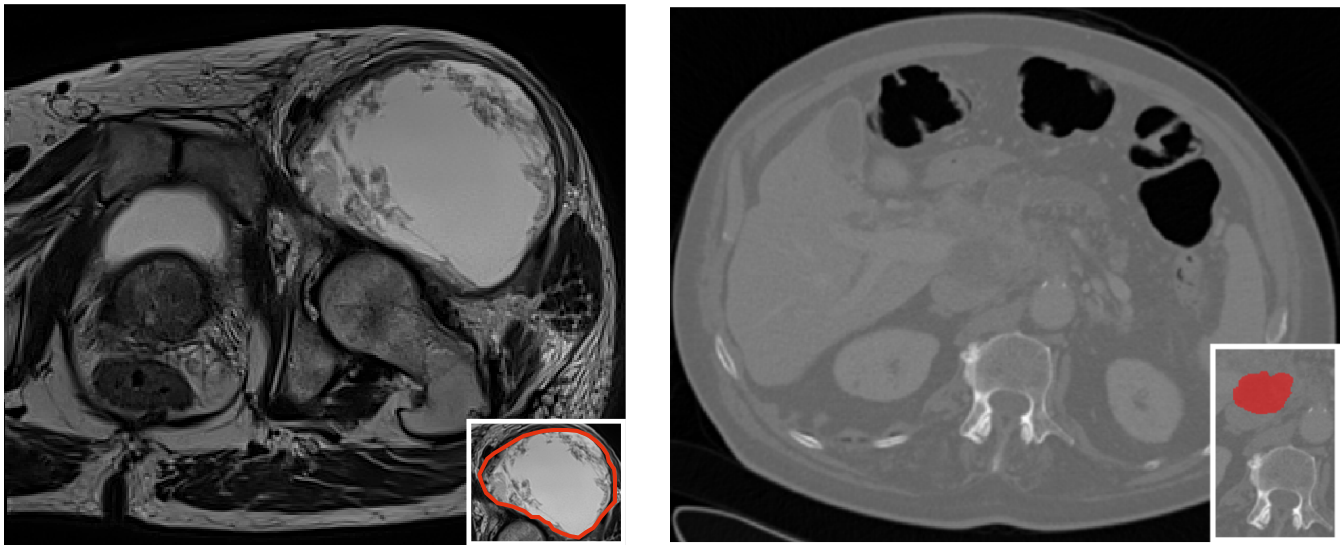
La région d'intérêt (ROI) ou volume d'intérêt en 3D (VOI) est un sous-ensemble des pixels/voxels d'une image médicale sur lesquels se concentre l'analyse. En oncologie il peut s'agir de n'importe quelle lésion (tumeur ou métastase), de son environnement (œdème, fibrose périphérique etc.), des cicatrices dues au traitement (après chirurgie par exemple), d'une zone (représentative) de tissus sains, voire du corps entier (seul le fond de l'image est exclu).

Le calcul des marqueurs radiomiques est pertinent sur image entière mais reste plus simple à réaliser et interpréter quand il est restreint à une zone d'intérêt. La délinéation est le processus d'extraction manuelle de ces zones et constitue le cœur de la pratique du radiologue en oncologie [GKH15]. Cette étape est également indispensable en radiothérapie où il est nécessaire de connaître précisément les limites d'une lésion ou des organes fragiles voisins (dits "à risque") pour envoyer la dose correcte à l'endroit souhaité.

Classiquement, le contourage est réalisé manuellement au cours de la routine clinique par un médecin ou physicien, avec un logiciel proposant les outils adéquats (type "crayon" pour les contours ou "pinceau" pour le remplissage). Un volume d'intérêt est dessiné coupe par coupe, en utilisant l'information qualitative apportée par les coupes voisines ou par d'autres modalités d'examens. Ce processus s'avère long et souvent fastidieux quand les marges d'une lésion sont peu visibles et/ou irrégulières (Fig. 1.1).

De plus, le résultat d'un contourage manuel est intrinsèquement sujet à des variations inter et intra-opérateurs [Cor+17b]. La figure 1.2 montre un exemple d'un même sarcome délinéé sur trois séquences différentes par un même radiologue. Les ROI 2D y sont visiblement différentes et les volumes entiers n'ont pas exactement la même forme.

1. Nous utiliserons indifféremment les termes descripteurs, caractéristiques, mesures, variables, indicateurs ou biomarqueurs.



(a) *source IRM : Institut Bergonié*

(b) *source CT : CHU Bordeaux*

FIGURE 1.1 – (a) Coupe axiale d'un sarcome des tissus mous de la cuisse en IRM T2. Le volume est grand et les bords de la tumeur sont nets, facilitant le contourage. (b) Coupe axiale de l'abdomen d'un patient atteint d'un cancer du pancréas en CT au temps portal. Le contraste est beaucoup moins prononcé entre la tumeur et les tissus sains et les marges sont indistinctes, ce qui complique la délinéation.

Par conséquent devenue un enjeu majeur de la dernière décennie, la segmentation automatique est actuellement un domaine de recherche très actif [LPXP00 ; RNZ18 ; AG18], faisant l'objet de publications nombreuses, de conférences dédiées voire de concours [Bra ; Aap]. De nombreux outils de segmentation semi-automatique ou automatique adaptés à des cas classiques (comme la segmentation de tumeurs du cerveau [Zhu+12]) ont été mis au point et intégrés aux logiciels. Cette aide à la segmentation peut simplement consister en une proposition d'un contours candidat, grâce à la sélection manuelle d'un point (graine / "*seed*") dans le VOI ou d'une zone grossière autour et à un paramétrage des algorithmes.

Il a été montré que la précision et l'efficacité des méthodes de segmentation semi-automatiques étaient meilleures que celles d'un contourage manuel [RV+12 ; Par+14]. Pour certains cas particuliers comme les tumeurs au cerveau (qui disposent de larges bases de données en ligne [Men+14]) ou celles du sein, la recherche s'oriente vers les solutions entièrement automatiques, grâce à des algorithmes d'apprentissage profond (voir [SWS17 ; Bn+18] en IRM, [Hai+19] en CT).

Qu'ils soient réalisés manuellement, semi-automatiquement ou automatiquement, les VOI peuvent être stockés coupe par coupe de deux façons :

— Soit enregistrés sous forme de listes de coordonnées des points du tracé des contours. Cette structure est la plus commune, retrouvée notamment dans les formats XML, JSON, RTStruct [2] et autres formats propriétaires des différents constructeurs de systèmes ;

2. Adaptation de la norme DICOM pour les contours en radiothérapie.

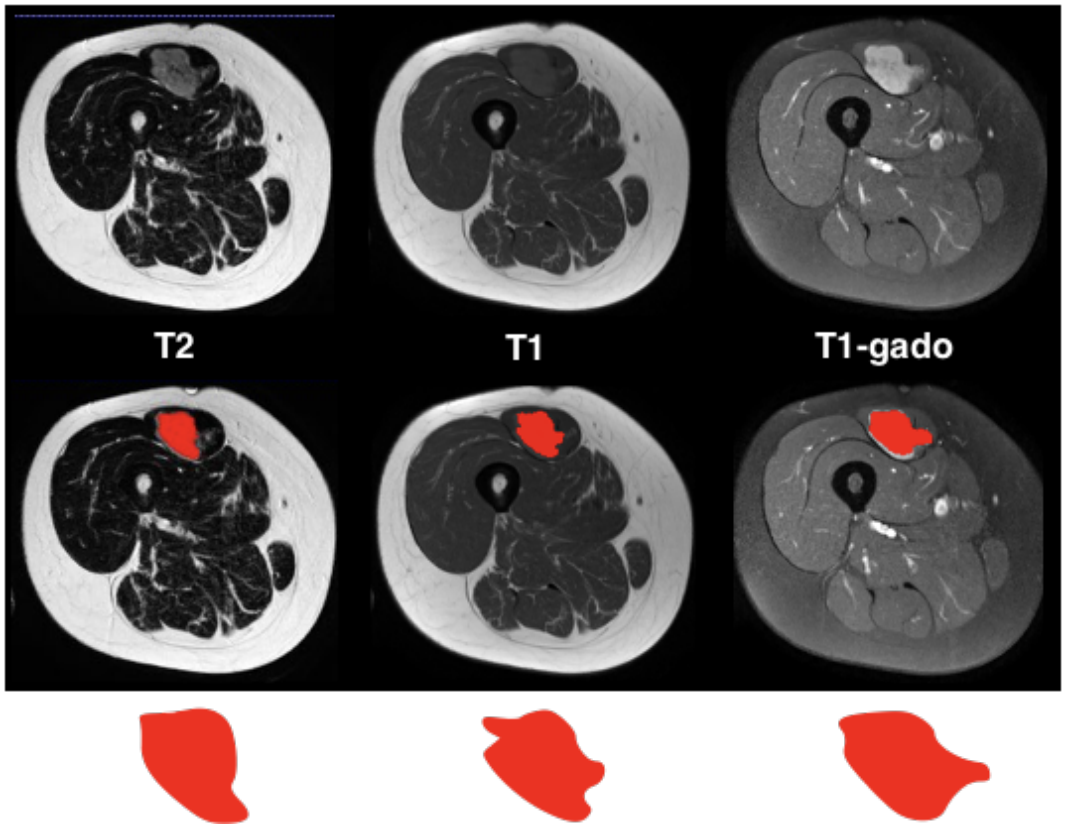


FIGURE 1.2 – Coupe axiale d'un sarcome de la cuisse contouré sur trois examens en T2, T1 et T1-gado par le même opérateur. La dernière ligne montre les ROI 2D agrandies, dont la forme change très visiblement. Le volume 3D contouré en T1 diminue de 2.6% seulement par rapport à celui du T2 et de 21.6% par rapport à celui contouré en T1-gado.
source IRM : Institut Bergonié

— Soit convertis en masques binaires (images à deux niveaux, un pour le VOI et un pour l'arrière-plan) et enregistrés sous forme de listes de niveaux de gris. Cette structure est générée par les logiciels de recherche type ITK-SNAP [Yus+06]. Les formats de stockage sont variés (JPEG, NII, VTK, DICOM).

La plupart des contours des études présentées dans ce manuscrit ont été réalisés par des radiologues ou radiothérapeutes avec le logiciel *Osirix* [RSR04] et exportés au format XML par le plugin *ExportROI*. Les contours ont été vérifiés plusieurs fois dans le temps et/ou par plusieurs expert.e.s.

1.1.2 Reconstruction 3D

Les contours stockés sous forme de listes de points sont ensuite convertis en masques dans un processus en deux étapes : regroupement des points/sommets en polygones de surface puis voxelisation de l'intérieur.

Dans notre cas, les points des contours ne sont pas disposés totalement aléatoirement : ils sont structurés, alignés par coupe (axe z), comme on peut le voir à gauche de la Fig. 1.4 et stockés dans leur ordre de tracé. Sans avoir besoin de générer un maillage complexe de la surface du volume, nous tirons parti de cette structure en reconstruisant le volume coupe par coupe. Les sommets sont reliés point par point de façon à créer un polygone pour chaque plan de coupe.

Les primitives sont ensuite converties en images 2D (principe détaillé Fig.1.3). Le support est une grille 2D de même origine, taille et pas d'espace que la coupe de référence (Fig.1.3c) pour que le masque obtenu lui soit parfaitement superposable. La grille intiale ne contient qu'une valeur de pixel par défaut, celle de l'arrière-plan. Elle est modifiée en déterminant pour chaque pixel s'il est situé dans le polygone (Fig.1.3e). C'est le "test de couverture" : si le résultat de la fonction de contours [3] de chaque arrête du polygone est positif pour un pixel, c'est qu'il est à l'intérieur de la ROI.

Il est donc marqué et on fait de même pour tous les autres pixels (Fig.1.3f). La zone à traiter est réduite à la boîte englobante du polygone pour éviter d'avoir à parcourir l'ensemble des pixels et ainsi gagner en temps de calcul. Nous effectuons cette opération avec la librairie ITK [Joh+13].

Les masques 2D (milieu Fig.1.4) sont ensuite extrudés (passés à la 3D, étendus sur une épaisseur d'1 voxel) perpendiculairement à l'orientation de l'image d'origine. Les coupes sont alors directement empilées en z pour donner le masque 3D final (Fig.1.4 à droite).

Cette méthode produit des masques au crénelage fort. Cependant, ces images n'ont pas pour objectif de constituer une représentation visuelle plaisante (voir Fig.1.5) mais d'être fidèles à la référence biologique. Nous n'utilisons donc pas de technique d'anti-crénelage (sur-échantillonnage, interpolation des valeurs etc.)

Les inconvénients principaux de la délinéation et de la reconstruction concernent les limites du VOI. Le premier problème est l'effet de volume partiel. Lorsqu'un voxel est positionné à la frontière entre différents tissus biologiques, le niveau de gris qui lui est attribué reflète

3. *Edge function* [Pin88].

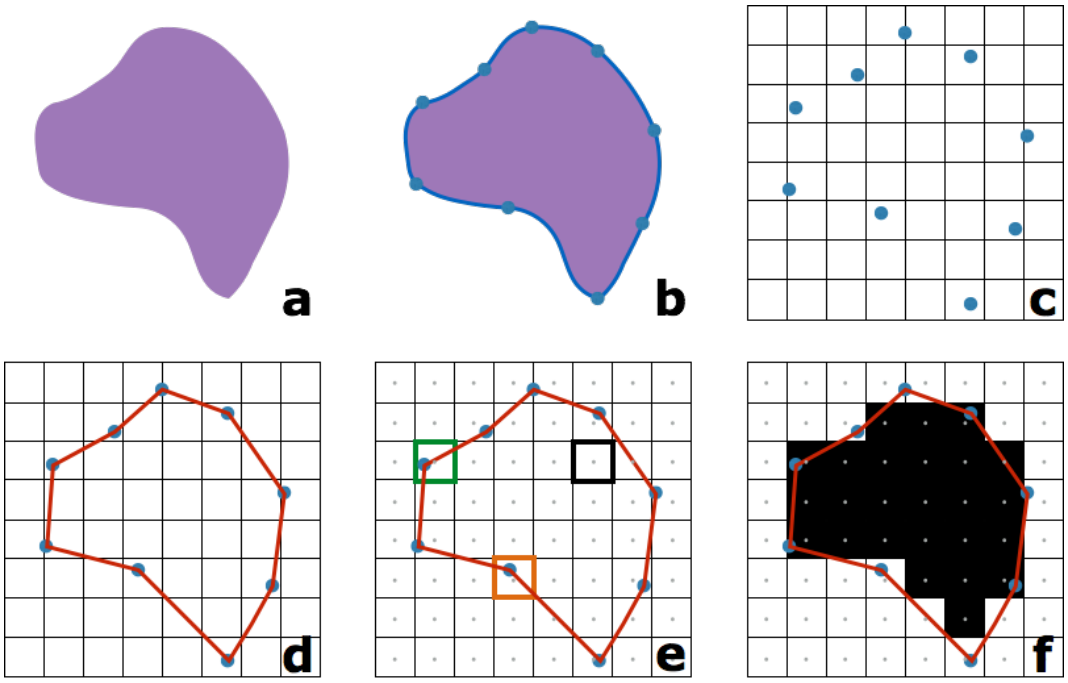


FIGURE 1.3 – Schéma de la reconstruction 2D d'une surface (a). Les points successifs du contours sont récupérés (b) et une grille d'espacement et d'origine donnés est choisie et positionnée (c). Les points servent à créer un polygone (d). Le test de couverture va déterminer si chaque voxel de la grille appartient ou non à la ROI. Ici, le pixel noir est entièrement dans la ROI, le centre du pixel vert aussi. Celui du pixel orange ne l'est pas (e). Le masque final est obtenu en remplissant les pixels concernés (f).

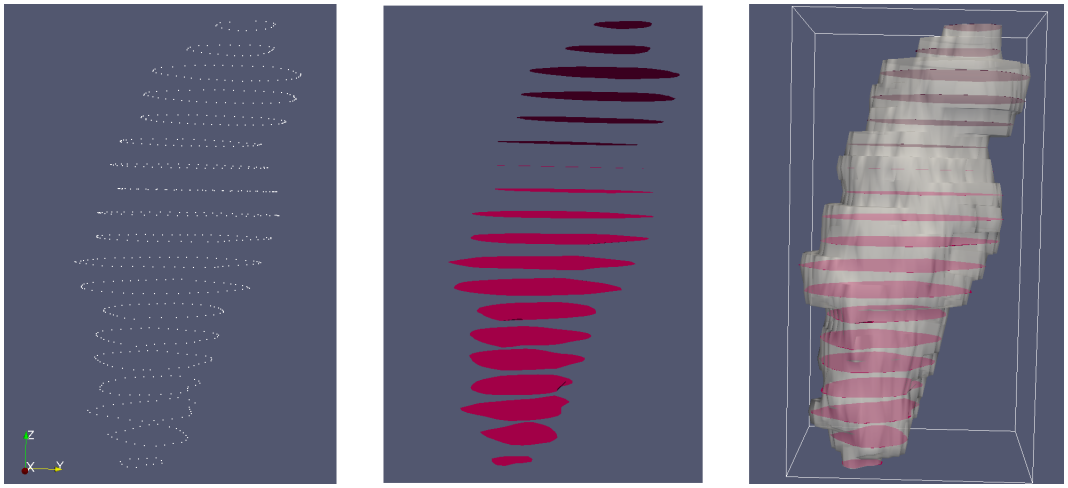


FIGURE 1.4 – (*à gauche*) Les listes de points sont organisées par ROI 2D. (*au milieu*) Cela permet de créer des polygones dans le plan et de remplir des coupes. (*à droite*) Le passage à la 3D se fait simplement par extrusion perpendiculaire.

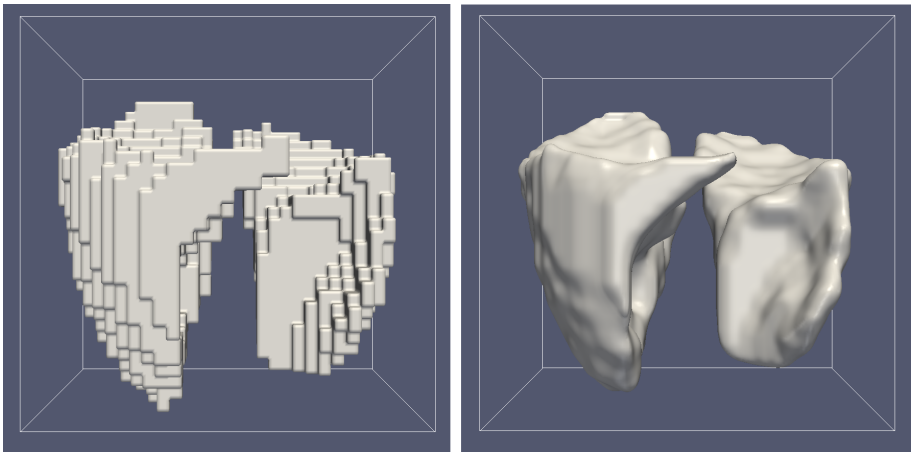


FIGURE 1.5 – VOI de poumons directement reconstruits en 3D (à gauche) ou préalablement lissés avec un filtre gaussien pour atténuer le crénelage.

la moyenne pondérée du signal des tissus sous-jacents [GBZB02]. Ces valeurs approximées peuvent conduire à des imprécisions de délinéation provoquant elles-même des erreurs de mesure.

En outre, plus la résolution de l'image initiale est grossière, plus les limites du VOI délinéé peuvent être décalées des contours réels de l'objet, sur une distance d'un demi voxel maximum. Le volume total peut être modifié : si par exemple l'objet d'intérêt est trop petit par rapport à la résolution de la grille, la mauvaise estimation de son volume sera proportionnellement beaucoup plus importante.

En gardant en tête ces limitations, les masques 3D peuvent directement être utilisés pour étudier la forme et le volume d'une lésion. Chaque voxel n'appartenant pas à l'arrière-plan constitue une brique élémentaire du volume total.

1.2 Descripteurs de la morphologie

La morphologie est la description des aspects géométriques d'un objet. Les indicateurs morphologiques discrets sont calculés sur le VOI.

La taille d'une tumeur ou autre lésion est souvent l'une des premières mesures d'intérêt puisqu'elle permet de déterminer le grade et d'assurer son suivi. La mesure des dimensions peut s'effectuer sur pièce chirurgicale après résection, mais est surtout utile quand réalisée dès le diagnostic et à chaque examen, à l'imagerie. On a notamment vu en introduction (section 2.3) que le critère RECIST 1.1 s'appuie sur une mesure du plus large diamètre de la tumeur à l'image.

La routine clinique inclue également des mesures du volume, généralement effectuées numériquement en comptant le nombre de voxels dans le masque et en multipliant ce nombre par le volume d'un voxel. Cette méthode naïve peut toutefois poser problème. Suivant le positionnement de la grille à la reconstruction, les voxels de la périphérie sont en réalité situés

seulement partiellement dans l'objet. Cela peut conduire à une légère erreur d'évaluation du volume.

Cette approximation est supposée négligeable pour des objets constitués d'un nombre de voxels suffisants (> 1000 [Zwa+16]). Dans le cas contraire, l'ordre de grandeur de la surestimation se rapproche de celui de la mesure. Certaines études excluent les VOI de moins de 3-5 cm^3 [Orl+14], et d'autres conseillent d'étudier et de reporter la corrélation entre le volume et les autres mesures radiomiques en dessous de $10cm^3$ de lésion mesurée sur une TEP [Hat+14] . Nous étudierons l'impact des volumes de taille réduite sur les mesures radiomiques en section 1.6.

Nous appliquons la méthode naïve de calcul, mais il est à noter que d'autres études ou logiciels préfèreront des mesures de volumes basées sur le maillage sous-jacent, constitué des coordonnées des centres des voxels.

Le calcul de l'aire de la surface est moins évident pour un objet 3D. La surface extérieure d'un objet peut être obtenue en sommant l'aire des faces du maillage de l'objet. La méthode de Mullikin et Verbeek [MV93], reprise par [Lin03] est basée sur les voxels et passe par deux étapes. D'abord, les voxels à la périphérie de l'objet sont détectés. Ensuite, les aires des faces qui constituent la frontière objet/arrière-plan sont ensuite pondérées suivant leur configuration dans l'espace et leur type de voisinage, puis sommées. La bibliothèque ITK a choisi de calculer les périmètres et surfaces avec la formule de Cauchy-Crofton [LL12], qui se base le nombre de points d'intersection entre la surface de l'objet et un ensemble de lignes droites [Li+03].

Le calcul de l'aire d'une portion de la surface d'un objet peut également être intéressant, par exemple en radiothérapie lorsqu'on veut estimer quel pourcentage de la surface d'un organe à risque ou d'une lésion a été atteint/manqué par les rayons [Hoc+16b].

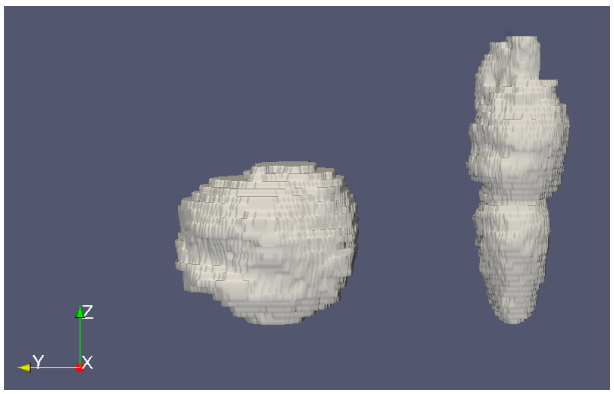


FIGURE 1.6 – VOI de deux sarcomes de haut grade. Le VOI de gauche a une élongation à 1.06, celui de droite a une valeur quasi trois fois supérieure

Autres indicateurs

Les dimensions classiques d'un objet, diamètre, aire, etc. ne sont pas les seuls indicateurs à extraire d'un masque. Nous résumons ici le sous-ensemble de mesures étudiées dans nos travaux.

Parmi cette liste, on note que les critères d'élongation et de planéité (*flatness*) reposent sur l'utilisation de l'analyse en composantes principales (ACP) pour déterminer l'orientation principale de la ROI. Pour un objet 3D, l'ACP donne ainsi trois vecteurs qui correspondent aux directions d'orientation d'une ellipsoïde de taille équivalente, ainsi que trois valeurs indiquant l'ampleur de l'extension de l'objet dans ces directions.

Soit **A** l'aire de la région d'intérêt, **V** son volume et λ_{majeur}, λ_{mineur}, $\lambda_{moindre}$ les valeurs propres (dans l'ordre d'importance) de l'ACP de la région cible. Les caractéristiques utilisées sont réunies Table 1.1.

Élongation	À quel point un objet est plus long que large et étendu. 1 : sphère - 0 : ligne ou point.	$\sqrt{\frac{\lambda_{second}}{\lambda_{largest}}}$
Planéité / flatness	À quel point un objet est plat comparé à sa longueur. 1 : sphère - 0 : objet plat.	$\sqrt{\frac{\lambda_{smallest}}{\lambda_{largest}}}$
Sphéricité	À quel point l'objet se rapproche d'une sphère compacte représentative. 1 : sphère parfaite	$\frac{\pi^{1/3}*(6*V)^{2/3}}{A}$
Rayon de la sphère équivalente	Rayon de la sphère ayant le même volume que l'objet.	
Diamètre de Feret	Plus grande distance mesurable entre deux lignes/plans parallèles tangent(e)s à l'objet. Comparable à la distance mesurée par le RECIST.	

TABLE 1.1 – Descripteurs quantitatifs de la morphologie d'un VOI : définitions et formules.

La Fig 1.6 donne un exemple de VOI de sarcomes dont les volumes ont un ordre de grandeur similaire mais dont l'élongation diffère fortement.

1.3 Descripteurs de l'intensité

L'intensité dans une image reflète la distribution statistique des niveaux de gris dans un volume d'intérêt (son histogramme). Sa quantification repose classiquement sur les quatre moments de la distribution : la moyenne, l'écart type, le coefficient d'asymétrie (*skewness*) et le coefficient d'aplatissement (*kurtosis*). D'autres paramètres de l'histogramme de distribution peuvent être pris en compte : médiane, percentile, intervalle entre les valeurs extrêmes, entropie, etc.

Soit $\bar{x}$ la moyenne de la distribution, s son écart type, n le nombre de niveaux de gris discrets et $p(x)$ l'histogramme normalisé. Nous calculons les descripteurs de l'histogramme listés Table 1.2 avec la bibliothèque *scipy*.

La figure 1.7 donne les caractéristiques et les histogrammes de deux VOI de sarcomes à l'IRM. Ces descripteurs suffisent à caractériser deux tumeurs très différentes d'aspect.

Les mesures d'intensité fournissent uniquement une information globale sur la région cible, sans prendre en compte leurs relations spatiales. Parfois considérées comme de la texture "de premier ordre", il ne s'agit pourtant pas à proprement parler de descripteurs de la texture.

Intervalle	Gamme des valeurs de pixels.	$max(x) - min(x)$
Kurtosis	Coefficient d'aplatissement. < 0 : distribution *aplatie*, > 0 : distribution *pointue*.	$\frac{\sum(x-\bar{x})^4/n}{s^4} - 3$
Skewness	Asymétrie autour de la moyenne. Le signe dépend de l'extrémité allongée.	$\frac{\sum(x-\bar{x})^3/n}{s^3}$
Entropie (de Shannon)	Évalue le caractère aléatoire et le désordre dans la dispersion de la distribution : une valeur de pixel fréquente est moins instructive qu'une valeur rare.	$-\sum p(x)logp(x)$

TABLE 1.2 – Descripteurs quantitatifs de l'intensité d'un VOI : définitions et formules.

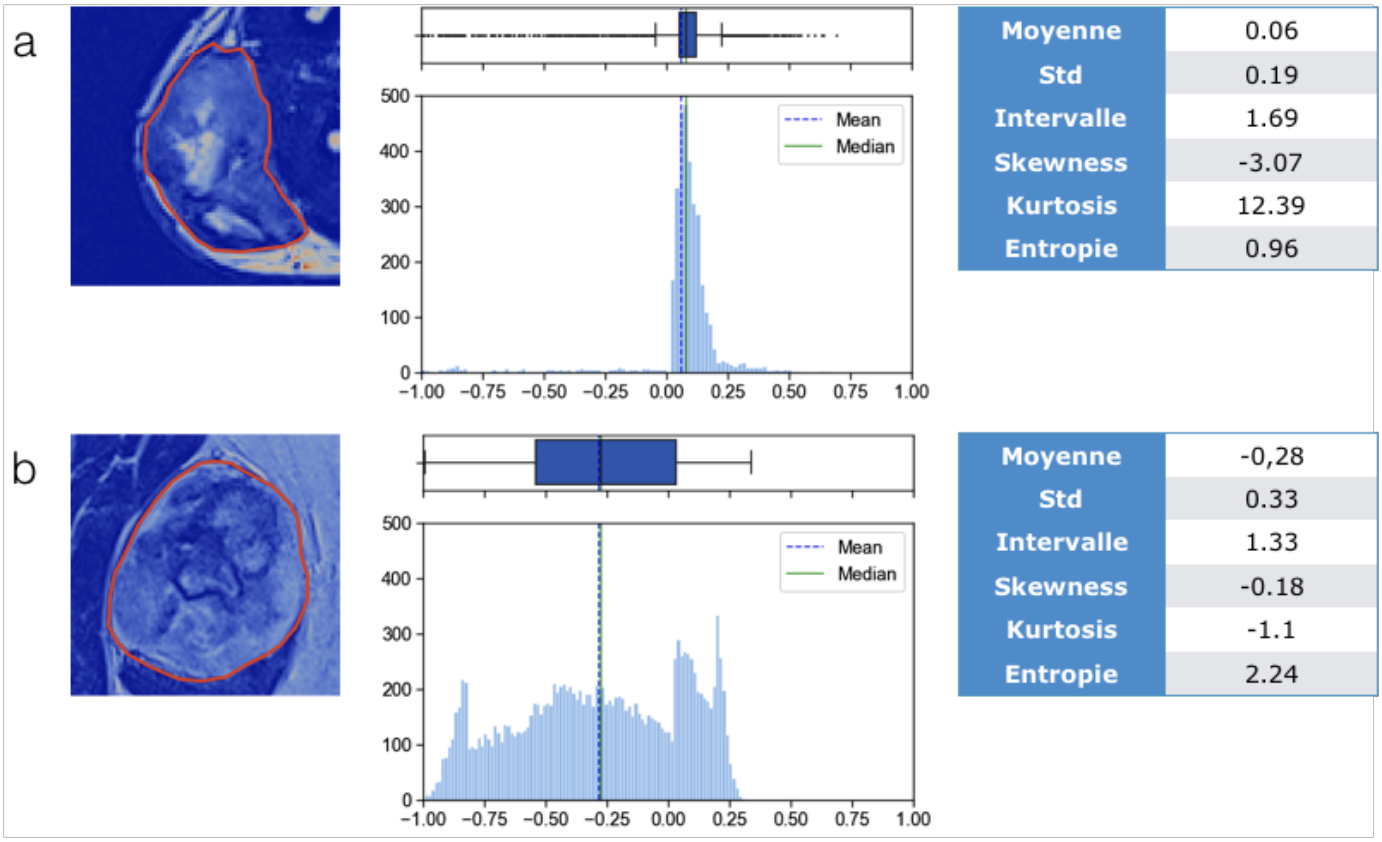


FIGURE 1.7 – Histogrammes de deux sarcomes et mesures associées, calculés sur IRM en T2. La coupe ayant le plus large diamètre est visualisée avec *Paraview* à titre indicatif. *source IRM : Institut Bergonié*

1.4 Descripteurs de la texture

La texture est une information sur les transitions, les variations spatiales d'intensité, comme le montrent les exemples de motifs de la Fig. 1.8).

La quantification de la texture présente dans une région d'intérêt dépend de l'échelle d'observation : l'hétérogénéité peut être décrite par des motifs fins ou plus grossiers. La direction des motifs est également un paramètre d'intérêt pour les tissus biologiques, surtout pour les lésions cancéreuses (arrangées autour d'un centre nécrotique ou le long des vaisseaux, fibres musculaires, os, etc.)

À noter, une texture peut être aussi bien considérée en 2D qu'en 3D.

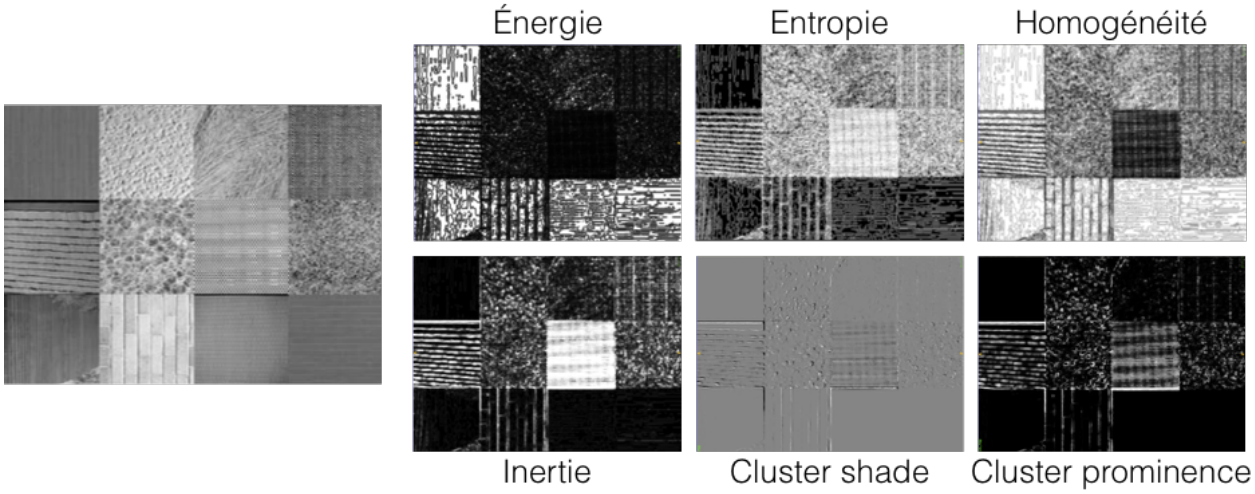


FIGURE 1.8 – Exemple de motifs de texture en 2D et cartes de textures associées.

1.4.1 Matrice de cooccurrence

La matrice de cooccurrence ou GLCM (***G**rey-**L**evel **C**ooccurence **M**atrix*) donne la distribution des différentes combinaisons des niveaux de gris existant dans l'image, pour une direction et une distance donnée.

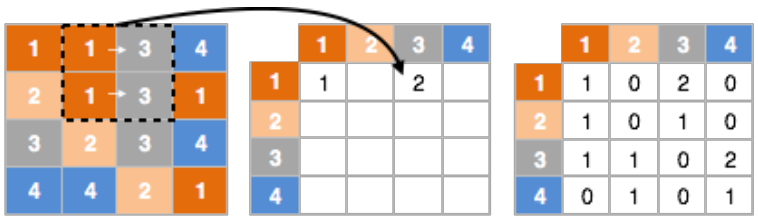


FIGURE 1.9 – Méthode de calcul d'une matrice de cooccurrence pour $d = 1$ et $\theta = 0^\circ$ (voisin direct à droite) à partir d'une image 2D à 4 niveaux de gris (*à gauche*). (*au milieu)* La valeur de la cellule indiquée par la flèche est la réponse à la question "combien de pixels ayant la valeur 3 sont des voisins directs à droite de pixels ayant la valeur 1 ?". (*à droite*) La matrice est remplie en répétant l'opération pour toutes les cellules.

Un exemple de calcul de la matrice est montré Fig. 1.9. Formellement, la GLCM est une matrice carrée de dimension $N \times N$, avec N le nombre de niveaux de gris de l'image (quatre dans l'exemple). Chaque élément de la matrice situé à une position (i, j) correspond à la fréquence d'apparition de l'arrangement d'un pixel de valeur i et d'un pixel de valeur j dans une disposition donnée, composée d'une direction θ et d'une distance d.

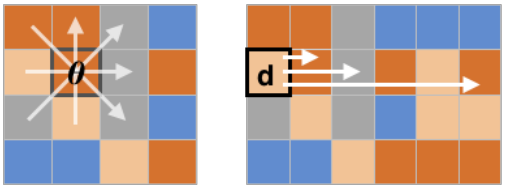


FIGURE 1.10 – Voisinage 2D d'un pixel : (*à gauche*) direction (angle), (*à droite*) distance (nombre de pixels).

Puisqu'à une distance d'un élément, un pixel a 8 voisins directs dans le plan et un voxel a 26 voisins directs dans l'espace, on dénombre un voisinage à respectivement 4 et 13 directions uniques (Fig. 1.10).

Remarques

Si N est grand, les matrices sont elles-même grandes et creuses. Il est préférable de discrétiser l'image au préalable de façon à obtenir un nombre réduit de classes de valeurs pour chaque ligne/colonne. Il faut alors choisir entre fixer ce nombre ou fixer la taille des classes.

Les déplacements le long de la diagonale de l'image sont souvent considérés comme ayant une norme entière au lieu de leur valeur réelle ($\sqrt{2}, 2\sqrt{2}$, ...) ce qui donne une description anisotrope. Cela vaut aussi lorsque les voxels eux-mêmes sont anisotropes. Il est alors possible d'utiliser une distance dans une unité réelle plutôt qu'une distance en nombre de voxels.

Les GLCM sont calculées pour une distance donnée. Les matrices de l'ensemble des directions choisies peuvent être additionnées ou moyennées en une seule qui restera ainsi invariante à la rotation. On remarque toutefois que les 4 et 13 directions proposées représentent un ensemble discret incomplet et peu exhaustif de l'ensemble continu des directions.

Pour une image 3D, il est possible de calculer une matrice par plan (texture 2D) ou une seule matrice pour tout le volume (texture 3D). Dans le cas 2D, les matrices de chaque coupe sont souvent fusionnées comme indiqué précédemment.

Des indicateurs sont ensuite calculés sur la ou les matrices obtenues.

1.4.2 Indicateurs de Haralick

Les matrices de cooccurrence sont des histogrammes en deux dimensions et les indicateurs de textures qui en sont extraites sont des descripteurs de texture de second ordre.

À partir d'une GLCM normalisée, avec $p_{i,j}$ la valeur de la GLCM à la ligne i et la colonne j et n le nombre de classes de la matrice on peut obtenir les caractéristiques de Haralick (d'après [HSD73]). Sont listés Table 1.3 les indicateurs utilisés dans nos études.

Inertie	*Contrast*	Mesure de la périodicité, du gradient. Les variations locales d'intensités sont quantifiées en favorisant les valeurs éloignées de la diagonale.	$\sum_{i=1}^{n}\sum_{j=1}^{n}(i-j)^2 p(i,j)$
Énergie (jointe)	*Angular second moment*	Mesure des motifs uniformes en sommant le carré des valeurs de la GLCM.	$\sum_{i=1}^{n}\sum_{j=1}^{n}(p(i,j))^2$
Entropie (jointe)	*Randomness*	Mesure du désordre dans la GLCM. Diminue quand la texture s'organise.	$-\sum_{i=1}^{n}\sum_{j=1}^{n} p(i,j) log_2(p(i,j)+\epsilon)$
Homogénéité (locale)	*Inverse different moment*	Mesure de l'homogénéité locale, avec des poids qui diminuent à partir de la diagonale.	$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{p(i,j)}{1+\lvert i-j\rvert^2}$
Cluster shade	*teinte de partition*	Mesure d'uniformité, élevée pour les motifs ayant des niveaux de gris peu nombreux mais très représentés.	$\sum_{i=1}^{n}\sum_{j=1}^{n}(i+j-\mu_x-\mu_y)^3 p(i,j)$
Cluster prominence	*dominante de partition*	Mesure d'asymétrie de la GLCM par l'analyse de la présence de pics près de la moyenne.	$\sum_{i=1}^{n}\sum_{j=1}^{n}(i+j-\mu_x-\mu_y)^4 p(i,j)$

TABLE 1.3 – Indicateurs de Haralick avec leurs noms alternatifs, leur définition et la formule associée. En gras, les désignations utilisées dans ce document.

Agrégation des indicateurs

Les indicateurs peuvent être calculés sur une ou plusieurs des matrices fusionnées ou distinctes évoquées précédemment. Ils peuvent ensuite eux-mêmes être moyennés. Notre méthode est basée sur l'implémentation d'ITK à savoir :

— les matrices sont sommées par coupe, ce qui donne une seule matrice pour chaque direction θ en 2D,
— les indicateurs sont calculées sur les quatre matrices résultantes correspondant au quatre directions du plan,
— les indicateurs sont ensuite moyennés.

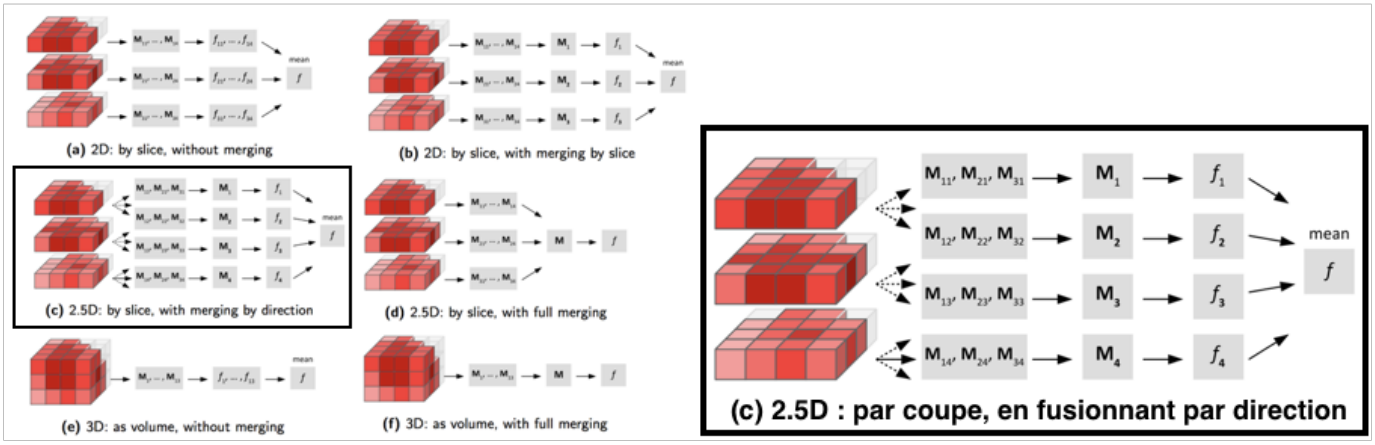


FIGURE 1.11 – Méthodes d'agrégation des caractéristiques calculées sur les matrices de cooccurrence. M_{dk} sont les matrices calculées pour une direction d dans une coupe k et f_{dk} sont les caractéristiques correspondantes. À droite, la méthode choisie. *source : d'après l'*IBSI *v9 [Zwa+16]*

Les différentes méthodes d'agrégation des indicateurs ont été listées par le recueil ***I**mage **B**iomarkers **S**tandardization **I**nitiative* [Zwa+16] et sont visibles Fig. 1.11. Notre approche y est référencée comme étant une méthode "2.5D" (*2.5D :avg, 2.5D :dmrg*).

1.4.3 Autres descripteurs de texture

Les matrices de cooccurrence ne sont qu'un sous-groupe d'opérateurs de texture non linéaires, les matrices d'intensité. D'autres approches sont également populaires.

Les GLRLM, *Gray-Level Run-Length matrices* [Gal74], décomptent les *runs* ou plages homogènes d'une image, c'est à dire la quantité de pixels consécutifs d'une direction donnée partageant le même niveau d'intensité. L'élément $p_{i,j}$ d'une GLRLM donne le nombre de plages homogènes ayant un niveau de gris i et une longueur j (en nombre de pixels). Comme pour les GLCM, les GLRLM sont calculées pour chaque angle θ séparément et la méthode d'agrégation peut être choisie.

Les GLSZM, *Gray Level Size Zone Matrix* [Thi+09], décomptent les zones d'intensité d'une image : le nombre de voxels connectés partageant le même niveau d'intensité. Deux pixels/voxels sont considérés connectés s'ils partagent une arête/face (leur distance en norme

infinie vaut 1). Une zone comprend respectivement 8 pixels ou 26 voxels. L'élément $p_{i,j}$ d'une GLSZM donne le nombre de zones ayant un niveau de gris i et une taille j (en nombre de pixels). Une seule matrice est calculée pour toutes les directions car elle ne dépend pas de la rotation.

La texture peut également être décrite par les motifs binaires locaux [OPM02], les propriétés des fractales [AKW08] ou les approches basées ondelettes [Dep+19], etc.

1.4.4 Cartes de textures

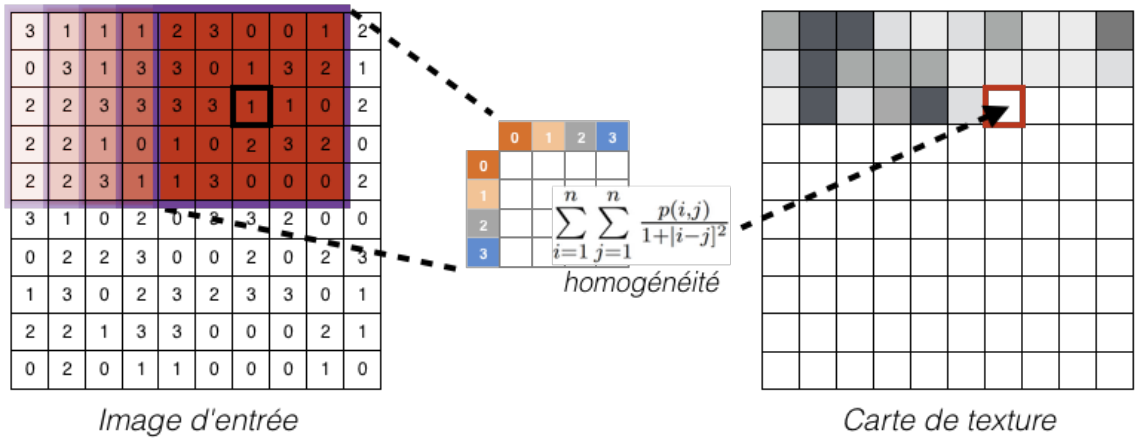


FIGURE 1.12 – Création d'une carte de texture d'une image à quatre niveaux de gris. Le kernel est un carré de 5 pixels de côté, qui glisse d'un pas de 1 pixel. À chaque pas, la GLCM est calculée pour les pixels couverts par la fenêtre. L'homogénéité obtenue est attribuée au pixel de la carte correspondant au pixel central du kernel.

Les caractéristiques de texture décrites précédemment sont calculées sur l'ensemble du VOI. Une description statistique de la texture à un niveau local est également possible. Le processus est résumé Fig. 1.12. Il suffit de définir un kernel dont les dimensions sont fixées et inférieures à celles du VOI. Ce masque cubique ou sphérique est centré sur un voxel et recouvre un voisinage de taille définie. On peut donc calculer un histogramme ou des matrices sur le sous-ensemble de voxels ainsi déterminé. Le voxel du centre se voit attribuer un vecteur contenant les valeurs des descripteurs à son emplacement.

Si on déplace le masque comme une fenêtre glissante, on obtient la distribution des indicateurs radiomiques sur l'ensemble de l'image ou du VOI : ce sont les cartes de texture. Une série d'exemples de cartes d'une tumeur du pancréas est donnée en Fig. 1.13. On voit que les clusters shade et cluster prominence montrent des motifs similaires, à des échelles de valeurs différentes. L'énergie et l'entropie, comme l'homogénéité et l'inertie, sont globalement inversement corrélées.

Chacun des voxels d'une carte en niveaux de gris représente la valeur de la caractéristique de texture à ce point pour un kernel donné. Plus cette fenêtre est petite, plus la description obtenue est locale. L'enjeu consiste donc à choisir :

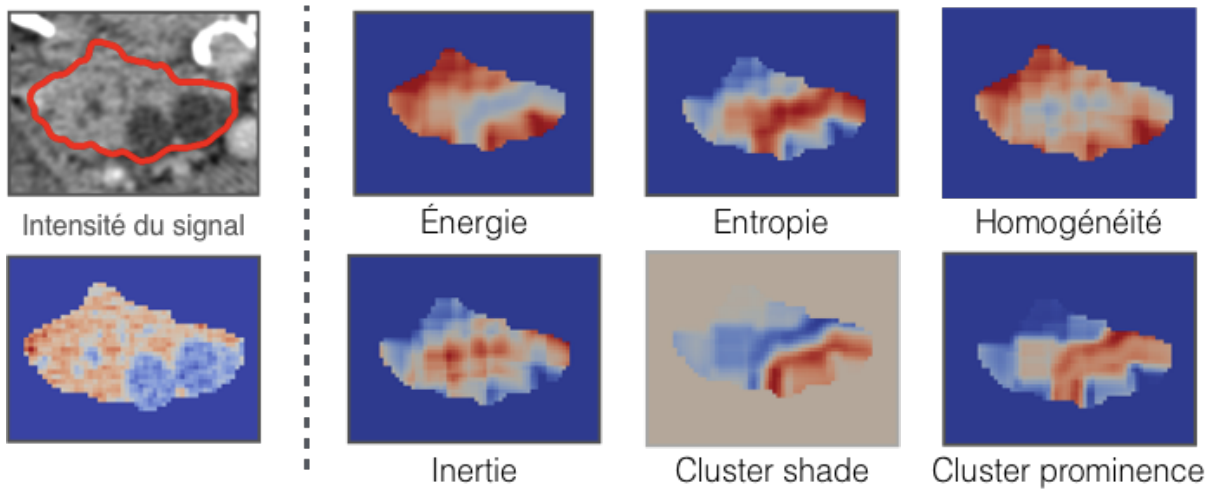


FIGURE 1.13 – Carte de textures GLCM d'une coupe de tumeur du pancréas, (voisinage à 5px, contraste accentué).

— les dimensions de la fenêtre, correspondant à la taille du voisinage et donc à la finesse de la résolution des indicateurs de texture calculés,
— l'amplitude du pas de déplacement de la fenêtre dans chaque dimension.

Le résultat donne une image adaptée à une évaluation visuelle qualitative de la texture locale, pratique pour repérer des compartiments dans une lésion par exemple. Les cartes de texture fournissent surtout des informations quantitatives supplémentaires pour chaque voxel, utilisables également en entrée d'un algorithme d'apprentissage.

1.5 Caractéristiques delta-radiomiques

L'analyse des biomarqueurs radiomiques est classiquement réalisée à un instant t donné, par exemple celui de l'examen de diagnostic. Cependant l'analyse de la variation des indicateurs au cours du temps peut donner une information sur le traitement et sur l'évolution de la maladie d'un patient. Elle complète voire se substitue aux indicateurs ponctuels. Ce type d'analyse longitudinale des descripteurs est qualifié de **delta-radiomique**.

L'évolution d'un marqueur X peut être calculée en terme de :

— différence absolue : $X_{t_1} - X_{t_0}$
— différence relative nette (ratio, pourcentage) * : $\frac{X_{t_1} - X_{t_0}}{X_{t_0}} (\times 100)$
— différence normalisée (dérivée en temps) : $\frac{X_{t_1} - X_{t_0}}{t_1 - t_0}$

Certaines études recommandent l'utilisation d'une différence normalisée lorsque le temps écoulé entre les deux examens étudiés varie fortement d'un patient à l'autre [Fav+17]. On notera toutefois que la notion de variation "forte" diverge entre les études. Cherezov et al. [Che+18] choisissent ainsi de garder un delta absolu car le temps entre les examens est considéré "régulier", avec un écart interquartile de 40 jours pour une moyenne de 375 jours entre les examens.

1.6 Étude de l'invariance des variables morphologiques à la reconstruction

1.6.1 Motivations

Une mesure est qualifiée d'invariante lorsque toute transformation ou opération sur la donnée d'entrée ne modifie pas sa réponse. Elle est alors dite insensible. Si la réponse est proportionnelle à l'amplitude de la transformation, elle est dite équivariante [DAKM17].

L'invariance à la rotation de l'image des caractéristiques radiomiques est un exemple de paramètre à étudier dans le contexte bio-médical car les tissus biologiques prennent des orientations variables. Les lésions cancéreuses ont notamment tendance à se développer le long des vaisseaux sanguins qui se présentent eux-même selon des directions diverses. Il est donc souhaitable de favoriser les caractéristiques insensibles aux rotations 2D comme 3D.

Lors de la reconstruction des volumes, le positionnement de la grille de discrétisation sur les objets influe sur l'aspect du VOI obtenu (voir exemple 2D Fig. 1.14). La résolution et l'orientation de la grille vont générer un crénelage qui peut biaiser certaines mesures de morphologie comme la surface extérieure. De même, des décalages de l'objet de moins d'un voxel sont susceptibles de fausser le volume final, d'autant plus si le VOI est de petite taille.

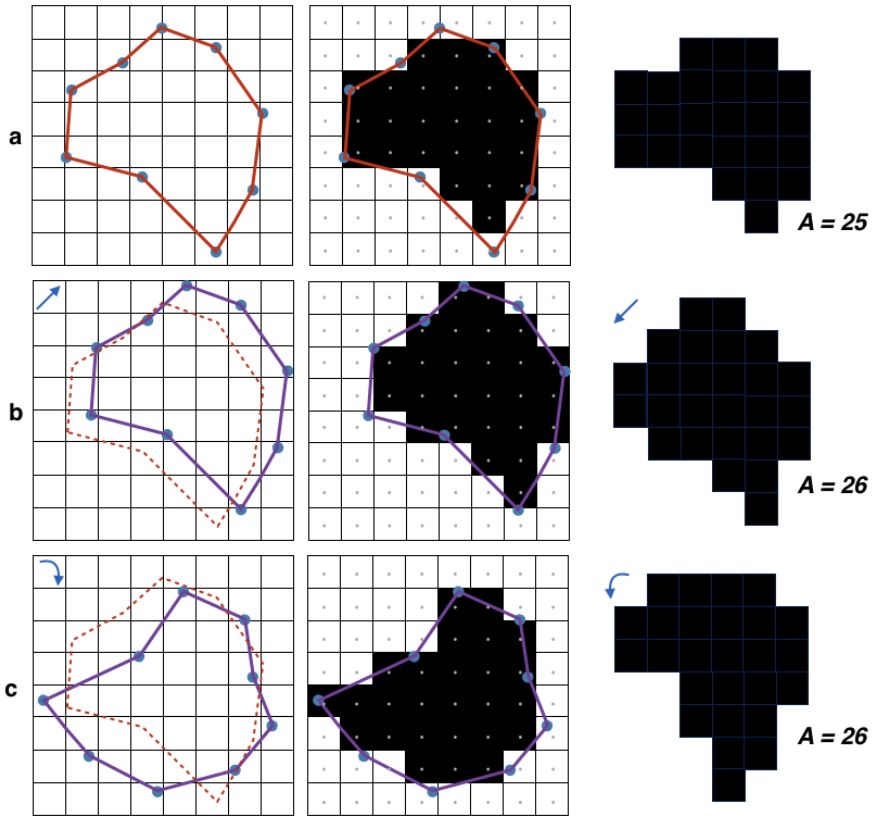


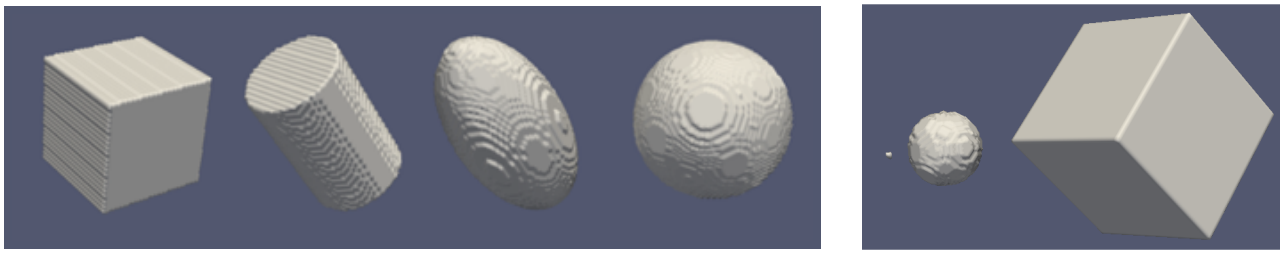
FIGURE 1.14 – (a) Contours et ROI, (b) reconstruite après translation de moins d'un voxel ou (c) rotation de la grille de reconstruction. Pour un pixel d'aire unitaire, les aires (A) des surfaces obtenues sont indiquées.

1.6.2 Matériel et méthode

Afin de mesurer l'ampleur de ce phénomène, nous analysons des jeux de données synthétiques de façon à reproduire les altérations mentionnées pour des objets géométriques (pleins) de morphologie connue (Fig. 1.15a). Ces ensembles de volumes ont été générés par D. Legland [LL12] pour évaluer l'efficacité de l'algorithme de Crofton pour l'estimation de la surface (voir section 1.2). Leurs dimensions sont rassemblées Table 1.4.

	Cube	Cylindre	Sphéroïde (prolate)	Tore	Sphère
Dimensions	c : 50	h : 60 r : 20	R : 40 r : 20	a : 30 b : 10	r ∈]0.5,12[

TABLE 1.4 – Dimensions des objets générés (unité arbitraire).



(a) (b)

FIGURE 1.15 – Visualisés sous *Paraview*, (a) les différents volumes générés (échelles arbitraires), (b) plus petite sphère, plus grande sphère et cube (échelles relatives).

Nous les utilisons pour comparer les différentes mesures de morphologie listées en section 1.2 avec les valeurs réelles théoriques de chaque objet. Les pourcentages d'erreur sont fournis. En outre, les caractéristiques calculées sont comparées à leur valeur théorique par un test de Student pour un échantillon unique ou un test de rang signé de Wilcoxon [4]. Une *p-value* < 0.05 est considérée significative.

1.6.3 Position de la grille de reconstruction

Nous utilisons un premier jeu de données pour évaluer l'effet du décalage de la grille et celui de la taille de l'objet. Il contient dix ensembles de 100 sphères pleines (boules) de rayon croissant, décalées dans l'espace d'un delta inférieur à 1 voxel choisi aléatoirement en x, y et z.

Résultats

Les pourcentages d'erreurs moyens sont visibles Fig. 1.16. Quel que soit le rayon de la sphère, seules les valeurs obtenues pour le volume et le rayon équivalent ne sont pas considérées comme étant significativement différentes de leur valeur théorique. L'impact de la reconstruction discrète et du décalage de la grille est donc réel sur une grande partie des caractéristiques calculées, même si en moyenne ces erreurs tendent à se compenser.

4. En fonction du résultat au test de normalité de la distribution de Shapiro-Wilk.

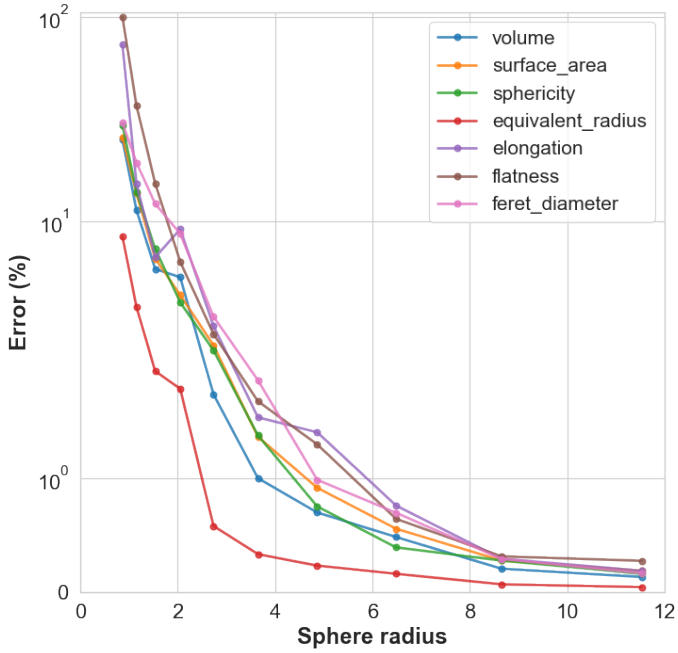


FIGURE 1.16 – Pourcentage d'erreur moyen des paramètres morphologiques mesurés sur les sphères reconstruire avec un décalage, en fonction de leur rayon réel (échelle symlog).

Pour l'ensemble des mesures, l'erreur de calcul moyenne issue des décalages diminue avec le rayon. Elle peut dépasser les 30% (sphéricité) voire 100% (flatness) pour les plus petits objets (volume < 5). Mais elle passe sous les 5% d'erreur pour les sphères de plus de 85 unités de volume. L'erreur est inférieure à 1% pour les objets de plus de 480 unités (rayon ≈ 5), hors élongation et flatness, pour lesquels il faut un peu moins du double (rayon ≈ 6).

Puisque les erreurs sont marginales au delà d'un certain volume qui est lui même réduit, nous pouvons considérer l'ensemble des caractéristiques morphologiques étudiées robustes au décalage de grille. On peut avancer que veiller à ce que les ROI sélectionnées dépassent ce volume suffit à rendre cette erreur négligeable.

1.6.4 Rotation de la grille de reconstruction

Le jeu de données utilisé pour étudier la robustesse des caractéristiques morphologiques à la rotation comprend des cubes, cylindres et sphéroïdes allongées en 135 exemplaires chacun, dont l'orientation 3D varie aléatoirement. Leur taille est fixe et choisie de façon à être suffisamment grande pour rendre négligeable l'erreur à la reconstruction observée précédemment avec les sphères (voir Fig. 1.15b).

Résultats

Au test de Wilcoxon/Student à un échantillon, le volume est la seule mesure pour laquelle

l'hypothèse de similarité entre médiane/moyenne et valeur théorique n'est jamais rejetée. Pour la sphéricité et le rayon équivalent, liés tous deux au volume, l'hypothèse n'est pas non plus rejetée dans le cas du cylindre et du sphéroïde. Ici encore, la discrétisation du volume sur des grilles de différentes orientation n'est pas sans effet sur les mesures de morphologie.

La Fig. 1.17 indique que les taux d'erreur ne dépassent pas 10% dans le pire des cas (9.17% d'erreur max pour le cube). Les erreurs moyennes, rassemblées Table 1.5, sont strictement inférieures à 3% pour le cube, à 1% pour le cylindre et à 0.5% pour le sphéroïde. L'aire de la surface et la sphéricité sont les caractéristiques les plus impactées par les variations d'orientation de l'objet. Le rayon équivalent et l'élongation sont les plus stables et ne dépassent jamais 1% d'erreur.

	Cube	Cylindre	Sphéroïde
volume	0.09 *(1.05)*	0.04 *(0.23)*	0.03 *(0.10)*
aire	2.79 *(8.40)*	0.93 *(4.07)*	0.06 *(0.23)*
sphéricité	2.88 *(9.17)*	0.91 *(4.23)*	0.05 *(0.19)*
rayon équivalent	0.03 *(0.35)*	0.01 *(0.08)*	<0.01 (0.03)
élongation	0.06 *(0.39)*	0.05 *(0.44)*	0.05 *(0.12)*
flatness	0.10 *(1.06)*	0.09 *(0.76)*	0.07 *(0.15)*
⌀de feret	1.67 *(2.55)*	0.42 *(1.21)*	0.44 *(1.02)*

TABLE 1.5 – Pourcentages d'erreurs moyens *(et max)* de calcul des paramètres de morphologie des objets construits à différentes rotations.

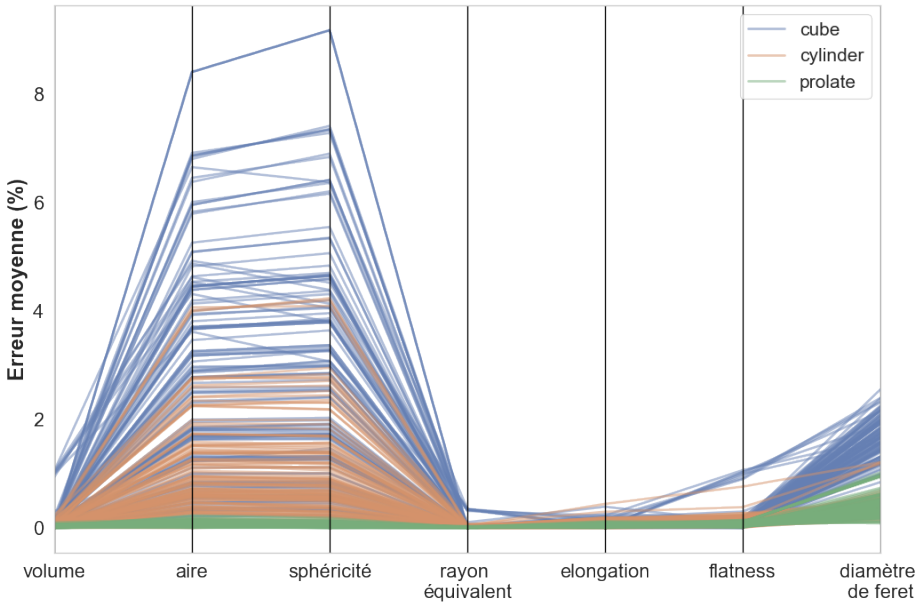


FIGURE 1.17 – Pourcentage d'erreur de calcul des paramètres de morphologie de l'ensemble des **cubes**, **cylindres** et **sphéroïdes** d'orientation aléatoire.

1.7 Discussion

Certains indicateurs ont été écartés de nos études car immédiatement dérivables (et donc hautement corrélés) aux autres. Ainsi la sphéricité (*sphericity*) est directement liée aux critères de compacité (*compactness*), d'asphéricité, de disproportion sphérique (*spherical disproportion*) et sera la seule mesure conservée de ce groupe. Les indicateurs dérivés du calcul des différentes boîtes englobantes de l'objet sont aussi redondants avec le diamètre de Feret.

Pour l'étude de l'intensité, nous avons également mis de côté les descripteurs superflus de l'histogramme : min et max (compris dans l'intervalle), percentiles (uniquement représentés par la médiane). Certains comme l'énergie sont directement proportionnels au volume. Le principe est le même pour la GLCM : un grand nombre des mesures possibles sont corrélées. Les caractéristiques que nous choisissons de calculer sont celles déterminées par Conners, Trivedi et Harlow [CTH84] comme étant discriminantes sans être redondantes.

Nous n'avons pas inclu d'autres types de caractéristiques de texture. Elles sont pourtant nombreuses à être décrites et utilisées dans la littérature [Zwa+16]. Nous avons préféré concentrer nos recherches sur un sous-ensemble restreint de caractéristiques relativement peu complexes et bien documentées. Nous facilitons ainsi l'interprétation biologique de nos résultats et il nous est possible de les comparer avec une majorité des études, comme celles listées par Yip et Aerts [YA16].

Nous invitons le lecteur intéressé à se tourner vers les bibliothèques et logiciels *pyradiomics* [Gri+17], *LIFEx* [Nio+18], *MaZda* [Szc+08] pour une implémentation et une documentation de ces autres paramètres.

La méthode de discrétisation de l'histogramme pour le calcul des matrices de texture est sujette à débat. Plusieurs études préfèrent fixer le nombre de classes de l'histogramme plutôt que la taille des classes. Des études sur les TEP ont a l'inverse montré que les caractéristiques sont plus reproductibles avec une taille fixe [Lei+15]. De cette façon, une variation d'un seul niveau discrétisé a la même signification quelle que soit la plage initiale de niveaux de gris.

L'initiative de standardisation des biomarqueurs radiologique [Zwa+16] rapporte qu'un nombre de classes fixe est conseillé pour les modalités d'image où l'intensité des pixels est relative et n'a pas de signification équivalente d'un examen sur l'autre. C'est le cas de l'intensité du signal des IRM. Cependant, si les niveaux de gris des IRM sont normalisés (donc rendus comparables), il est toujours recommandé de calculer des classes de taille fixe [com16]. Récemment, plusieurs études dédiées à l'IRM ont effectivement obtenu des mesures plus stables avec une taille fixe [GO+18 ; Dur+19]. De fait, nous utilisons donc également la méthode de discrétisation à classes de taille fixe lorsque les images sont normalisées.

Reste le problème de la valeur à choisir pour la taille des classes. La documentation de *pyradiomics* rapporte qu'une valeur donnant un nombre de classes entre 30 et 150 permet généralement d'éviter d'obtenir une matrice creuse et produit des caractéristiques reproductibles et performantes [Tix+11]. Nous avons suivi ce principe en obtenant 30 à 50 classes sur les VOI manipulés dans les analyses que nous présentons dans les chapitres suivants.

Dans notre étude de l'invariance à la reconstruction, les différences entre mesures de forme s'avèrent significatives selon les tests statistiques. Cependant, les erreurs restent en pratique

d'ordre de grandeur relativement faible.

En outre, les volumes théoriques discrétisés pour l'occasion ne sont pas tous représentatifs des VOI obtenus en oncologie. Le cube, qui obtient les pires résultats, est également l'objet le moins susceptible d'être retrouvé en pratique, alors que le sphéroïde, beaucoup plus réaliste pour décrire une tumeur (cf. Fig. 1.6), s'en sort le mieux.

Les résultats sur le jeu de données de sphères de tailles variables permettent de confirmer une hypothèse. La translation dans l'espace de moins d'un voxel des points de contours d'un VOI induit des différences plus ou moins nettes de l'aspect de l'objet reconstruit. Pourtant, plus il est de taille importante, plus les modifications de son aspect à la reconstruction sont négligeables au regard du volume total. Les erreurs de mesure de la morphologie d'un objet seront donc proportionnellement plus conséquentes à mesure de la diminution de son volume.

Plus concrètement, on distingue une chute plus significative (plus de 5 %) de la fiabilité de la majorité des mesures pour les sphères de volume inférieur à 85 voxels unitaires. Aussi, nous tachons dans nos études d'exclure les VOI de volume inférieur à cette valeur pour assurer la robustesse des caractéristiques extraites.

L'étude méthodologique proposée se concentre sur les descripteurs radiomiques morphologiques, classiquement moins étudiés que ceux de texture ou d'intensité [Orl15 ; Lei+15 ; DAKM17 ; Dur+19]. Elle peut toutefois leur être étendue, à condition d'utiliser un véritable jeu de données IRM. Là aussi, peu d'études se concentrent sur l'invariance des descripteurs radiomiques à l'IRM. Dans leur analyse comparative de 2018, Traverso et al. n'en recensent qu'une seule respectant les critères d'éligibilité [Tra+18]. Elle concerne uniquement les entropies de l'histogramme et de la matrice de cooccurrence [Gua+16].

Comme dans l'étude [Bol+18], on peut imaginer translater plusieurs fois l'ensemble des points de contours des VOI sur une distance inférieure à un voxel choisi aléatoirement, puis comparer les textures et histogrammes ainsi obtenus. On s'attend à ce que la corrélation entre les différentes grilles de reconstruction reste très forte. Cette méthode peut d'ailleurs être employée pour calculer des descripteurs moyens, plus robustes.

Conclusion

La discrétisation des images, le contourage des VOI, leur reconstruction et les mesures sont autant de variables sources d'approximations à prendre en compte et à minimiser, que ce soit par les choix d'implémentation (agrégation des GLCM, discrétisation des examens, etc.), ou le design de nos études (volume minimum des VOI, nombre et choix des descripteurs calculés, etc.)

Après avoir défini les outils méthodologiques nécessaires pour reconstruire et caractériser un VOI de façon fiable sur un examen, nous souhaitons intégrer les informations de plusieurs images pour décrire et analyser une cohorte d'observation et répondre à une problématique clinique. Une transformation des valeurs et taille de voxels des examens est alors nécessaire pour que les lésions et autres régions d'intérêt soient comparables d'une image à l'autre. En gardant en tête les spécificités de l'IRM, nous allons donc étudier les contraintes apportées par ces traitements sur les caractéristiques radiomiques.

Chapitre 2

Pré-traiter les IRM pour des caractéristiques fiables et robustes

Sommaire

Les valeurs de pixels d'une image peuvent être normalisées comme c'est le cas pour les CT avec les unités Hounsfield ou la valeur de fixation normalisée (***S**tandard **U**ptake **V**alue*) pour les TEP. Ce n'est pas le cas de l'IRM.

Le traitement des IRM a pour but de :
- corriger les défauts de l'image, qu'ils soient propres à la technique d'imagerie en elle-même ou à un examen précis,
- normaliser pour assurer une interprétation invariante des caractéristiques d'une image à l'autre,
- faire ressortir les caractéristiques de l'intensité.

Il est nécessaire de jouer aussi bien sur la valeur des voxels que sur leur organisation dans l'espace (dimensions, éventuellement origine et orientation si un recalage est prévu). On distingue également les traitements effectués indépendamment sur chaque IRM des traitements visant à en harmoniser plusieurs entre eux.

2.1 Correction des artefacts

L'IRM, comme les autres techniques d'imagerie, est concernée par la survenue éventuelle dans l'image de faux motifs n'ayant pas de réalité anatomique : les artefacts. Ils peuvent être dus aux mouvements du patient (le plus courant, l'acquisition de l'IRM étant relativement lente), au champ magnétique ou encore à l'échantillonnage du signal. L'utilisation de séquences rapides permet de limiter l'impact du mouvement du patient, au détriment de la résolution spatiale de l'image et du rapport signal-bruit.

2.1.1 Bruit

Le bruit en IRM provient principalement de l'agitation thermique des protons dans le corps du patient, à l'origine d'émissions parasites, et dans une moindre mesure du système de mesure (bruit "électronique"). il est supposé aléatoire et ricien [1]. Le rapport signal-bruit est égal au ratio entre l'intensité moyenne du signal et l'écart-type du bruit mesuré dans l'arrière-plan. Il est plus favorable au signal quand la taille des voxels augmente.

Une étude de Yang et al. [Yan+18] analyse l'impact de l'ajout d'un bruit gaussien à des fantômes en IRM sur les valeurs des caractéristiques radiomiques. Elle suggère au final que les paramètres d'acquisition et de reconstruction IRM ont une influence proportionnellement plus grande. D'autre part, Diaz et al. [Dia+11] suggèrent que les algorithmes classiques de réduction du bruit (filtre gaussien, ondelettes, algorithmes de diffusion anisotropes etc.) floutent les limites des tissus, voire pour certains introduisent des artefacts.

Dans le cadre de nos travaux, les experts radiologues des différentes études réalisées ont considéré le bruit des examens utilisés comme étant faible à négligeable. Pour toutes ces raisons, nous choisissons de ne pas intégrer de méthode de correction du bruit à notre chaîne de traitement des IRM.

2.1.2 Biais

L'imagerie RMN est également caractérisée par l'apparition éventuelle d'un phénomène d'inhomogénéité en intensité à basse fréquence. L'image subit une variation spatiale lisse : un même tissu peut alors avoir des intensités différentes suivant sa localisation dans l'image. Ce phénomène est appelé dérive d'intensité dans l'espace, ou plus simplement, **biais**. Il est principalement dû au scanner IRM dont le champ magnétique diminue à mesure que l'on s'éloigne du centre de l'aimant, ainsi qu'à l'hétérogénéité des antennes de réception. Il est tout spécialement visible sur les vieux systèmes d'imagerie [Jun+05]. Également lié à la

1. Le bruit électronique suit une loi gaussienne [BW61]. L'IRM acquiert un nombre complexe avec une partie réelle et une partie imaginaire et son bruit suit donc une loi de Rice [HP96].

puissance du champ magnétique, il sera plus fort dans le cas des machines à haut champ (7T et plus).

Si le biais a peu de chances d'influencer l'interprétation qualitative des résultats, il peut en revanche diminuer les performances des algorithmes de recalage ou de segmentation automatique qui y sont très sensibles [Jun+05]. Aussi, de nombreuses méthodes de correction du biais en intensité ont été développées, comme les mélanges de gaussiennes, le filtrage des basses fréquences ou le calcul de gradients dans des zones homogènes [SZH17 ; RP14].

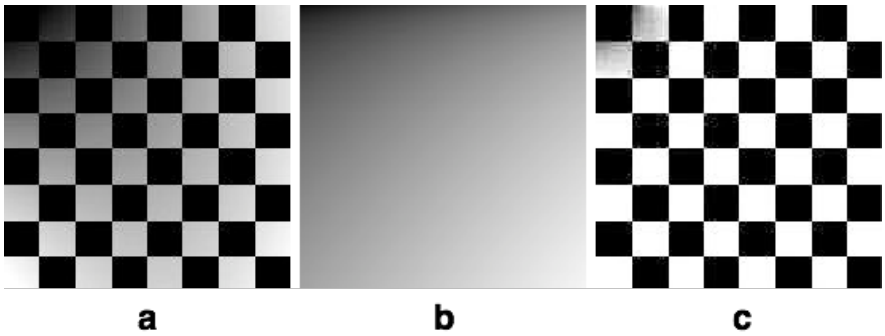


FIGURE 2.1 – (a) Image de damier corrompue par un biais d'intensité. (b) Surface représentant le champ de dérive d'intensité. (c) Résultat de la division de l'image *(a)* par l'image *(b)*. *source : Juntu et al. [Jun+05]*

L'impact b du biais est multiplicatif (voir Fig. 2.1) et peut être modélisé par l'équation 2.1, avec y et x les valeurs observées et réelles au voxel i, et n le bruit additif en i (peut éventuellement être négligé).

$$y_i = b_i x_i + n_i \tag{2.1}$$

Dans le cadre de cette thèse, nous nous intéressons à une méthode populaire basée sur l'analyse de l'histogramme, la méthode N3 (***N**onparametric **N**onuniformity **N**ormalization* de Styner et al. [SZE02]). Elle présente l'avantage d'être automatisable car elle ne requiert que peu de paramètres. Elle est également libre de droit (sources proposées par le McConnell Brain Imaging Centre, Montreal Nuerological Institute, McGill). Arnold et al. [BA+01] ont comparé six algorithmes dont le N3 sur des images fantômes et réelles provenant de systèmes à 1.5T et 3T. Le N3 a montré les performances les plus stables sur tous les scénarii étudiés.

Dans la méthode N3, le biais est assimilé à un filtre passe-bas [2]. Il est approché par une distribution gaussienne dont la moyenne est nulle et la variance connue. L'idée derrière le N3 est que l'histogramme de l'IRM est une version floue de l'histogramme réel, obtenue après convolution avec l'histogramme de b. L'algorithme inverse ce processus en appliquant une déconvolution de Wiener et en extrait un champ de polarisation lisse. L'opération est en outre plus efficace lorsqu'elle procède par itérations.

Dans le cadre de ces travaux, nous avons utilisé l'implémentation dite *N4*, proposée par la bibliothèque ITK. Il s'agit d'une version améliorée introduisant entre autre la possibilité de corriger le biais à plusieurs résolutions.

2. Qui supprime les fréquences hautes de l'image.

Nous extrayons le biais sur l'image entière. Dans la mesure où l'algorithme utilise le logarithme de l'image pour rendre le biais additif et simplifier la correction, des images sans valeurs négatives sont requises. Nous réalisons donc un décalage préalable adéquat des valeurs d'intensités. Les paramètres par défaut de l'algorithme sont majoritairement conservés à l'exception du nombre de niveaux d'ajustement, fixé empiriquement à 3 au lieu de 1. Nous divisons le nombre d'itérations par défaut, 50, entre ces trois niveaux, l'algorithme étant gourmand en temps de calcul.

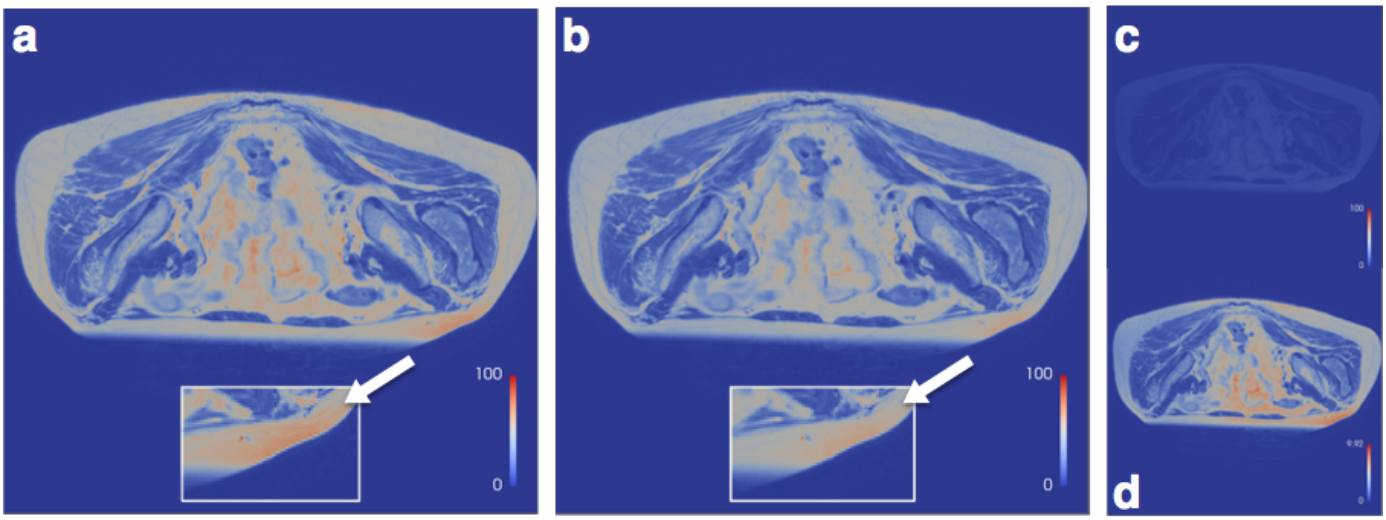


FIGURE 2.2 – (a) IRM en T2 du bassin, après ré-échelonnage entre 0-100 pour supprimer les valeurs négatives. (b) Après correction du biais, on perçoit visuellement de légers changements. (c) Carte des erreurs absolues. (d) Carte des erreurs, contraste renforcé pour la visualisation. La dérive spatiale progressive est bien visible de gauche à droite.
visualisation : Paraview

Un exemple de résultat de la correction sur une IRM du bassin est visible Fig. 2.2. Le zoom sur les sous-images (a) et (b) montre une zone d'intensité clairement atténuée. Ailleurs, même si la correction est assez difficile à voir à l'œil nu sur la différence réelle (c), un gradient est clairement visible de gauche à droite de la sous-image au contraste amélioré (d). La gamme de niveaux de gris des pixels de différence représente 9% de la gamme de l'image avant correction ce qui est n'est pas négligeable et confirme le besoin d'une correction.

2.2 Normalisation de la taille des voxels

Une sensible variation dans la taille des pixels (en x et y) et l'épaisseur des coupes (z) est clairement attendue, puisque les jeux de données sont acquis sur des systèmes divers, par des opérateurs multiples et à des moments différents. Cette variabilité inter-patients (voire inter-examens, chez un même patient) porte préjudice à l'interprétabilité des caractéristiques extraites ainsi qu'à la reproducibilité de leur étude statistique. Il faut donc normaliser les dimensions inter-patients pour rendre les comparaisons pertinentes entre les résultats. Vallières et al. [Val+15] ont cependant observé que le choix des dimensions auxquelles on adapte

les voxels des IRM ou des TEP est le paramètre ayant le plus grand impact sur la valeur prédictive d'un modèle radiomique de rechute métastatique des sarcomes.

D'autre part, au sein d'une même image, les voxels sont généralement anisotropes : si les résolutions en x et y sont quasi-systématiquement identiques, la résolution en z diffère régulièrement des deux autres. Le calcul de la texture, qui prend en compte le voisinage des voxels, peut donc être impacté. De plus Shafiq et al. [SUH+18] montrent que l'extraction de caractéristiques de texture et de forme bénéficient grandement de résolutions unitaires (voxels cubiques, pour faciliter les calculs) et égales (pour des analyses 3D non biaisées).

Nous cherchons donc à avoir des grilles comparables d'un examen à l'autre et des voxels isotropes. Ce processus est le ré-échantillonnage.

Principe du ré-échantillonnage

Le ré-échantillonnage spatial de la taille des voxels consiste à transformer une image en changeant le nombre de voxels dans une direction donnée tout en maintenant sa taille physique (= changement de résolution). On parle de sur-échantillonnage et de sous-échantillonnage pour une image dont on augmente ou diminue la résolution respectivement.

Les intensités des voxels de l'image rééchantillonnée sont déterminées par une fonction d'interpolation choisie en fonction du type de valeur à ré-échantillonner et du compromis entre le temps de calcul et la qualité souhaitée [APKET83].

Les voxels en niveaux de gris peuvent ainsi être interpolés par des fonctions de complexité croissante :

- interpolation linéaire : le voxel de l'image de sortie utilise une moyenne pondérée des centres des voxels les plus proches dans l'image d'entrée. En 3D, on parle d'interpolation trilinéaire. Si le voisinage est considéré dans le plan seulement, l'interpolation est bilinéaire et ne concerne que les 4 plus proches voisins,
- interpolation cubique : tient compte des centres des 16 pixels les plus proches dans le plan (bicubique) ou des 64 voxels dans l'espace (tricubique),
- interpolation polynomiale de degré supérieur.

Les masques peuvent simplement être réchantillonnés en utilisant la valeur du plus proche voisin pour conserver le caractère binaire des voxels. Ce type d'interpolation peut dans le pire des cas décaler l'image jusqu'à un demi-voxel [APKET83]. Il est également possible d'utiliser un interpolateur linéaire ou cubique sur un masque à condition que l'image soit ensuite seuillée pour redevenir binaire. L'*IBSI* recommande l'une ou l'autre méthode au choix [Zwa+16].

La Fig. 2.3 montre que le choix du seuil a un impact sur la forme et les dimensions de l'objet final. L'aire est utilisée pour comparer les ROI reconstruites après interpolation linéaire puis seuillage. Le sous-échantillonnage, grossier, provoque de plus grands changements des résultats (amplitude d'aire potentielle de 4). On constate également que la forme et la position de la ROI changent dans le cas (d) sur-échantillonné.

La méthode du voisin le plus proche peut aussi être employée sur les images en niveaux de gris, au risque de donner un aspect visuel rugueux au résultat. L'homogénéité à échelle très locale peut alors être impactée.

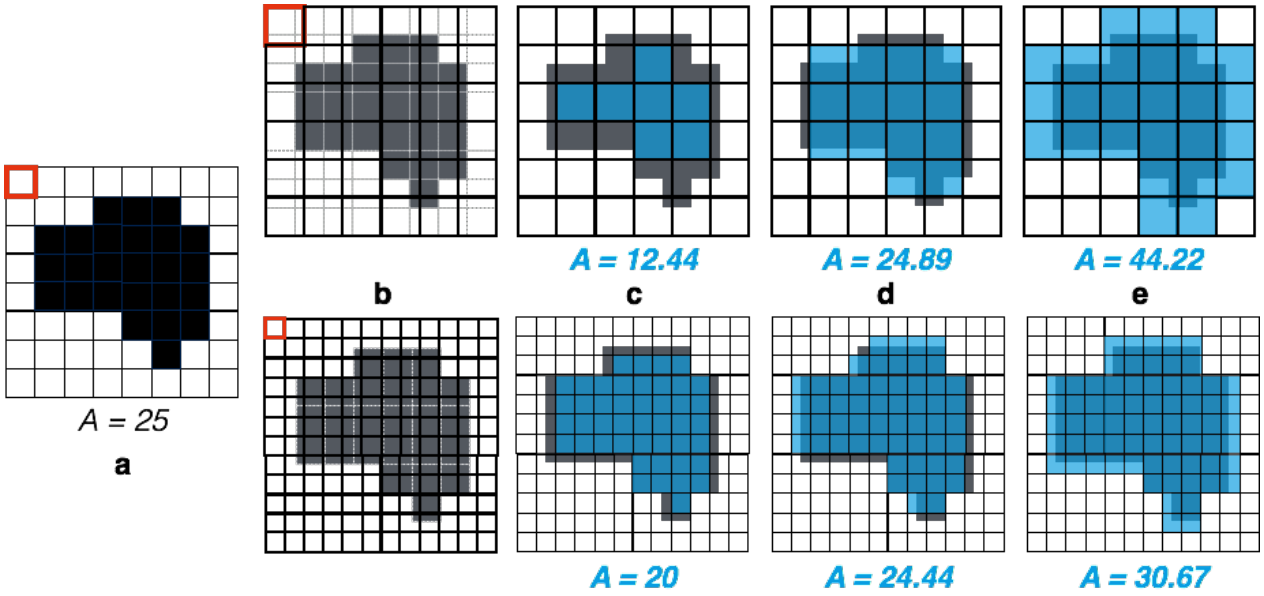


FIGURE 2.3 – Résultat du ré-échantillonnage de la taille des pixels d'un masque 2D et aire **A** associée. La nouvelle grille **(b)** est sous-échantillonnée (en haut, taille du pixel $\times\frac{4}{3}$) ou sur-échantillonnée (en bas, taille du pixel $\times\frac{2}{3}$) par rapport aux pixels de taille initiale unitaire **(a)**. Les valeur de la ROI initiale (en gris) sont interpolées linéairement puis seuillées avec une tolérance minimale **(c)**, intermédiaire (0.5) **(d)** et maximale **(e)** pour donner une nouvelle ROI (en bleu).

Les tests empiriques réalisés sur nos jeux de données n'ont pas montré d'intérêt significatif d'une complexification de la fonction d'interpolation au delà d'une fonction cubique. Nous utilisons donc les interpolations par plus proches voisins, trilinéaire et tricubique, que nous nommerons respectivement *nearest*, *linear* et *cubic* par la suite.

2.3 Normalisation de la valeur des voxels

La normalisation des intensités est préconisée pour les modalités d'images dont la valeur des niveaux de gris n'a pas de lien avec une quantité physique absolue, comme c'est le cas des IRM, exprimées dans une unité arbitraire (cf section 2.1). L'intensité des voxels exprime une *intensité du signal* qui n'a pas de signification fixe y compris entre deux images issues d'un même protocole, d'une même région du corps et d'un même patient sur le même scanner [NKUZ00].

Pour l'analyse, les caractéristiques de texture doivent pourtant décrire uniquement des motifs et non des altérations liées à l'acquisition. Certains marqueurs dépendent en partie de la luminosité ou du contraste global de l'image. L'homogénéisation des valeurs d'intensité sur IRM est nécessaire pour pouvoir faire une comparaison pertinente des caractéristiques radiomiques entre les examens.

Beaucoup d'études se sont intéressées à la normalisation d'IRM du cerveau. Une étude de 2004 [CSM04] fait exception en comparant la texture à l'IRM de fromages suisses, extraite

après trois types de normalisations simples visant à conserver une moyenne, un maximum ou un interval similaire. Plus tard Bergeest et al [BJ08] comparent les qualités respectives de cinq méthodes plus complexes basées sur les histogrammes (distances, points de référence etc.), les particularités du cerveau, ou la divergence de l'entropie relative.

Shah et al. ont effectué une revue très complète des différentes méthodes de normalisation des IRM du cerveau [Sha+11]. Ils évaluent notamment un des cinq algorithmes comparés dans l'étude de Bergeest, la méthode de normalisation par alignement d'histogramme (***H**istogram **M**atching*) de Nyul et Udupa [NKUZ00]. Cette méthode présente plusieurs avantages. Même si d'autres algorithmes ont depuis montré des performances légèrement supérieures, elle est robuste et fournit des images de bonnes qualité. Rapide à implémenter et à exécuter [BJ08], elle est également populaire dans les études radiomiques [Zac+09 ; Sha+11 ; Li+17 ; Bak+17 ; Rat+18] et largement documentée.

2.3.1 Principe de l'alignement d'histogrammes

L'alignement d'histogrammes normalise les niveaux de gris d'une image en se basant sur un histogramme de référence. La technique proposée par Nyul et Udupa a été conçue pour normaliser efficacement des IRM de parties du corps similaires. Il s'agit d'un alignement linéaire par morceau entre l'image à normaliser et une échelle standard de référence (*standard scale*).

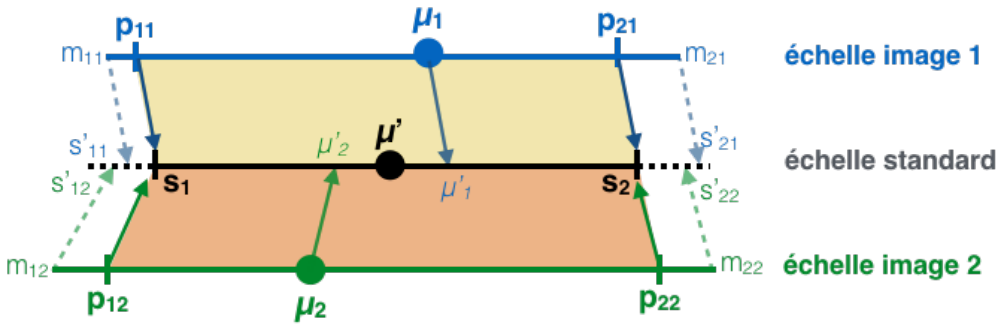


FIGURE 2.4 – Alignement d'histogrammes, phase d'entrainement. Les bornes s_1 et s_2 de l'échelle standard sont préalablement fixées. Les histogrammes de deux images sont utilisés pour calculer un point de repère μ' sur l'échelle standard.

La première étape est un entrainement qui consiste à mettre en place l'échelle standard. Il s'effectue à partir d'une série d'images de référence issues d'une même partie du corps. Ces images sont préalablement seuillées à leur valeur moyenne pour retirer les intensités correspondant à l'arrière-plan. L'analyse de l'histogramme de ces i images fournit des **points de repère** ou **points de comparaison** :

— les percentiles minimum p_{1i} et maximum p_{2i}

— N percentiles μ_{Ni} choisis régulièrement le long de l'histogramme.

La figure 2.4 montre un cas simple d'entrainement avec deux images et un seul point de repère supplémentaire, la médiane μ_i. L'échelle de référence représente un histogramme artificiel dont les bornes maximales, s_1 et s_2 sont préalablement choisies. On fait correspondre ces bornes aux percentiles p_{1i} et p_{2i} des images d'entrainement de façon à trouver pour

chacune une fonction linéaire. Cette fonction permet de calculer les équivalents des médianes μ_i sur l'échelle, les valeurs μ'_i. La moyenne des μ'_i donne le point de référence μ' de l'échelle standard.

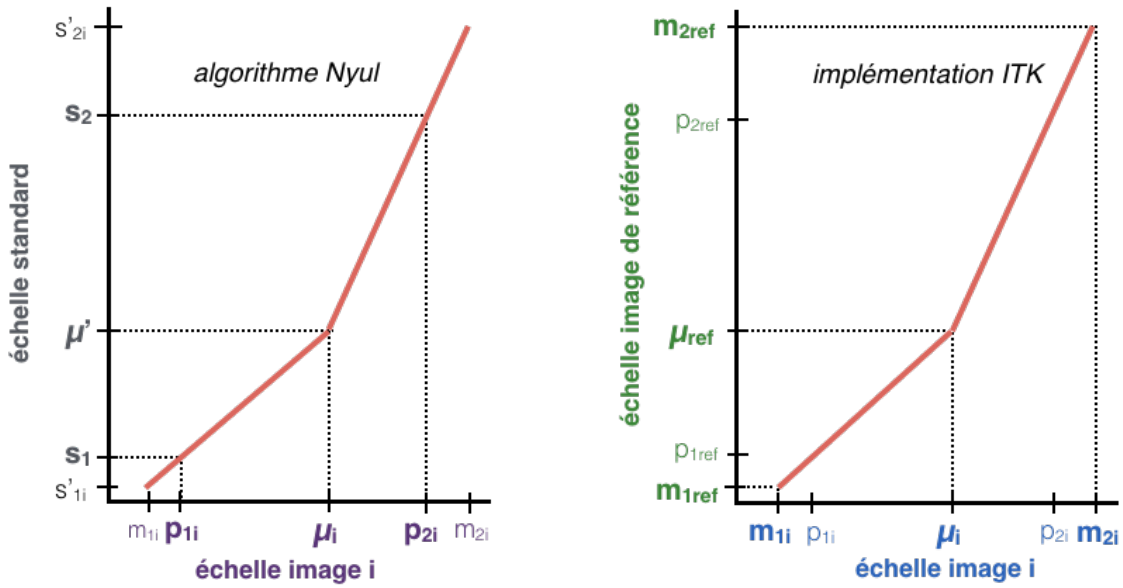


FIGURE 2.5 – Alignement linéaire par morceaux des niveaux de gris d'une image i sur une référence avec un point de répère. (*gauche*) Version originale de l'algorithme. La référence est l'échelle standard. On lui fait correspondre les premiers et derniers percentiles p_{1i} et p_{2i} et la médiane μ_i de l'image à normaliser. (*droite*) Version implémentée dans ITK. La référence est l'histogramme d'une image unique. On lui fait correspondre les intensités min (m_{1i}) et max (m_{2i}) et la médiane μ_i de l'image à normaliser.

Les valeurs s_1, s_2 et μ' permettent ensuite de transformer l'histogramme de nouvelles images une fois associées à leur p_{1i}, p_{2i} et μ_{Ni} . On obtient dans ce cas deux fonctions linéaires permettant de transformer par morceau les valeurs d'intensité des images à normaliser (voir Fig. 2.5 à gauche). Elles sont ainsi alignées sur l'histogramme de l'échelle standard.

2.3.2 Nos choix d'implémentation

Nous avons utilisé l'implémentation de la librairie ITK (*HistogramMatchingImageFilter*). Contrairement à ce qui est sous-entendu dans la documentation, le code ne suit pas exactement l'algorithme de Nyul dans la mesure où il s'agit d'une simplification pour une échelle standard constituée d'une unique image de référence. La figure 2.5 à droite montre que cette version définit les $N + 1$ morceaux de l'histogramme d'une image i obtenus à partir de N quantiles et de ses deux valeurs extrêmes m_{1i} et m_{2i}. Elle les aligne ensuite directement sur les quantiles et extrema correspondants de l'image de référence. Il n'y a donc pas de phase d'apprentissage ni de création d'une échelle standard, ni d'utilisation des percentiles.

La Fig. 2.6 montre la normalisation de l'IRM d'un patient atteint d'un sarcome. Nous utilisons ici des histogrammes à 100 classes et un seul point d'alignement en plus des extrema.

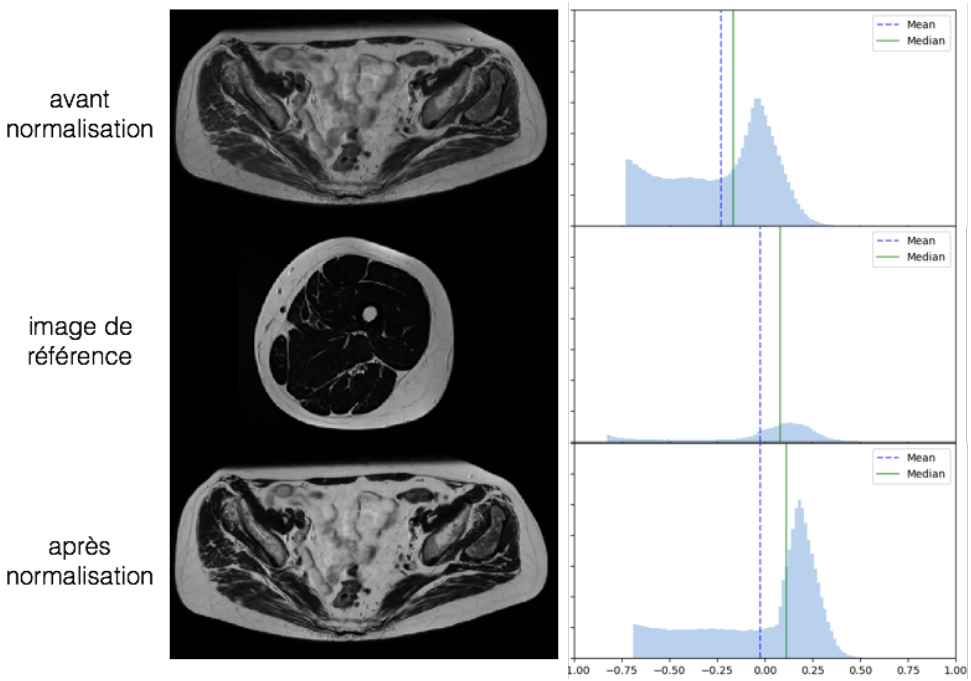


FIGURE 2.6 – IRM T2 du bassin d'un patient atteint d'un sarcome, image de référence (cuisse d'un patient sain) et IRM normalisé par alignement par morceaux de leurs histogrammes respectifs. La médiane est représentée en vert et la moyenne en pointillés bleus (les voxels de l'arrière-plan sont exclus). *source IRM : institut Bergonié*

La soustraction des voxels de l'arrière-plan est activée, la moyenne et la médiane représentées sont celles du reste des voxels. Le pic de l'histogramme initial est décalé vers la droite de façon à correspondre au pic de l'histogramme de référence.

Les images de référence potentielles que nous avons rassemblées proviennent généralement de la même modalité d'IRM d'un patient sain sur une région du corps similaire. Si aucune image de ce type n'est disponible pour une étude donnée, nous sélectionnons une image aléatoirement parmi la cohorte étudiée. Nous retirons les niveaux de gris correspondant aux lésions de son histogramme pour ne pas fausser l'alignement. Dans l'exemple proposé ici et dans l'ensemble du projet *sarcomes* décrit au chapitre 5, nous avons choisi l'IRM de la cuisse d'un patient sain pour référence. Les membres inférieurs sont en effet l'emplacement majoritaire des tumeurs de cette cohorte (58.5%).

Nous avons par la suite empiriquement fixé le nombre de points d'alignement basés sur les quantiles à 2, des valeurs supérieures améliorant peu la correspondance avec l'histogramme de l'image de référence. La figure 2.7 montre en effet le même exemple d'IRM normalisé avec 1, 2 ou 10 points de repère. À mesure de l'augmentation du nombre de points, on constate sur l'histogramme du VOI l'éloignement des deux principaux modes et l'apparition d'un troisième pic, traduit visuellement par une légère hausse de contraste.

Remarque Même lorsqu'ils sont normalisés avec l'alignement d'histogramme, les niveaux de gris obtenus ont une unité arbitraire qui n'a pas de signification biologique. Les valeurs

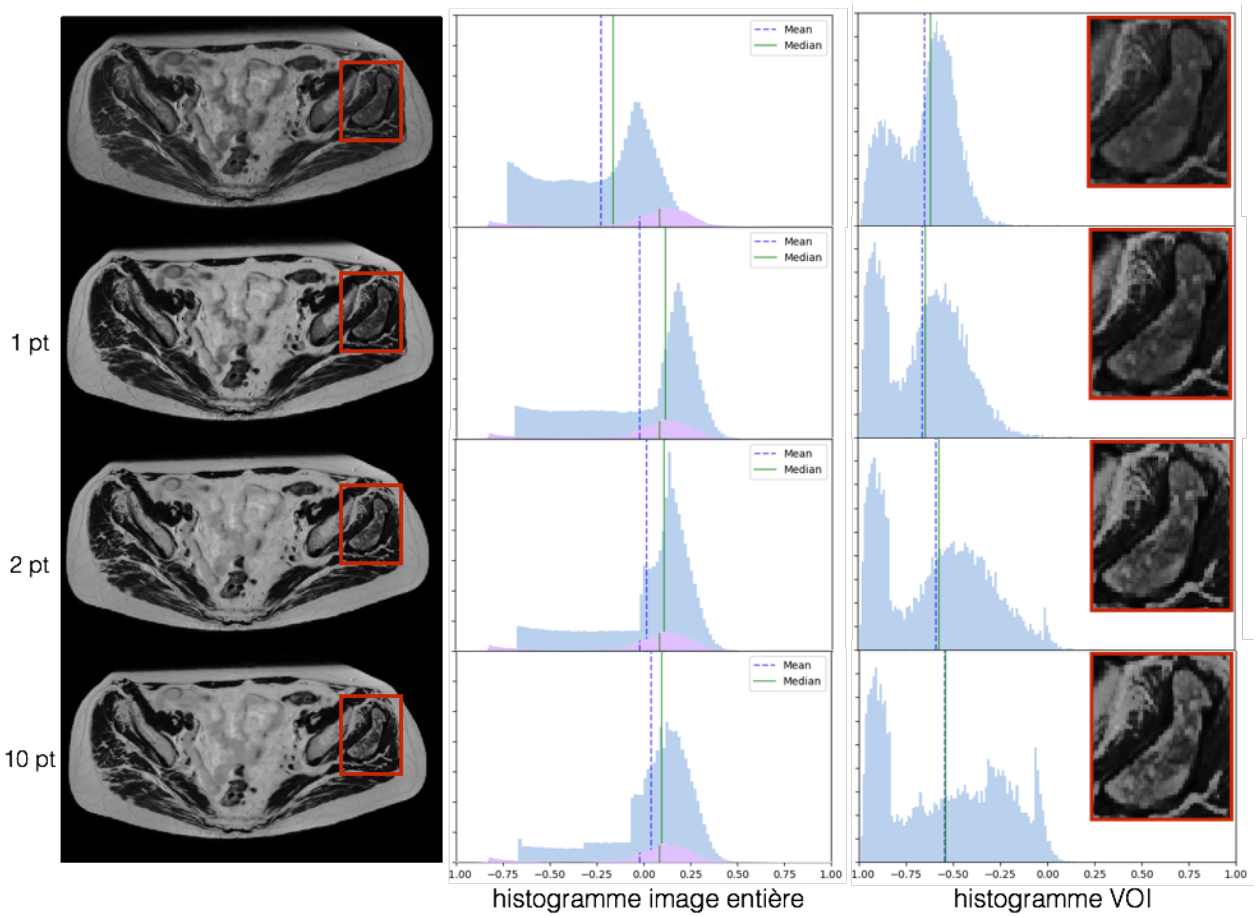


FIGURE 2.7 – IRM normalisée par alignement d'histogrammes à 1, 2 et 10 points de repère. L'histogramme de l'image de référence est visible en rose. Les histogrammes de la partie droite montrent l'impact de la normalisation sur le VOI de la tumeur (encadrés rouges).

ont simplement la même signification pour un même tissu d'un examen à l'autre.

2.4 Étude de l'invariance des variables radiomiques aux prétraitements

Le chapitre 1 met en évidence la dépendance légère des mesures de forme et de texture aux paramètres de reconstruction (positionnement et orientation dans l'espace de discrétisation). Dans leur état de l'art, [Tra+18] rapportent qu'elles dépendent également des pré-traitements effectués sur les images d'entrée, comme le ré-échantillonnage ou les filtres. Toutefois, les études de robustesse aux traitements rassemblées concernent très majoritairement le CT et la TEP [Fav+16 ; BE+17 ; Bog+16 ; Zha+14] et se concentrent plus rarement sur les problématiques liées à l'IRM comme la normalisation intra- et inter-examens.

Nous souhaitons à présent vérifier si les caractéristiques choisies dans nos études sont robustes aux traitements que nous avons choisis. Nous cherchons également quels choix d'implémentation ou de paramétrage des algorithmes de pré-traitement permettent de réduire leur impact sur la répétabilité de l'extraction.

2.4.1 Impact du ré-échantillonnage sur les descripteurs de forme d'objets géométriques

Dans un premier temps, nous utilisons le sous-ensemble de volumes artificiels vu en section 1.6.2 pour déterminer l'influence des choix de ré-échantillonnage sur les caractéristiques de forme.

Méthode. Ces images binaires (masques) 3D ont des voxels de taille isotrope unitaire [1,1,1]. Nous les ré-échantillonnons avec un interpolateur *nearest* selon différents pas d'espace dont les valeurs en x, y et z sont choisies de façon à couvrir plusieurs types d'interpolation :

- le sous-échantillonnage et le sur-échantillonnage,
- le ré-échantillonnage isotrope 2D *(*$[1, 1, 1]$ *devient* $[x, x, 1]$*)* ou isotrope 3D *(*$[1, 1, 1]$ *devient* $[x, x, x]$*)*,
- le ré-échantillonnage à opération entière (les dimensions sont multipliées ou divisées par 2) ou décimales (dimensions $\pm 10\%$).

On obtient 2^3 pas d'espace différents à tester. Les valeurs obtenues pour les caractéristiques morphologiques du jeu de données artificiel sont comparées à leur valeur analytique théorique. L'erreur obtenue inclut donc également l'erreur de reconstruction.

Encore une fois les caractéristiques calculées sont comparées à leur valeur théorique par un test de Student pour un échantillon unique ou un test de rang signé de Wilcoxon.

Résultats. Le test statistique indique que la différence est significative pour l'ensemble des objets, pas d'espace et caractéristiques de forme testés, à l'exception du volume et du rayon équivalent de la sphère.

Les pourcentages d'erreur moyens obtenus pour les quatre types de volumes et chaque taille de voxel sont visibles Fig 2.8 et Fig. 2.9. Les moyennes et les erreurs maximales du sphéroïde sont données à titre d'exemple Table 2.1 pour chaque caractéristique et Table 2.2 pour chacune des trois catégories de ré-échantillonnage.

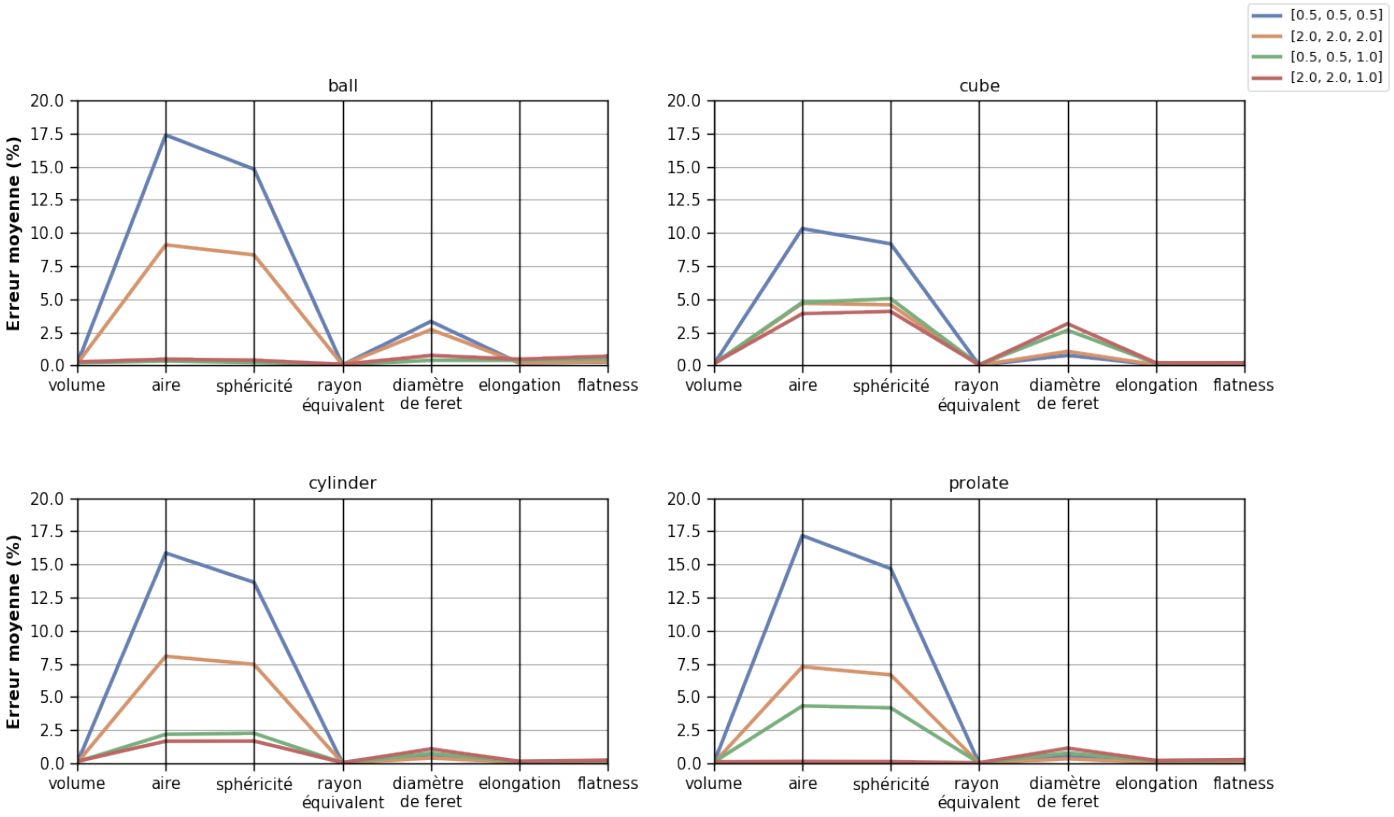


FIGURE 2.8 – Pourcentage d'erreur absolue moyen sur les paramètres de morphologie après rééchantillonnage avec voxels de taille +100% (en **rouge** ou en **orange**) ou -50% (en **vert** ou en **bleu**) dans respectivement 2 ou 3 dimensions.

On fait essentiellement quatre constatations.

1. **Volume, rayon de la sphère de volume équivalent, élongation et flatness** sont les mesures qui **dévient le moins** de leur valeur théorique quelles que soient les conditions (moins de 1% d'erreur en moyenne, moins de 3% au max). Tous les descripteurs de la morphologie restent en moyenne sous la barre des 5% pour le sphéroïde.
2. **En moyenne, tous descripteurs confondus, le sous-échantillonnage** (orange et rouge) **impacte moins la morphologie de l'objet** que le sur-échantillonnage (bleu et vert). Mais si on regarde le détail, **les descripteurs non liés à l'aire de la surface sont au contraire moins altérés par un sur-échantillonnage**.
3. **Plus la taille des voxels est modifiée, plus l'erreur est grande** (on note l'échelle divisée par deux de la seconde figure par rapport à la première).

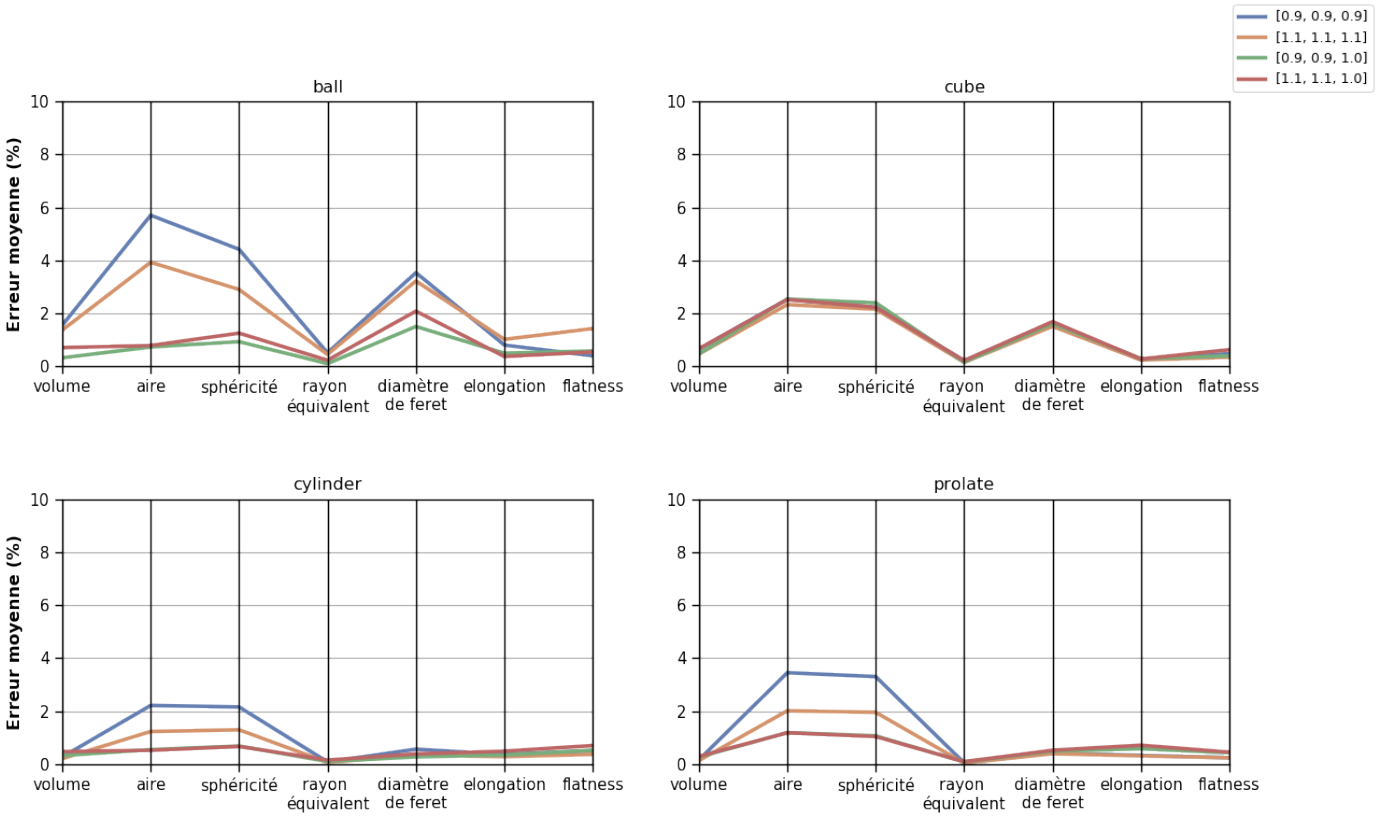


FIGURE 2.9 – Pourcentage d'erreur absolue moyen entre les paramètres de forme après ré-échantillonnage avec voxels de taille +10% (en **rouge** ou en **orange**) ou -10% (en **vert** ou en **bleu**) dans respectivement 2 ou 3 dimensions.

SPHEROIDE	[0.5,0.5,0.5]	[0.5,0.5,1.0]	[0.9,0.9,0.9]	[0.9,0.9,1.0]	[1.1,1.1,1.1]	[1.1,1.1,1.0]	[2.0,2.0,2.0]	[2.0,2.0,1.0]	*moyenne*
volume	**0.07** ***(0.18)***	**0.07** ***(0.18)***	0.16 (*0.39*)	0.13 (*0.38*)	**0.30** ***(0.83)***	0.27 (*0.87*)	0.11 (*0.38*)	0.09 (*0.26*)	0.15
aire	**17.1** ***(18.3)***	7.2 (*12.8*)	3.45 (*3.92*)	2.01 (*2.99*)	1.19 (*1.65*)	1.19 (*2.36*)	**0.14** ***(0.50)***	4.32 (*7.55*)	**4.56**
sphéricité	**14.6** ***(15.5)***	6.6 (*11.4*)	3.29 (*3.58*)	1.95 (*2.79*)	1.05 (*1.52*)	1.07 (*2.12*)	**0.12** ***(0.42)***	4.18 (*6.97*)	4.10
r. équivalent	**0.02** ***(0.06)***	**0.02** ***(0.06)***	0.05 (*0.13*)	0.04 (*0.13*)	**0.10** ***(0.27)***	0.09 (*0.29*)	0.04 (*0.12*)	0.03 (*0.09*)	**0.05**
⌀ de féret	0.52 (*0.92*)	**0.34** ***(0.90)***	0.43 (*1.44*)	0.41 (*0.96*)	0.53 (*1.33*)	0.50 (*1.29*)	**1.15** ***(2.12)***	0.78 (*2.08*)	0.58
élongation	0.11 (*0.25*)	**0.10** ***(0.23)***	0.30 (*0.96*)	0.28 (*0.71*)	**0.72** ***(1.27)***	0.59 (*1.20*)	0.21 (*0.61*)	0.18 (*0.50*)	0.31
flatness	0.15 (*0.40*)	**0.14** ***(0.39)***	0.23 (*0.72*)	0.23 (*0.63*)	**0.46** ***(1.30)***	0.43 (*1.28*)	0.27 (*0.57*)	0.22 (*0.53*)	0.27
moyenne	**4.65**	2.07	1.13	0.72	0.62	0.59	**0.29**	1.4	1.43

TABLE 2.1 – Détail des pourcentages d'erreur absolue moyens *(et max)* du calcul des paramètres de morphologie des sphéroïdes ré-échantillonnées à différents pas.

SPHEROIDE	sous-échantillonnage	sur-échantillonnage	2D	3D	±10 %	±100 %
% d'erreur moyen	0.73 %	2.14 %	1.20 %	1.67 %	0.77 %	2.1 %

TABLE 2.2 – Pourcentages d'erreur absolue moyens du calcul des paramètres de morphologie des sphéroïdes par catégorie de réchantillonnage.

4. L'**erreur de ré-échantillonnage sur la forme est en moyenne plus grande lorsqu'il est effectué dans toutes les dimensions** (bleu et orange) que lorsque qu'un des axes n'est pas modifié (vert et rouge), pour les 4 volumes et dans toutes les conditions.

2.4.2 Impact du ré-échantillonnage sur les descripteurs radiomiques de données réelles

Dans un second temps, nous utilisons un ensemble de patients issus de la cohorte de sarcomes du chapitre 5 pour estimer l'ampleur des modifications provoquées par le ré-échantillonnage sur l'ensemble des descripteurs radiomiques, pour des données réelles. Le jeu de données rassemble 50 IRM en T2 du tronc ou des membres effectués au diagnostic ainsi que les VOI associés.

Méthode. Soit $[a, b, c]$ l'espacement en 3D des IRM. Pour l'ensemble des examens utilisés, $a = b$ et $a < c$ (avec $a \in [0.31 - 1.09]$ [3] et $c \in [3.45 - 8.8]$, en mm). Nous comparons les descripteurs extraits avant et après ré-échantillonnage pour quatre types d'espacement :

- $[a, a, a]$: voxels isotropes, volume sur-échantillonné le long des coupes selon les dimensions des pixels *(ré-échantillonnage 1D, grande augmentation du nombre de coupes)*,
- $[c, c, c]$: voxels isotropes, plans sous-échantillonnés selon l'épaisseur d'une coupe *(ré-échantillonnage 2D, beaucoup moins de pixels par coupe)*,
- $[1, 1, 1]$: voxels isotropes de taille unitaire *(ré-échantillonnage 3D, plus de coupes, moins de pixels par coupe)*,
- $[1, 1, c]$: pixels isotropes unitaires dans le plan, épaisseur de coupe conservée *(ré-échantillonnage 2D, un peu moins de pixels par coupe)*.

Nous comparons en sus les interpolations *linear*, *cubic* et *nearest* sur les descripteurs de texture et d'intensité. Il est à noter que nous ré-échantillonnons les masques (VOI) en utilisant systématiquement l'interpolateur *nearest*, nous ne faisons donc pas cette comparaison sur les indicateurs de forme.

Le nombre d'images ré-échantillonnées obtenu s'élève donc à 4×3.

Les textures sont calculées dans le plan et agrégées en 2.5D comme expliqué en section 1.4.2, pour $d = 1$. Elles sont obtenues à partir de matrices de cooccurrence calculées avec *un nombre de classes fixé* à 30 (puisque les images ne sont pas ici normalisées, cf. section 1.7).

3. $a >= 1$ pour un unique IRM.

Résultats. Nous calculons le pourcentage d'erreur moyen provoqué par le ré-échantillonnage pour chaque variable et nous dissocions les résultats obtenus pour les descripteurs de la forme (Fig. 2.10), de l'intensité (Fig 2.11) et de la texture (Fig. 2.12).

Descripteurs de forme

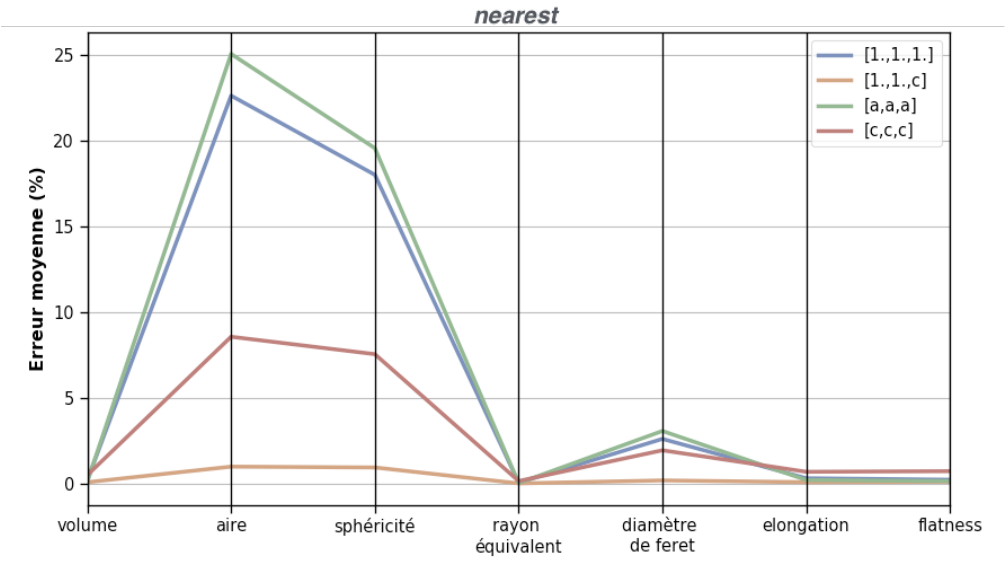


FIGURE 2.10 – Pourcentages d'erreur absolue moyens sur les paramètres de forme après ré-échantillonnage des VOI de sarcomes. Les deux sous-échantillonnages ([1,1,c] et [c,c,c]) restent sous la barre des 10% d'erreur pour toutes les mesures.

Les descripteurs de forme sont les mieux préservés. L'impact du ré-échantillonnage sur la forme des VOI est similaire à celui constaté section 2.4.1 sur la forme des objets géométriques. Le sur-échantillonnage (ici représenté par [a,a,a] et dans une moindre mesure par [1,1,1]) provoque toujours les plus grandes disparités et l'aire et la sphéricité sont toujours plus sensibles. C'est le ré-échantillonnage anisotrope par coupe ([1,1,c]) qui préserve le mieux les formes, avec un pourcentage d'erreur toujours $< 1\%$ quel que soit le descripteur.

Descripteurs de l'histogramme

À l'inverse des descripteurs de forme, c'est le sur-échantillonnage qui préserve le mieux l'histogramme pour l'interpolateur *nearest*. En revanche pour les interpolateurs cubiques et linéaires, c'est plutôt le [1,1,c] qui obtient les meilleurs résultats et ne monte jamais au-dessus de 10% d'erreur. La figure 2.11 rassemble le comportement respectif des trois interpolateurs pour ce pas d'espace. L'interpolation *nearest* y est la moins susceptible de modifier les valeurs d'intensité ($< 2\%$) tandis que la linéaire dépasse les 5% d'erreur pour l'intervalle, le skewness et le kurtosis.

Descripteurs de la texture

Alors que l'erreur maximale de mesure s'élève à 25% maximum pour les descripteurs de forme et d'histogramme (tous types de ré-échantillonnage confondus), on constate que l'erreur sur la texture peut être considérable, jusqu'à approcher les 70% en [c,c,c] voire à dépasser les 1000% pour l'inertie.

Avec l'interpolateur *nearest*, [a,a,a] donne des mesures de texture similaires à plus de 99% car le plus proche voisin est choisi dans le plan, non modifié par l'interpolation. Dans les cas *linear* et *cubic*, les valeurs dans le plan sont tout de même amenées à être modifiées par l'interpolation 3D, expliquant des erreurs plus grandes. Les plus gros changements sont obtenus avec [c,c,c], le sous-échantillonnage du plan de plus grande envergure.

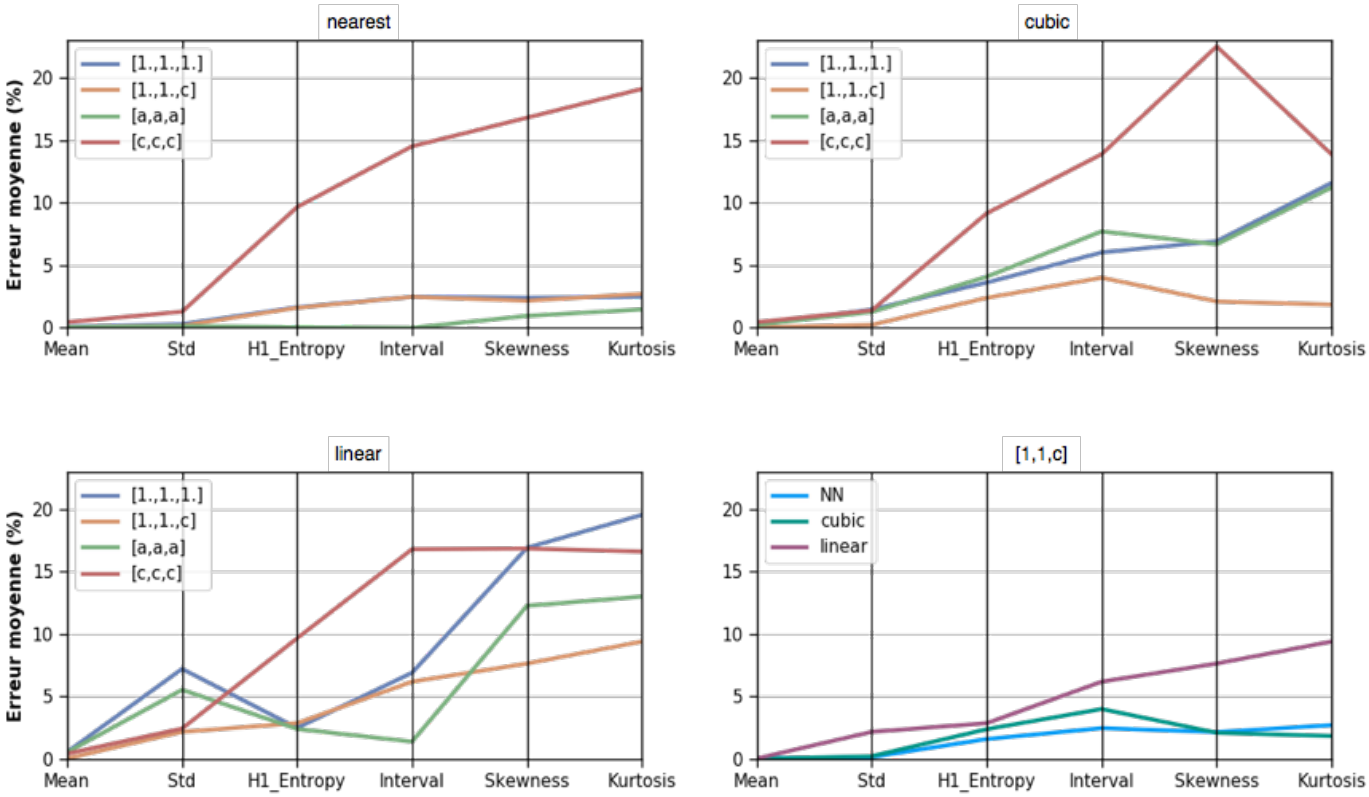


FIGURE 2.11 – Pourcentages d'erreur absolue moyens sur les paramètres d'intensité après ré-échantillonnage des VOI de sarcomes. [1,1,c] transforme le moins les descripteurs pour *linear* et *cubic* et maintient une erreur $< 3\%$ avec *nearest*. [c,c,c] obtient les pires performances pour tous les interpolateurs. L'interpolation, par nature, ne change pas la moyenne de l'intensité.

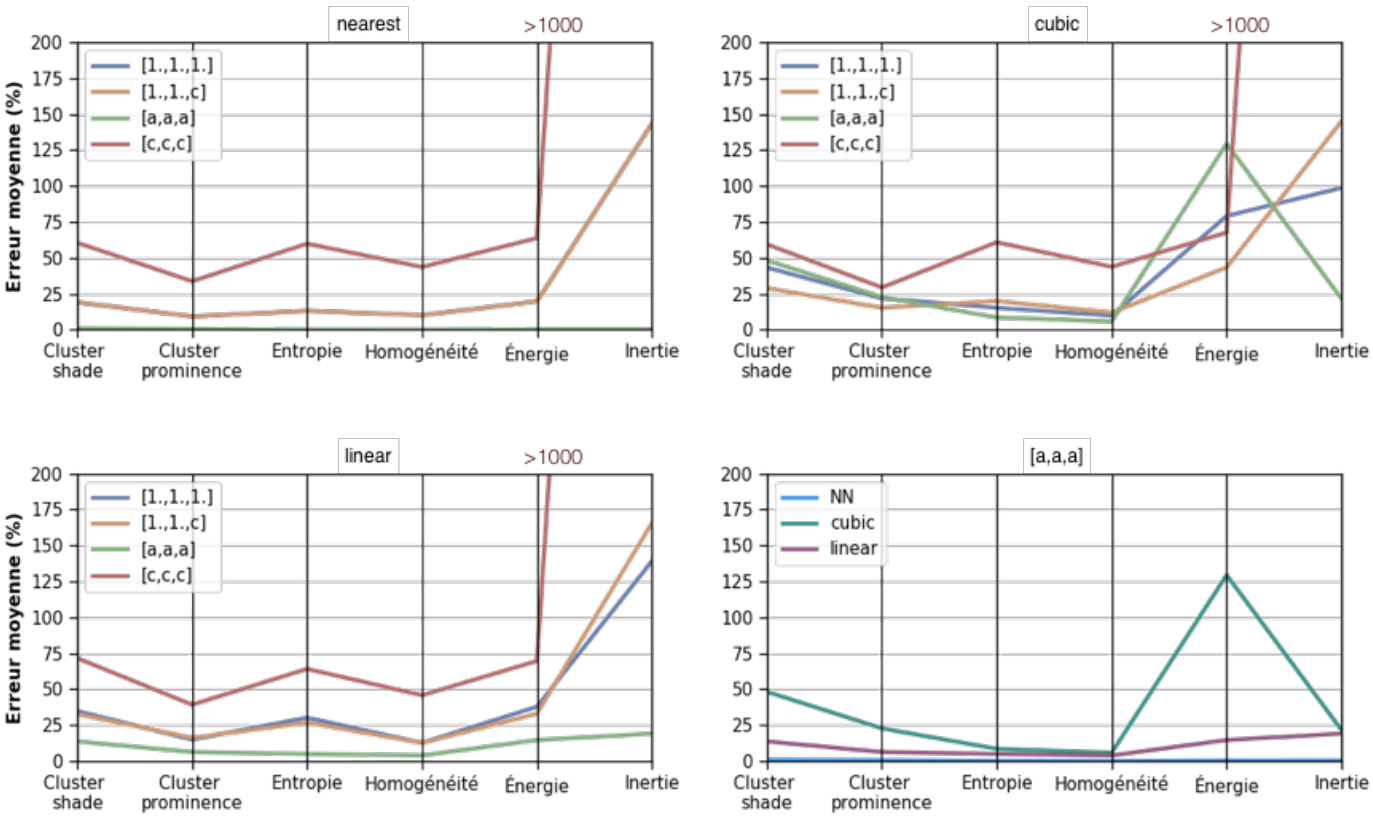


FIGURE 2.12 – Pourcentages d'erreur absolue moyens sur les paramètres de texture après ré-échantillonnage des VOI de sarcomes. L'erreur est grande ($> 10\%$) pour tous les interpolateurs et tous les descripteurs, sauf avec le sur-échantillonnage [a,a,a]. En dehors de ce pas d'espace en *nearest*, l'énergie et l'inertie sont excessivement impactées. À noter, en *nearest*, [1,1,1] et [1,1,c] sont confondus.

Remarques

Le ré-échantillonnage modifie le nombre et les valeurs de voxels du VOI. Son impact reste faible à modéré sur les descripteurs de forme et d'intensité. Dans le pire des cas, changer le pas de la grille va modifier la valeur des kurtosis et skewness qui sont dépendants du nombre de voxels [Rim14]. L'aire de la surface du VOI est sensible à la taille des voxels, mais également à la position de la grille de l'image (nous en avons vu plusieurs exemples en section 1.1).

Les descripteurs de texture sont très sensibles au ré-échantillonnage, en raison des changements locaux de la distribution des valeurs provoqués par les interpolations linéaires et cubiques. De plus, modifier le pas d'espace revient à étudier la texture à différentes échelles. Autrement dit, utiliser plusieurs résolution d'une même image (pyramides d'images) pour extraire la texture équivaut à utiliser plusieurs valeurs du déplacement d dans la matrice de cooccurrence [RAM87].

On note que [a,a,a], un ré-échantillonnage œuvrant uniquement sur la taille et le nombre de coupes, provoque des modifications des descripteurs de texture avec les interpolateurs *linear* et *cubic*. Pourtant, les matrices de cooccurrence choisies n'analysent que des voisinages en 2D. Ce phénomène est dû au caractère 3D des interpolations testées. À l'inverse, comme nos images ont des tailles de coupe bien supérieures à la résolution 2D des pixels, le plus proche voisin d'un voxel sera toujours sur le même plan que lui : l'erreur en *nearest* est négligeable, (uniquement imputable à un éventuel décalage d'un voxel sur les bords). L'agrégation du résultat de chaque coupe cumule donc des séries de matrices identiques.

Nous n'utiliserons pas l'interpolation *cubic*, qui a un temps de calcul beaucoup plus élevé que les autres (voir Annexe B.1) et n'améliore pas spécialement la stabilité des descripteurs.

2.4.3 Impact de la normalisation inter-examens sur les descripteurs de l'intensité et de la texture

Nous sélectionnons à nouveau le jeu de donnée Sarcome de façon à tester l'influence de l'alignement d'histogramme sur les caractéristiques de texture et d'intensité. Comme le principe de l'algorithme est précisément de transformer l'histogramme d'une image, ces descripteurs sont naturellement modifiés (c'est même le but recherché). Nous regardons donc les modifications éventuelles provoquées par deux alignements, le second visant à retrouver l'histogramme d'origine.

Méthode. Le protocole de traitement est détaillé Fig. 2.13. Nous choisissons pour référence un IRM en T2 de la cuisse, sans tumeur. Il est préalablement ré-échelonné linéairement entre -1 et 1 pour en fixer l'échelle, tout comme les IRM de sarcomes à normaliser (A_2). Nous calculons une première fois les descripteurs de texture/intensité de leur VOI.

Cette fois, nous *fixons la taille des classes* des matrices de cooccurrence, à 0.05. Les textures sont calculées dans le plan pour un voisinage d=1.

Un alignement d'histogramme (à 100 classes) normalise chaque observation sur la référence avec 1, 2 ou 10 points de comparaisons. L'histogramme de l'image ainsi obtenue (A_3) est à nouveau aligné avec l'histogramme de sa version non transformée (A_2) pour un nombre de classes et de points identique. Les descripteurs calculés à partir du VOI de l'image finale sont comparés aux valeurs obtenues avant traitement.

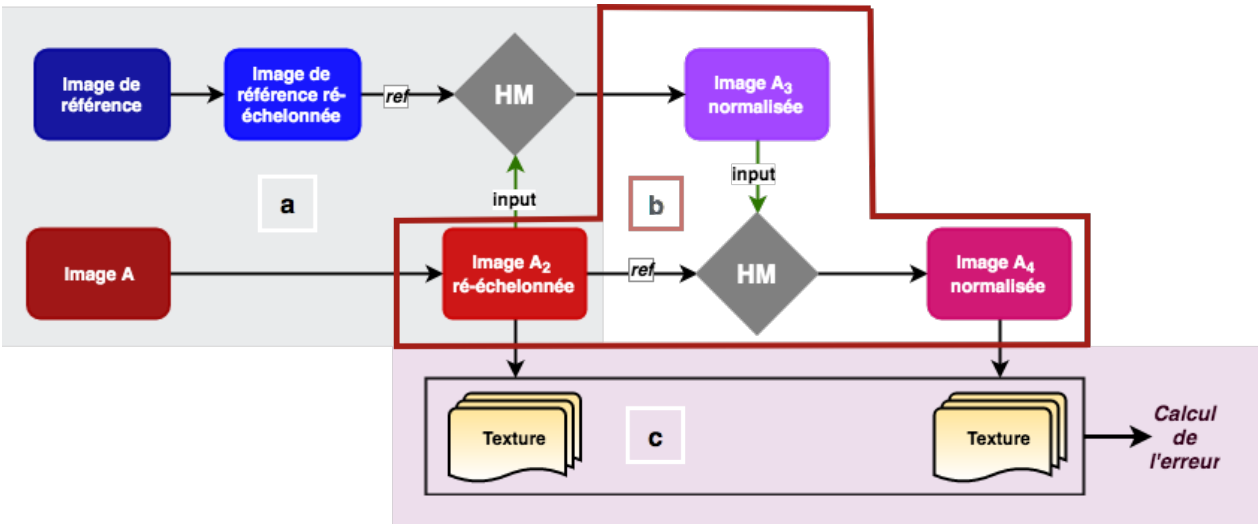


FIGURE 2.13 – Méthode d'analyse de l'impact de la normalisation sur les descripteurs de texture et d'intensité. (a) Une image à tester A et une image de référence sont ré-échelonnées entre -1 et 1. L'alignement d'histogrammes (HM) est utilisé deux fois : d'abord sur l'image A obtenue (A_2) en fonction de l'histogramme de la référence, (b) ensuite sur A normalisée (A_3) avec A_2 pour référence. (c) La différence entre la texture et l'histogramme de A_2 et A_4 est ensuite calculée.

Résultats. Le détail de l'erreur moyenne par caractéristique et par nombre de points de repère est donné en Annexe B.1. On constate au final que l'impact du choix du nombre de points d'alignement est assez faible, ou en tout cas s'avère difficile à évaluer car il diffère selon le descripteur considéré.

Seule une partie des caractéristiques reste stable après la double normalisation (Fig. 2.14 en bas, différence de mesure moyenne <10%). L'intervalle maximum des VOI ne change presque pas puisque les images ont été préalablement ré-échelonnées de façon identique entre deux bornes et que les valeurs extrêmes font partie des points d'alignement. L'entropie de l'histogramme et l'écart type restent également stables.

À l'inverse, les skewness et kurtosis diffèrent beaucoup malgré la seconde normalisation. Comme nous l'avons vu dans l'exemple Fig. 2.7, la première normalisation a modifié la symétrie et l'aplatissement de l'histogramme pour les faire correspondre à ceux de l'image de référence. La seconde n'a potentiellement pas totalement rétabli leur aspect initial.

L'homogénéité et l'entropie sont les caractéristiques de texture les plus robustes alors que l'énergie est fortement impactée. L'inertie est beaucoup moins sensible à la normalisation qu'au ré-échantillonnage.

2.5 Discussion

Du contourage des zones d'intérêt aux traitements de l'image en passant par la reconstruction des volumes, la chaîne de traitement de l'IRM s'avère être une accumulations de légères approximations dont les erreurs s'additionnent jusqu'au moment d'extraire les caractéristiques radiomiques.

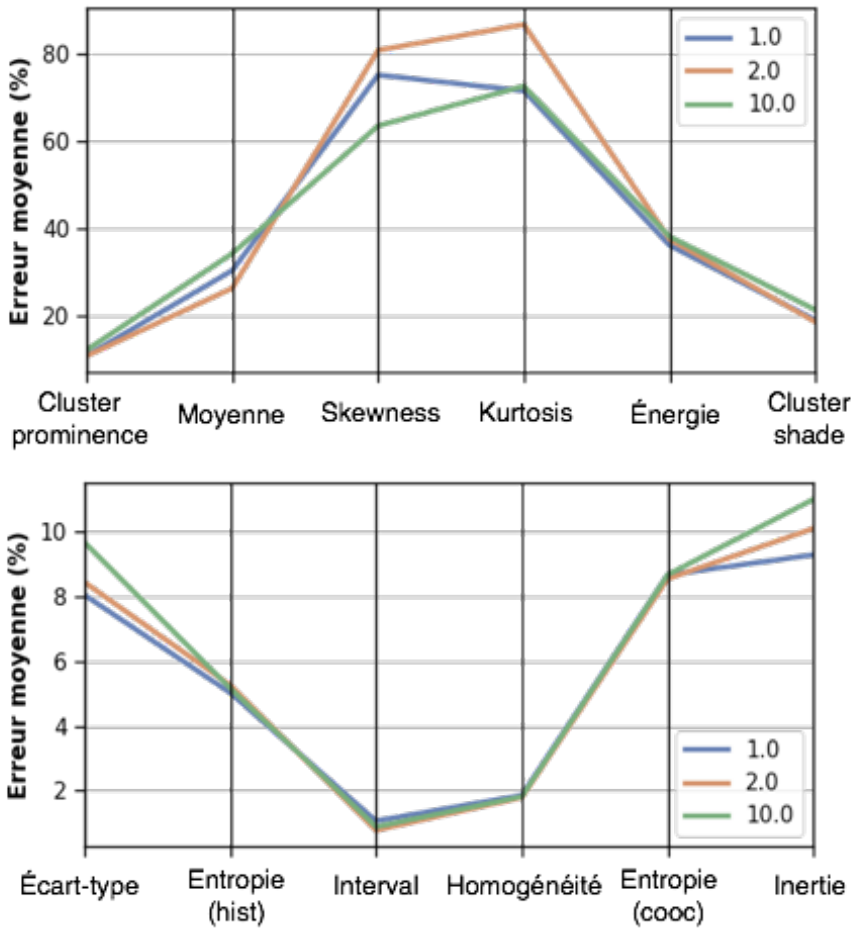


FIGURE 2.14 – Pourcentage d'erreur moyen sur les paramètres d'intensité et de texture après normalisation des niveaux de gris selon la méthode décrite Fig. 2.13. Les échelles sont différentes en haut et en bas.

L'harmonisation des protocoles d'imagerie et la constitution de cohortes issues d'une même machine permettrait de limiter dans une certaine mesure l'ampleur des variations constatées entre les examens. Une méthode spécifique de compensation de l'effet de différents protocoles a notamment été développée pour l'imagerie CT par [Orl+18]. En pratique pour l'IRM, cela reste plus difficilement faisable et ne supprime pas le besoin de pré-traiter l'image.

L'impact moyen de ces traitements sur les descripteurs radiomiques extraits est plus élevé et doit faire l'objet de plus d'attention que la reconstruction. C'est particulièrement vrai pour les caractéristiques de texture, dont les valeurs sont fortement modifiées après traitement : les différentes manipulations de la taille et de l'intensité des voxels suppriment en effet certaines variations locales ou au contraire renforcent des détails [Tra+18], ce qui se répercute directement sur les matrices de cooccurrence. L'interpolation linéaire induit notamment un lissage qui dénature la texture locale. Quelques marqueurs y semblent plus robustes, comme l'entropie et l'homogénéité.

Pour les descripteurs de l'histogramme, c'est l'écart-type et l'entropie qui varient le moins, tous traitements confondus.

Les indicateurs de la morphologie sont les plus robustes. La plupart résistent mieux au sur-échantillonnage, même si leur évolution est proportionnelle au changement de taille des voxels. Le volume, le rayon de la sphère équivalente ou encore l'élongation font partie des critères les plus invariants. L'impact du sur-échantillonnage sur l'aire de la surface et ses dérivés est en revanche plus fort.

Ces conclusions recoupent assez bien celles rassemblées par [Tra+18] pour d'autres modalités d'imagerie. Les mesures de forme sont également considérées parmi les plus robustes par les études listées par [**Limkin2019**].

Une étude très récente de [Why+19] sur l'impact du ré-échantillonnage sur les caractéristiques radiomiques en TEP semble en revanche fournir une conclusion différente de la nôtre sur l'interpolateur à privilégier. Les interpolations *linear* et *cubic* y sont les plus robustes. Nous notons toutefois que l'intégralité des ré-échantillonnages effectués pour cet article comprennent un sur-échantillonnage des voxels dans le plan (en plus d'un sur-échantillonnage des coupes). Nous n'avons pour notre part pas testé cette configuration, ce qui nous empêche une comparaison totale.

En suivant l'évolution des descripteurs plutôt que leur valeur brute, nous supposons que l'utilisation de marqueurs delta-radiomiques, lorsqu'elle est possible, permettrait de minimiser l'impact des traitements. Une évaluation du comportement des delta-radiomiques IRM selon le protocole de traitement des examens reste donc à mener pour confirmer cette hypothèse. À l'heure actuelle, les études méthodologiques se concentrent peu sur l'évolution des paramètres radiomiques (un exemple trouvé s'intéresse uniquement à la différence entre Δ-mesures 2D et 3D [HJ+19]).

On pourrait également envisager d'utiliser différents traitements d'images optimisés pour chaque type de caractéristiques radiomiques. Une image sous-échantillonnée est par exemple préférable pour les caractéristiques d'intensité ou l'aire de la surface, quand la stabilité des mesures de texture et des autres paramètres de forme est maximisée avec un sur-échantillonnage. Il faudrait donc idéalement mettre en place deux chaînes de traitement aboutissant à deux

versions transformées de l'image initiale. Pour étudier la texture à différentes échelle, il est plutôt envisageable de ré-échantillonner plusieurs fois avec des pas d'espace différents. Ces méthodes vont toutefois à l'encontre de la simplicité d'application de la chaîne de traitement et entraînent des coûts de stockage et de temps de calcul non négligeables, surtout pour les sur-échantillonnages.

Nous avons globalement essayé de limiter l'ampleur des pré-traitements à appliquer aux IRM. Le meilleur compromis en terme d'erreur de ré-échantillonnage s'avère être le sous-échantillonnage dans le plan 2D unitaire ([1,1,c]) : il impacte le moins les descripteurs de forme et d'intensité et rend la texture comparable d'un examen à l'autre. Dans le cadre de nos travaux et suite à ces mesures, nous avons donc choisi de ne pas modifier l'épaisseur des coupes et de redéfinir le pas d'espace en 2D seulement. Les pixels de taille unitaire [1,1] sont intéressants car ils sont identiques pour tous les patients et permettent d'évaluer plus facilement l'échelle de la texture.

Ce choix a un effet sur l'extraction des caractéristiques de texture : les voxels n'étant pas isotropes, il n'est pas recommandé de calculer des descripteurs 3D. Nous devons donc remplir les matrices de cooccurrence coupe par coupe, dans les quatre directions du plan (= paramètres de texture 2.5D, cf section 1.4.2).

Notre étude de robustesse reste limitée et une longue liste de paramètres peut encore faire l'objet d'une analyse similaire.

Nous n'avons pris en compte qu'une seule technique de normalisation et un seul de ses paramètres, le nombre de points de référence. En outre, Shinohara et al [TS+14] jugent que l'alignement d'histogramme a tendance à échouer à préserver toutes les caractéristiques biologiques d'une image et supprime certaines variations informatives entre sujets. Ils proposent une chaîne de traitement adaptée, incluant également le traitement de la dérive d'intensité pour le cas particulier des IRM du cerveau.

L'impact de la correction du biais par l'algorithme N4 n'a d'ailleurs pas été étudié ici. Il faut pour cela ajouter un faux biais multiplicatif aléatoire à une série d'images de test, pour le corriger ensuite avec N4. La texture et l'histogramme pourraient alors être comparés avant l'ajout du biais et après sa correction. Il serait également intéressant de constater l'impact seul du ré-échelonnage entre -1 et 1. Nous n'avons pas non plus analysé l'impact de l'interpolation linéaire et cubique sur les masques.

Enfin, il est possible que l'ordre des pré-traitements lui-même influence les caractéristiques radiomiques obtenues. Un exemple donné figure 2.15 montre une même image traitée avec ré-échantillonnage, correction N4 et alignement d'histogramme appliqués dans deux ordres différents. Si la différence s'avère difficile à voir à l'œil nu, les histogrammes obtenus sont bien différents.

Ce processus de vérification est long mais nécessaire à terme. Toutefois, nos résultats actuels suffisent à cerner les précautions élémentaires à prendre pour conduire une analyse clinique et à valider nos contributions spécifiques (voir chapitres 3, 5 et 6).

Au final, calculer l'impact des prétraitements sur la mesure des caractéristiques radiomiques n'est qu'une première étape. Il est par exemple attendu que les normalisations inter et intra-examens modifient l'intensité de façon globale et locale, puisqu'elles ont pour objectif

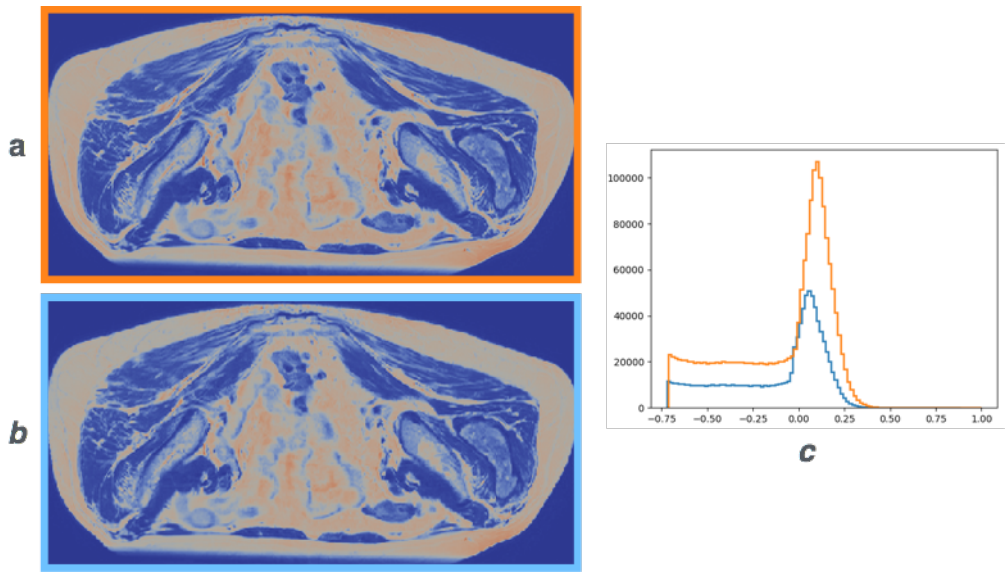


FIGURE 2.15 – IRM du bassin en T2 après deux séquences différentes de traitement.
(a) HM (1 pt) + *nearest* [1,1,c] + N4 + ré-échelonnage [-1,1]
(b) N4 + ré-échelonnage [-1,1] + HM (1 pt) + *nearest* [1,1,c]
(c) Histogrammes de **a** et **b**, seuillés avant la moyenne.
(*source : Institut Bergonié, visualisation : Paraview*)

d'obtenir des niveaux de gris ayant la même signification d'un examen à l'autre. Il serait plus pertinent de regarder directement en quoi ces changements influencent l'analyse et les modèles statistiques construits à partir des textures, comme ont pu le faire [Val+15]. L'objectif de l'analyse radiomique reste en effet de caractériser les lésions dans le but d'acquérir une compréhension plus profonde et plus précoce du devenir des patients. Dans leur état de l'art, [Tra+18] indiquent qu'ils n'ont dénombré que 11 études analysant la valeur prédictive des caractéristiques extraites, sans forcément y ajouter l'impact des traitements de l'image. Aucune de ces études ne concerne l'IRM.

Bilan. Le traitement des IRM s'avère indispensable pour réduire les artefacts de l'image, effectuer des comparaisons entre examens et calculer des caractéristiques radiomiques fiables. L'objectif de ce chapitre n'est pas d'être parfaitement exhaustif dans les tests d'algorithmes et de leur paramètres, ni de fournir une solution parfaite en toutes situations. Nous avons souhaité montrer différentes stratégies permettant de compenser les biais de l'imagerie RMN et donner les arguments permettant de choisir les éléments de notre chaîne de traitement grâce à une série d'exemples simples.

Un nombre conséquent d'études ne précisent pas de quels traitements ont pu faire l'objet les examens qu'elles ont utilisés. De notre côté, nous avons mis en place le pipeline de traitement visible figure 2.16.

L'étape suivante consiste à étudier le potentiel prédictif des caractéristiques radiomiques de l'IRM.

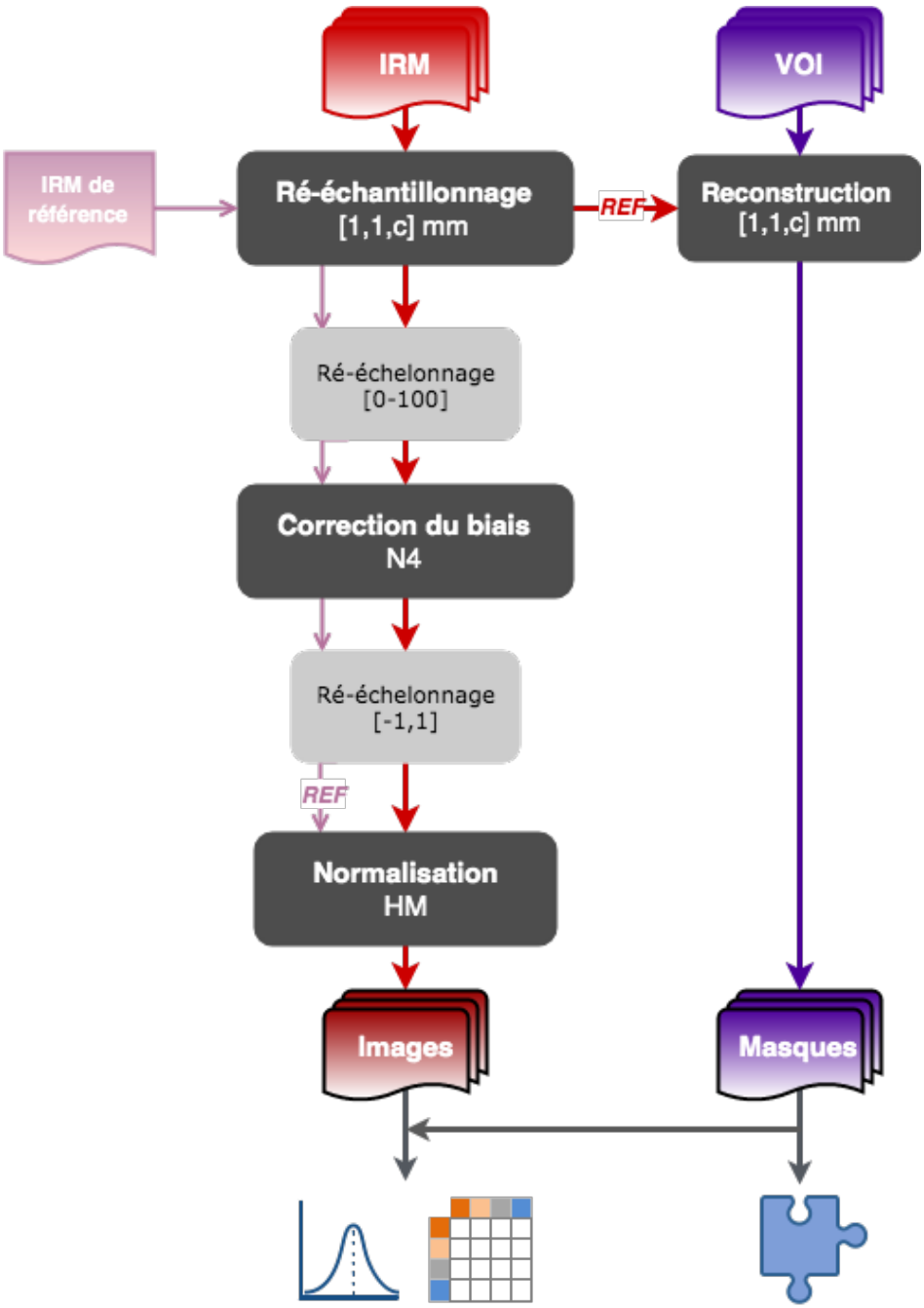


FIGURE 2.16 – Chaîne de traitement des IRM jusqu'à l'extraction des caractéristiques radiomiques.

Chapitre 3

Caractériser la survie sans évènement du carcinome du canal anal avec des biomarqueurs de texture

Sommaire

Les sections suivantes présentent une traduction des travaux effectués en collaboration avec Arnaud Hocquelet, Thibaut Auriac et Hervé Trillaud, radiologues, ayant fait l'objet d'une publication dans le journal European Radiology intitulée *Pre-treatment magnetic resonance-based texture features as potential imaging biomarkers for predicting event free survival in anal cancer treated by chemoradiotherapy* [Hoc+18].

Après un bref rappel des notions statistiques utilisées, nous verrons comment nous avons employé l'intensité et la texture du signal IRM pour caractériser les carcinomes et leur survie sans progression.

3.1 Cas clinique : le carcinome du canal anal

Le carcinome épidermoïde du canal anal (CECA) est une pathologie rare représentant 1 à 2% des tumeurs du tube digestif. Ayant un taux de métastases bas, le CECA subit généralement un traitement locorégional, mais la chimio-radiothérapie (CRT) est devenu le traitement initial standard [Gra98 ; Nor+10 ; Gun+12 ; DJ+13]. Ainsi après CRT, la survie à 5 ans dépasse les 80% même si un sous-groupe de patients a un pronostic moins favorable avec progression de la tumeur [Goh+10 ; HOU+17]. Bien que le bilan d'extension se base sur l'IRM (stadification locale), elle n'a pas vraiment fourni de facteurs pronostics en ce qui concerne la survie sans progression [Goh+10 ; Gun+12 ; Koc+16]. La CRT n'est par conséquent pas ajustée sur les observations IRM initiales et la même dose de radiation est employée quel que soit le grade de la tumeur.

L'objectif de l'étude est d'évaluer le potentiel prédictif des caractéristiques de texture en IRM pré-CRT sur la survie sans évènements après CRT pour les CECA sans métastases. Nous avons pour cela traité les IRM et leurs VOI fournis par l'hôpital Haut l'Évêque et extrait les caractéristiques de texture et d'intensité décrivant chaque lésion. Nous rapportons ici également les résultats de l'analyse statistique effectuée par les cliniciens.

3.2 Pré-requis : les méthodes d'analyse de la survie

L'analyse de survie est une approche statistique visant à évaluer le temps nécessaire à l'apparition d'un évènement. En médecine, le temps de survie est compris entre la réponse au traitement et la survenue d'un évènement : progression, rechute locale ou à distance, ou encore décès du patient.

La probabilité de survie, donnée par la fonction de survie $S(t)$, est la probabilité qu'un individu survive/ne présente pas d'évènement entre le diagnostic et un temps t donné. Le risque $h(t)$, est la probabilité qu'un individu suivi jusqu'à t présente un évènement à ce moment là.

Les analyses de survie en oncologie utilisent essentiellement les méthodes suivantes.

3.2.1 Estimation de Kaplan-Meier

L'estimation de la survie de Kaplan-Meier (KM) [KM58] est une méthode non paramétrique d'évaluation de la probabilité de survie à partir de durées de survie observées. Soit un instant t_i, n_i le nombre de patients sans évènement juste avant t_i, et e_i le nombre d'évènements survenus à t_i. La probabilité de survie $S(t_i)$ est donnée par l'équation 3.1.

$$S(t_i) = S(t_{i-1})(1 - \frac{e_i}{n_i}) \tag{3.1}$$

$S(t)$ est une série de marches horizontales décroissantes qui ne change de valeur qu'à la survenue d'un évènement. L'avantage de la méthode KM est qu'elle peut prendre en compte les données censurées, c'est à dire les patients perdus de vue avant la fin de la période d'étude ou n'ayant pas subit l'évènement. Dans ce cas, n_i est le nombre de patients sans évènements moins le nombre de cas censurés à t_i.

Les courbes de Kaplan-Meier représentent la probabilité KM en fonction du temps. Le test Mantel-Haenzel (*log-rank*) [KLP11] est un test évaluant si le cumul des écarts entre les courbes KM de deux groupes ou plus est du au hasard.

L'estimation KM et le *log-rank* pour la comparaison de groupes sont des analyses univariées : elles décrivent la survie en fonction d'un seul paramètre d'étude et ignorent l'impact des autres. Ce paramètre ne peut être qu'une variable catégorielle ce qui s'avère limitant.

3.2.2 Modèle de Cox

Le modèle de Cox est une des méthodes les plus utilisées en médecine pour modéliser les données de survie [CJG15]. Il s'agit d'un modèle de régression permettant d'évaluer l'effet d'une ou plusieurs covariables sur la survie (sans évènements) des patients. Autrement dit, il permet d'évaluer comment chaque facteur influence le taux de risque à un instant t donné.

Soit n covariables x_i chacune associée à un coefficient β_i. Le risque h est donné à chaque instant t par l'équation 3.2.

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i) \tag{3.2}$$

$h_o(t)$ est la fonction représentant le risque de base. C'est la seule partie de l'équation qui varie en fonction du temps. Les coefficients β_i sont supposés constants. Le modèle de Cox est donc semi-paramétrique. $\exp(\beta_i)$ est le rapport de risque (**H**azard **R**atio).

- Si $HR > 1$, la covariable augmente le risque,
- Si $HR < 1$, la covariable diminue le risque,
- Si $HR = 1$, la covariable n'a pas d'effet.

Le modèle de Cox impose des contraintes fortes. Il assume la notion de risques proportionnels : si un individu a un risque deux fois plus élevé qu'un autre à un t donné, ce risque restera double à n'importe quel autre instant t.

3.3 Matériel et méthode

3.3.1 Patients et suivi

Les patients traités pour un carcinome épidermoïde du canal anal par CRT à l'hôpital universitaire de Bordeaux ont été inclus entre 2007 et 2014, de façon à garder un temps de suivi de deux ans minimum (fenêtre de 80% des évènements). Sont exclus les patients ayant des métastases ou présentant un examen IRM de mauvaise qualité. La cohorte regroupe 28 patients. Leur suivi moyen est de 58 mois *[51-66]*.

Un évènement est défini par une progression locale ou distante de la tumeur, ou le décès du patient. La progression est diagnostiquée par biopsie, examen de la pièce chirurgicale ou preuve évidente de la présence de métastases du foie ou des poumons au TEP-scan. Durant le suivi, huit progressions sont apparues (2 locales, 2 distantes, 4 distantes et locales), en moyenne en 6.8 mois *[4-10]*. La survie sans évènements à 5 ans est de 71%. Six patients sont décédés des suites d'une progression.

Les patients sont examinés une fois par semaine minimum, évalués deux mois après la fin du traitement puis tous les quatre mois pendant deux ans et tous les six mois pendant trois ans de plus. Chaque patient dispose d'une imagerie CT (poitrine, abdomen, bassin) et d'une IRM pelvienne dans les 6 mois suivant la fin du traitement, puis tous les ans. D'autres examens (biopsie, échographie endorectale ...) sont effectués pour vérifier la présence d'une éventuelle progression locale ou distante avant chirurgie.

3.3.2 IRM

Les images ont été acquises au cours de la routine clinique avec un système IRM à 1.5T. Elles incluent trois séquences T2W sans suppression de graisses dans le plan axial, coronal et sagital du canal anal ainsi qu'une une séquence T1W dynamique avec produit de contraste.

Les analyses radiologiques ont été réalisées indépendamment par deux radiologues séniors spécialisés dans l'imagerie du bassin, puis validées par consensus. Le grade est évalué sur le plus grand diamètre de la tumeur.

3.3.3 Contourage et prétraitements

L'analyse de la texture est effectuée sur les séquences T2, faciles à acquérir, moins susceptibles de présenter des artefacts et disponibles pour chaque patient. De plus, le ratio signal/bruit y est élevé, tout comme la résolution spatiale et le contraste des tissus mous du sphincter anal. Un VOI 3D y est contouré manuellement coupe par coupe sur l'intégralité du volume observé de la tumeur.

Les images sont ré-échantillonnées de manière à toutes présenter des pixels isotropes de taille $1 \times 1mm^2$ dans le plan xy. La taille des voxels en z est conservée et donc différente pour chaque examen (avec $(x = y) \in$ [0.47mm, 0.88mm] et $z \in$ [3.3mm, 5.0mm]).

Nous ré-échelonnons également les images entre 0 et 1 pour obtenir des intensités d'échelles comparables.

3.3.4 Analyse quantitative de la texture

La distribution de l'intensité du signal RM est décrite par des indicateurs de texture de premier ordre calculés sur le VOI : moyenne, écart type, intervalle minimum-maximum. La symétrie, le désordre et le comportement aux limites sont évalués respectivement avec les coefficients de skewness, entropie et kurtosis [Hoc+16a].

La présence de motifs visuels est quantifiée avec les indicateurs de texture de second ordre dérivés de la matrice de cooccurrence (GLCM). Sont calculés ici l'inertie, l'entropie de second ordre, l'énergie, l'homogénéité locale, le cluster shade et le cluster prominence.

En raison du caractère anisotrope des voxels dans leur troisième dimension (et de leur résolution modérée en z), nous calculons les GLCM dans le plan (angle et distance). Dans le cadre de cette étude, la composante directionnelle des caractéristiques de texture est moyennée pour les angles $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$, afin de supprimer les redondances. Les résultats sont obtenus et analysés séparément pour des voisinages dans le plan distants de d=$\{1, 2, 5\}$ pixels.

Enfin, nous établissons des cartes de texture pour chaque coupe 2D individuelle. Pour cela des GLCM sont générées à chaque pixel sur un voisinage de 6 pixels de rayon et les indicateurs correspondants sont calculés.

3.3.5 Analyse statistique

Nous donnons la moyenne et l'écart type des variables continues et les variables catégorielles sont exprimées en pourcentages. Pour chaque distance de voisinage on étudie la corrélation entre les paramètres de texture et la durée de survie sans progression. Les analyses univariées sont effectuées avec des modèles à risques proportionnels de Cox.

Pour chaque variable, le groupe de patients est dichotomisé ("bonne" vs "mauvaise" survie) afin de réaliser une analyse de type Kaplan-Meier. Les groupes sont comparés avec un test de *log-rank*. Le seuil de chaque caractéristique est choisi en sélectionnant la *p-value* la plus basse au test de façon à optimiser leur séparation.

Les facteurs pronostics significatifs sont ajustés pour les covariables cliniques importantes (l'âge, le genre et le grade de la tumeur) avec le modèle de Cox multivarié. L'intervalle de confiance du rapport des risques instantanés et la *p-value* correspondante sont calculés avec des méthodes *bootstrap* (100 échantillons). Le c-index est calculé afin d'évaluer la capacité de discrimination du modèle prédictif [1].

Seulement huit évènements ont été constatés durant les années de suivi, ce qui empêche toute analyse multivariée incluant plusieurs paramètres de texture.

3.4 Résultats

Du point de vue des données cliniques et biologiques, le genre s'avère être la seule covariable associée à la survenue d'un évènement de façon significative : la survie à 5 ans est de 82% pour les femmes contre 33% pour les hommes (p = 0.026). Dans notre population, l'âge, le stade T ou N, ou encore le grade de la tumeur ne sont pas significativement associés.

Pour les caractéristiques de texture, le modèle de Cox montre que les paramètres skewness, cluster shade à d=1 et cluster prominence à d = 1, 2 et 5 sont des facteurs pronostics statistiquement significatifs en analyse univariée. Les résultats sont regroupés dans Table 3.1.

Après ajustement sur l'âge, le genre et le grade de la tumeur, seuls les paramètres skewness et cluster shade à d=1 se révèlent associés de façon significative à la survenue d'un évènement. Le c-index du modèle ajusté est de 0.846, (avec un intervalle de confiance à 95% de $[0.697 - 0.993]$ et $p < 0.001$) pour le skewness, et c-index = 0.851 ($95\% CI = [0.708 - 0.994]$, $p < 0.001$) pour le cluster shade à d=1.

La survie sans progression est associée aux paramètres dichotomisés suivant : skewness, cluster shade (toutes distances), cluster prominence à d=1 et entropie à d=5 (voir Table 3.2). Des skewness, cluster shade ou cluster prominence (d=1) plus faibles sont liés à une survie

1. Voir section 6.5.1 pour plus de détails. Un indice de 0.5 est caractéristique d'une absence de valeur prédictive tandis qu'à 1.0 la séparation des patients est parfaite.

Texture parameter	HR	CI	p-value
Age	0.982	0.927-1.040	0.543
Male gender	**5.226**	1.309-21.2	**0.019**
Tumour size sup 5 cm	1.315	0.328-5.269	0.698
T-stage T4 (vs. T1-2-3)	2.14	0.533-8.581	0.283
N-stage N+ (vs. N0)	2.966	0.364-24.2	0.309
TNM grade 3-4 (vs. 1-2)	2.413	0.297-19.6	0.41
Circumferential tumour extant sup 2/3	1.873	0.417-8.418	0.413
Mean signal	4.155	0-1e09	0.791
SD	0	0-5e14	0.32
Entropy	1.217	0.123-12	0.867
Skewness	**0.434**	0.12-0.924	**0.03**
Kurtosis	0.877	0.582-1.025	0.18
Interval	0.381	0-75	0.642
Cluster shade d1	**0.851**	0.618-0.946	**0.025**
Cluster shade d2	**0.814**	0.455-0.968	**0.055**
Cluster shade d5	**0.793**	0.470-0.993	**0.077**
Cluster prominence d1	0.986	0.962-0.995	0.034
Cluster prominence d2	**0.981**	0.949-0.994	**0.031**
Cluster prominence d5	**0.973**	0.931-994	**0.031**
Energy d1	0.011	0-6.835	0.188
Energy d2	0.017	0-59	0.239
Energy d5	0.005	0-27	0.175
Homogeneity d1	1.268	0-2697	0.952
Homogeneity d2	4.107	0-22e3	0.596
Homogeneity d5	3.234	0-5e3	0.613
Entropy d1	1.053	0.290-2.844	0.887
Entropy d2	0.993	0.381-2.59	0.982
Entropy d5	1.181	0.461-3.026	0.583

TABLE 3.1 – Résultats de l'analyse univariée pour la survie sans évènement avec un modèle de Cox sur les variables cliniques et les variables de texture.

Texture parameter	Threshold value	Number of patients above and below threshold value	5-year event-free survival	p-value (log-rank)
Mean signal	0.182	≤ 12	83 %	0.25
		> 16	62.5 %	
Std	0.036	≤ 10	70 %	0.69
		> 18	72 %	
Skewness	**0.588**	≤ 15	53 %	**0.019**
		> 13	92.3 %	
Entropy	1.419	> 17	63 %	0.45
		≤ 11	76 %	
Kurtosis	1.972	≤ 13	44 %	0.057
		> 15	87 %	
Interval	0.376	≤ 15	67 %	0.65
		> 13	77 %	
Cluster shade_d1	**3.855**	≤ 17	53 %	**0.0098**
		> 11	100 %	
Cluster shade_d2	**2.198**	≤ 16	56 %	**0.036**
		> 12	92 %	
Cluster shade_d5	**1.594**	≤ 16	53 %	**0.026**
		> 12	92 %	
Cluster prominence_d1	**46.07**	≤ 15	53 %	**0.034**
		> 13	92 %	
Cluster prominence_d2	30.85	≤ 13	54 %	0.083
		> 15	87 %	
Cluster prominence_d5	22.5	≤ 17	59 %	0.081
		> 11	90 %	
Energy_d1	0.166	≤ 17	59 %	0.061
		> 11	91 %	
Energy_d2	0.149	≤ 18	61 %	0.098
		> 10	90 %	
Energy_d5	0.192	≤ 27	69 %	0.087
		> 1	100 %	
Homogeneity_d1	0.746	≤ 10	90 %	0.12
		> 18	61 %	
Homogeneity_d2	0.679	≤ 10	90 %	0.12
		> 18	61 %	
Homogeneity_d5	0.639	≤ 17	59 %	0.087
		> 11	90 %	
Entropy_d1	3.134	≤ 10	90 %	0.098
		> 18	61 %	
Entropy_d2	3.318	≤ 10	90 %	0.098
		> 18	61 %	
Entropy_d5	**3.365**	≤ 9	100 %	**0.03**
		> 19	58 %	

TABLE 3.2 – Résultats de l'analyse Kaplan-Meier sur le seuil optimal pour chaque paramètre de texture de la tumeur pré-CRT.

sans progression diminuée ($p = 0.019$, 0.009, et 0.034 respectivement). À l'inverse, c'est une entropie plus élevée qui est associée à un mauvais pronostic.

3.5 Discussion

Ces résultats illustrent l'apport potentiel de l'analyse de texture des IRM avant la CRT pour la prédiction de l'évolution des carcinomes. Nous avons constaté que les skewness et cluster shade (d=1) de la tumeur sont des facteurs indépendants de la survenue d'un évènement ayant une bonne valeur prédictive (c-index > 0.8), même après ajustement des covariables (âge, grade, sexe).

Ahmed et al. [Ahm+13] ont étudié l'apport des paramètres de texture IRM dans la prédiction de la réponse à la chimiothérapie du cancer du sein. Ils ont mis en évidence qu'un cluster shade faible est relié à une invasion des ganglions lymphatiques plus importante et à un pronostic moins favorable. Song et al. [Son+16] ont découvert que les patients ayant un cancer du poumon non à petites cellules sous traitement par inhibiteur de la tyrosine-kinase ont un risque de progression tumorale double lorsque le cluster prominence de la tumeur est bas. De même, Coroller et al. [Cor+15] ont montré qu'à un cluster prominence faible est associé un plus grand nombre de métastases distantes pour l'adénocarcinome pulmonaire. Comme expliqué en section 1.4.2, des cluster shade et prominence bas décrivent des motifs aux niveaux de gris peu représentés. Nous supposons que ces deux paramètres reflètent le désordre d'une architecture intra-tumorale composée de motifs différents, tous révélateurs d'une tumeur agressive : zones de nécrose, de prolifération cellulaire, de néo-angiogenèse importante etc.

Comme d'autres études avant [Cor+15 ; Goh+11 ; Ng+12], nous avons observé qu'une entropie élevée est associée à un pronostic défavorable. L'entropie de l'histogramme lorsqu'elle est forte indique effectivement que la tumeur est globalement hétérogène et donc probablement plus maligne et agressive [Cui+11 ; Gan+10]. Nketiah et al. [Nke+16] ont trouvé que l'entropie (sur T2W) est associée à un score de Gleason [2] plus élevé chez les patients atteints de cancer de la prostate. L'hétérogénéité de la tumeur peut révéler des foyers de nécrose provoqués par une vascularisation tumorale insuffisante. Ce phénomène est connu pour entraîner une baisse de la sensibilité de la CRT [Höc+96 ; GRWW17] : la nécrose s'oppose à l'administration intra-tumorale de la chimiothérapie et l'hypoxie diminue la sensibilité aux radiations. L'hétérogénéité de la tumeur est aussi corrélée à son métabolisme du glucose, évalué par TEP-CT [Gan+11].

Actuellement, il est recommandé d'utiliser la même dose de radiations quel que soit le grade de la tumeur. Cependant, puisque les paramètres de texture IRM semblent à même d'identifier les patients ayant un risque de progression plus élevé, il devient utile d'adapter le protocole de CRT. Les patients à haut risque pourrait alors bénéficier d'une irradiation à plus forte dose. Ainsi, l'étude de Muirhead et al. [MPH14] suggère qu'une réduction de dose de 5Gy aux premiers stades d'une tumeur réduit son contrôle local à 2 ans de 98% à 95%.

2. Échelle d'évaluation des cellules cancéreuses pour la classification histologique du cancer de la prostate.

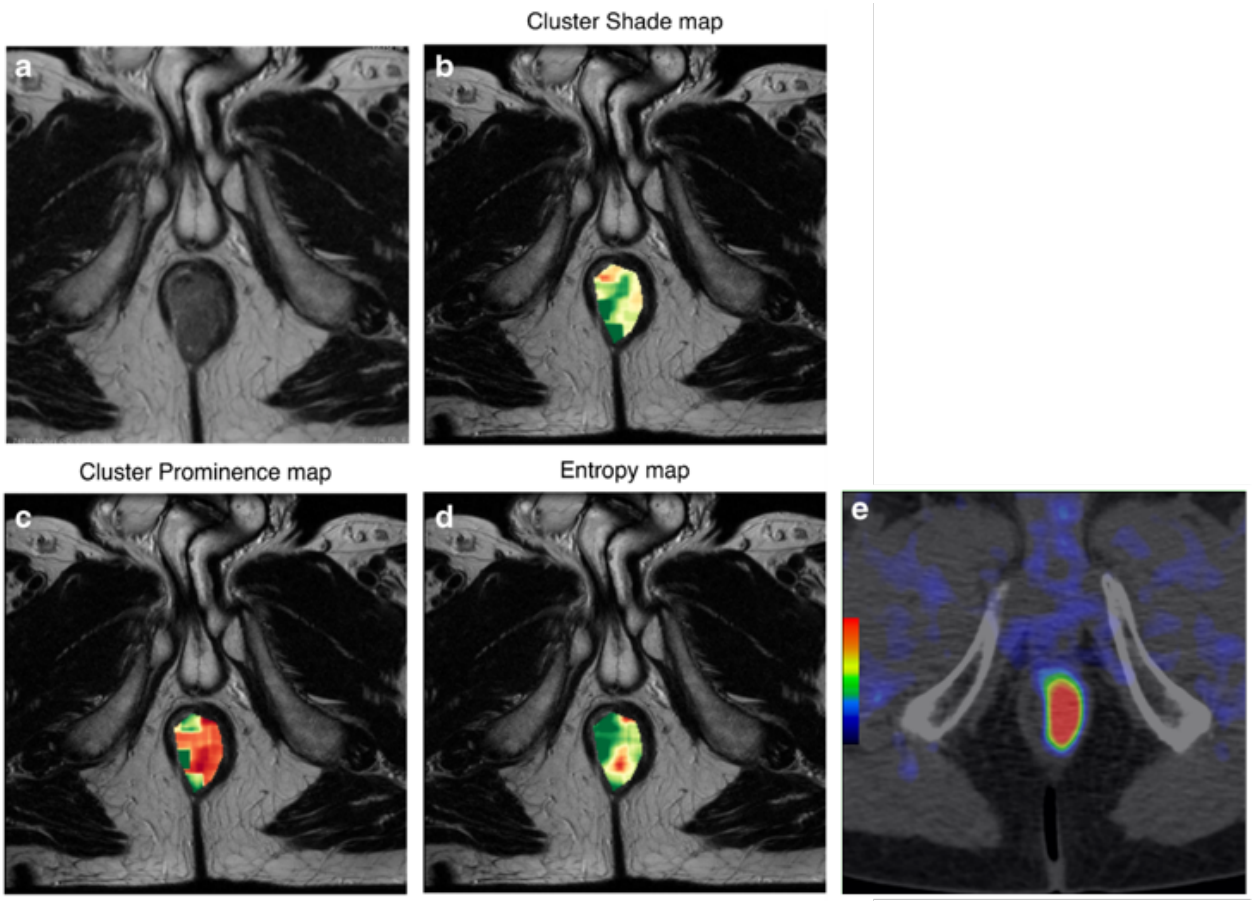


FIGURE 3.1 – Corrélation entre les paramètres d'une tumeur et sa progression locale diagnostiquée sur TEP-CT. (a) Carcinome du canal anal en IRM T2. Les cartes de texture montrent une hétérogénéité intra-tumorale avec une section à gauche potentiellement à haut risque de progression d'après (b) le cluster shade faible (d=1), (c) le cluster prominence faible et (d) l'entropie de l'histogramme élevée. (e) La progression sur la section gauche de la tumeur est confirmée 6 mois après la fin de la CRT : la zone est hautement métabolique (SUVmax = 12.7) sur TEP.

Dans les derniers stades, une augmentation de 5Gy améliore au contraire le contrôle à 2 ans de 50% à 80%.

Une autre façon d'améliorer le contrôle local d'une tumeur sans augmenter l'exposition aux radiations des tissus sains serait d'utiliser l'information de l'hétérogénéité intra-tumorale. C'est le principe du *dose-painting* : la dose prescrite est appliquée en suivant les caractéristiques biologiques de la tumeur [JH05]. L'analyse de la texture à l'IRM nous permet pour cela de calculer des cartes de texture, visibles Fig. 3.1. Ces cartes pourraient être recalées avec un CT-scan de planning de dose afin de définir des volumes cibles pour la radiothérapie [Dif+17 ; Lel+14]. Le *dose-painting* pourrait alors être utilisé pour intégrer un boost d'irradiation visant les zones de la tumeur les moins sensibles. Des études supplémentaires sont bien sûr indispensables pour évaluer l'utilité des paramètres de texture IRM dans de telles stratégies.

Limitations

Des améliorations méthodologiques du traitement des IRM et de l'analyse sont à prévoir. Il serait d'une part nécessaire d'ajouter les corrections de biais N4 et la normalisation par alignement d'histogrammes à notre chaîne de traitement (cf chapitre 2). L'homogénéisation des niveaux de gris des images notamment est susceptible de modifier la valeur voire le rang des descripteurs de texture.

D'autre part, une étude de la corrélation entre les volumes et les caractéristiques de texture est également nécessaire car certaines des VOI délinéés sont de dimensions faibles.

La principale limitation de cette étude vient de la taille très limitée de la cohorte étudiée et du peu d'évènements de progression dénombrés. Ce dernier point est d'ailleurs la raison du regroupement des progressions locales et à distance, au lieu de réaliser deux analyses distinctes. Des études plus approfondies seraient donc requises pour trouver des caractéristiques plus spécialement associées avec les rechutes locales ou distantes car leur traitement est différent. Ainsi, un patient à haut risque de reprise locale bénéficiera d'une irradiation à plus forte dose, quand un patient à risque de rechute métastatique profitera mieux de la chimiothérapie. Il serait intéressant de compléter ce jeu de données avec des images en provenance d'un autre centre, voire par une étude prospective.

Le manque de données pose également la question du sur-apprentissage dans notre analyse. Ce risque a toutefois été minimisé en utilisant des méthodes bootstrap pour calculer les intervalles de confiance et les *p-values*. Nos résultats sont également en adéquation avec ceux d'autres études évaluant la texture d'autres types de tumeur pour prédire la survenue d'évènement ou la survie globale.

Enfin, la taille réduite de la cohorte empêche de mener une étude multivariée plus poussée. Bien que d'interprétation directe, l'estimation de Kaplan-Meier oblige à dichotomiser les variables et la régression de Cox impose des conditions d'application restrictives. Les méthodes d'apprentissage statistique en revanche sont adaptées à la modélisation de relations inconnues et complexes (non linéaires) entre des covariables nombreuses et une variable cible. Elles feront l'objet de la seconde partie de ce manuscrit.

Deuxième partie

Évaluation et prédiction de l'évolution clinique

Chapitre 4

De l'analyse à la prédiction avec l'apprentissage statistique

Sommaire

Dans la partie précédente, nous avons vu comment caractériser des lésions et trouver des relations entre les variables. Nous souhaitons à présent aller au delà des modèles d'inférence statistique et créer des modèles capables de fournir des prédictions sur des données inconnues.

Pour apporter une réponse aux problématiques cliniques qui nous ont été posées, nous allons d'abord définir les outils qui composent notre chaîne d'apprentissage, du traitement des données à l'analyse de la prédiction.

4.1 Principes et vocabulaire

Avant de décrire nos choix d'implémentation, les concepts utilisés dans ce chapitre sont définis dans cette section.

4.1.1 L'apprentissage statistique

D'après Arthur Samuel, pionnier de l'apprentissage statistique (***M**achine **L**earning*, défini en 1959 par [Sam63]), le ML est une catégorie de modèles mathématiques qui permet à un ordinateur d'apprendre à partir de données d'entrée et de s'améliorer sans être programmé explicitement. Autrement dit (Fig. 4.1), un *input* est fourni à un algorithme d'analyse qui va donner une prédiction (*output*) et de nouvelles données d'entrée permettront de le mettre à jour.

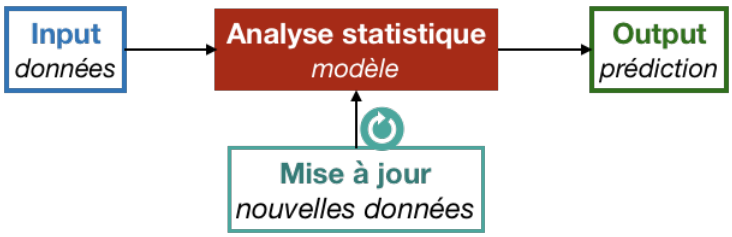


FIGURE 4.1 – L'apprentissage statistique est parfois considéré comme une boîte noire.

On distingue deux grands principes en ML : l'apprentissage supervisé et le non supervisé (Fig. 4.2).

En apprentissage supervisé, les données d'entrée fournies au système sont labellisées : chaque observation vient avec sa prédiction réelle étiquetée. L'objectif de la phase d'entrainement est de trouver une fonction de modélisation qui relie suffisamment bien l'*input* (x) et l'*output* (Y) des données connues. La phase de test effectue la prédiction de l'*output* d'une nouvelle donnée à partir du modèle entraîné.

Le ML supervisé est lui-même divisé en deux catégories.

— la Classification, où l'*output* est une catégorie. Ex : le grade d'une tumeur.

— la Régression, où l'*output* est de type continu. Ex : le temps écoulé avant la rechute d'un patient.

Lorsque les données d'entrée n'ont pas d'étiquette, on fait de l'apprentissage non supervisé : on cherche les relations entre les observations sans a-priori. Ce type d'algorithme est par exemple utilisé pour le *clustering* ou dans les systèmes de recommandation.

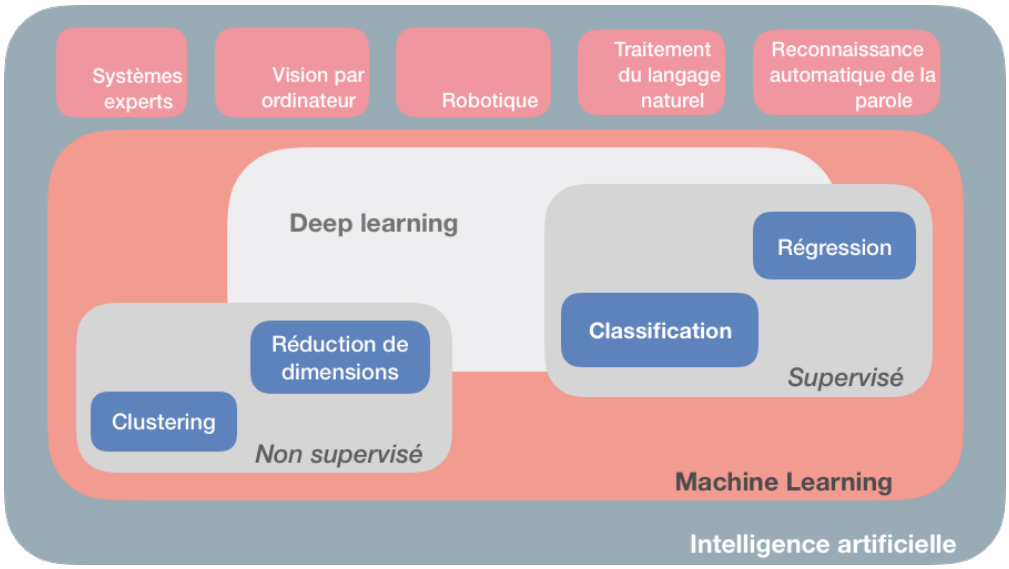


FIGURE 4.2 – Les catégories d'approches en IA et en *machine learning*.

Quand les données sont acquises de façon séquentielle, voire en temps réel, on passe au ML par renforcement. Des agents apprennent seuls en interaction avec leur environnement en ayant pour unique a-priori un système de récompenses et de pénalités à maximiser/minimiser.

Nos études sont des problèmes supervisés de classification binaire, c'est à dire à seulement deux catégories. L'étude sur la rechute du pancréas cherche à séparer les patients qui rechutent de ceux qui ne rechutent pas (cf. section 3.2). L'étude sur les sarcomes de haut grade vise à prédire qui aura une bonne réponse ou un mauvaise réponse à un traitement donné (cf. section 5.1.4).

En ce qui concerne l'étude sur la rechute métastatique du sein, il s'agit à la base d'une étude de régression consistant à trouver la durée de survie sans rechute métastatique des patients. Mais la classification permet de simplifier le problème en séparant les patients à rechute longue et courte grâce à un seuillage empirique (cf. section 6.1).

Sauf si précisé, les explications données dans le reste du présent chapitre concerneront donc les modèles de classification binaire.

4.1.2 Les données d'entrée et la sortie d'un modèle de classification

Un même jeu de données contient souvent des variables de natures différentes [Wan17] : continues, binaires, catégorielles, ordinales, graphes, cartes ou images ... Il peut s'agir de données indépendantes ou non, comme les séries en temps ou les séries spatiales. Quel que soit leur type, les entrées des modèles sont fournies sous la forme d'une matrice de variables.

La sortie d'un modèle de classification est une variable discrète. Lorsqu'elle est binaire, elle est encodée en fonction de la présence ou de l'absence de l'évènement étudié (observation positive ou négative). On parle de jeu de donnée équilibré lorsque la distribution des deux classes est similaire et de jeu déséquilibré dans le cas contraire.

4.1.3 L'erreur de prédiction

Un modèle de prédiction n'est jamais parfait et présente toujours des différences non nulles entre la valeur prédite et la valeur réelle. L'erreur de prédiction est de la forme :

$$\text{erreur totale} = \text{biais} + \text{variance} + \text{erreur incompressible}$$

Le biais est la part de l'erreur totale due à un modèle trop simplifié. Par conséquent, un modèle avec un biais élevé est trop simple et on considère que les paramètres choisis pour décrire les observations ne suffisent pas à modéliser leurs relations. L'erreur de prédiction est grande pour le jeu d'entrainement et le jeu de test.

A l'inverse, un modèle qui a une variance élevée est un modèle trop complexe pour pouvoir généraliser correctement ses prédictions à de nouvelles données. Il est parfaitement spécialisé sur le jeu d'entrainement mais fonctionne mal sur le jeu de test.

Il faut donc trouver un compromis afin de modéliser de façon pertinente et précise les données d'entrainement sans tomber dans la sur-interprétation du bruit (les *outliers*). En d'autres termes, on cherche à minimiser le sous-apprentissage (lié au biais) comme le sur-apprentissage (lié à la variance).

Certains type d'algorithmes de ML seront plus ou moins sujet à l'un ou l'autre problème (voir Juntu et al. qui propose une méthodologie de comparaison de sept algorithmes [Jun+11]). On évalue leurs performances en traçant l'erreur du modèle sur le jeu d'entrainement et sur celui de test en fonction de la quantité de données d'entrainement : c'est la courbe d'apprentissage (Fig. 4.3)).

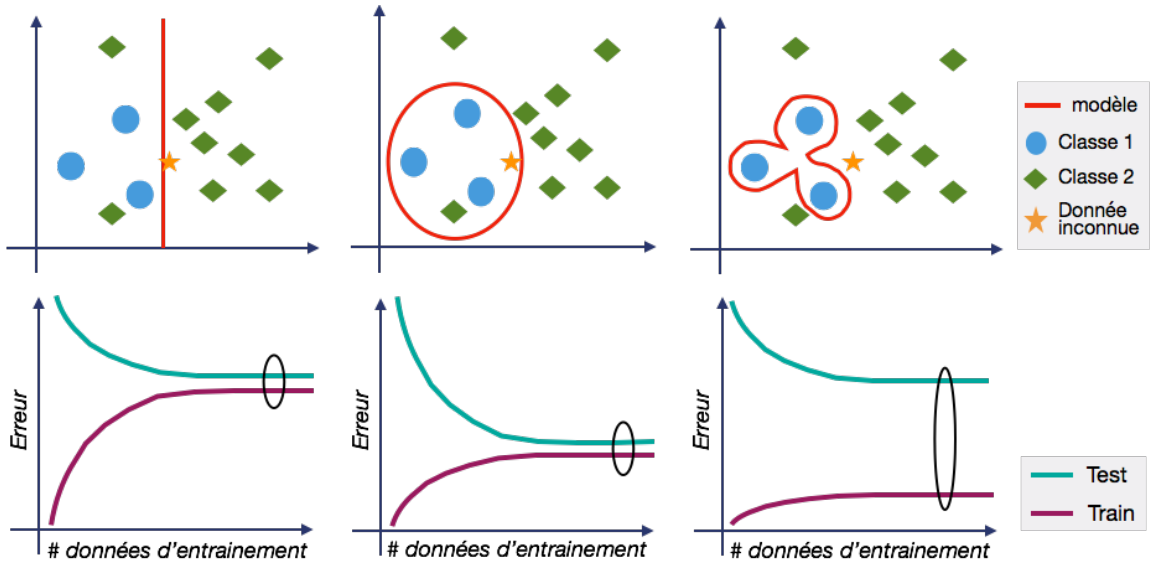


FIGURE 4.3 – Représentation schématique de trois modèles de classification à deux variables (en haut) et courbes d'apprentissages associées (en bas). À gauche, un cas de sous-apprentissage, à droite de sur-apprentissage, au milieu un compromis apte à généraliser tout en restant assez complexe pour modéliser le comportement des données.

Le biais est important lorsque les scores d'entrainement et de test convergent mais restent faibles : l'augmentation de la quantité de données n'améliore pas vraiment la performance du

modèle (Fig. 4.3, à gauche). Il est conseillé dans ce cas d'augmenter le nombre de paramètres pour complexifier le modèle ou de diminuer son terme de régularisation. Si l'écart entre l'erreur d'entrainement (faible) et celle de test se réduit à mesure de l'ajout de données, c'est qu'il y a sur-apprentissage (Fig. 4.3, à droite). La variance pourrait diminuer en augmentant la quantité de données, en simplifiant le modèle ou en augmentant la régularisation des paramètres.

La courbe d'apprentissage est un outil important pour évaluer la quantité de données nécessaires à la construction d'un modèle d'apprentissage fiable.

4.2 Préparation des données d'entrée

4.2.1 Nettoyage

Tout processus d'apprentissage commence par une phase de nettoyage des données, souvent la plus longue étape de traitement [FIC19]. Elle consiste à filtrer les données inutiles, corriger les anomalies, compléter les valeurs manquantes et transformer chaque variable au format et à l'échelle souhaitée. L'ensemble de ces étapes évite de contaminer l'analyse par de fausses informations et prévient une partie des erreurs de modélisation.

Compléter les données manquantes La plupart des collections rassemblant des données cliniques sont incomplètes, que ce soit en raison de patients perdus de vue / suivis sur plusieurs établissements, de mesures non effectuées, de traitements interrompus etc... La cohorte $C_{9.25}$ de l'étude de la rechute métastatique du cancer du sein (cf. section 6.2.1) en présente une proportion assez élevée, visible Fig. 4.4. Seize paramètres sur quarante ont jusqu'à 5% de données manquantes. La taille de la tumeur primitive à l'évaluation clinique n'est pas renseignée pour plus de 12% des patients.

Les algorithmes d'apprentissage comme les méthodes statistiques multivariées fonctionnant rarement avec des valeurs manquantes, il est nécessaire de compléter ces inconnues pour éviter d'avoir à supprimer une grande proportion d'observations [RTA09].

Les méthodes d'attribution des valeurs varient souvent en fonction du type ou de la signification de la variable. Pour les variables continues, le plus évident est d'utiliser la valeur moyenne du jeu de données. La médiane est moins perturbée par les valeurs extrêmes et les éventuelles *outliers*. Pour les variables catégorielles ou binaires, on choisit le mode, ou une variable par défaut définie par les cliniciens. Sur des données longitudinales il est possible d'inférer les valeurs manquantes grâce à d'autres points dans le temps.

Des méthodes plus sophistiquées comme l'algorithme *MissForest* [SB12] peuvent ainsi gérer les différents types de variables simultanément et effectuer des inférences non linéaires en utilisant des forêts aléatoires (cf. section 4.10). Nous utiliserons cet algorithme quand les observations manquantes sont plus nombreuses mais préférerons la médiane dans les autres cas car le temps de calcul de *MissForest* est élevé.

Corriger les observations anormales Les valeurs à la marge (*outliers*) dans un jeu de données peuvent aussi bien contenir une information clé que représenter une véritable erreur susceptible de biaiser un modèle. La détection non supervisée des valeurs aberrantes passe

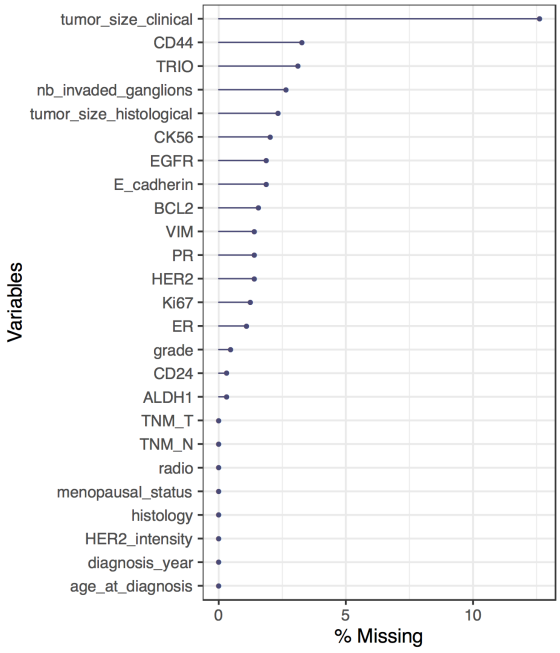


FIGURE 4.4 – Pourcentage de données manquantes pour les 25 paramètres du jeu de données sur le cancer du sein. *source : Chiara Nicolò*

par la visualisation (distribution, box-plot, etc., cf. Fig. 4.5) et impose la plupart du temps de réduire les dimensions du jeu de données (cf. section 4.2.3) voire d'analyser l'écart-type des caractéristiques individuellement. Hodge et Austin [HA04] passent en revue les méthodes automatiques de détection supervisée. Elles recourent à l'étude de la proximité entre les points du jeu de données (par exemple avec un algorithme kNN, cf. section 4.3.3). Les 'anomalies' sont supposées apparaître éloignées de la distribution des données "normales".

Le caractère erroné ou informatif des outliers ne peut souvent être déterminé qu'en fonction du contexte ou en testant les modèles en leur absence (si le volume de données est suffisant).

4.2.2 Transformation

Variables continues Un jeu de données cliniques, radiologiques et radiomiques contient des caractéristiques dont les unités et les échelles varient parfois amplement : âge d'ordre de grandeur maximal 10^2 ou nombre de cycles de chimiothérapie d'ordre 10^1, sphéricité entre 0 et 1, *cluster shade* à valeurs négatives, etc.

Les algorithmes d'apprentissage qui fonctionnent en calculant des distances euclidiennes entre les observations s'en retrouvent fortement impactés car les attributs y ont un poids proportionnel à l'ampleur de leurs valeurs. D'autre part, la descente de gradient opérée durant l'étape d'optimisation de certaines méthodes s'avère instable et lente sur des données d'ordres de grandeur non homogènes [MB05]. La solution est la mise à l'échelle des données.

La normalisation (ou *min-max scaling*) recale le vecteur de valeur entre 0 et 1. La standardisation (*standard scaling* ou *z-score standardization*) le recalcule de façon à ce que les

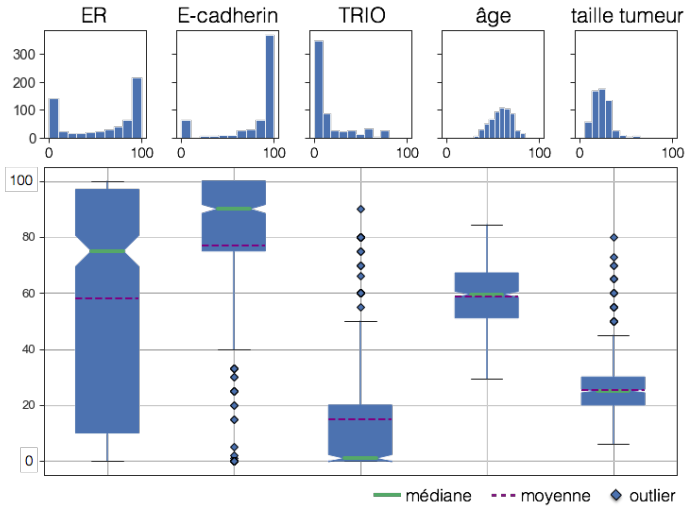


FIGURE 4.5 – Histogrammes et diagrammes en boîte de 5 attributs de la cohorte $C_{9.25}$. Les *outliers* indiqués par le box-plot sont susceptibles d'apporter de l'information.

données soient centrées autour de 0 avec un écart-type de 1 (éq. 4.1). C'est la méthode la plus commune et la moins risquée [Gru15] et celle que nous avons appliquée sur nos données lorsqu'un recalage de valeurs était nécessaire.

$$x_i = \frac{x_i - moyenne(x)}{std(x)} \tag{4.1}$$

On note toutefois que le *standard scaling* suppose que la variable à transformer est normalement distribuée. En cas de distribution non gaussienne ou de données clairsemées, il est préférable d'échelonner en fonction de la valeur maximale absolue (MaxAbsScaler) qui recale les valeurs entre -1 et 1 tout en préservant leur dispersion et en conservant les valeurs nulles. Des tests informels sur nos jeux de données n'ont pas spécialement modifié les résultats en utilisant cette méthode. Une étude complémentaire plus poussée serait tout de même nécessaire pour vérifier si un changement de méthode pourrait bénéficier à nos résultats.

Variables catégorielles Certains algorithmes comme les arbres de décision peuvent fonctionner avec des variables catégorielles (attributs représentant des classes de valeurs non ordonnées, ex : histotype d'une tumeur, localisation de la lésion, etc). Pour les autres, elles doivent le plus souvent être binarisées avant d'être utilisées.

Elles peuvent être regroupées manuellement et transposées en entiers. C'est par exemple le cas de la variable "*menopausal_status*" du jeu de données du cancer du sein, réduite à un booléen (patiente ménopausée ou non). Dans le jeu de données sarcomes, les catégories d'évolution de l'œdème sont rassemblées en "absence, diminution, disparition" vs "stabilité, augmentation". Les valeurs des variables ordinales peuvent également être groupées par modalités. Par exemple, les grades 1 et 2 d'une tumeur seront rassemblés car considérés globalement moins sévères que le grade 3.

Si le regroupement en deux classes simplifie trop la variable, on peut lui préférer un

encodage plus complexe. Le *one-hot-encoding* est la technique la plus utilisée [PPP17]. Elle consiste à créer une nouvelle variable booléenne pour chaque modalité existante d'une caractéristique. Ainsi le "traitement" d'un cancer du sein, variable représentant une liste de valeurs parmi N chaînes de caractères, devient une série de N variables booléennes : "chimiothérapie", "radiothérapie", "hormonothérapie".

4.2.3 Sélection et réduction des dimensions des données

Sélectionner les variables

La sélection des variables vise à choisir un sous-ensemble limité de caractéristiques qui soit suffisamment pertinent pour l'analyse statistique et la construction de modèles de prédiction efficaces.

Les raisons pour supprimer une partie des attributs sont nombreuses :

1. Simplifier les modèles améliore leur capacité à généraliser et permet d'éviter le sur-apprentissage.
2. L'interprétation des résultats est facilitée avec moins de variables.
3. Lorsque le nombre de variables est très grand, opérer une sélection réduit les probabilités d'obtenir par hasard une corrélation entre une variable et la valeur à prédire.
4. On réduit le temps d'exécution et/ou éventuellement l'espace en mémoire utilisé (particulièrement important en fouille de données massives).

On ne discutera pas ici de l'exclusion manuelle de variables relevant de l'étape de nettoyage (ex : variables qualitatives à la reproductibilité faible, variables collectées hors sujet ...)

Les méthodes de sélection de variables visent à identifier automatiquement les attributs qui ne contribuent pas à la prédiction, parce qu'ils sont redondants ou inutiles. Nous avons principalement appliqué deux types de sélection.

Filtres Filtrer les variables consiste à effectuer un test statistique univarié (de type χ_2, Wilcoxon ou test de corrélation) entre les attributs et la variable à prédire. Les caractéristiques sont ensuite classées selon le score obtenu et les meilleures sont sélectionnées. Les tests univariés évaluent l'impact de chaque variable indépendamment et ne dépendent pas d'un classifieur. Ils s'avèrent souvent être les premières méthodes à tester grâce à leur efficacité de calcul et à leur faible tendance au sur-apprentissage [**Parmar2015**]. Ils peuvent même permettre de déterminer la quantité optimale de variables de façon exhaustive en les rajoutant une à une.

Régularisation Ces méthodes étudient l'impact de chaque attribut sur le score de prédiction du modèle pendant sa construction. Elles introduisent une pénalisation qui favorise la diminution de la complexité de l'algorithme testé.

On peut citer ElasticNet, compromis entre les méthodes Ridge et Lasso qui combine les régularisations L1 et L2 pour diminuer ou supprimer l'influence d'une variable dans le modèle [ZH15]. Le niveau de pénalisation est un paramètre à régler déterminant car il conditionne le nombre de variables restantes. On note qu'un groupe de variables corrélées sera entièrement

sélectionné ou recalé par ElasticNet, contrairement à Lasso qui ne gardera qu'une variables (voir [Par+19] pour plus de détails). L'algorithme est pensé à l'origine pour les travaux de régression mais peut être adapté à la classification [LJ10].

Réduire les dimensions

Bien qu'elle vise également à réduire le nombre de paramètres d'un jeu de données, la réduction de dimensions diffère de la sélection des caractéristiques en créant de nouveaux attributs, combinaisons linéaires des précédents. On lutte ainsi contre la multicollinéarité (variables mesurant la même information).

C'est également une façon de prévenir la "malédiction de la dimension" : les points de données sont de plus en plus espacés à mesure qu'augmente le nombre de variables. Pour extraire de l'information en conservant des résultats statistiquement significatifs, la quantité de données doit alors elle aussi augmenter exponentiellement. Il est généralement plus simple de réduire le nombre de paramètres.

L'analyse en composantes principales (ACP) est l'approche la plus classique de réduction des dimensions. Cet algorithme non supervisé crée de nouvelles caractéristiques non corrélées ordonnées de façon à maximiser la "variance expliquée" : le composant principal est à l'origine de la majorité de la variance du jeu de données, le second explique la deuxième cause de variance, etc.

L'ACP doit impérativement être précédée d'une normalisation des attributs car la transformation dépend de leur échelle. Dans le cas contraire, les caractéristiques d'ordre de grandeur plus élevé dominent inévitablement la composition des composants principaux.

La contraction des dimensions à 2 ou 3 permet en bonus de visualiser le comportement des données sur un graphe. L'inconvénient principal est la difficulté à donner du sens aux nouvelles dimensions, ce qui complique l'analyse des résultats. Il est également difficile d'optimiser le nombre de composants principaux à conserver sans recherche exhaustive.

Dans la pratique, si l'ACP a montré un intérêt relatif dans nos travaux préliminaires sur le pancréas, les résultats sur l'étude de la réponse au traitement des sarcomes ou la rechute métastatique du sein n'ont pas été concluants.

L'analyse linéaire discriminante (LDA) est l'équivalent supervisé de l'ACP. On note que pour les données catégorielles, une analyse en composantes multiples (AMC) devrait être préférée. Dans le cas de données mixtes contenant à la fois des variables numériques et catégorielles, les démarches classiques ne fonctionnent donc plus et il est préférable d'utiliser une approche comme la *PCAMix* qui combine ACP et AMC [Cha+11].

La réduction de dimensions non linéaire, qui rassemble des méthodes comme les autoencodeurs ou UMAP, est également un domaine de recherche actif qui n'a pas encore fait l'objet de nos travaux.

4.2.4 Gérer les jeux de données asymétriques

Les études de classification sont régulièrement confrontées à des problèmes liés au déséquilibre de la distribution des classes et nos études cliniques ne font pas exception. Les

métriques classiques d'évaluation des résultats sont biaisées et ne décrivent pas réellement la performance des modèles de prédiction. Certains types de modèles eux-même auront systématiquement tendance à favoriser la classe dominante.

Plusieurs types de méthodes ont été développés pour résoudre ce problème. Il est possible de prendre en compte l'asymétrie des classes dans les modèles eux-mêmes grâce à une matrice de coût : on pondère le déséquilibre en donnant un poids plus fort aux erreurs faites sur la classe minoritaire. Nous avons choisi cette option lorsque cela était possible.

Lorsque les résultats sont insuffisants, on peut aller plus loin en ré-échantillonnant une classe ou l'autre (Fig. 4.6) pour rétablir l'équilibre et améliorer les performances de classification.

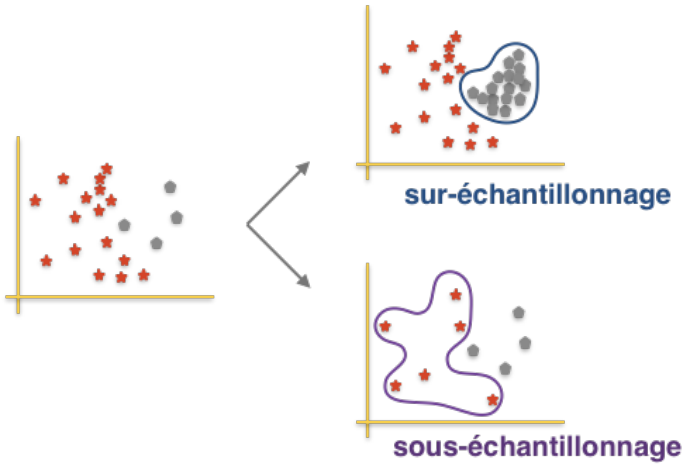


FIGURE 4.6 – Les deux principes de ré-échantillonnage des jeux de données déséquilibrés.

Sous-échantillonner la classe majoritaire

La méthode la plus simple consiste à retirer aléatoirement des observations de la classe majoritaire jusqu'à atteindre la proportion souhaitée.

D'autre méthodes plus avancées permettent de supprimer certaines observations de la classe dominante de façon plus ciblée :

— en retirant seulement les observations majoritaires dont le plus proche voisin appartient à la classe minoritaire, de façon à creuser l'espace entre les deux classes [Tom76]. La distribution de la classe majoritaire est alors altérée.

— en effectuant un *clustering* de la classe majoritaire et en conservant uniquement les observations centroïdes. Cette méthode conserve la distribution de la classe mais peut conduire à supprimer des *outliers* informatifs. Elle rajoute un hyperparamètre à optimiser, le nombre de clusters.

Dans tous les cas la perte d'information peut être rédhibitoire pour les jeux de données de taille modeste à faible.

Sur-échantillonner la classe minoritaire

Le sur-échantillonnage vise à répliquer aléatoirement les observations de la classe minoritaire pour en renforcer le signal. On peut procéder simplement à une duplication avec remise car l'implémentation est facile et le coût numérique limité. Cependant, la copie à l'identique de valeurs parfois non représentatives favorise le sur-apprentissage.

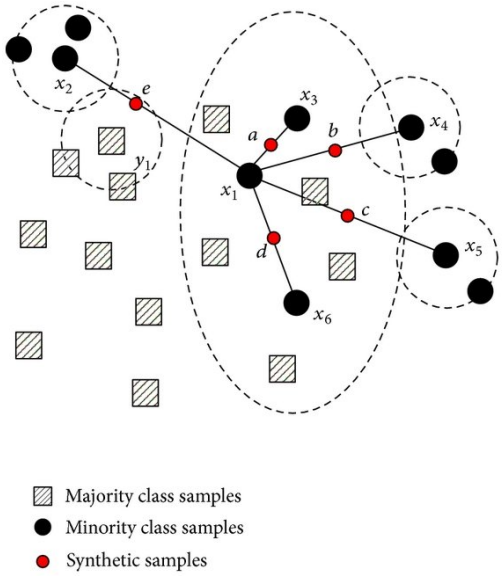


FIGURE 4.7 – Principe de l'algorithme SMOTE. *source : Hu et al. [HL13]*

Pour éviter les points identiques, on préfèrera augmenter les données avec des observations synthétiques. L'algorithme SMOTE [Cha+02] est la méthode la plus populaire. Elle consiste à créer de nouvelles observations, combinaisons linéaires entre des points cibles choisis au hasard et leurs plus proches voisins. Une dose d'aléatoire est ajoutée grâce à un terme multiplicatif tiré entre 0 et 1.

Ici le sur-apprentissage est moindre. En revanche lorsque les plus proches voisins réels d'une observation ne sont pas de la classe minoritaire, l'algorithme n'en tient pas compte. Cela peut augmenter le chevauchement des classes et ajouter du bruit.

Les sur- et sous-échantillonnages des données peuvent être combinés de façon à conserver une cohorte de taille raisonnable tout en limitant le sur-apprentissage.

4.3 Algorithmes de classification étudiés

Les algorithmes de classification supervisés sont nombreux et variés, tant dans leur approche théorique que dans les problèmes spécifiques qu'ils adressent ([FD+14] cite 17 familles distinctes). Nous limitons cette section à la présentation des algorithmes utilisés dans nos études et à la comparaison de leurs qualités et défauts respectifs.

Nous utilisons les implémentations proposées par la bibliothèque Python *scikit-learn* [Ped+11].

4.3.1 La régression logistique

L'algorithme de régression logistique est une régression adaptée à la classification binaire. La probabilité d'un élément d'appartenir à une classe y est supposée suivre une loi logistique : on utilise une sigmoïde (Fig. 4.8) pour associer une probabilité à une prédiction. La probabilité H_0 pour un vecteur de variables X est donnée équation (4.2).

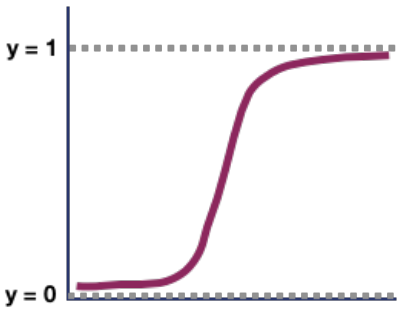


FIGURE 4.8 – Fonction sigmoïde.

$$H_0(X) = \frac{1}{1 + e^{-(\beta_0 + \beta 1 X)}} \tag{4.2}$$

Les valeurs obtenues appartiennent à l'intervalle [0,1]. La fonction de coût (eq. (4.3) est basée sur le principe du maximum de vraisemblance et est minimisée par descente de gradient [AC19].

$$\text{coût}(H_0(X), y) = \begin{cases} -\log(H_0(X)) & \text{if y} = 1 \\ -\log(1 - H_0(X)) & \text{if y} = 0 \end{cases} \tag{4.3}$$

Pour prédire à quelle classe appartient un point, un seuil doit être établi. La probabilité estimée est classée en fonction de son positionnement par rapport au seuil. On peut également déterminer une limite de décision non linéaire.

La régression logistique favorise une variance faible, notamment si on rajoute un terme de régularisation L2 (attention dans ce cas à normaliser l'échelle des variables). Facile à implémenter, paramétrer et interpréter, c'est une bonne première méthode à tester avant de complexifier le modèle de prédiction. Toutefois, elle est plus robuste quand il n'y a pas d'attribut corrélé ou inutile et s'adapte globalement mal à l'augmentation du nombre de variables. Lorsque les données ne sont pas séparables linéairement, il vaut mieux tester les arbres de décision.

4.3.2 Les forêts aléatoires

L'algorithme des forêts aléatoires (***R**andom **F**orests*) a été développé en 2001 par Leo Breiman [Bre01]. C'est une méthode dite ensembliste basée sur le vote de plusieurs éléments de base simples, les arbres de décision.

Arbre de décision

Un arbre est un graphe orienté dont les arcs sont les branches, les sommets sont les nœuds (issus d'une seule branche) et les extrémités des branches sont les feuilles. Chaque nœud d'un arbre de décision constitue un test sur une des variables [SL91]. L'ensemble des tests partitionne le jeu d'entrainement en sous-groupes, eux mêmes divisés par les nœuds suivants (Fig. 4.9a).

A chaque nœud, l'objectif est de déterminer quelle variable doit être choisie et quel test doit lui être appliqué pour découper le sous-ensemble. Dans le cas d'une variable quantitative le test revient à déterminer un seuillage optimal. Ces deux choix s'effectuent de façon à

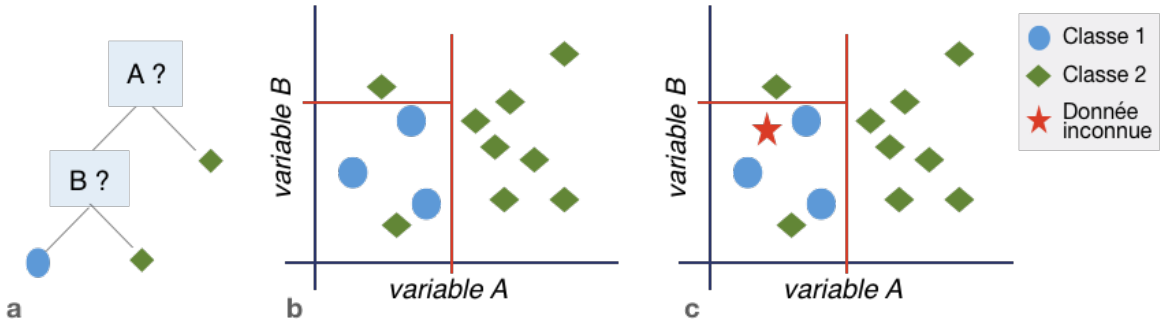


FIGURE 4.9 – Arbre de décision. (a) Construction. (b) Entrainement. (c) Prédiction d'une donnée inconnue.

maximiser un critère d'homogénéité ou à minimiser un critère d'entropie des groupes ainsi générés. Par exemple, l'indice de diversité de Gini [Gin21] pour un sous-ensemble renvoie la probabilité p_i de mal étiqueter un élément i choisi au hasard, si on le classe aléatoirement en suivant la distribution des étiquettes C du sous-ensemble (eq. 4.4).

$$G = \sum_{i=1}^{C} p_i * (1 - p_i) \tag{4.4}$$

Il est nécessaire de déterminer des conditions pour stopper le partitionnement et éviter le sur-apprentissage avec de multiples branches à une feuille. On fixe donc des seuils qui sont des hyperparamètres de l'algorithme : pourcentage d'homogénéité satisfaisant, taille minimale d'un sous-groupe, profondeur maximale de l'arbre, nombre total maximal de subdivisions, etc.

Quand l'arbre est construit, chacune de ses feuilles se voit assigner une prédiction et une probabilité. La première est une étiquette qui correspond à la valeur majoritairement représentée dans son groupe d'appartenance (vote). La seconde est la proportion d'éléments positifs du groupe.

Lors de la prédiction d'une nouvelle observation, ses réponses successives aux tests des nœuds lui font parcourir une sélection d'embranchements. La classe prédite et la probabilité de la prédiction sont celles de la feuille qu'elle atteint au terme du parcours.

Un classifieur constitué d'un arbre seul est limité car très sensible au bruit dans un jeu de données. Il aura une nette tendance au sur-apprentissage.

Forêts

L'algorithme des forêts aléatoires tire profit de la construciton d'une suite d'arbres de décision (Fig. 4.10). Chaque arbre est différent car développé à partir d'un sous-ensemble de variables et d'un sous-ensemble de données tirées aléatoirement avec remise. Le nombre de variables et d'échantillons conservés par arbre peut être optimisé et comme souvent, il n'existe pas vraiment de consensus quant à la façon de déterminer ces valeurs optimales [1].

1. *In practice the best values for these parameters will depend on the problem, and they should be treated as tuning parameters.* [Has09]

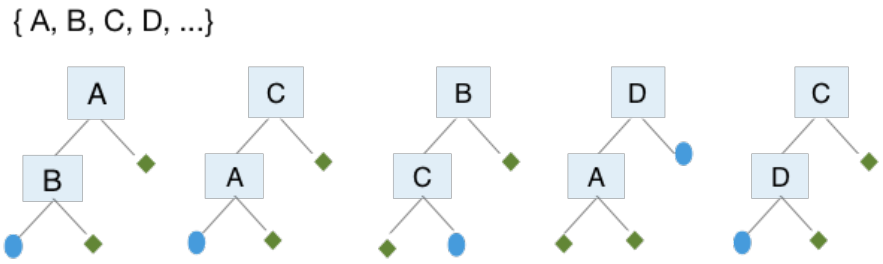


FIGURE 4.10 – Une forêt aléatoire est un ensemble d'arbres votant.

Pour chaque nouvelle observation à tester, la RF renvoie une prédiction déterminée par les votes de chacun de ses arbres pondérés par les probabilités calculées. La probabilité de la prédiction finale est définie par la moyenne des probabilités obtenues par les arbres.

Utiliser une collection d'arbres comme dans les RF rend la méthode plus robuste au bruit. Chaque arbre ne voit qu'une sous-partie des échantillons et de leurs variables, ce qui garantit que le classifieur final est moins sujet au sur-apprentissage. En contrepartie, les modèles RF sont plus compliqués à analyser.

Les forêts peuvent donner une estimation fiable de l'importance relative de chaque caractéristique dans la prédiction en moyennant leur participation respective à l'optimisation de l'homogénéité des branches. On note toutefois que les variables corrélées entre elles ont autant de chance d'être sélectionnées les unes que les autres et en cela, voient leur importance relative diminuer [GMSP13].

4.3.3 Les k plus proches voisins

La méthode des k plus proches voisins (***k-N**earest **N**eighbours* [CH67]) est un algorithme de classification supervisé (également utilisable pour la régression). Basés sur un principe intuitif, les kNN permettent de modéliser des relations non-linéaires entre plusieurs points. Aucune étape d'entrainement spécifique n'est requise.

Pour chaque nouvelle observation à classer, l'algorithme cherche les k points déjà étiquetés dont les coordonnées sont les plus proches (distance euclidienne) dans l'espace des caractéristiques (Fig 4.11). Ces "plus proches voisins" votent alors pour assigner une classe à la nouvelle observation.

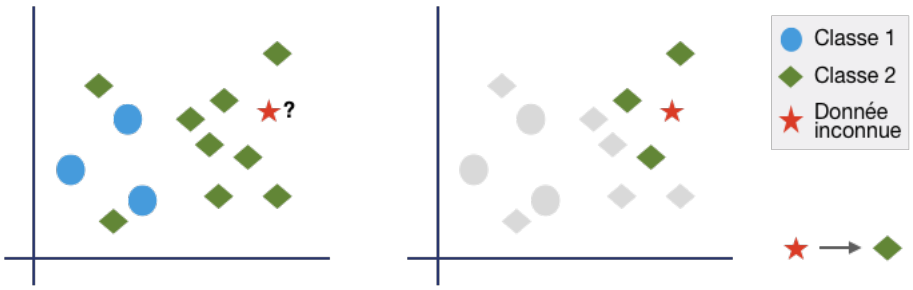


FIGURE 4.11 – Méthode des kNN avec k = 3 pour deux variables.

Le vote d'un point étiqueté est pondéré par l'inverse de sa distance avec le point à classer.

En revanche cette méthode n'applique aucune pondération sur les différentes variables de l'espace de description des points. Elles sont donc toutes considérées d'égale importance dans la recherche des plus proches voisins. Ce défaut est amplifié si les valeurs des variables ont des ordres de grandeur différents et n'ont pas été normalisées entre elles (voir section 4.2.2).

On note qu'en cas de jeu de données non équilibré, la classe la plus fréquente aura statistiquement plus de chances de se trouver parmi les plus proches voisins et donc d'orienter fortement le vote. Cette caractéristique est en partie contrebalancée par la pondération des distances, mais il reste assez fréquent d'éprouver des difficultés à obtenir une prédiction de la classe minoritaire dans les cas d'asymétrie poussée. La figure 4.12 montre un exemple où la classe dominante pèse sur la prédiction : la donnée inconnue est assignée à la classe 2 potentiellement à tort.

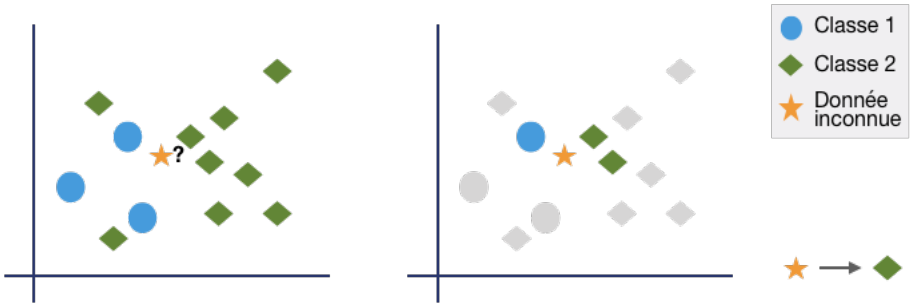


FIGURE 4.12 – Nouveau cas ambigu : la classe dominante influe sur le résultat.

4.3.4 Les séparateurs à vastes marges

Les séparateurs à vastes marges ou machines à vecteurs de support (**S**upport **V**ector **M**achine) sont basés sur le principe de classification à marges maximales [VL63]. Cet algorithme vise à trouver l'hyperplan optimal pouvant séparer un jeu de données en deux groupes. L'hyperplan optimal est celui dont les distances avec les observations sont maximales : il réduit ainsi au mieux l'erreur de généralisation (Fig. 4.13). Les marges sont définies par l'espace entre l'hyperplan et les points les plus proches, les points de support. L'hyperplan ne dépend donc que de ces points particuliers (s'ils bougent, l'hyperplan aussi).

Dans les cas où les données ne sont pas séparables linéairement, on utilise une adaptation de l'algorithme qui autorise un nombre (limité) de mauvaises classifications : le *soft margin classifier*. Cette méthode passe par l'augmentation artificielle de l'espace de dimensions avec une fonction noyau (*kernel*) qui re-décrit les données de façon à les rendre linéairement séparables.

Les SVM sont réputées compétentes même avec un nombre de dimensions élevé voire supérieur au nombre d'observations [GBG01]. Cependant, les résultats sont difficiles à interpréter en cas d'utilisation d'un *kernel* car les données ont été transformées.

4.3.5 Les réseaux de neurones

Nous avons jusque là introduit des algorithmes d'apprentissage statistique classiques. Nous terminons cette présentation des méthodes de classification par les réseaux de neurones

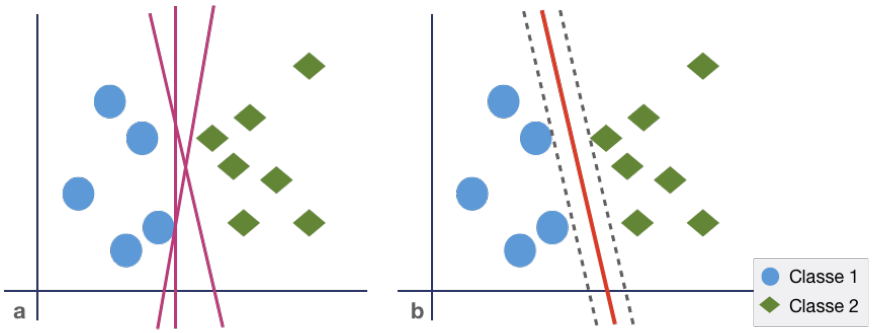


FIGURE 4.13 – SVM sur deux variables. (a) Le jeu de données peut être divisé par une infinité de plans. (b) En rouge, l'hyperplan donnant les marges maximales.

artificiels (**A**rtificial **N**eural **N**etwork), à la base de la famille des algorithmes d'apprentissage profonds (*deep learning*). L'un des plus simples, le perceptron multi-couches, est représenté Figure 4.14.

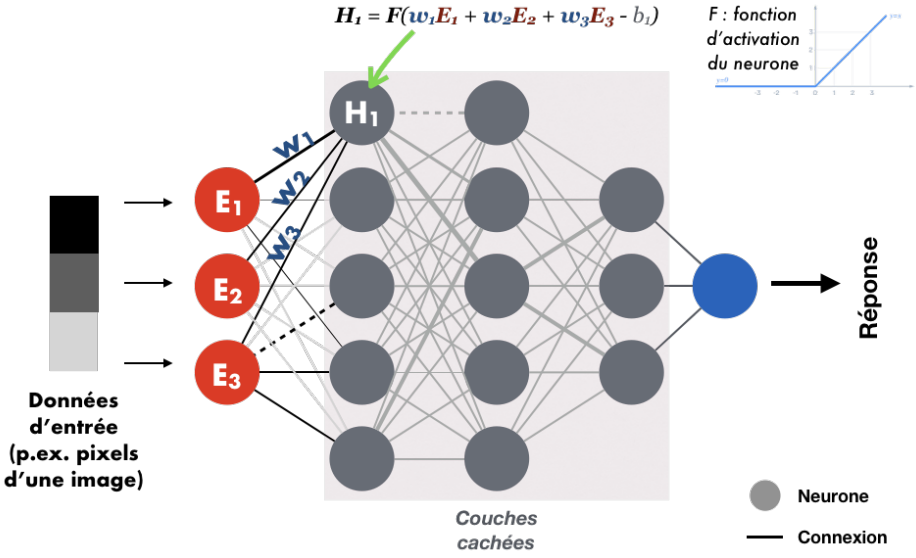


FIGURE 4.14 – Réseau de neurones à 3 couches cachées. Les données des neurones d'entrée sont envoyées dans la succession des nœuds des couches cachées. Chaque nœud opère une combinaison pondérée non linéaire des valeurs de la couche précédente avant de renvoyer le résultat au suivant.

Un neurone artificiel (ou nœud) est une fonction mathématique non linéaire dont l'entrée et/ou la sortie est connectée à d'autres neurones, de façon à former un réseau organisé en couches. La valeur de chaque nœud est une combinaison linéaire pondérée des valeurs entrantes. L'objectif de l'apprentissage est d'estimer la valeur des coefficients de façon à retrouver la bonne sortie. Les données d'apprentissage constituent chacune un neurone de la couche d'entrée et la prédiction finale est renvoyée par le neurone de sortie.

Les méthodes d'apprentissage profond montrent une puissance supérieure à celle des algorithmes d'apprentissage classique, au prix d'un grand nombre de paramètres à estimer,

d'une structure complexe bien plus difficile à analyser et d'une quantité décuplée de données nécessaires à l'apprentissage.

4.4 Stratégie d'apprentissage

Le schéma 4.1 résume l'ensemble de la chaîne de traitement et d'apprentissage comme une boîte noire où entrent des données et sortent des prédictions. Le processus n'est pourtant pas totalement opaque. La mise en place d'une chaîne de traitement maîtrisée entre les *input* et les *output* aide à comprendre et à améliorer les prédictions.

Nous présentons ici le pipeline d'apprentissage que nous avons construit, résumé Fig. 4.15.

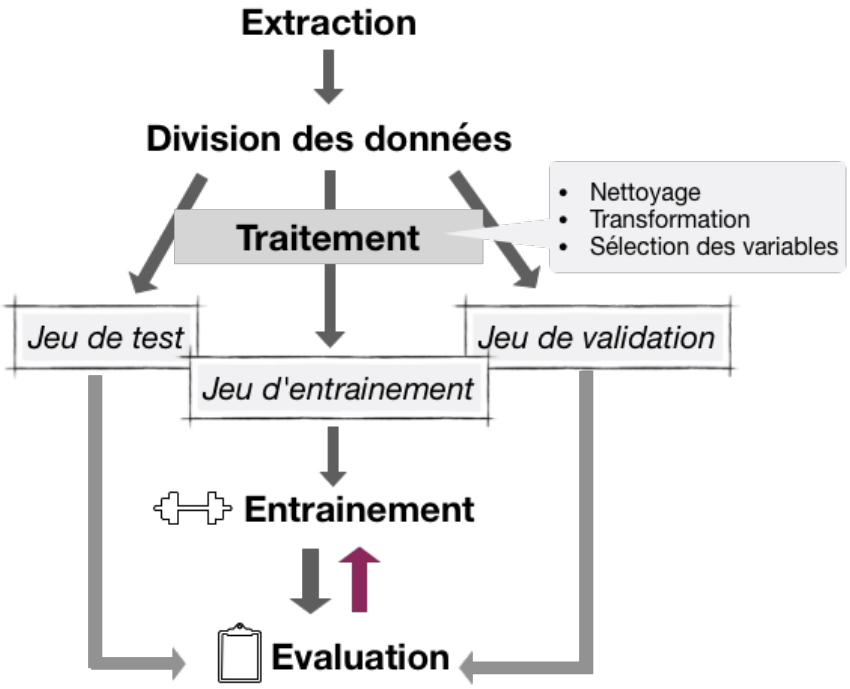


FIGURE 4.15 – De l'extraction des données à la validation du modèle : la chaîne de traitement.

4.4.1 Du jeu de données à la validation de la prédiction

L'objectif fondamental du ML est de construire un modèle dont la précision est basée sur la qualité de son schéma de prédiction sur des données inconnues. On peut simuler la prédiction de données inconnues en divisant le jeu de données existant en sous-ensembles d'apprentissage et d'évaluation. À la phase d'extraction vue aux chapitres 1 et 2 (préparation des données brutes et extraction des caractéristiques) et à la phase d'observation, de nettoyage et de transformation des observations, on ajoute donc une étape de ségrégation des données.

La division peut s'effectuer de façon aléatoire, ou temporelle au fur et à mesure de l'acquisition. Il est également utile de rééchantillonner plusieurs fois les données, avec remise (*bootstrap*) ou non (validation croisée, détaillée section 4.4.2).

L'apparition d'un jeu de données d'entrainement entraîne une modification du processus de traitement et de sélection des variables et de leurs valeurs (cf. section 4.2.1 et 4.2.2). Le pré-traitement ne peut être appliqué d'un bloc sur toutes les données sous peine **d'inclure l'information des données de test** dans les statistiques extraites pour ces transformations (moyenne, médiane etc.) Les données de test sont alors implicitement apprises par le modèle alors qu'on souhaite simuler la prédiction de données complètement inconnues.

La complétion des données manquantes se fait donc à partir des valeurs fournies par les données d'entrainement exclusivement : calcul de la médiane, entrainement des forêts de *MissForest*, etc. De la même façon le jeu de test est normalisé par la méthode de *standard scaling* avec les moyennes et écart-types du jeu d'entrainement uniquement.

Si une sélection des variables ou une réduction des dimensions est souhaitée, là encore seules les données d'entrainement doivent être prises en compte dans les calculs nécessaires au choix ou à la construction des caractéristiques.

Note : À l'heure actuelle le module PCAMix *évoqué en section 4.2.3 est disponible en R et peut donc être utilisé pour transformer les données en amont du processus d'apprentissage. En revanche pour l'employer au sein du pipeline comme nous venons de le décrire, il nécessite une transcription préalable en Python qui n'a pas encore été réalisée.*

4.4.2 La validation croisée

La validation croisée ou *cross-validation* est un principe utilisé en statistique pour estimer la fiabilité d'un modèle [Zie03]. Il se base sur l'échantillonnage systématique de l'ensemble du jeu de données, pour estimer sa capacité à généraliser ses prédictions sur des données inconnues. On limite ainsi le problème de sur-apprentissage. La *cross-validation* est généralement utilisée quand il n'y a pas d'ensemble de validation indépendant disponible.

On regroupe sous ce terme deux principaux types d'algorithmes.

"k-fold cross-validation" où le jeu de données de taille N est découpé en k échantillons de tailles similaires. On effectue plusieurs cycles d'apprentissage / test en testant sur un des échantillons et en apprenant sur les $k-1$ autres. De cette façon, chaque donnée sera forcément testée une unique fois et utilisée pour construire un modèle $k-1$ fois. La figure 4.16 montre la succession des étapes.

"leave-one-out cross validation" est l'équivalent d'une *k-fold cross validation* pour un k égal à la taille du jeu de donnée. Seul un élément est placé dans le jeu de test à chaque tour. La méthode *leave-p-out* est une généralisation de ce principe avec p élèments dans le jeu de test [Cel08]. Ces deux approches sont exhaustives, au prix d'un temps de calcul plus long que la méthode précédente.

La validation croisée peut être répétée plusieurs fois avec des découpages aléatoires du jeu de données pour donner plus de puissance à l'apprentissage. Les résultats de chaque itération sont moyennés (ou combinés par un vote) pour estimer la performance de prédiction du modèle tout en diminuant la variance.

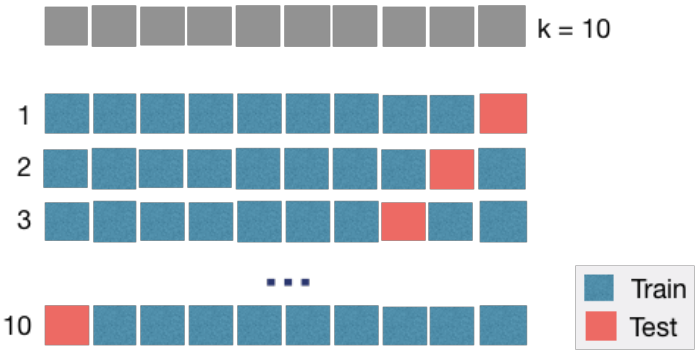


FIGURE 4.16 – *k-fold cross-validation* pour k = 10.

En cas de jeu de données déséquilibré, on préfère stratifier les échantillons pour éviter les tirages malencontreux : chacun d'entre eux contient alors une proportion d'observations positives et négatives similaire à celle du jeu de données complet. À noter, la stratification n'est pas possible avec la méthode *leave-one-out*.

La valeur k est un paramètre critique à déterminer et en pratique elle est souvent choisi empiriquement. Généralement comprise entre 5 et 10 dans la littérature, il est également envisageable de choisir un diviseur entier de la taille du jeu de données ou de considérer k comme un hyperparamètre à régler automatiquement [Ang+12].

4.4.3 Ajuster les hyperparamètres

L'ensemble des algorithmes présentés nécessite une quantité plus ou moins importante de paramètres à choisir avant d'itérer le processus d'apprentissage. Ce sont les hyperparamètres : nombre de voisins des kNN, profondeur maximale des arbres des RF, quantité maximum de mauvaises classification tolérée par les SVM etc. Leur réglage va évidemment affecter la précision de prédiction ou la durée de l'entrainement et les valeurs par défaut ne sont pas toujours optimales.

Un ajustement manuel peut s'avérer fastidieux et peu efficace. On a donc recours à une approche systématique ou aléatoire de tests parmi le sous-ensemble de possibilités de l'espace des hyperparamètres. Dans le cas exhaustif, l'algorithme génère toutes les combinaisons possibles : c'est le *grid-search*. Elles sont testées sur un jeu de données indépendant (quand le volume de données disponible le permet), ou sur un sous-ensemble du jeu de données d'entrainement (si elles sont plus rares.) Les observations des jeux de test ne sont jamais incluses, sous peine d'introduire trop de connaissances dans le modèle et de biaiser les performances. La meilleure combinaison est déterminée par une métrique (fiabilité, AUROC ... cf. section 4.5) moyennée après validation croisée.

Pour diminuer les temps de calcul, on peut fouiller l'espace des hyperparamètres aléatoirement. Nous privilégions tout de même le *grid-search* dans nos études car le nombre d'algorithmes à tester est raisonnable et que l'opération n'est à effectuer qu'une fois. Nous choisissons généralement une validation croisée à trois itérations pour limiter le temps de calcul.

La validation croisée, comme celle utilisée pour optimiser les combinaisons d'hyperparamètres, est validée par un score à optimiser. On peut s'intéresser à l'erreur de prédiction ou aux probabilités, par exemple en étudiant l'aire sous la courbe ROC. Les nombreuses métriques possibles font l'objet de la section suivante.

4.5 Estimer l'erreur de prédiction

La validation des résultats des modèles de classification s'effectue au moyen de scores qui mesurent la qualité des prédictions binaires. Une partie des métriques est basée sur le décompte des vrais/faux négatifs ou positifs dans les résultats (Table 4.1) :

		Valeur prédite +	**Valeur prédite** -
Valeur réelle	+	**T**rue **P**ositive	**F**alse **N**egative
Valeur réelle	-	**F**alse **P**ositive	**T**rue **N**egative

TABLE 4.1 – Matrice de confusion

D'autres métriques évaluent plutôt les valeurs correspondant aux probabilités d'appartenir à la classe positive de chaque observation. La prédiction binaire est issue d'un seuillage de cette probabilité. On peut utiliser un seuil à 0.5 quand la distribution des classes est similaire entre les ensembles d'entrainement et de test, comme c'est le cas en validation croisée stratifiée.

4.5.1 Calcul de la fiabilité

La fiabilité (ou *accuracy*[2]) est la mesure la plus simple, qui décrit la proportion de bonnes prédictions sur le total des éléments à prédire.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

On s'attend à avoir une fiabilité forte sur le jeu de données d'apprentissage, signe d'un *fitting* correct du modèle. Cependant, un score élevé sur les données de test ne suffit pas à confirmer la bonne performance du classifieur car cette métrique ne tient pas compte des

2. Le terme *accuracy* peut aussi se traduire en français par le mot "précision". Une confusion est donc possible avec le score "*precision*" lui même traduit de la même façon (cf. section 4.5.4).

fausses prédictions. C'est surtout vrai lorsque le jeu de données n'est pas équilibré : le score ne fait que refléter la distribution sous-jacente des classes, c'est le "paradoxe de l'*accuracy*" [VAPM14]. Il est possible de pondérer le résultat par l'inverse de la distribution de chaque classe.

En cas de validation croisée, moyenner le score de chaque itération ou le calculer sur la prédiction finale de chaque élément revient au même.

4.5.2 Courbe de calibration

On teste la pertinence des probabilités renvoyées en étudiant leur calibration : un modèle est dit bien calibré si pour l'ensemble des éléments ayant obtenu une probabilité p d'appartenir à la classe positive, une proportion p d'entre eux est effectivement positive. En d'autres termes, en prenant le cas d'un pronostic clinique : si on observe 100 patients à qui un classifieur prédit 80% de chances de rechuter, on s'attend à ce qu'il y ait environ 80 d'entre eux qui présentent effectivement un évènement. Si le modèle testé suit ce principe, il est dit "bien calibré".

Une courbe de calibration représente les probabilités prédites (y) en fonction des probabilités réelles (x) (ex. Fig. 4.17). Ces dernières sont calculées pour chaque sous-ensemble d'observations partageant le même score. Si le jeu de données est trop petit pour avoir suffisamment de scores identiques pour former ces sous-ensembles, on le découpe en classes de scores similaires. Les éléments réellement positifs et négatifs sont comptés : leur ratio donne la probabilité réelle du sous-ensemble. De ce fait, un déséquilibre dans la répartition des classes du jeu de donnée sera répercuté sur sa courbe de calibration.

Un modèle parfaitement calibré est représenté par la diagonale $y = x$.

Une courbe de calibration montre le biais d'un classifieur et pas la qualité de ses prédictions : un modèle non biaisé n'est pas un modèle parfait. Si on donne à tous les éléments appartenant à la classe positive une probabilité de 0.6 et à tous ceux de la classe négative une probabilité de 0.4, la discrimination est parfaite mais la calibration très mauvaise.

La qualité de la calibration est résumée par le score de Brier (BS) [GW50], qui calcule la différence entre la probabilité prédite et la valeur réelle pour toutes les observations. L'objectif est donc de minimiser ce score.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(\text{valeur\_prédite} - \text{valeur\_réelle})^2 \tag{4.6}$$

4.5.3 Courbe ROC

La valeur de retour (0 ou 1) d'un modèle est déterminée à partir des probabilités en fixant un seuil de discrimination permettant de les interpréter. Ce seuil peut être choisi en fonction du problème à optimiser, notamment le compromis entre faux positifs et faux négatifs. Si le seuil de discrimination est élevé, il est plus compliqué d'obtenir des positifs, vrais ou faux. S'il est bas, le nombre de prédictions négatives diminue.

La fonction efficacité du récepteur, ou courbe ROC (***R**eceiver **O**perating **C**haracteristic curve*) utilise les probabilités obtenues par les modèles pour évaluer et comparer leur efficacité

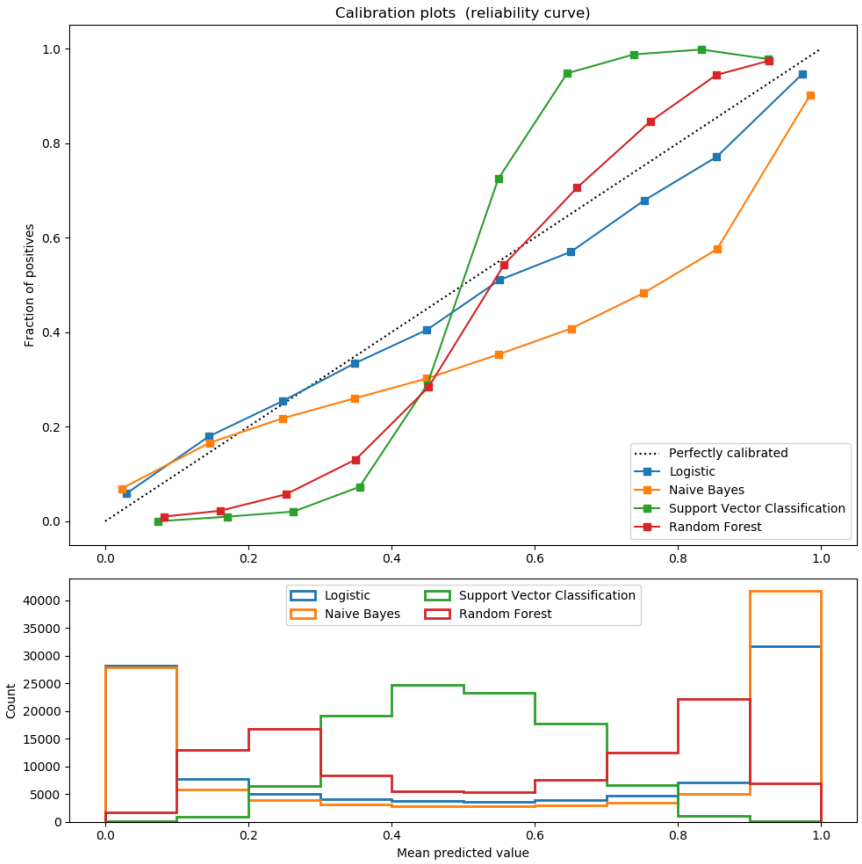


FIGURE 4.17 – Exemples de courbes de calibration. La régression logistique donne des probabilités très bien calibrées. Les RF renvoient rarement des probabilités qui valent 0 ou 1 puisqu'elles moyennent celles de plusieurs arbres. *source : scikit learn*

de classification lorsque ce seuil de discrimination varie. La courbe ROC trace le taux de vrais positifs (***T**rue **P**ositive **R**ate*) en fonction du taux de faux positifs (***F**alse **P**ositive **R**ate*) (Fig. 4.18).

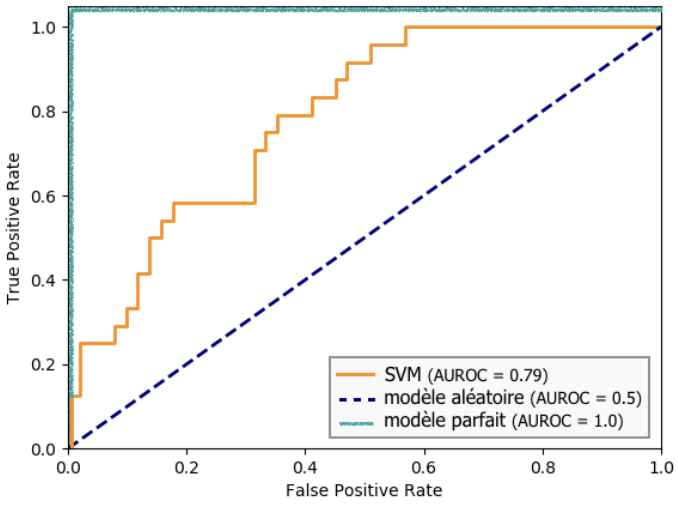


FIGURE 4.18 – Courbes ROC d'une classification avec un modèle SVM, avec un modèle purement aléatoire et avec un modèle parfait.

Le premier est également appelé sensibilité :

$$TPR = \frac{TP}{TP + FN} = sensitivity \tag{4.7}$$

Le second est lié à la spécificité :

$$FPR = \frac{FP}{FP + TN} = 1 - \frac{TN}{FP + TN} = 1 - specificity \tag{4.8}$$

En d'autres termes, la courbe représente le taux de réussite (cibles correctement atteintes) en fonction du pourcentage de fausses alertes (dommages collatéraux).

L'aire sous la courbe (***A**rea **U**nder the **ROC** curve*) synthétise la qualité du modèle quel que soit le seuil. Les bons classifieurs sont supposés renvoyer en moyenne une probabilité beaucoup plus élevée pour un élément positif choisi aléatoirement que pour un élément négatif. Ils sont représentés par des courbes approchant le coin supérieur gauche du graphe et leur AUROC converge vers 1. La diagonale du graphe a une AUROC = 0.5 et représente un modèle purement aléatoire, non informatif.

Dans le cas d'une cross-validation, il y a deux moyen de calculer l'AUROC totale :

1. En construisant une seule ROC issue de l'ensemble des probabilités individuelles obtenues avec tous les sous-ensembles.
2. En construisant un nombre k de ROC pour chacun des k sous-ensembles de validation croisée et en moyennant les k AUROC obtenues.

La première méthode présente un inconvénient. Si les probabilités issues de sets différents sont analysées ensemble, il faut que les classifieurs construits à partir de chaque sous-ensemble soient tous calibrés de façon identique. En effet, si les probabilités obtenues n'ont pas la même signification selon le sous-ensemble sur lequel est construit le modèle, leur combinaison en une seule ROC peut donner des AUROC sous-estimées [FS10]. C'est le cas par exemple des classifieurs qui ne renvoient pas une vraie probabilité mais plutôt un score (RF, SVM etc.). Nous utiliserons donc l'AUROC moyennée de la solution 2.

On note que l'aire sous la courbe ROC peut induire en erreur en présentent une vue d'autant plus optimiste des performances d'un modèle que le jeu de données est déséquilibré [DG06 ; SR15]. Ces courbes sont donc à analyser avec attention voire à compléter avec d'autres mesures pour éviter les interprétations incorrectes.

Pour terminer, nous soulignons que la calibration d'un modèle ne change pas sa courbe ROC. La ROC essaie de discriminer au mieux les classes prédites en vérifiant que les probabilités obtenues sont bien ordonnées. La courbe de calibration vise à discriminer les probabilités en vérifiant qu'elles sont représentatives de la distribution réelle des observations entre les deux classes.

4.5.4 Précision et rappel

Les dernières métriques d'évaluation des modèles de classification que nous présentons sont la précision (*precision*, **PR**) et le rappel (*recall*, **RE**) La précision décrit la quantité de cas réellement positifs parmi l'ensemble des cas prédits positifs par le modèle ($\rightarrow$ *quelle proportion d'éléments sélectionnés est pertinente ?*) On lui donne aussi le nom de valeur prédictive positive (***P****redictive* ***P****ositive* ***V****alue*).

$$précision = \frac{TP}{TP + FP} = PPV \tag{4.9}$$

Le rappel (ou *recall*, TPR, Sensibilité) est la proportion de cas prédits positifs parmi l'ensemble des cas réellement positifs ($\rightarrow$ *quelle proportion d'élément pertinents a été sélectionnée ?*)

$$rappel = \frac{TP}{TP + FN} = TPR = sensitivity \tag{4.10}$$

Il existe plusieurs manières de combiner les deux scores pour l'analyse.

* Le score F1 est leur moyenne harmonique.

$$F1 = 2 * \frac{précision * rappel}{précision + rappel} \tag{4.11}$$

Contrairement à l'AUROC, [FS10] conseillent de calculer le F1 directement sur l'ensemble des prédictions (TP, TN, ...) individuelles obtenues à la fin de la validation croisée plutôt que de moyenner les F1 obtenus à chaque *fold*.

* Les courbes *précision-rappel* (PR) affichent la précision en fonction du rappel pour des seuils de discrimination différents, comme le fait la courbe ROC (Fig. 4.19).

Un modèle non instructif a une courbe définie par une ligne horizontale, la constante y (éq. 4.12).

$$y = (TP + FN)/(TP + FN + TN + FP) \tag{4.12}$$

Plus la courbe dépasse y et se rapproche de la ligne $f(x) = 1$, plus le modèle est informatif.

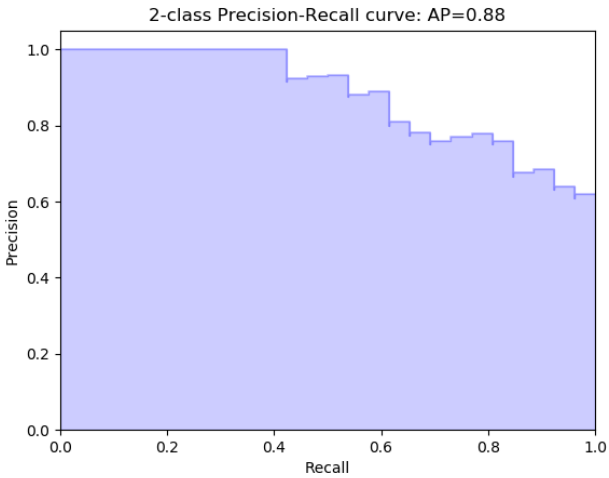


FIGURE 4.19 – Courbe précision-rappel d'un classifieur SVM linéaire. *source : scikit-learn*

Saito et al [SR15] insistent sur l'apport des courbes précision-rappel et des métriques associées dans l'évaluation des jeux de données déséquilibrés, alors que les courbes ROC leur sont très souvent préférées dans les publications en médecine, biologie et bioinformatique. Ces deux mesures se focalisent sur la prédiction correcte de la classe positive alors que l'AUROC inclue également l'information associées aux vrais négatifs (le FPR).

On peut synthétiser le graphe PR avec l'aire sous la courbe comme pour la ROC ou calculer sa précision moyenne (***A**verage **P**recision*). Il s'agit de la somme des moyennes pondérées des précisions obtenues à chaque seuil n utilisé pour la courbe PR :

$$AP = \sum_n (rappel_n - rappel_{n-1}) * précision_n \tag{4.13}$$

Un classifieur aléatoire a une AP égale à la proportion d'éléments positifs dans le jeu de données. Une bonne prédiction donne une AP proche de 1.

Cette section termine la présentation des principales méthodes de traitement des données, d'apprentissage statistique et de validation des résultats que nous utilisons dans nos études cliniques. Le chapitre suivant expose un premier cas d'application de classification pour l'évaluation précoce de la réponse au traitement néo-adjuvant des sarcomes de haut grade.

Chapitre 5

Prédire la réponse au traitement des sarcomes avec des marqueurs delta-radiomiques IRM

Sommaire

Dans ce chapitre seront présentés les travaux effectués sur la prédiction de la réponse au traitement de sarcomes de haut grade traités par chimiothérapie néo-adjuvante grâce aux caractéristiques radiomiques. Cette étude a été réalisée en association avec Amandine Crombé, radiologue à l'Institut Bergonié et doctorante dans l'équipe Inria MONC. Elle a fait l'objet d'une publication dans le journal Journal of Magnetic Resonance Imaging intitulée *T2-Based MRI Delta-Radiomics Improve Response Prediction in Soft-Tissue Sarcomas Treated by Neoadjuvant Chemotherapy* [Cro+18b].

Ce chapitre s'attachera entre autres à présenter la contribution de la thèse dans le traitement des images, l'extraction des descripteurs radiomiques et la prédiction par apprentissage statistique.

5.1 Contexte : les sarcomes des tissus mous de haut grade

5.1.1 Sarcomes

Les sarcomes des tissus mous (***S**oft **T**issu **S**arcoma*) des membres et du tronc constituent un groupe hétérogène de tumeurs mésenchymateuses rares et malignes ayant une structure complexe. Elles se développent à partir des tissus conjonctifs et de soutien, muscles, graisses, vaisseaux sanguins et lymphatiques, nerfs, tendons. On les retrouve souvent dans les membres, parfois dans l'abdomen ou le thorax [Bre05]. Les sarcomes surviennent aussi bien chez l'adulte que chez l'enfant, à raison d'environ 4000 nouveaux cas diagnostiqués par an en France [Hon+15].

FIGURE 5.1 – Plusieurs types et localisations de STS, observés en T2.// *source : Institut Bergonié*

Un sarcome de haut grade est formé de cellules cancéreuses peu différenciées ou indifférenciées, peu semblables aux cellules normales. Il se développe rapidement et est plus susceptible de se propager.

5.1.2 Protocole de traitement

Le traitement repose sur la réussite de l'exérèse chirurgicale R0, éventuellement accompagnée de traitements adjuvants. La radiothérapie post-opératoire vise ainsi à diminuer le taux de récidive : même après une chirurgie optimale, des métastases apparaissent chez 35% des patients [AP+85].

Dans deux études cliniques de phase 3 [1], la survie globale et la survie sans évènement ont été améliorées en traitant les patients avec une chimiothérapie néoadjuvante à base d'anthracyclines (***N**eoadjuvant **A**nthracycline **C**hemotherapy*), ce qui a amené à repenser le traitement standard pour les STS de haut grade avancés localement [Sap+17 ; Gro+17 ; Iss+10]. Son objectif est de réduire le volume tumoral des tumeurs non opérables pour permettre une résection R0 ou R1 [2].

Toutefois, la chimiothérapie est un traitement lourd parfois mal supporté par les patients. Elle est uniquement utilisée dans le cas des STS de haut grade pour lesquels la chirurgie est impossible (si elle est dangereuse ou qu'elle entraîne une perte de fonction) ou complexe. On cherche donc à connaître l'efficacité de la NAC pour chaque patient.

5.1.3 Évaluation de la réponse au traitement

L'appréciation de la réponse tumorale au traitement par NAC se fait classiquement après la résection et se base sur l'étude de la réponse histologique de la pièce opératoire. Il s'agit de compter la quantité de cellules tumorales viables, de cellules fibreuses et de cellules nécrosées sur coupes histologiques (Fig. 5.2a).

Les patients sont considérés "bons" répondeurs à la NAC si le pourcentage de cellules tumorales viables du volume total est strictement inférieur à 10%. Bien que non validé pour les STS, ce seuil s'est révélé pertinent dans une récente étude visant à prédire la survie globale [3] (Fig. 5.2b, [Cou+17]).

L'appréciation de la réponse sur pièce histologique présente toutefois plusieurs inconvénients. La quantification des cellules est visuelle et approximative et la pièce chirurgicale est souvent légèrement incomplète (perte de fluides ou de cellules nécrosées). On obtient donc un résultat imprécis que l'on confronte à un seuil de décision en partie arbitraire. La chronologie du processus reste cependant son défaut majeur : l'effet de la NAC sur la tumeur n'est connu qu'**après** opération. Cela s'avère paradoxal puisqu'il peut être souhaitable d'utiliser cette information pour rediriger un patient mauvais répondeur vers une thérapie moins lourde au plus vite. Une biopsie en cours de traitement pourrait théoriquement être envisagée pour estimer le résultat plus tôt. Il s'agit néanmoins d'une fausse piste puisque le décompte serait alors obtenu sur un échantillon de la lésion, encore plus incomplet que la pièce chirurgicale [Lon12]. Encore plus problématique, l'obtention de l'échantillon est un processus invasif et douloureux. Une démarche éthique, adaptée aux STS et applicable suffisamment tôt dans la routine clinique est donc nécessaire.

La méthode la plus communément utilisée pour évaluer la réponse au traitement est le critère RECIST 1.1, qui repose sur le suivi du plus grand diamètre de la tumeur à l'image-

1. Phase d'un essai clinique correspondant à la comparaison d'un nouveau traitement au traitement standard.
2. Exérèse complète sans marges.
3. Pour les STS, ce seuil a été établi par analogie avec celui utilisé pour classer les ostéo-sarcomes.

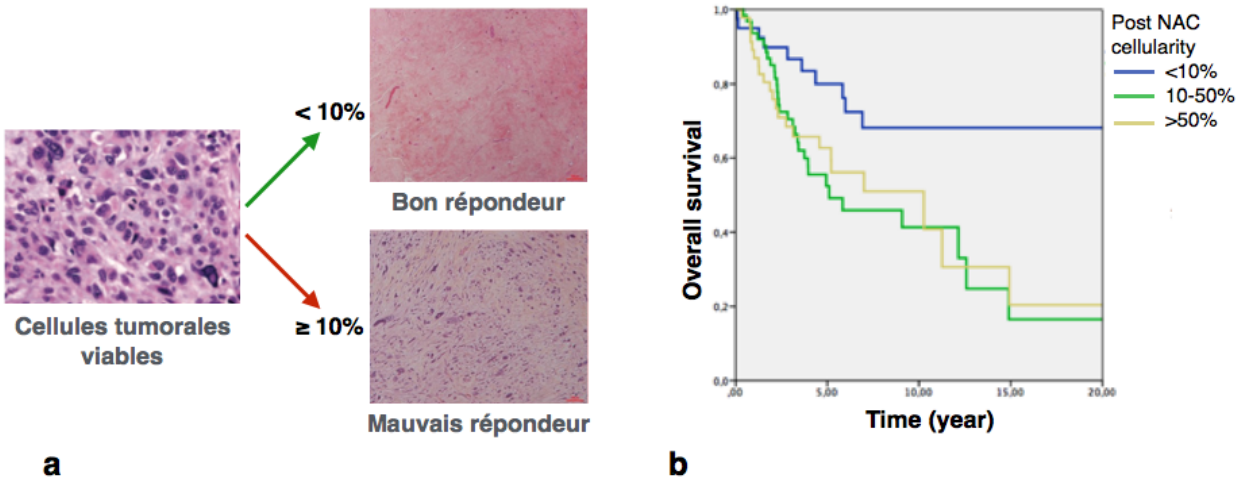


FIGURE 5.2 – Analyse de la réponse histologique au traitement des STS. (a) Coupes histologiques sur STS de haut grade. Les cellules tumorales viables sont visibles en violet sombre. (b) Courbes Kaplan-Meier de la survie globale en fonction du taux de cellules tumorales viables post-NAC. *source : [Cou+17]*

rie médicale (cf introduction, section 2.3). Pourtant, s'il s'agit d'un indicateur standardisé reproductible et facile à mesurer, de nombreuses critiques lui sont adressé. Les principales concernent le caractère arbitraire et invariant au type de maladie des seuils de classification ainsi que la non pertinence du critère pour les traitements ne provoquant peu ou pas de diminution de taille. La correspondance entre les seuils et la survie globale du patient n'a pas non plus été validée [MN+12 ; Hon+15].

Dans le cas des STS, le changement de diamètre de la tumeur ne semble pas être un indice du pronostic oncologique. La Fig. 5.3 montre des exemples de faux positifs et faux négatifs obtenus par l'évaluation du plus grand diamètre. Le critère RECIST est incapable d'intégrer l'information apportée par les modifications architecturales et les altérations vasculaires de la tumeur, qui surviennent avant la modification de la taille.

D'autres solutions ont été proposées dans la littérature. La tomographie d'émission de positons au [18F]-fluorodéoxyglucose (TEP-FDG), le critère de Choi modifié et l'IRM améliorée par contraste dynamique (DCE-MRI) ont montré des résultats encourageants. Ils sont encore sujets à débat : des études ont par exemple mis en évidence que la précision du critère de Choi ou l'évaluation DCE-MRI dépendent du délai d'acquisition de l'image après injection de l'agent de contraste [Cro+18a].

5.1.4 Particularités des STS en IRM et solution proposée

De part leur structure complexe, hétérogène et changeante à l'IRM, les sarcomes se prêtent particulièrement bien aux études de texture et de morphologie : le potentiel des caractéristiques radiomiques est prometteur. Une étude a ainsi montré leur l'intérêt dans la détermination du grade des STS à la microbiopsie [Cor+17a]. Hayano et al. ont mis en évidence l'association entre les paramètres de texture sur CT-scan, la néo-angiogenèse et la survie globale pour des STS sous radiothérapie et bevacizumab [Hay+15 ; Tia+14]. Plus

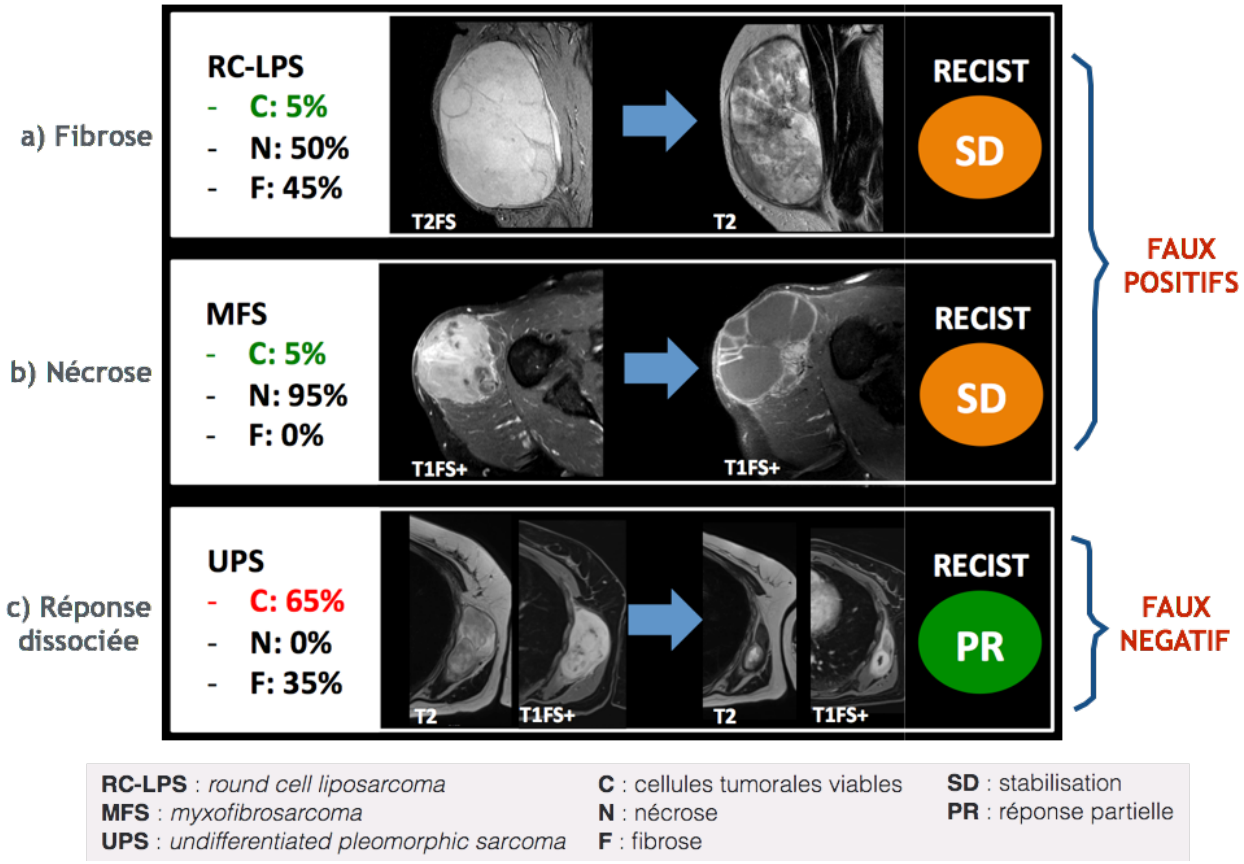


FIGURE 5.3 – Le critère RECIST échoue parfois à correctement évaluer la réponse. (a) et (b) Les cellules tumorales se fibrosent et se nécrosent massivement mais le volume global n'évolue pas. Le critère RECIST conclut que la réponse au traitement est mauvaise. (c) La réponse est dissociée : la lésion d'intérêt décroit et le patient est considéré bon répondeur, alors que d'autres lésions s'agrandissent. *source : d'après Amandine Crombé, JFR 2018.*

récemment, des caractéristiques de textures extraites de l'IRM et de la TEP-FGG ont permis d'identifier dès le diagnostique les tumeurs agressives susceptibles de provoquer une rechute métastatique [Val+15].

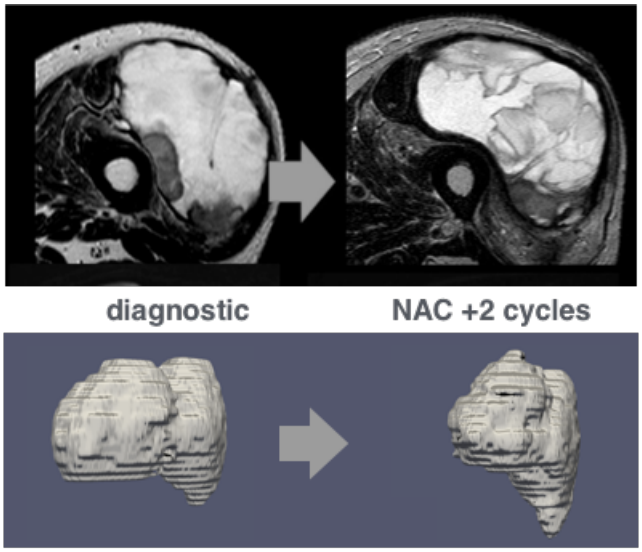


FIGURE 5.4 – Coupe axiale (plus grand diamètre) d'un STS de la cuisse en T2 et son VOI après reconstruction, au diagnostic et après deux cycles de NAC. Le plus grand diamètre axial change peu, mais la tumeur s'allonge le long des muscles et son hétérogénéité augmente.

Pourtant, à ce jour et à notre connaissance, l'application des radiomics à la NAC n'a pas été étudiée. La NAC provoque dans les STS une grande variété de processus nécrotiques et fibrotiques, de saignements ou d'obstruction des vaisseaux, d'apparition de composantes résistantes etc. Ces altérations induisent des changements morphologiques et texturaux visibles à l'IRM dès les premiers cycles de chimiothérapie. La Fig. 5.4 montre un exemple de ces changements sur un patient après seulement deux cycles. Les modifications de la texture et de la morphologie de la tumeur y sont clairement visibles.

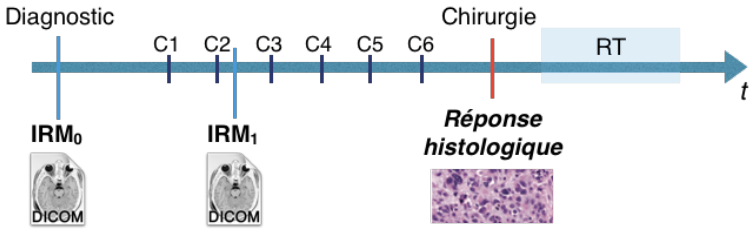


FIGURE 5.5 – Chronologie du traitement des STS. La réponse histologique est tardive et une évaluation précoce à C2 serait plus utile.
(C1, ..., C6) : cycles de NAC. RT : radiothérapie.

A partir de l'imagerie effectuée en routine clinique, c'est à dire des examens IRM réalisés au diagnostic et après le deuxième cycle de NAC, ce projet propose donc de quantifier

l'évolution de l'hétérogénéité et de la forme des tumeur à l'IRM afin de prédire rapidement leur réponse au traitement (Fig. 5.5).

5.2 Matériel

5.2.1 Critères d'inclusion des patients

Ont été inclus tous les patients soignés à l'Institut Bergonié entre Juin 2007 et Juin 2017 pour un STS de grade 3 des membres ou du tronc sans métastase et éligibles à un traitement de chimiothérapie néo-adjuvante à base d'anthracyclines (NAC). Les critères d'inclusion supplémentaires sont :

- une tumeur mesurable à l'IRM,
- 4 à 6 cycles de NAC menés à terme,
- la présence d'un IRM de référence (diagnostic) réalisé 28 jours maximum avant le premier cycle de NAC, noté IRM_0 ou MRI_0 par la suite,
- la présence d'un IRM réalisé entre le deuxième et le troisième cycle de NAC, noté IRM_1 ou MRI_1 par la suite,
- la présence d'une réponse histologique estimée sur pièce opératoire par un pathologiste spécialisé (la référence).

La réponse histologique est considérée bonne si les cellules tumorales viables représentent une quantité strictement inférieure à 10% du total. Les patients ayant une bonne réponse seront par la suite désignés bon répondeurs ou *Good-HR* et mauvais répondeurs ou *Poor-HR* dans le cas contraire.

5.2.2 Paramètres d'acquisition de l'imagerie

Les images ont été acquises au cours de la routine clinique avec différents systèmes IRM à 1.5T. Les paramètres de la machine (bobine, champ de vue etc.) ont été réglés en fonction de la localisation de la tumeur. Les IRM incluent des séquences 2D T2-WI turbo-spin echo (TSE) sans suppression de graisses et T1-WI avant et après injection de gadolinium. Quatre-vingt treize examens (72%) ont été effectués sur un Magnetom AERA, (Siemens Healthineers) avec les paramètres d'acquisitions suivant : TR/TE=6860/120, résolution (2D) = 1x1mm^2, épaisseur de coupe = 4mm.

5.2.3 Caractéristiques du jeu de données final

Nous obtenons une cohorte de 65 patients ayant leur IRM_0 et IRM_1 en T2 ainsi qu'une liste de leurs caractéristiques radiologiques démographiques et sémantiques (voir Annexe B.2). Ces caractéristiques ont été évaluées indépendamment par Amandine Crombé et Michelle Kind, radiologues à l'institut Bergonié (étude de la variabilité inter-observateurs). Elles ont également fait l'objet d'une seconde lecture par A. Crombé 1.5 mois après (étude la variabilité intra-observateur).

Elles comprennent notamment pour chaque IRM :

- le plus large diamètre de la tumeur en mm, son évolution relative et le statut RECIST correspondant,

— le pourcentage du volume de la tumeur constitué de fibrose et/ou de nécrose, classé en trois catégories 0%, $< 50\%$ et $> 50\%$,
— l'évolution de la marge de la tumeur en T1, classée en "bien définie ou mieux définie" contre "mal définie ou moins bien définie",
— l'évolution de l'œdème périphérique en T2, classée en "absence ou diminution" contre "stable ou augmentation",
— l'évolution de la prise de contraste en T1, classée en "absence ou diminution" contre "stable ou augmentation".

5.3 Méthode

5.3.1 Post-traitement des IRM

Un contourage coupe par coupe de la tumeur entière est effectué manuellement sur les IRM T2-WI par A. Crombé, radiologue sénior, avec le gestionnaire de ROI du logiciel Osirix [RSR04].

Nous ré-échantillonnons les coupes par interpolation bi-linéaire pour obtenir un pixel isotrope dans le plan, de résolution $1 \times 1mm^2$ en 2D. Le troisième axe n'est pas transformé pour diminuer l'impact de l'interpolation de grille sur l'extraction des caractéristiques radiomiques.

Nous corrigeons les intensités en T2 de façon à supprimer le biais avec l'algorithme N3 après ré-échelonnage entre des valeurs positives (0 à 100). On note cependant de très faibles valeurs d'artefacts. Un ré-échelonnage supplémentaire entre -1 et 1 est réalisé après coup.

Nous effectuons une normalisation inter-patients par alignement d'histogramme à deux points de référence avec des histogrammes à 100 niveaux de discrétisation. L'examen de la cuisse d'un volontaire sain fourni par l'Insitut Bergonié nous sert de référence, une fois pré-traité selon la même procédure.

5.3.2 Extraction des caractéristiques radiomiques

Les caractéristiques radiomiques extraites sont au nombre de 31 : 6 marqueurs d'intensité, 18 de texture et 7 de forme. Les caractéristiques de texture sont déterminées à partir des matrices de cooccurrence que nous construisons :

— Après discrétisation des niveaux de gris du VOI en classes de taille fixe (égale à 0.05 [4]) entre ses valeurs extrêmes. On obtient entre 15 et 40 classes,
— Pour des voisinages 2D uniquement (angles $\theta = [0^\circ, 45^\circ, 90^\circ, 135^\circ]$). Les valeurs de chaque coupe sont fusionnées dans une seule matrice, mais les caractéristiques calculées sont moyennées sur les quatre orientations (agrégation 2.5D),
— Pour des voisinages de 1, 2 et 5 voxels, ce qui donne trois matrices au final.

Pour chaque caractéristique d'un patient, nous calculons la différence absolue entre les deux IRM, ce qui nous donne des Δ-radiomiques.

4. Unité arbitraire après rééchelonnage entre -1 et 1.

5.3.3 Analyse par apprentissage statistique

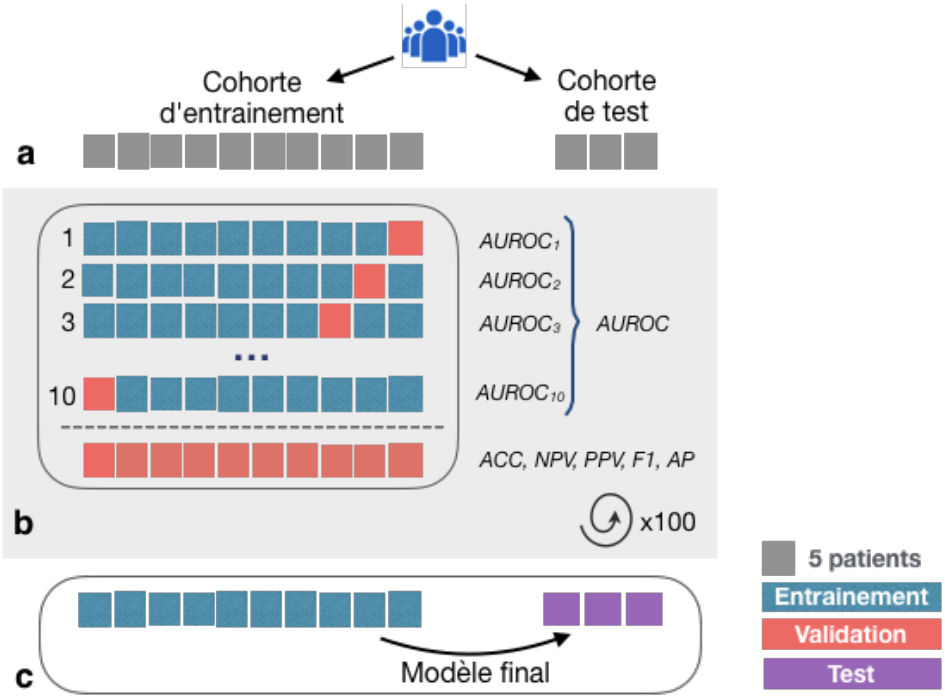


FIGURE 5.6 – Stratégie d'apprentissage et de test. (a) Les données sont séparées en deux cohortes distinctes à 50 et 15 patients. (b) La cohorte d'entrainement est utilisée pour valider les modèles par validation croisée (sur 100 répétitions). Les scores sont calculés à chaque cycle ou directement sur la prédiction de chaque patient. (c) Les modèles sont entraînés une dernière fois sur l'ensemble des 50 patients puis testés sur la seconde cohorte.

Gestion du jeu de données

Afin de mettre en place et de valider les modèles de prédiction, nous divisons le jeu de donnée en deux (Fig. 5.6a). Une cohorte d'entrainement/validation est constituée pour la construction des modèles (50 patients inclus de Juin 2007 à Juin 2016). Une cohorte de test est réservée pour un contrôle final des performances (15 patients de Juin 2016 à Juin 2017, pour lesquels la fin du suivi a eu lieu après le début du projet).

Analyse statistique univariée classique

Des comparaisons entre bons et mauvais répondeurs histologiques dans la cohorte à 50 patients sont effectuées avec des tests de Student ou de Mann-Whitney [5] pour les variables continues. Pour les variables ordinales ou catégorielles, nous déterminons l'association avec la réponse au traitement par des tests du χ^2 et de Fischer. La corrélation entre les caractéristiques est calculée avec le test de rang de Spearman. Tous les tests sont bilatéraux et nous fixons le seuil de signification à $p < 0.05$.

5. En fonction de leur résultat au test de normalité (Shapiro-Wilk).

Réduction du nombre de variables

Nous sélectionnons la meilleure caractéristique au regard de l'analyse univariée (plus faible *p-value*) pour chaque catégorie : caractéristiques sémantiques, mesures de texture/intensité et mesures de forme. Cet ensemble forme une signature à trois variables peu corrélées entre elles. Elle est comparée aux signatures obtenues avec une sélection par régression de type ElasticNet ou avec l'ensemble des caractéristiques significatives à l'analyse univariée.

Algorithmes de classification utilisés

Nous implémentons plusieurs méthodes de classification grâce à la bibliothèque *scikit-learn* : forêts aléatoires (RF), k-plus proches voisins (kNN), machines à vecteurs de support (SVM) et régression logistique (LR). Nous avons préalablement optimisé les paramètres de ces algorithmes sur la cohorte d'entrainement par *grid search* à trois échantillons de validation croisée, en fonction de l'aire sous la courbe ROC.

Stratégie de validation croisée et de test des modèles

Dans un premier temps, nous construisons des modèles avec les différentes signatures et nous les testons par validation croisée à 10 échantillons sur la cohorte à 50 patients uniquement (Fig. 5.6b). Cette *cross-validation* est stratifiée de façon à compenser l'asymétrie de la répartition des patients entre *Good-HR* et *Poor-HR*.

A chaque étape de la validation croisée, les valeurs manquantes du jeu de données sont inférées en utilisant la médiane des mesures des données d'entrainement seulement. De même, nous normalisons toutes les variables par *standard scaling* avec la moyenne et l'écart type des données d'entrainement.

Nous répétons la validation croisée 100 fois pour chaque algorithme, en mélangeant aléatoirement la répartition des patients dans les échantillons. C'est aussi l'occasion de faire varier la graine d'initialisation aléatoire des RF. Comme conseillé par [FS10], l'aire sous la courbe ROC ($AUROC$) est moyennée sur l'ensemble des cycles de validation croisée et les 100 répétitions, tandis que le score *F1*, la valeur prédictive positive (PPV, équivalente à la précision), la valeur prédictive négative (NPV), la sensibilité (SEN, équivalente au rappel) et la spécificité (SPE) sont calculées avec le nombre final de vrais/faux positifs/négatifs. La fiabilité moyenne de prédiction (ACC), le score moyen d'entrainement (*Train*) et l'*average precision* (AP, obtenu avec les probabilités finales de chaque patient) sont également reportés.

Dans un second temps, nous répétons une dernière fois ce pipeline en utilisant les mêmes méthodes de pré-traitement (valeurs manquantes, *scaling*) (Fig. 5.6c). L'objectif est de valider les algorithmes, hyperparamètres et signatures sélectionnés. Nous entrainons les modèles sur l'ensemble de la cohorte à 50 patients et nous les testons sur la cohorte à 15. Les mêmes scores sont reportés pour ce test final.

Dans l'objectif de vérifier la pertinence d'une signature courte à trois variables, l'étape **b** est réitérée en augmentant le nombre de caractéristiques utilisées. Nous ajoutons les variables dans le modèle une à une, dans l'ordre croissant de leur *p-value* à l'analyse univariée.

5.4 Résultats de l'analyse univariée

Caractéristiques patients

Sur 65 patients (dont 27 femmes, avec un âge moyen de 57.9 ± 12.8 ans), 16 (24.6%) sont des bons répondeurs. Les histotypes les plus fréquents sont les sarcomes indifférenciés (50.8%) suivis par les sarcomes myogéniques (leiomyosarcomes and rhabdomyosarcomes, 20%). La majorité est localisée en profondeur (93.8%) dans les membres inférieurs (58.5%). Trente-deux patients (33.8%) n'ont reçu que quatre cycles de NAC au total.

Le jeu de donnée a été réparti en deux cohortes d'entrainement et de test de 50 patients (dont 11 bons répondeurs, soit 22%) et 15 patients (dont 5 bons répondeurs, soit 33%) respectivement. Aucune différence statistique significative n'a été relevée entre ces deux cohortes au niveau des caractéristiques épidémiologiques au diagnostic.

Aucune association n'a été trouvée entre les caractéristiques épidémiologiques mesurées au diagnostic et la réponse histologique.

Caractéristiques radiologiques standards

Le plus large diamètre (%LD) au diagnostic est significativement plus grand chez les bons répondeurs ($146 \pm 66mm$ contre $110 \pm 51mm$, avec $p = 0.038$). Le changement relatif du LD à l'évaluation précoce est également significativement différent entre les bons et mauvais répondeurs ($\breve{}11.2 \pm 20.8\%$ contre $2.9 \pm 19.5\%$, avec $p = 0.027$). Cependant, le statut selon le critère RECIST 1.1 n'est lui-même pas associé avec la réponse histologique ($p = 0.112$) puisque d'après ce critère, la plupart des bons et mauvais répondeurs sont classés comme étant stables (81.3% et 79.6%, respectivement).

Parmi l'ensemble des caractéristiques radiologiques sémantiques, seul le Δ-*Edema* (évolution de la quantité d'œdème péri-tumorale) est associé avec la réponse ($p = 0.003$) tout en ayant une bonne reproductibilité inter- et intra-évaluateur (0.637 et 0.769, respectivement). Ces résultats sont détaillés en Annexe B.3.

On note que des données sont manquantes pour quelques patients pour la définition des marges et la prise de contraste péritumorale en raison de la présence de défaillances légères dans le protocole d'acquisition IRM (acquisition incomplète de l'œdème ou acquisition sur des plans différents entre IRM_0 et IRM_1).

Caractéristiques radiomiques

Dans la cohorte d'entrainement, une analyse univariée montre que l'évolution de quatre caractéristiques d'intensité (premier ordre) et cinq de morphologie sont significativement associées avec la réponse : $\Delta\_Histogram\_Entropy$ ($p = 0.004$), Δ_Std ($p = 0.005$), $T1\_Histogram\_Entropy$ ($p = 0.016$), $\Delta\_Elongation$ ($p = 0.018$), $T0\_Feret\_Diameter$ ($p = 0.019$), $\Delta\_Equivalent\_radius$ ($p = 0.041$), $\Delta\_Flatness$ ($p = 0.044$), $T1\_Kurtosis$ ($p = 0.044$), $\Delta\_Feret\_Diameter$ ($p = 0.047$). Les résultats pour les caractéristiques Δ-radiomiques sont détaillés en Annexe B.4.

L'étude de la matrice de corrélation des caractéristiques significatives, présentée Fig. 5.7 montre des corrélations significatives fortes (valeur absolue du coefficient de Spearman > 0.5)

entre une majorité des caractéristiques de texture de premier ordre. L'évolution de l'œdème est corrélée à certains descripteurs morphologiques, mais pas l'élongation.

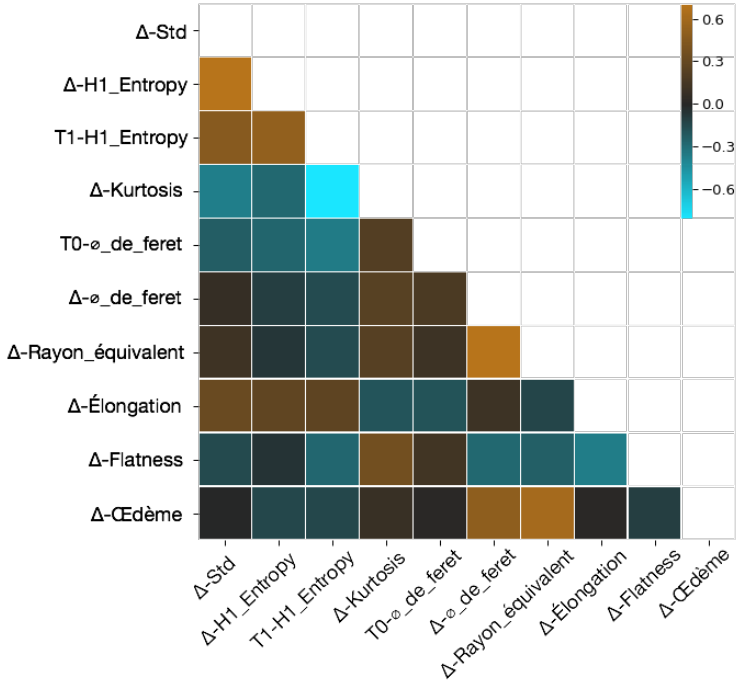


FIGURE 5.7 – Matrice de corrélation des caractéristiques radiomiques significatives à l'analyse univariée.

Puisque les *p-values* les plus basses ont été obtenues pour $\Delta\_Histogram\_Entropy$ pour les caractéristiques de texture et $\Delta\_Elongation$ pour les caractéristiques morphologiques, la signature initiale inclut donc ces deux variables avec le Δ_Edema.

5.5 Résultats de l'apprentissage statistique

La Table 5.1 rassemble les performances des différents algorithmes pour leurs hyperparamètres optimaux et la signature à 3 variables en validation croisée. Les résultats du test final sont visibles Table 5.2. Les hyperparamètres obtenus par le *grid-search* pour trois variables sont disponibles en Annexe B.5. Sauf si explicitement précisé, les résultats suivants concernent la signature à trois variables.

5.5.1 Validation croisée

Les courbes de calibration des modèles sont visibles Fig. 5.8 et les courbes d'apprentissage Fig. 5.11.

Sur le jeu de 50 patients, la plus haute fiabilité après validation croisée est obtenue pour les RF (84.8%), suivi par la LR (83.1%), les kNN (80.5%) et les SVM (70.1%). En comparaison, le changement relatif du plus grand diamètre (RECIST 1.1) seul donne la fiabilité la plus faible avec 76% de patients correctement prédits. Par ordre décroissant, les scores AUROC

sont de 0.87 pour la LR, 0.85 pour les RF, 0.80 pour les kNN, 0.66 pour le critère RECIST 1.1 et 0.65 pour les SVM. La PPV des modèles est forte, jusqu'à 0.89 pour les RF, mais le déséquilibre de la distribution des réponses augmente probablement ce score.

Les classifieurs RF, puis LR obtiennent donc les meilleurs résultats. Aussi, nous comparons directement leurs performances à celles du critère RECIST Fig. 5.9. La moyenne de leurs courbes ROC est visible Fig. 5.10a. Leurs courbes précision-rappel (Fig. 5.10b) ont une précision moyenne supérieure à la limite informative (> 0.78, voir section 4.5.4). Ces trois figures montrent que les deux modèles dépassent largement les performances du critère RECIST sur la cohorte à 50 patients. L'exception est le score de spécificité. Le taux de faux positifs est modérée en RF et faible avec le RECIST.

	AUROC	ACC	SEN	SPE	PPV	NPV	F1	AP	Train
RF	0.85 ± 0.04	$\mathbf{0.85} \pm 0.03$	0.92 ± 0.02	0.57 ± 0.11	$\mathbf{0.89} \pm 0.03$	$\mathbf{0.68} \pm 0.07$	$\mathbf{0.9} \pm 0.02$	0.94	0.96
LR	$\mathbf{0.87} \pm 0.04$	0.83 ± 0.003	0.92 ± 0.01	0.49 ± 0.05	0.87 ± 0.01	0.65 ± 0.03	$\mathbf{0.9} \pm 0.007$	**0.95**	0.85
kNN	0.80 ± 0.04	0.80 ± 0.01	$\mathbf{0.97} \pm 0.01$	0.21 ± 0.05	0.81 ± 0.009	0.67 ± 0.07	0.89 ± 0.007	0.89	-
SVM	0.65 ± 0.07	0.7 ± 0.04	0.82 ± 0.04	0.27 ± 0.07	0.8 ± 0.02	0.3 ± 0.08	0.81 ± 0.03	0.84	0.94
RECIST	0.66	0.76	0.57	**0.91**	0.67	0.21	0.62	0.89	-.

TABLE 5.1 – Scores moyens et écarts-type de la classification de la réponse des 100 validations croisées sur 50 patients.

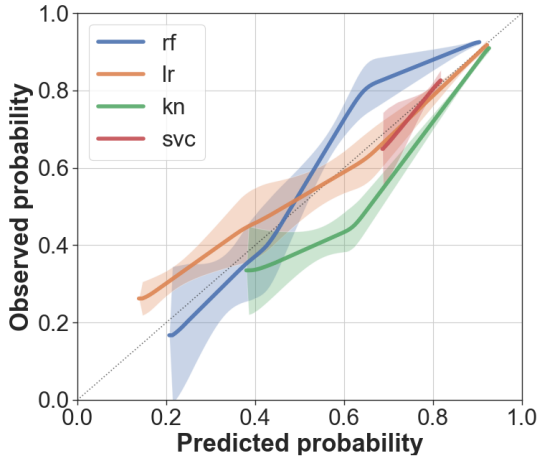


FIGURE 5.8 – Calibration moyenne des 100 modèles construits en validation croisée sur 50 patients.

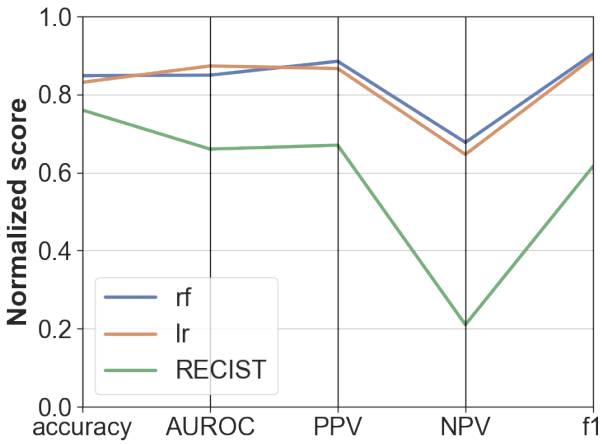
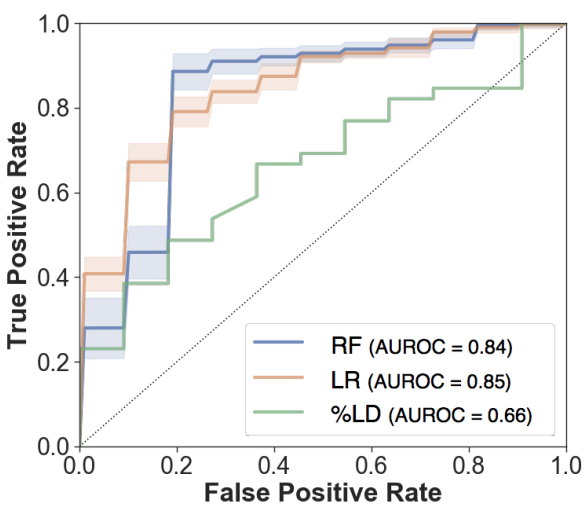


FIGURE 5.9 – Scores normalisés en validation croisée sur 50 patients des modèles RF, LR et de la prédiction du critère RECIST.

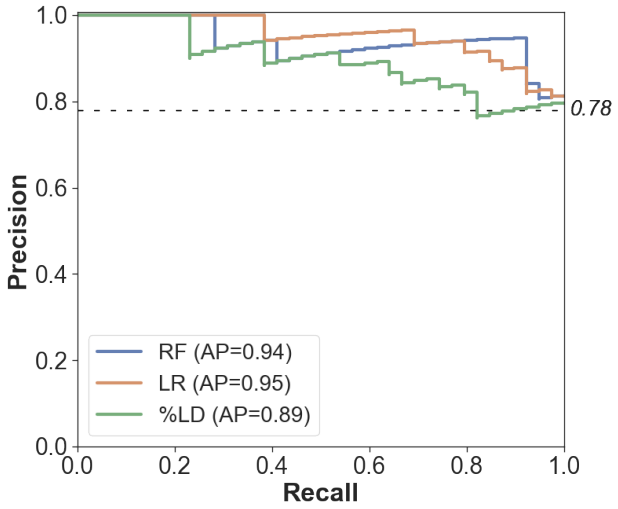
5.5.2 Nombre de variables du modèle

Nous nous intéressons ensuite à l'impact du nombre de variables dans le modèle le plus performant en moyenne, les RF.

La Fig. 5.12 montre les différents scores avec un nombre de variables croissant sur la cohorte à 50 patients. On constate que la fiabilité et l'AUROC ne sont pas améliorées en



(a) Courbes ROC (moyennes + écart-type des 100 cycles en LR et RF).



(b) Courbe précision-rappel des moyennes des probabilités obtenues par individu.

FIGURE 5.10 – Fiabilité des probabilités prédites par les modèles RF et LR pour 100 cycles de validation croisée, comparé aux probabilités dérivées de l'évolution du plus grand diamètre (%LD), sur la cohorte à 50 patients. Les lignes en pointillés représentent la limite en dessous de laquelle les modèles sont non informatifs.

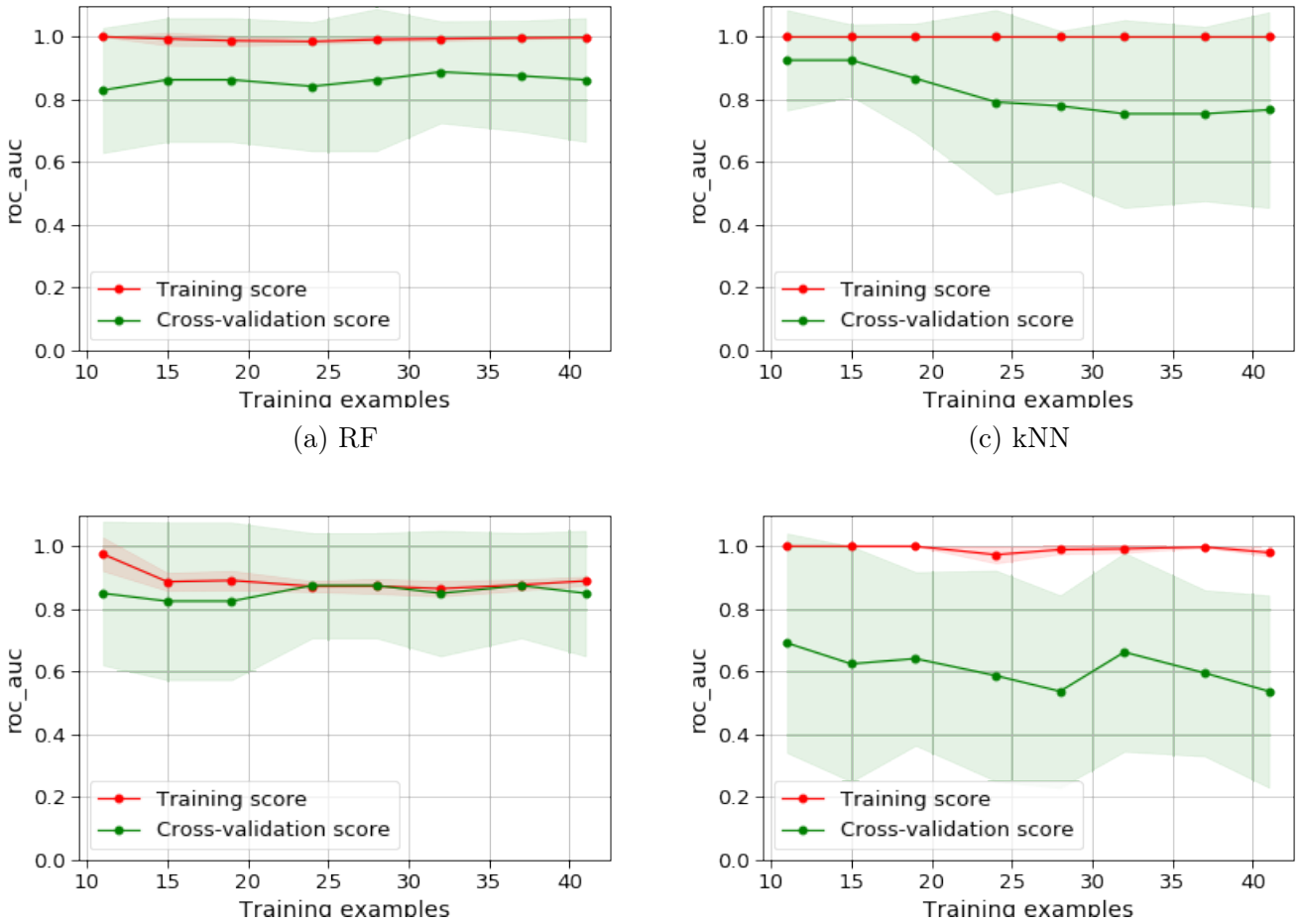
ajoutant plus de variables que les trois sélectionnées. De même, les PPV (= précision) et NPV tendent à diminuer. La sensibilité (= rappel) augmente lentement, au détriment d'une chute franche de la spécificité. La baisse de précision à peu près proportionnelle à l'augmentation du rappel donne un score F1 relativement constant. Le score AP est également stable.

Le modèle comprenant toutes les caractéristiques significatives ne dépasse jamais les performances du modèle à 3 variables.

Dans un second temps, nous testons la réduction de dimensions avec ElasticNet. À partir des 103 variables initiales, la régularisation sélectionne six caractéristiques d'impact supérieur à la moyenne : $\Delta\_Histogram\_Entropy$, $\Delta\_Elongation$, $\Delta\_Equivalent\_radius$, Δ_Edema, $T1\_Kurtosis$, $T1\_Histogram\_Entropy$. La sélection est identique quelle que soit la composition du jeu d'entrainement fourni par la validation croisée. Ces variables font toutes partie des caractéristiques significatives à l'analyse univariée et incluent les trois caractéristiques que nous avons favorisées jusqu'à présent. Utilisée par les RF ou la LR, cette sélection n'améliore aucun des scores obtenus précédemment (voir Tables en Annexe B.6).

5.5.3 Validation finale

Avec l'intégralité du jeu de 50 patients pour entrainer le modèle, les performances ont globalement diminué (Fig. 5.13). La précision de prédiction des 15 patients du jeu de validation est de 77.3% pour les RF, 66.7% pour la LR et 73.3% pour la classification avec le critère RECIST 1.1 seul (Table 5.2). Ce dernier conserve la meilleure AUROC (0.72) (Fig. 5.14).



(a) RF (c) kNN

(b) LR (d) SVC

FIGURE 5.11 – Courbes d'apprentissage des modèles RF, LR, SVC et kNN sur la cohorte de 50 patients. Le score évalué est l'AUROC. Les marges colorées représentent l'écart-type. Ces résultats dénotent d'un sur-apprentissage fort des kNN et SVC. La LR n'est pas améliorée par l'ajout de patients et souffre peut être d'un biais.

	AUROC	ACC	SEN	SPE	PPV	NPV	F1	AP	Train
RF	0.63 ± 0.03	$\mathbf{0.77} \pm 0.03$	$\mathbf{1.0} \pm 0.01$	0.32 ± 0.10	$\mathbf{0.75} \pm 0.03$	$\mathbf{0.99} \pm 0.06$	$\mathbf{0.85} \pm 0.02$	0.71	0.96
LR	0.56	0.67	0.9	0.2	0.69	0.50	0.78	0.76	0.84
kNN	0.55	0.67	1.0	*0*	0.67	*0*	0.80	0.72	-
SVM	0.42	0.47	0.7	*0*	0.58	*0*	0.64	0.70	0.94
RECIST	**0.72**	0.73	0.9	**0.4**	**0.75**	0.67	0.82	**0.80**	-

TABLE 5.2 – Scores moyens de la classification de la réponse des tests finaux sur 15 patients.

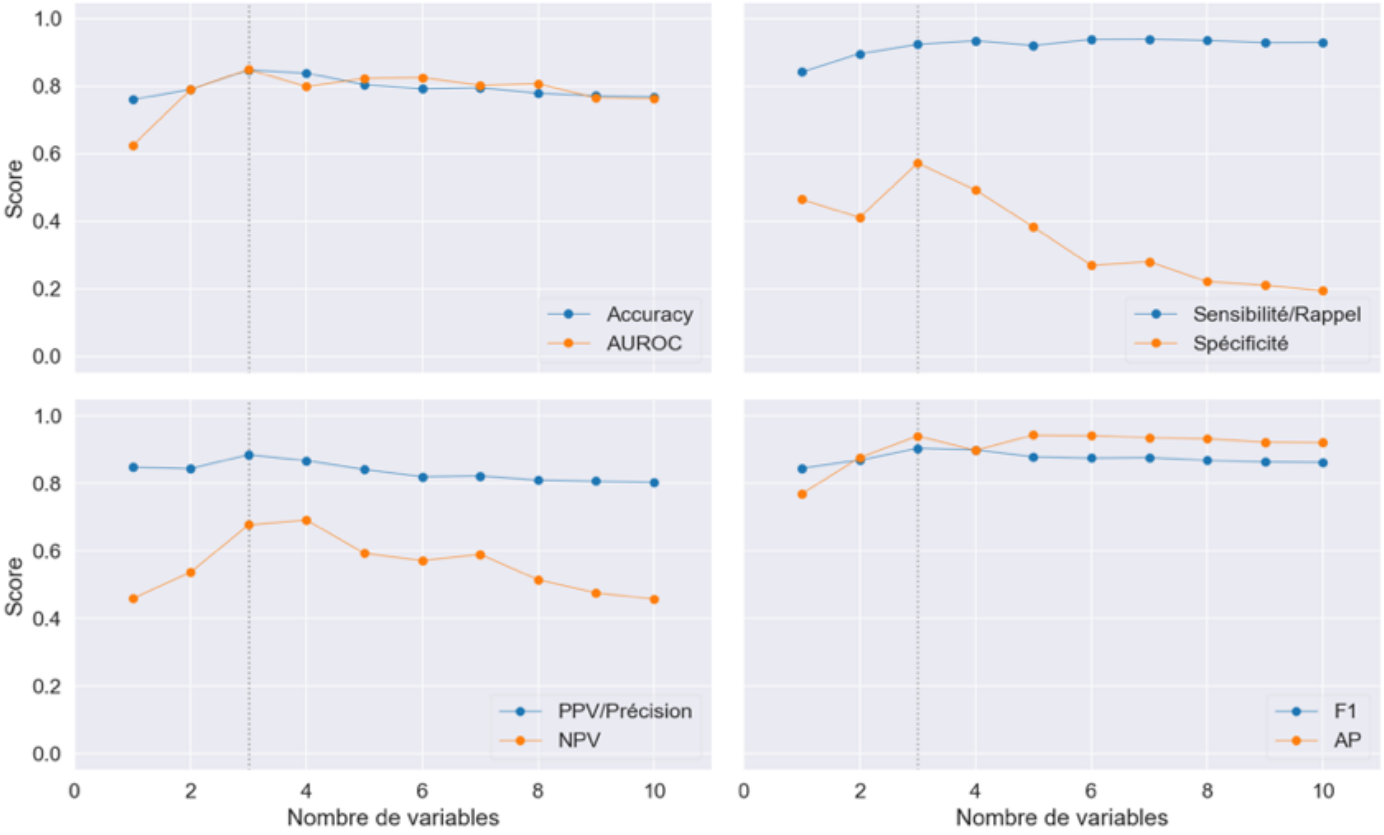


FIGURE 5.12 – Scores du modèle RF en fonction du nombre de caractéristiques apprises. Les variables encore non utilisées dans la signature à 3 caractéristiques (représentée par la ligne verticale) sont ajoutées une par une dans l'ordre croissant de leur *p-value*, jusqu'à $p = 0.05$.

Les RF obtiennent toujours les meilleurs scores pour une grande partie des métriques, avec de très bons résultats en F1 (0.85) et NPV (0.99). L'*average precision* de la LR lui est supérieure de 5 points (Fig. 5.15), mais en comparaison, le modèle a un score d'entrainement assez faible (0.84) (sous-apprentissage probable). Les autres scores sont également beaucoup moins élevés. On note que les kNN et SVM prédisent systématiquement des rechutes (spécificité et NPV nulles).

La spécificité de l'ensemble des méthodes est encore plus basse qu'en validation croisée, ce qui est problématique puisqu'on ne souhaite pas interrompre un traitement s'il est efficace (conséquence des faux positifs).

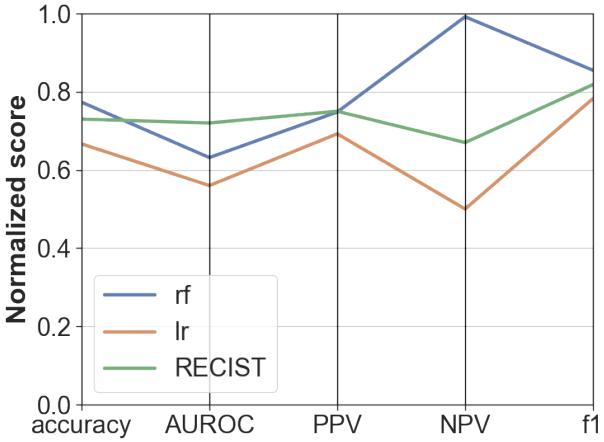


FIGURE 5.13 – Scores normalisés des test finaux sur 15 patients pour les deux meilleurs algorithmes testés et la prédiction RECIST.

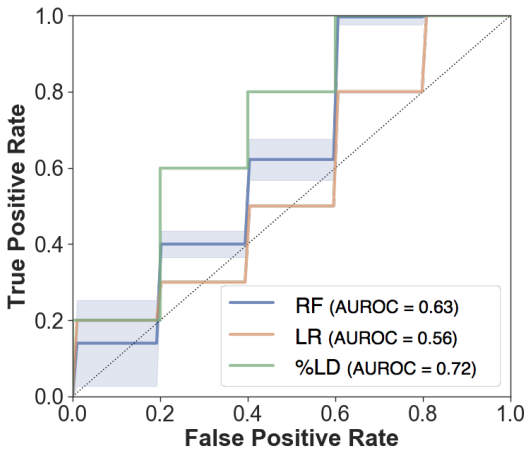
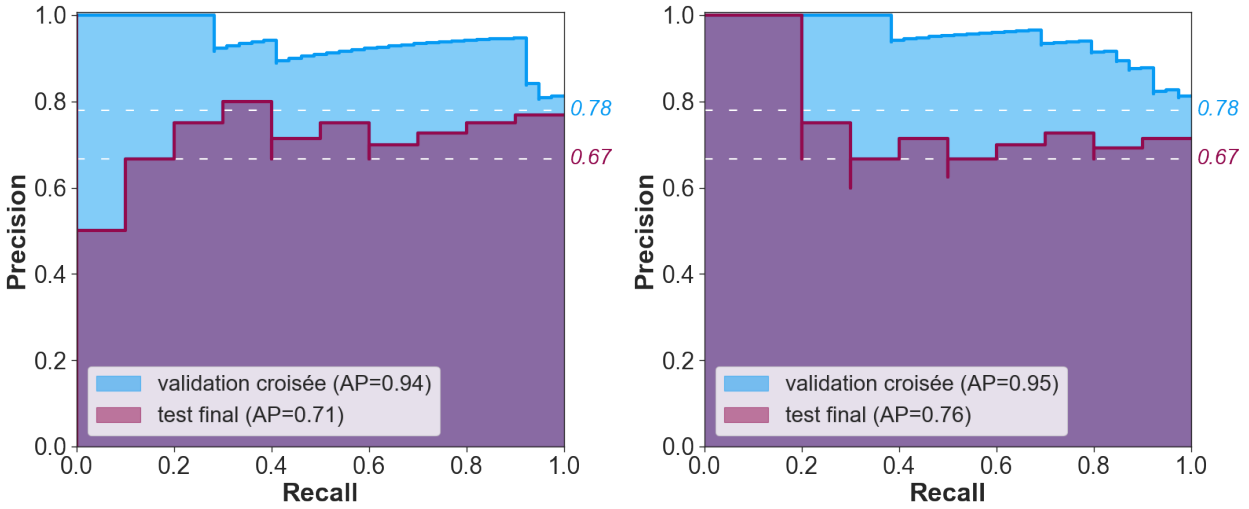


FIGURE 5.14 – Courbes ROC finales des modèles RF (moyenne + écart-type), LR et du plus large diamètre (%LD).

La table 5.5.3 montre la probabilité moyenne prédite par les algorithmes pour chacun des 15 patients. On remarque que si l'ensemble des dix patients de mauvais pronostic est correctement prédit avec les RF, 3/5 patients de bon pronostic ont plus de 70% de chances d'être mal classifiés. Les autres algorithmes prédisent également mal ces trois patients, entre autres erreurs. En comparaison, seuls deux des patients de bon pronostic du jeu d'entrainement (moins de 20%) ont le même problème.



(a) RF (b) LR

FIGURE 5.15 – Courbes précision-rappel de la moyenne des probabilités obtenues pour les patients du jeu de validation croisée et ceux du jeu de test final. Les lignes en pointillés représentent leur limite informative respective.

ID	Classe	RF	LR	SVC	KNN	RECIST
92	0	0.34	0.23	0.8	0.54	1
93	1	0.97	0.97	0.85	1	0
96	1	0.78	0.45	0.81	0.75	1
97	1	0.86	0.98	0.85	1	1
98	0	0.91	0.69	0.77	0.91	1
99	1	0.58	0.85	0.71	0.8	1
100	1	0.64	0.87	0.72	0.86	1
101	0	0.97	0.93	0.85	1	1
102	1	0.93	0.62	0.79	0.9	1
103	1	0.96	0.87	0.76	0.97	1
104	1	0.68	0.76	0.7	0.75	1
106	0	0.49	0.85	0.78	0.77	0
107	0	0.71	0.97	0.82	0.92	0
108	1	0.58	0.71	0.8	0.78	1
109	1	0.97	0.96	0.87	1	1

TABLE 5.3 – Probabilités moyennes prédites des 15 patients de la cohorte de test. 0 = *Good-HR*, 1 = *Bad-HR*. Erreur de prédiction de plus de 50%, de plus de 80%.

5.5.4 Analyse des patients de la cohorte de test final

Valeur ajoutée du modèle radiomique : exemples

La figure 5.16 illustre la valeur ajoutée du modèle final pour deux patients ayant une maladie stable selon le critère RECIST 1.1 : l'un a au final un mauvais pronostic, l'autre un bon pronostic.

Cas 1 (a) Un homme de 76 ans présente un rhabdomyosarcome pléomorphe profond de grade 3 à l'épaule. Après deux cycles de chimiothérapie, la tumeur est stable selon le critère RECIST 1.1. Pourtant, elle montre une augmentation de la quantité d'œdème périphérique (flèche blanche), une forme stable et une entropie de l'histogramme stable également. Aussi, et en désaccord avec le critère RECIST, le modèle RF final prédit une mauvaise réponse histologique. Cette évaluation précoce est confirmée sur la pièce opératoire (70% de cellules tumorales viables résiduelles).

Cas 2 (b) Un homme de 50 ans présente un sarcome pléomorphe indifférencié profond de grade 3 au genou. Après deux cycles de chimiothérapie, la tumeur est considérée stable par le critère RECIST 1.1. L'œdème périphérique décroit ostensiblement (flèche blanche), son volume se rétracte, l'entropie de l'histogramme diminue. Le modèle RF final prédit lui aussi une bonne réponse histologique, ce qui est confirmé sur pièce opératoire (5% de cellules résiduelles).

Cas particuliers

Une analyse rétrospective des faux positifs obtenus par le modèle RF révèle des cas de tumeurs hémorragiques massivement nécrotiques, des profils de répondeurs tardifs au traitement et des cas de liposarcomes myxoïdes, extrêmement hétérogènes.

Ainsi, la quantification de la texture des tumeurs massivement nécrotiques a été biaisée par la présence à l'IRM_0 de larges caillots de sang hétérogènes et par leurs modifications à l'IRM_1. Dans le cas des répondeurs tardifs, même l'analyse de modalités d'imagerie supplémentaires n'est pas suffisante pour deviner une bonne réponse au traitement après deux cycles de NAC seulement. C'est uniquement à l'évaluation pré-chirurgicale que des changements fibro-nécrotiques intenses et une diminution du SUVmax indiquent enfin que le patient a répondu favorablement à son traitement.

Les quelques exemples de la Fig. 5.17 illustrent ces valeurs aberrantes (*outliers*).

Cas 3 (a) Un patient de 52 ans présente à la cuisse gauche un sarcome pléomorphe indifférencié profond de grade 3. Des caillots sanguins et des cloisons fibreuses sont mélangées aux cellules nécrosées (flèche blanche), seules de petites excroissances tumorales sont visibles contre la paroi tumorale. Les changements de texture de la tumeur sont donc essentiellement provoqués par des modifications de la structure et du signal du compartiment nécrotique. **(b)** Ce patient mal classé a bénéficié d'une TEP au diagnostic et après deux cycles de NAC. Ces examens montrent une forte diminution du SUVmax (de 8.16 à 3.94, –51.7%), signe d'une chimiothérapie efficace.

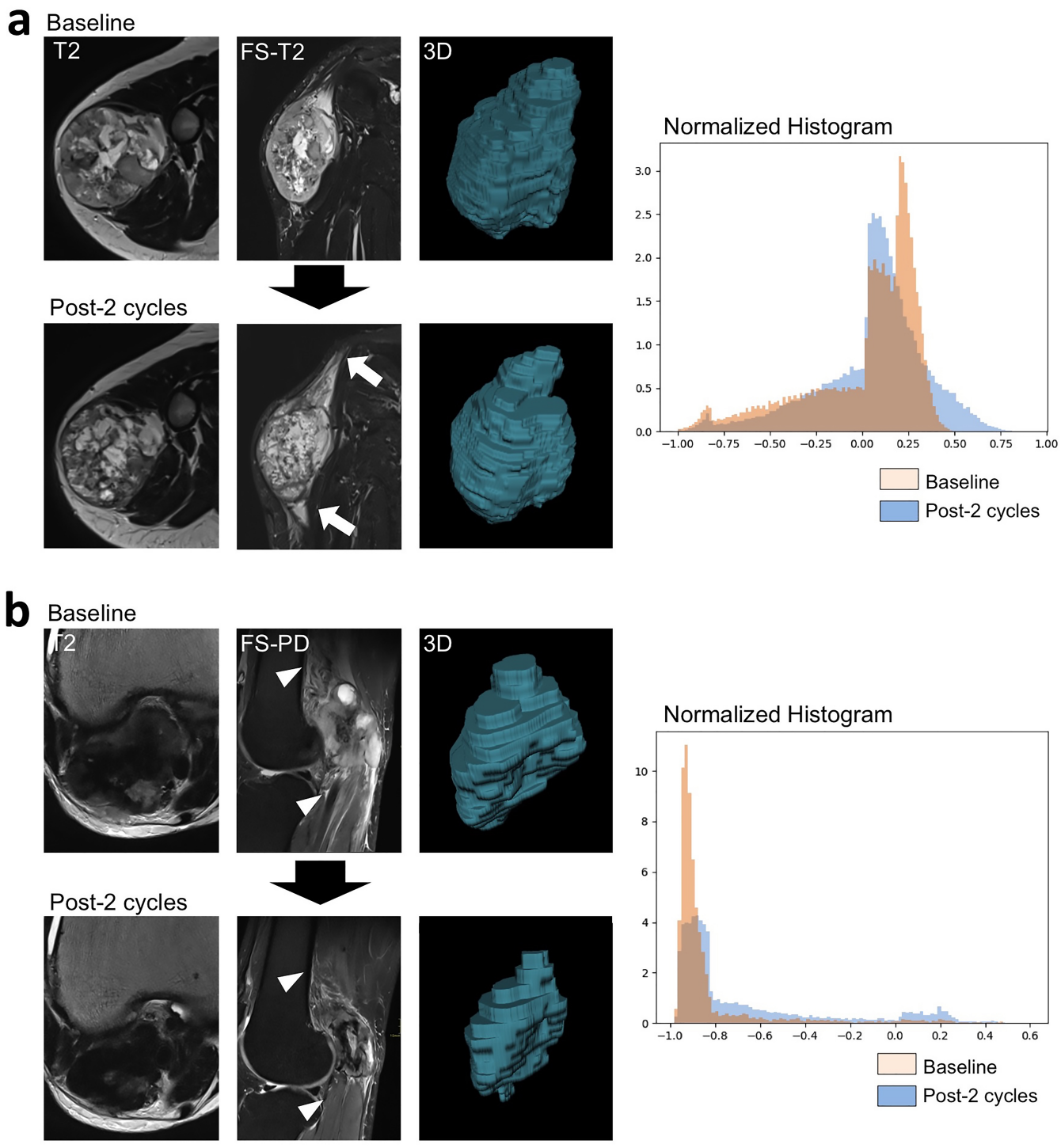


FIGURE 5.16 – Valeur ajoutée du modèle RF final pour la prédiction précoce de la réponse au traitement des STS.

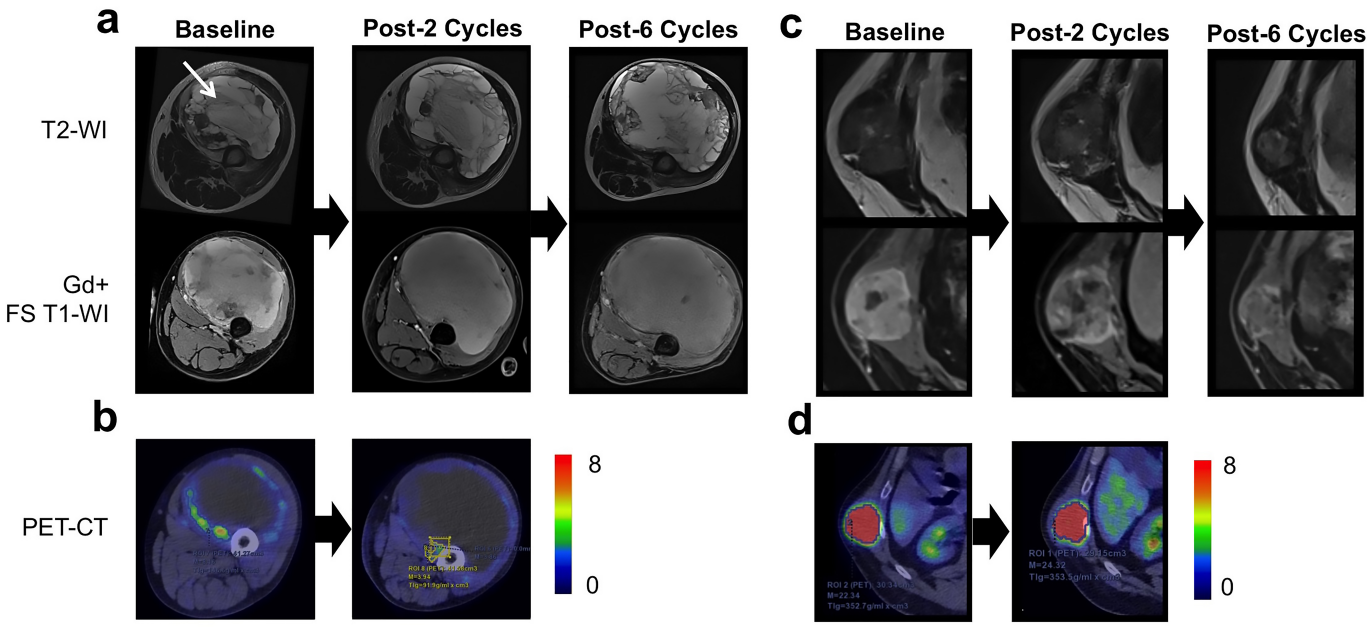


FIGURE 5.17 – Cas particuliers classés mauvais répondeurs par erreur par nos modèles.

Cas 4 (c) Exemple d'un profil de type répondeur tardif : un homme de 66 ans présente un rhabdomyosarcome pléomorphe profond de la sangle abdominale de grade 3. Aucun changement manifeste n'est identifiable visuellement à l'évaluation précoce. **(d)** À l'inverse du cas 1, la TEP montre ici une augmentation paradoxale du SUVmax après quatre cycles (de 22.34 à 24.32), alors que le patient est un bon répondeur au traitement. La réponse histologique n'est finalement décelable qu'après quatre cycles.

5.6 Discussion

Dans cette étude nous avons développé et évalué des modèles radiomiques pour prédire la réponse histologique de STS pendant la NAC grâce aux changements à l'IRM en T2 entre le diagnostic et l'évaluation précoce. Notre meilleur modèle est obtenu avec une classification par des forêts aléatoires sur trois variables issues de l'analyse de la morphologie, de l'hétérogénéité et des tissus environnant des STS. Ce modèle est plus efficace que le critère RECIST 1.1, avec 84.5% de précision de prédiction en validation croisée. Il obtient également les scores les plus hauts quand il est validé sur une cohorte indépendante. Cependant, ces derniers résultats ont mis en évidence des cas particuliers requérant une description supplémentaire.

Les performances de notre meilleur modèle prédictif sont comparables voire plus hautes que celles obtenues dans d'autres études avec d'autres biomarqueurs.

Il est toutefois nécessaire de rester prudent lorsque l'on fait des comparaisons entre différentes méthodologies. On note des différences aussi bien dans les seuils pour caractériser une "bonne réponse histologique" que dans les protocoles de chimiothérapies et dans les périodes auxquelles sont réalisées les examens radiologiques.

Stacchiotti et al. ont étudié le critère de Choi pour prédire une bonne réponse pathologique établie à moins de 10% de cellules viables sur les pièces opératoires de 37 patients. Ils ont obtenu un score de prédiction de 74.1%, (14/22 répondeurs partiels selon Choi sont effectivement des bons répondeurs, et 6/6 non répondeurs selon Choi sont effectivement des mauvais répondeurs) [Sta+09 ; Sta+12]. Dans une autre étude, une diminution de 30.5% de la prise de contraste entre deux IRM ayant un délai optimisé d'acquisition après injection a fourni un score de 82.8% de bonnes prédictions [Cro+18a]. Sur une série rétrospective de 23 patients, une évaluation qualitative de plusieurs paramètres d'imagerie de diffusion DCE-MRI a donné au mieux une AUROC de 0.833 [Sol+15]. Enfin à l'évaluation précoce sur ^{18}F-FDG-PET-CT, une étude prospective sur 50 patients obtient une AUROC de 0.83 pour une diminution de minimum 35% du SUV_{max} [RB+08 ; Ben+09].

Nos trois variables ($\Delta\_Histogram\_Entropy$, $\Delta\_Elongation$ et Δ_Edema) sont d'autant plus intéressantes que leur évolution peut trouver une explication biologique. L'association entre une diminution de la quantité d'œdème et une bonne réponse au traitement n'est par exemple pas surprenante. L'œdème périphérique est un signe distinctif des STS de haut grade et est associé à la présence de cellules tumorales satellites. Une NAC efficace doit logiquement entrainer une diminution du nombre de ces cellules satellites et donc une diminution du signal irrégulier à l'IRM autour des STS.

Une diminution du nombre de cellules de la tumeur, dégénérant en tissus fibronécrotiques, peut expliquer le changement de forme de la lésion, dont les bords vont se rétracter le long des fibres musculaires. Une tumeur réagissant bien au traitement voit donc effectivement son élongation augmenter. En outre, les processus fibronécrotiques conduisent à un signal plus large en intensité. L'histogramme s'aplatit et son entropie se modifie elle aussi.

Les performances des modèles sont moins bonnes sur le jeu de test final. La spécificité notamment est mauvaise, ce qui conduirait à interrompre un traitement pour un patient chez qui il est efficace. Nous l'expliquons par la taille réduite de cette cohorte, bien plus que celle du jeu de validation croisée et par la présence de cas particuliers, dans une proportion sans équivalence dans le jeu d'entrainement. Ces *outliers* sont systématiquement mal prédits par tous les modèles. Une analyse rétrospective attentive de ces tumeurs a identifié des profils de "répondeurs tardif" et de "lésions massivement nécrosées". Ces dernières sont délicates à capter à l'imagerie et leurs changements morphologiques pendant le traitement sont compliqués à interpréter visuellement. L'évaluation avec le critère RECIST 1.1 est aussi biaisée car il mesure essentiellement la nécrose et non les changements des composantes viables de la tumeur. L'IRM dynamique et l'imagerie de diffusion sont complexes à utiliser parce que les tissus tumoraux viables ne représentent plus que de petites excroissances fixées à la paroi de la tumeur et incluses dans une large masse hémorragique. Dans notre cas, la TEP montre correctement une bonne réponse histologique selon le critère de 35-38% de diminution du SUV_{max} [Ben+09].

Ces observations sur la cohorte de test fournissent un aperçu des variables à ajouter dans de futurs modèles de prédiction. La division de notre jeu de données en deux cohortes indépendantes nous a ainsi permis de prendre du recul sur la composition de la réponse des STS et d'envisager d'ajouter des variables venant d'autres modalités d'imagerie pour améliorer les modèles. Des premiers tests avec des variables évaluant qualitativement la pré-

sence de compartiments massivement nécrotiques ou de compartiment de cellules résistantes sont en cours. L'évolution (qualitative) de la taille du compartiment résistant apparaît ainsi significativement différente entre bons et mauvais répondeurs ($p = 0.006$) sur le jeu de 50 patients, mais n'a pas encore permis d'améliorer les performances de prédiction.

D'un point de vue méthodologique, il est possible de vérifier l'influence des cas particuliers sur le modèle en mélangeant les deux cohortes de façon à ré-équilibrer leur répartition Cette méthode présente un inconvénient. La re-répartition des 65 patients va déplacer des observations de l'ex-cohorte de développement dans la nouvelle cohorte de test, ce qui revient à y transférer une partie des connaissances acquises lors du premier développement (pendant le *grid-search*, la sélection des variables et de la construction des modèles). Or, cette cohorte n'est pas supposé influencer la construction des modèles. La conséquence est une sur-performance possible sur le jeu de test. Une redistribution des données demande donc théoriquement de recommencer l'intégralité de la procédure décrite dans ce chapitre, sans apriori. Il est classiquement plutôt recommandé d'acquérir une toute nouvelle cohorte de test en vérifiant son adéquation avec la distribution de la cohorte de développement.

De façon intéressante, nos meilleurs modèles se basent sur une signature courte tirée de séquences sans injection de produit de contraste. Corino et al. ont eux aussi montré qu'une combinaison courte de quatre variables tirées de séquences de diffusion fournissent la prédiction la plus juste du grade histologique des STS [Cor+17a]. Le meilleur modèle pour la prédiction de rechutes métastatiques pulmonaires de STS se base aussi sur quatre variables chez Vallières et al [Val+15]. Dans leur étude comme dans la nôtre, l'ajout de caractéristiques supplémentaires n'a pas amélioré la qualité de prédiction. Notre méthode de sélection des variables, basée sur l'analyse univariée avec un test de Wilcoxon, est également ressortie comme étant l'une des plus fiables et des plus performantes une fois associée aux RF selon une étude comparative étendue [**Parmar2015**].

Dans un contexte controversé sur les effets à long terme du gadolinium des agents de contraste, on pourrait imaginer une stratégie d'imagerie où seuls les cas particuliers connus ou bien les patients ayant une probabilité de réponse intermédiaire à la première évaluation seraient expertisés dans un second temps sur IRM dynamique / imagerie de diffusion.

Notre étude comporte quelques limitations. La principale est probablement la taille relativement modeste de la cohorte et son caractère rétrospectif. Pourtant, il s'agit quand même d'une des séries IRM des STS les plus larges à ce jour, avec des patients ayant reçu un traitement de référence similaire et homogène et ayant deux examens chacun. De plus et bien que modeste, cette cohorte s'est avérée suffisamment conséquente pour permettre de mettre quelques patients de côté pour une validation indépendante du processus de construction du modèle.

La cohorte étudiée inclut uniquement des patients présentant les caractéristiques épidémiologiques associées à un des pires pronostics (lésion profonde avec un plus large diamètre supérieur à 5cm) [MC+96]. Aussi les données épidémiologiques n'ont pas été ajoutées au modèle, aucune n'étant associée significativement avec la réponse de la tumeur.

Dans un second temps, on note que les protocoles d'imagerie n'ont pas été réalisés dans le but d'une étude radiomique. Nous avons donc choisi d'utiliser les séquences T2 (T_2-

WI) pour calculer les marqueurs. Le T2 présente l'avantage principal d'être la séquence la plus régulièrement disponible parmi les patients suivis pour un STS, avec des paramètres d'acquisition identiques pour toute la période d'étude. En outre, elle permet d'observer une large gamme de changements morphologiques pendant le traitement. Le T_2-WI peut capter les processus fibrotiques et nécrotiques (respectivement une diminution et une augmentation du signal d'intensité en T2). C'est une séquence qui a déjà montré de bons résultats dans les études de texture pour d'autres types de tumeurs [Don+17 ; Hen+17 ; Nke+16 ; Gne+16] comme pour l'étude présentée au chapitre 3 [Hoc+18]. À l'inverse, les séquences en T1 dans notre série comportent des examens aux protocoles d'acquisition hétérogènes : 2D TSE, 3D *gradient echo recalled imaging*, techniques de suppression des graisses différentes, agents de contraste différents. Le délai d'acquisition après injection de l'agent de contraste n'y est pas standardisé alors qu'il peut avoir un impact significatif sur la quantification de l'évolution de l'hétérogénéité.

Nous n'avons pas pu comparer directement les performances de notre modèle à celles obtenues avec le critère de Choi modifié pour l'IRM. Ce critère est défini à partir de deux séquences T1[6]. Malheureusement, la plupart des patients de notre cohorte n'ont pas eu le même type de séquence T1 avant et après injection. Une étude antérieure a démontré que la précision du critère modifié de Choi dépend du délai d'acquisition après injection [Cro+18a]. Ici, ce délai n'a pas été contrôlé et le calcul de ce critère n'a donc pas pu être systématisé.

L'évolution de l'œdème périphérique est un des meilleurs facteur prédictifs de cette étude. Même s'il y a consensus entre deux radiologues, son estimation est à l'heure actuelle uniquement qualitative, les séquences d'évaluation n'étant pas standardisées. De futures études devront donc inclure une quantification automatique du volume de l'œdème et de son évolution. L'ajout d'autres modalités d'imagerie est prometteuse, mais aurait clairement réduit le nombre d'examens de cette cohorte. Nous avons donc décidé de garder une seule séquence informative pour maximiser la taille du jeu de données. Cependant, ce point met à nouveau en évidence le besoin de standardiser le protocole d'acquisition IRM des STS.

Un autre moyen de décrire quantitativement l'œdème serait d'inclure les marges dans les contours de la tumeur, ou d'en segmenter les contours à part comme le font [Val+15] pour étudier la stabilité de leur modèle. Même grossièrement délinées, des marges permettraient dans notre cas d'extraire une information de texture de l'environnement direct de la tumeur. Braman et al. [Bra+19] ont ainsi montré qu'une signature radiomique de la périphérie de tumeurs du sein était plus informative qu'une signature intra-tumorale pour la réponse à la NAC. Une façon simple d'obtenir une première estimation de la marge serait d'étendre le masque de la tumeur par dilatation ou avec une carte de distance.

Les descripteurs de la texture issus de la matrice de cooccurrence n'ont en revanche pas fourni une information profitable au développement du modèle, contrairement à leur étude au chapitre 3. Il est encore difficile d'en connaître la cause exacte. La problématique biologique et le jeu de données ne sont peut être pas décrits correctement par ces variables. D'autres plus complexes pourraient apporter une meilleure information au modèle, aux dépends de l'interprétabilité. Les causes méthodologique, comme l'impact de la normalisation ou la corrélation au volume, demanderont aussi une étude plus approfondie.

6. Soustraction d'un CE_T_1-WI et d'un TE-WI avant injection.

Nous avons décidé de restreindre le nombre de caractéristiques radiomiques extraites, même si de nombreuses autres auraient pu être calculées (autres matrices de texture, ondelettes, analyse fractale [DAKM17]). De cette manière nous avons limité le risque de trouver des caractéristiques pertinentes par chance au vu de la taille limitée de notre jeu de données. Ce nombre réduit de variables a aussi facilité l'interprétation biologique des résultats. L'ensemble de marqueurs analysé est classiquement utilisé dans un grand nombre d'études [CTH84] et implémenté dans des bibliothèques libres.

Pour des raisons similaires, nous n'avons pas intégré d'algorithmes d'apprentissage profond à cette étude et nous nous sommes concentrés sur les algorithmes d'apprentissage classiques. Ces derniers sont également plus simples à interpréter, ce qui s'avère plus pertinent pour les analyses biologiques. En outre, comme nous l'avons précédemment évoqué, la réponse au traitement semble mieux décrite par la forme de la tumeur et le volume de son œdème (deux descripteurs associés à la morphologie) que par des caractéristiques de texture classiques. Or, les techniques d'apprentissage profond comme les CNN ont largement tendance à favoriser l'information de texture [Gei+19].

Une dernière remarque concerne la variable à prédire, la réponse histologique. Cette valeur est utilisée en routine clinique comme estimation intermédiaire de l'efficacité immédiate de la NAC et donc du pronostic du patient [Cou+17]. Il n'est donc pas question de l'observation directe de l'information d'intérêt (évènement, rechute ou décès). Il s'agit en outre d'une mesure semi-quantitative, donc potentiellement subjective et biaisée. L'objectif ultime serait donc plutôt de construire un modèle prédictif de la survie sans évènement. La difficulté réside là encore dans le jeu de données, qui ne comporte pas assez de patient suivis suffisamment longtemps à l'heure actuelle : 41 patients (63%) de plus de deux ans, 26 (40%) de plus de cinq ans.

5.7 Segmenter automatiquement l'œdème pour le

L'une des limites principales de notre étude de la réponse au traitement est le caractère qualitatif de l'appréciation d'une des variables les plus informatives, l'évolution de la quantité d'œdème. L'évaluation quantitative de cette variable passe en premier lieu par la détection et la segmentation de l'œdème à l'image.

Le contourage manuel de l'œdème est complexe car il s'infiltre dans les muscles et se confond avec d'autres structures. Il est aussi particulièrement long, surtout en 3D, en raison de ses multiples ramifications. Aussi, une approche automatique ou semi-automatique est nécessaire pour systématiser la segmentation de ce volume sur toute une cohorte.

L'estimation numérique du volume de l'œdème péritumoral nécessite d'acquérir une modalité d'imagerie pertinente et d'y délinéer un VOI relativement précis. Les séquences T2-STIR (***S**hort **TI** **I**nversion **R**ecovery*) permettent de l'observer en hypersignal. L'œdème des STS y est caractérisé par de multiples infiltrations en flammèches le long des muscles et des vaisseaux sanguins où se développe le sarcome (Fig. 5.18).

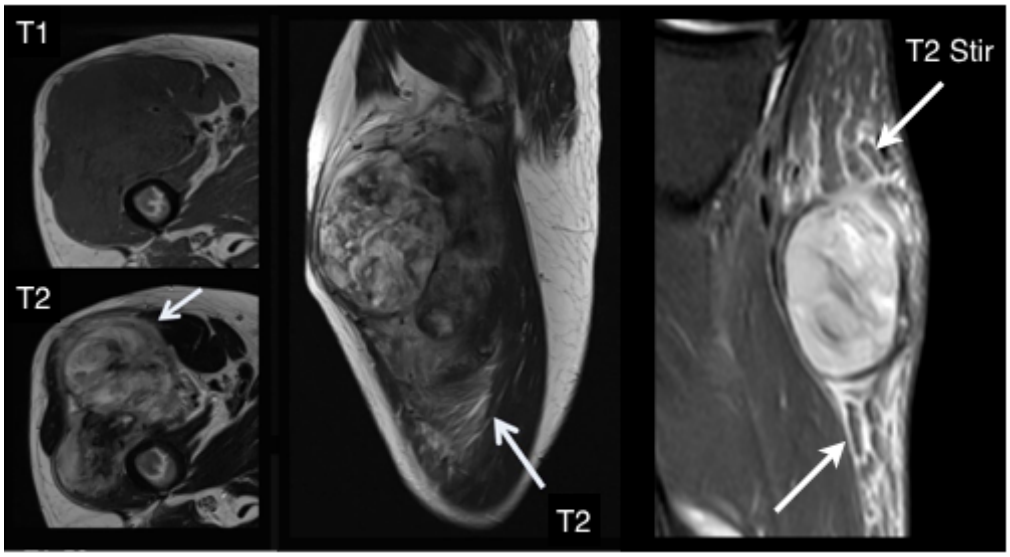


FIGURE 5.18 – Œdème péri-tumoral en T1, T2 et STIR. *source : d'après [Cro17]*

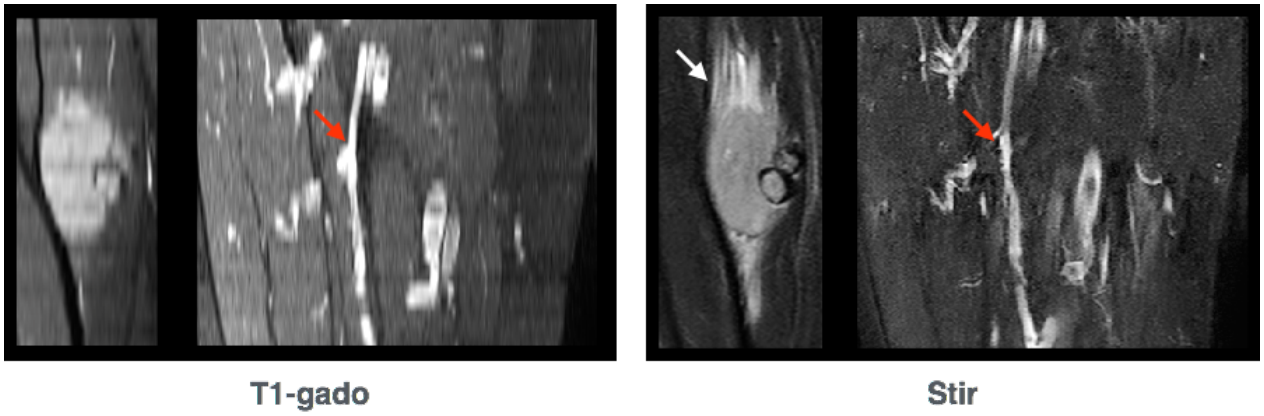


FIGURE 5.19 – L'oedème est en hypersignal en STIR, les vaisseaux le sont en STIR et en T1-gado.

La figure 5.19 montre cependant que si les séquences STIR sont sensibles dans la détection de l'œdème, elles lui sont également peu spécifiques. Les vaisseaux sanguins y sont eux aussi visibles en hypersignal, ainsi que les lésions. À l'inverse, d'autres séquences comme le T1-gado mettent aussi en évidence les vaisseaux, mais pas l'œdème. Une approche multi-modale pourrait donc permettre de récupérer uniquement l'œdème par un système d'opérations logiques.

Notre objectif est donc de pouvoir prédire automatiquement quels voxels appartiennent à l'œdème des patients traités par NAC à partir d'un petit nombre d'examens disposant à minima d'une séquence STIR, d'un T1-gado et d'un masque de référence.

5.7.1 Matériel et méthode

Avec un nombre de patients aussi réduit, impossible de construire des réseaux de neurones de convolution avec l'image entière.

En revanche, le pipeline d'apprentissage statistique présenté en section 5.3.3 est réutilisable à condition de redéfinir ce qu'est la donnée d'entrée. Ici, la prédiction doit aboutir à une image, un masque binaire indiquant quels voxels appartiennent à l'œdème. C'est donc pour chaque voxel que l'on prend une décision. Aussi durant les phases d'apprentissage et de test, un voxel correspond à une observation, elle-même constituée des différentes informations disponibles à son emplacement : valeurs des différentes modalités, distance à la tumeur, mesures locales de texture, etc.

Données IRM et contours

Trois séries d'IRM de membres atteints d'un STS, comprenant une séquence coronale ou sagittale pondérée T2-STIR, une séquence axiale pondérée T1-gado et une T2 classique sont extraites rétrospectivement (voir Fig. 5.20). Les trois séquences ont été réalisées en l'espace de 30-40 min environ pour chaque patient et ne présentent pas de décalage entre elles.

Les contours de l'œdème sont délinéés manuellement sur chaque T2-STIR par la radiologue A. Crombé et constituent notre référence. Les contours de la tumeur sont également récupérés, ainsi qu'un volume supplémentaire encadrant grossièrement l'ensemble pour restreindre la zone d'étude, faciliter le travail des algorithmes et accélérer les calculs.

Post-traitement des IRM. Pour obtenir un même nombre de voxels alignés sur les différents masques et modalités, une série de traitements des images est nécessaire.

Les contours de l'œdème, de la tumeur et de la zone restreinte sont effectués sur le STIR qui est donc choisi comme modalité de référence. Nous découpons les images T2 et T1-gado selon les dimensions de la boîte englobante du STIR, puis nous les ré-échantillonnons en 3D selon la taille de ses voxels ([0.78, 0.78, 4] pour tous les patients) et selon son orientation (coronale ou sagittale). Enfin, nous ne conservons au final que les voxels correspondant à ceux de la zone d'étude restreinte.

Un des trois patients est choisi au hasard comme référence de l'alignement d'histogrammes permettant de normaliser les examens des deux autres. Chaque modalité est normalisée séparément.

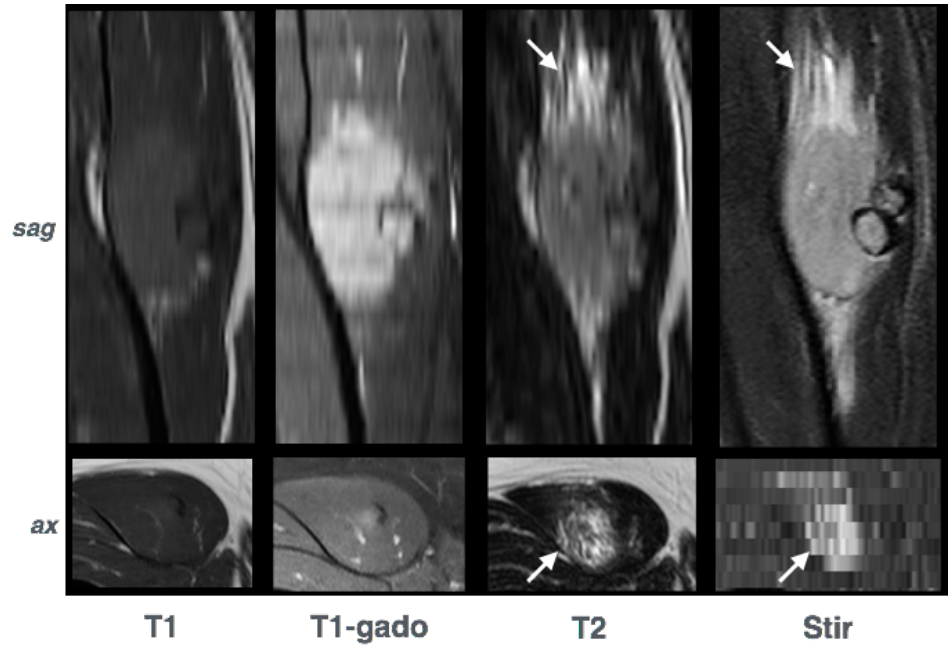


FIGURE 5.20 – Coupes IRM sagittales et axiales de l'œdème péri-tumoral d'un sarcome de la cuisse, quatre modalités.

Pour chaque patient, l'ensemble des examens est acquis sur un temps très court. Après double vérification manuelle de leur alignement (incluant celle de la radiologue) il a été conclu que les séquences n'ont pas besoin d'être recalées les unes par rapport aux autres.

Caractéristiques mesurées

Pour constituer les vecteurs de chaque observation, nous récupérons d'abord les valeurs normalisées de chaque voxel en T2, en T2 Stir et en T1-gado. Comme nous ne voulons pas segmenter de voxels intra-tumoraux, nous ajoutons la valeur du masque de la tumeur au vecteur (Fig. 5.21).

Nous construisons également des caractéristiques dérivées des images initiales.

Nous cherchons à éviter de sélectionner des voxels situés dans des vaisseaux sanguins ou qui appartiennent à de l'œdème non lié à la tumeur. Il est possible d'éliminer ces derniers en ajoutant l'information de la carte de distance euclidienne signée (ou distance de Maurer [QR03]) par rapport aux contours de la tumeur.

Nous formons l'hypothèse que l'information de texture locale est utile pour décrire les voxels constituant les filaments d'œdème. Les mesures de texture locales sont fournies par le calcul de cartes issues de la GLCM et de la GLRM (cf. section 1.4.3) sur chacune des séquences (cf. section 1.4.4). Plusieurs distances ont été testées pour générer ces cartes coupe par coupe. Nous avons empiriquement conservé une fenêtre 2D de 5x5 pixels et fixé la taille des classes des matrices à dix niveaux de gris. Nous ajoutons également les composantes 2D du gradient d'image de chaque coupe pour obtenir une information directionnelle, afin d'essayer de capter l'allongement des filaments le long des fibres musculaires.

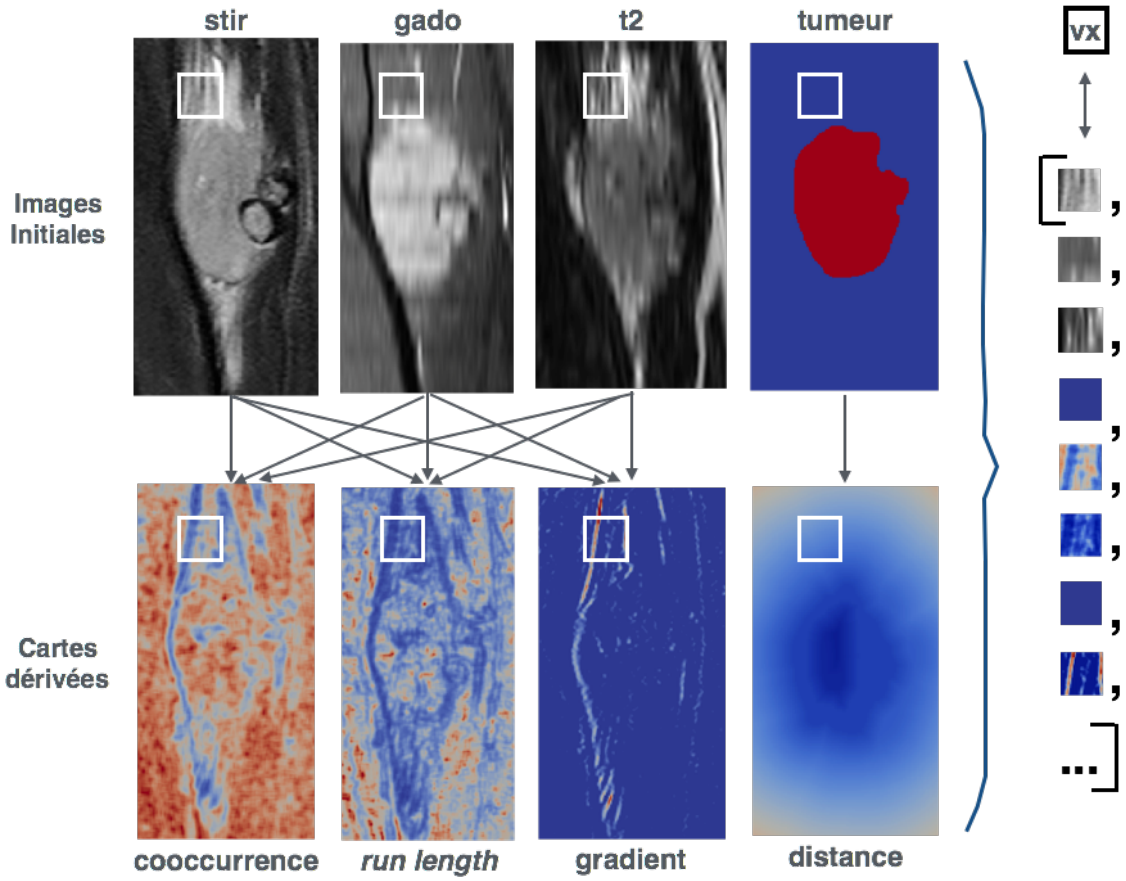


FIGURE 5.21 – A chaque centre d'un voxel est associé un vecteur contenant la valeur de chacune des images et cartes à cette position. Un vecteur = une observation.

Apprentissage statistique

Nous utilisons deux types de classifieurs :

— les forêts aléatoires (120 arbres, minimum 3 observations par feuille),

— le réseau de neurones le plus simple, le perceptron (cf. section 4.3.5). Les paramètres par défaut de l'implémentation de *scikit-learn* sont conservés (1 seule couche cachée de 100 neurones et fonction d'activation ReLU).

Nous constituons trois jeux de données d'entrainement différents, chacun composé de la somme des vecteurs d'une nouvelle combinaison de deux patients. Les vecteurs du troisième patient servent de jeu de données de test. Le cycle apprentissage-test est donc effectué trois fois de façon à tester chaque patient une fois (validation croisée).

Nous effectuons la sélection des caractéristiques avec ElasticNet et composons deux signatures. La première est mixte et consiste à laisser chaque cycle d'apprentissage utiliser sa propre sélection de variable sur son jeu de données d'apprentissage. La seconde est une signature commune composée de l'intersection des trois sélections de caractéristiques.

Les attributs sont normalisés par *standard scaling* avec les moyennes et écarts types des valeurs d'entrainement.

Nous convertissons les résultats prédits pour chaque voxel en cartes de probabilité et en masques de prédiction. Les mêmes métriques que pour l'étude de la réponse au traitement sont relevées, selon les mêmes conditions.

5.7.2 Résultats

L'œdème occupe respectivement 15.96%, 15.3% et 13.47% des voxels de la zone restreinte des trois patients, ce qui constitue une léger déséquilibre dans la distribution des classes.

Le nombre de variables calculées s'élève à 120. La sélection de variables par ElasticNet a fait ressortir 17, 18 et 20 variables sur les trois jeux d'entrainement (= signature mixte). Sept caractéristiques ont systématiquement été sélectionnées par la méthode de régularisation (voir Table 5.4). La signature commune ainsi composée ne contient que le masque de la tumeur et des caractéristiques de l'intensité ou de la texture issues du STIR. Les mesures issues de l'imagerie T2 ou T1-gado n'ont en effet jamais été sélectionnées plus de deux fois. Les cartes de gradient et de distance à la tumeur sont totalement rejetées.

Nombre de sélections	**Intensité du signal**	**Masque et distance**	**Cartes de texture : cooccurrence**	**Cartes de texture : *run lengts***
3/3	**stir**	masque tumeur	Énergie, Corrélation de Haralick, Homogénéité	GLN, LRLGE
2/3	**gado**, **t2**		Énergie, Corrélation de Haralick	RLN, LGRE, LRHGE, HGRE, SRHGE, LRLGE, LRHGE, GLN, GLN,

TABLE 5.4 – Variables sélectionnées par ElasticNet sur les trois jeux de données d'entrainement.

La Table 5.5 rassemble les scores atteints pour la signature à 7 variables et la signature mixte. Les meilleures performances sont obtenues pour le réseau de neurones (NN) à 7 variables avec une AUROC à 0.93, un score F1 à 0.65 et un score AP à 0.73. Les moins bons scores sont également donnés par un réseau de neurones, construit avec la signature mixte.

Modèle	**Nb variables**	**ACC**	**AUROC**	**PR**	**RE**	**F1**	**AP**
RF	17-20	0.89	0.91	0.73	0.41	0.52	0.66
RF	7	0.90	0.92	0.70	**0.62**	0.64	0.69
NN	17-20	0.88	0.90	0.71	0.39	0.49	0.63
NN	7	**0.91**	**0.93**	**0.75**	0.60	**0.65**	**0.73**

TABLE 5.5 – Scores moyens de la détection de l'œdème pour les forêts aléatoires et les réseaux de neurones. Les modèles sont construits avec les variables sélectionnées par ElasticNet sur chacun des trois jeux d'apprentissage, ou avec les sept variables communes aux trois.

Ces résultats sont confirmés qualitativement par l'analyse des cartes de probabilités et des masques obtenus, avec un exemple Fig. 5.22. Visuellement, la zone couverte par les probabilités > 0 est correctement placée. Toutefois les modèles mixtes (20 variables pour ce patient en test) sont bien moins efficaces que les modèles à 7 variables. Il y a probablement sur-apprentissage, l'information apportée par le masque de la tumeur ne ressort pas et de l'œdème est faussement détecté à l'intérieur.

Le résultat du modèle NN à 7 variables, qui est le plus fidèle à la référence, est repris en plus gros plan figure 5.23. Nous nous apercevons que des erreurs de segmentation demeurent. D'abord, quelques vaisseaux sanguins (cercles blancs) sont inclus à l'intérieur du masque. Ce n'est pas surprenant puisque les 7 variables n'incluent pas l'information du T1-gado. Ensuite, des voxels appartenant à l'œdème sont régulièrement oubliés (cercle pointillé orange). Globalement, la sensibilité du modèle n'est pas optimale et des voxels sont omis pour les trois patients. Enfin, de l'œdème est détecté le long des fibres musculaires sur tout le pourtour de la tumeur (rectangles violets) alors qu'il n'est pas inclus dans le masque de référence. Il peut cette fois s'agir d'un défaut de délinéation sur l'image réelle : les outils grossiers de contourage pourraient avoir empêché l'inclusion d'une couche aussi fine de pixels sur chaque coupe.

Comme la sensibilité (ou rappel, score le plus faible) semble augmenter lorsque le nombre de variables diminue, nous avons tenté de comparer ces résultats à une signature simple à 4 variables uniquement composée de l'intensité du signal des trois modalités disponibles et du masque de la tumeur. Nous avons également essayé d'améliorer l'exclusion des vaisseaux sanguins, toujours visibles avec la signature STIR, en ajoutant l'intensité du signal en T1-gado à la signature commune pour obtenir une signature à 8 variables.

Les résultats de ces deux tests pour le réseau de neurone, visibles en Annexe B.8 ne dépassent pas les performances précédentes, ni quantitativement, ni qualitativement.

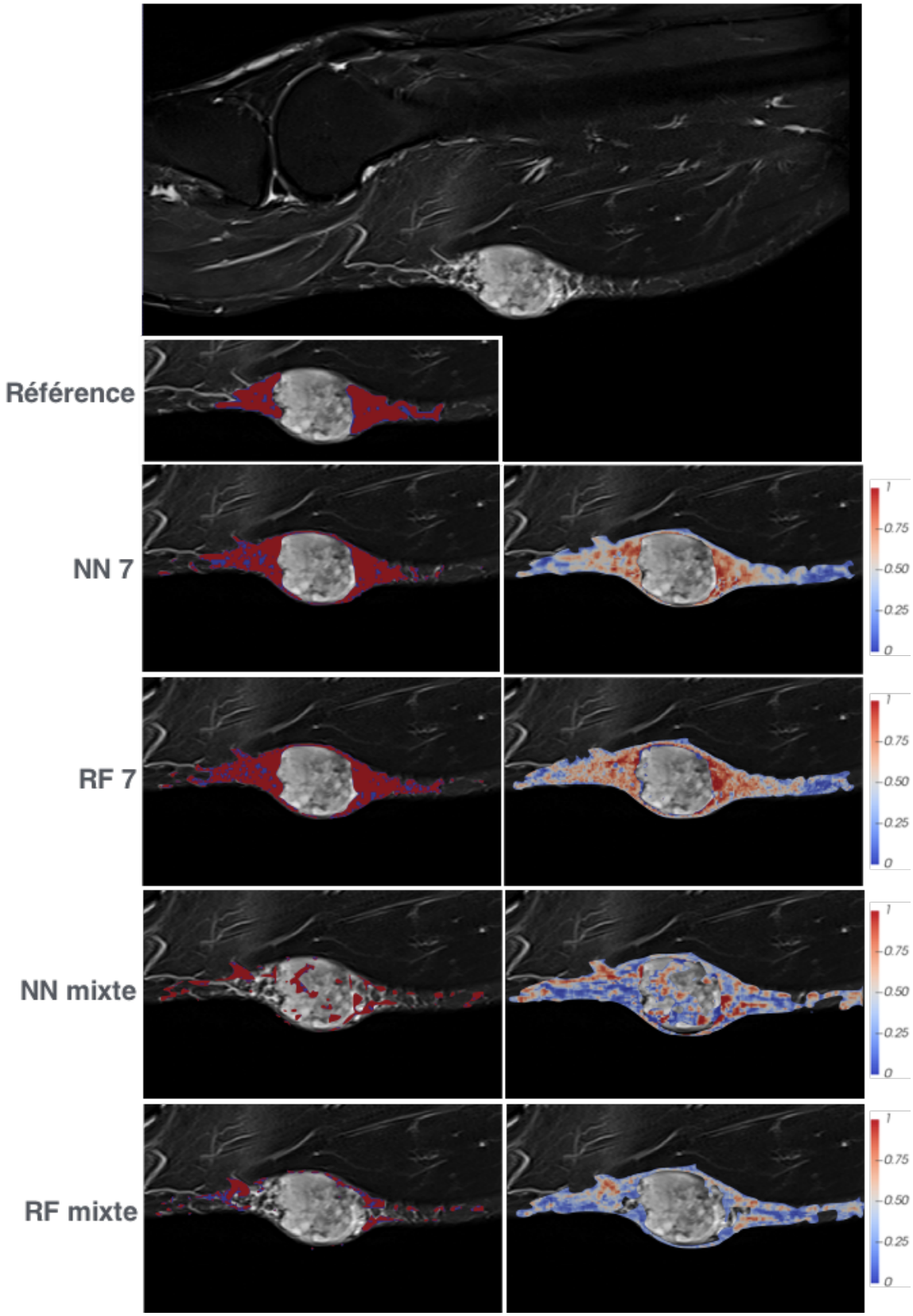


FIGURE 5.22 – Résultat de segmentation sur une coupe. À droite, les cartes de probabilités, à gauche, les masques (seuillage de la carte à P=0.5). *(images orientées à 90 °)*

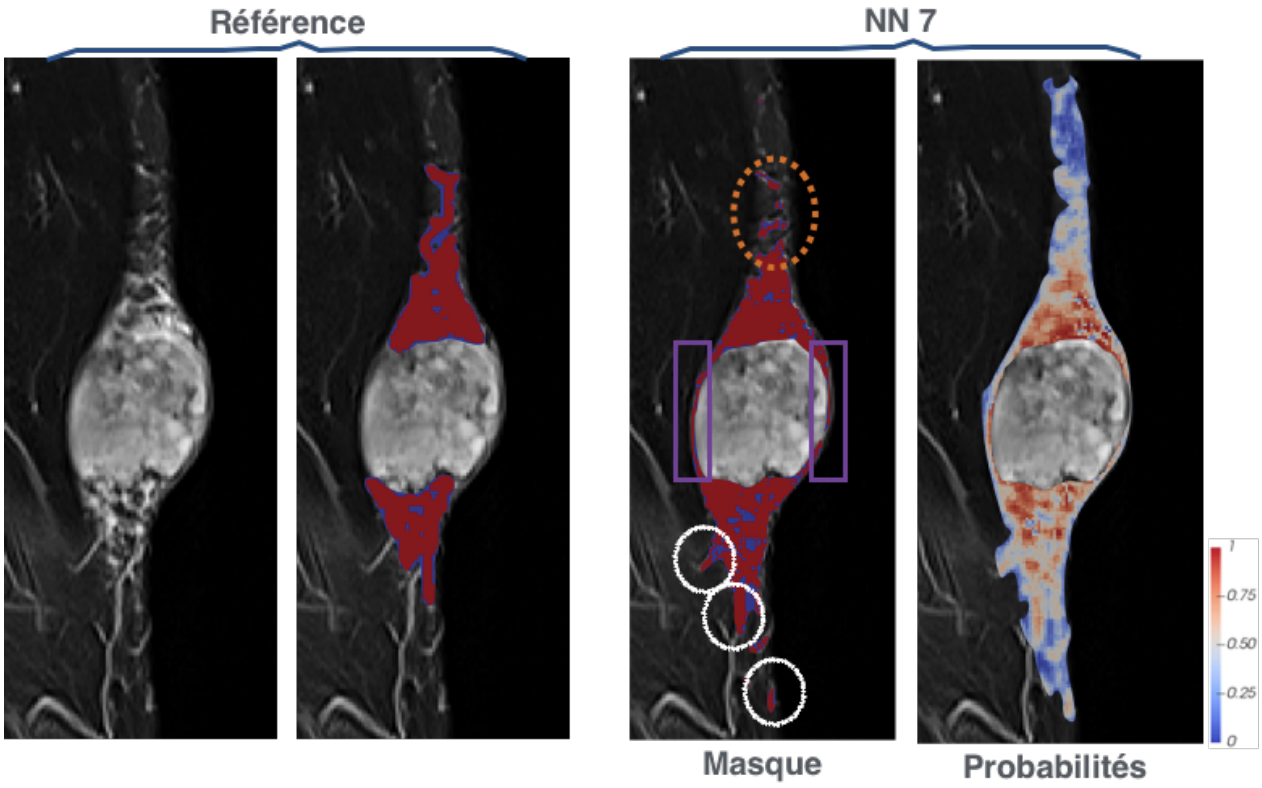


FIGURE 5.23 – Meilleure segmentation obtenue, proposée par les réseaux de neurones avec la signature à 7 variables. On y distingue trois différences principales avec la référence : des vaisseaux sanguins sont inclus dans le masque (blanc), des zones sont non continues (orange), de l'œdème est détecté sur le pourtour de la tumeur (violet).

5.7.3 Discussion

Dans l'objectif de quantifier l'évolution du volume d'œdème péritumoral, nous en avons étudié la segmentation automatique par voxel avec des méthodes classiques d'apprentissage statistique. Les résultats préliminaires obtenus en adaptant notre méthode de classification et nos études radiomiques à une donnée image avec seulement trois patients sont prometteurs. Nous avons pour cela récupéré trois types de séquence IRM et en avons dérivé des cartes de texture. Un petit nombre de descripteurs des séquences STIR sortent du lot et donnent de bons résultats. Les zones couvertes par les VOI obtenus par un réseau de neurones artificiel simple s'avèrent correctement positionnées et une grande partie des motifs de l'œdème est retrouvée.

Il reste encore difficile d'éviter d'inclure les vaisseaux sanguins dans la segmentation puisque leur signal est similaire à celui de l'œdème en STIR. L'ajout d'autres modalités pour palier au problème s'avère encore peu utile à l'heure actuelle. Les informations issues du T1-gado ou du T2 sont ponctuellement sélectionnées par ElasticNet sans qu'un descripteur ne se démarque ni n'améliore les scores. L'apport de l'approche multimodale n'est donc pas encore confirmé, sans toutefois que son intérêt puisse être écarté car elle a fait ses preuves dans d'autres études en offrant de meilleures performances pronostiques que les approches à une séquence [Li+17].

Dans l'ensemble, l'ajout des autres variables testées et les signatures longues conduisent directement au sur-apprentissage, malgré la grande quantité de voxels participant à l'entrainement des algorithmes.

Les scores d'apprentissage de la signature à 7 variables sont élevés mais il est difficile de savoir si les métriques sont totalement adaptées au problème. La fiabilité, très haute ($> 90\%$), n'est pas pertinente au vu de la quantité de voxels du fond, faciles à classer. Le déséquilibre de la distribution des voxels peut également fausser la courbe ROC et fournir une AUROC trop optimiste. Le rappel est relativement faible alors que concrètement, ce sont plutôt les faux positifs des vaisseaux sanguins (et donc la précision) qui posent problème.

A court terme, ces résultats pourraient être améliorés en retravaillant le pré-traitement des images. Le ré-échantillonnage des séquences entre elles est peut être à revoir, en préférant par exemple le sur-échantillonnage du STIR en axial et du T1-gado en sagital, de façon à favoriser la robustesse des descripteurs de texture (cf chapitre 2) et à renforcer la qualité de l'information apportée par le T1-gado .

On note également que le STIR est une modalité au ratio signal-bruit plus défavorable que celui des modalitées étudiées jusqu'à présent. L'utilisation de techniques de débruitage pourrait donc cette fois se révéler profitable, aussi bien pour aider à la délinéation que pour renforcer la stabilité des indicateurs locaux de texture des cartes.

Enfin, le caractère pixélisé des prédictions obtenues pourrait être amélioré en généralisant notre approche par voxel avec une approche par patches.

L'inconvénient majeur d'une majorité des études recourant à l'apprentissage statistique est le manque de données et celle-ci ne fait pas exception à la règle. L'approche par voxels permet de contourner ce problème en fournissant artificiellement un grand nombre d'observations (de l'ordre de 10^5 par examen). Cela ne permet toutefois pas forcément aux modèles construits

de généraliser suffisamment, ni de compenser les spécificités propres à ces trois patients. La première chose à faire serait donc d'ajouter quelques examens de façon à créer une vraie cohorte d'apprentissage variée, pour optimiser les algorithmes, limiter l'impact des erreurs de segmentation et au final limiter le sur-apprentissage.

En attendant et bien que populaires actuellement, les approches d'apprentissage profond par réseaux de neurones de convolution ne sont pas une option. Les CNN demandent en effet un nombre conséquent d'images pour l'apprentissage, d'autant plus lorsqu'elles sont 3D. Étant donné la prévalence relativement faible des sarcomes et les difficultés à harmoniser les cohortes entre les centres, il nous paraît ainsi difficile d'arriver à rassembler un jeu d'examens STIR suffisant pour satisfaire les exigences minimales d'un CNN.

Le besoin conséquent en données de ce type d'algorithmes pâtit encore plus de notre second problème, qui est le coût excessif de l'obtention de la référence. Particulièrement longue, laborieuse et complexe, la délinéation des contours réels de l'œdème souffre également de l'effet de volume partiels. Même réalisés par un.e radiologue expert.e, la référence est donc sujette à encore plus d'erreurs que d'ordinaire et borne les performances maximales de l'apprentissage. L'acquisition des labels est donc à la fois la raison pour laquelle nous souhaitons automatiser l'opération et le frein principal à la création d'une cohorte d'entrainement. L'apprentissage supervisé montre ainsi ses limites. Les solutions d'améliorations passent potentiellement par le développement de nouvelles approches non supervisées comme le clustering [Mor+18] ou les auto-encodeurs..

Le manque de données actuel demande à être compensé par l'ajout de connaissances au modèle. Nous avons ainsi pu constater l'intérêt des caractéristiques de texture dérivées des matrices *run lengths* et nous pouvons envisager que d'autres descripteurs pourraient se substituer avec succès aux variables inutiles testées ici. Nous pensons par exemple que les filtres de Gabor pourraient améliorer la prise en compte de l'information directionnelle de l'œdème [Sub+13] : il s'agit de filtres orientés qui mettent en évidence des textures et zones homogènes d'une image et informent sur son contenu énergétique dans la direction du filtre choisi.

Nous rappelons toutefois que l'objectif final n'est pas d'obtenir un contourage parfait, mais de remplacer l'estimation qualitative de l'évolution de la quantité d'œdème par une évaluation quantitative, afin d'améliorer la prédiction de la réponse au traitement. Pour le moment, les volumes mesurés sur les masques prédits sont d'ordres de grandeur similaires à ceux des références (surtout sur les cartes de probabilité), mais l'erreur commise reste encore importante (voir détails par patient en Annexe B.9).

Même sans étendre le nombre d'examens étiquetés, il serait possible de vérifier la pertinence des estimations de volume sur quelques patients de la cohorte d'entrainement de la réponse au traitement, à condition d'avoir un examen STIR au diagnostic et un autre après deux cycles de NAC. Nous imaginons pour cela apprendre l'information des trois patients ayant une référence pour prédire l'œdème des examens non étiquetés. Les volumes obtenus peuvent alors être mesurés aux deux temps et leur différence comparée à la valeur qualitative ('stable', 'diminution' etc.) fournie par les radiologues.

Conclusion

Ces résultats préliminaires indiquent qu'une approche delta-radiomique appliquée aux STS avec traitement NAC est réalisable, fournit des informations utiles pour prédire la réponse histologique après deux cycles seulement et améliore les résultats du critère RECIST 1.1 avec une signature radiomique simple. Une optimisation du modèle reste encore nécessaire avec une validation sur une cohorte plus grande, le test d'autres critères radiomiques, d'autres modalités d'imagerie, voire d'autres caractéristiques omiques.

Chapitre 6

Évaluer le temps de rechute métastatique du cancer du sein par apprentissage statistique ou modélisation

Sommaire

6.1 Contexte : cancer du sein et rechute métastatique

Le cancer du sein est le cancer le plus fréquent et la seconde cause de décès lié au cancer chez la femme [LSDMJ19]. La majorité des cancers du sein sont diagnostiqués à un stade précoce avec une lésion localisée et opérable [Noo+18]. Cependant, environ 20 à 30 % des patientes rechutent à distance après la chirurgie [LK00], ce qui suggère que des micrométastases non détectables sont déjà présentes. Leur prévalence est estimée à 90% des cas. La plupart des patientes subissent donc un traitement adjuvant, adapté à la classification moléculaire de leur tumeur (hormonothérapie, chimiothérapie, thérapie ciblée [Sen+13]). Pourtant, certaines patientes sont probablement sur-traitées et n'auraient pas dû recevoir

de traitement adjuvant aussi poussé car les toxicités associées peuvent être sévères. Pour diminuer le risque de rechute tout en évitant les effets secondaires nocifs et une procédure coûteuse, le traitement adjuvant doit être personnalisé et le nombre de cycle adapté. Une meilleure prédiction du risque d'évènements métastatiques pourrait en cela être cruciale.

Plusieurs outils pronostics ont été développés pour quantifier le risque de rechute précoce des patientes. Les outils en ligne qui calculent la probabilité de survie individuelle et le risque de rechute (*Adjuvant !* [Moo+09], modèle *PREDICT* [Wis+10]) sont basés sur des analyses multivariées et intègrent des variables cliniques comme l'âge, la taille de la tumeur, le grade histologique, les récepteurs hormonaux et l'atteinte ganglionnaire. Les tests génomiques comme *MammaPrint* et *Oncotype DX* utilisent les signatures d'expression des gènes d'échantillons de tumeur pour classer les patientes à faible ou haut risque de rechute [Duf+17]. Ces outils sont majoritairement basés sur des modèles statistiques comme les modèles à risque proportionnels de Cox (cf. section 3.2.2). La majorité d'entre eux se focalise sur la survie globale plutôt que sur le risque de rechute.

D'un autre côté, les modèles d'apprentissage statistique sont de plus en plus utilisés, traditionnellement pour des tâches de classification et de régression. L'adaptation des modèles ML à l'analyse de survie est plus récente, avec le développement de méthodes comme ElasticNet pour les modèles de Cox [Noa+11] ou l'adaptation des forêts aléatoires à la survie. D'autres encore utilisent l'apprentissage profond pour la prédiction de la survie à partir de grands jeux de données génomiques [You+17]. A l'heure actuelle, peu d'études emploient l'apprentissage statistique pour l'analyse de la survie sans évènement du cancer du sein [You+17 ; Kim+12 ; DWK05].

Les approches mécanistiques, où l'on injecte de la connaissance biologique dans le modèle, jouent un rôle de plus en plus important dans la recherche contre le cancer. Elles ont permis d'analyser le développement de la tumeur et des métastases ainsi que la réponse ou la résistance au traitement [ALM15 ; Bar+15]. Par contre, si un nombre conséquent de modèles mathématiques décrivent les dynamiques métastatiques [KTV85 ; Ret+97 ; New+12], peu d'entre eux ont été développés dans un but prédictif. La quantité limitée de mesures longitudinales est probablement en cause.

Objectif En collaboration avec Chiara Nicolò et Sébastien Benzekry, ce projet vise à évaluer la validité d'un modèle mécanistique de développement des métastases du cancer du sein (Fig. 6.11). L'objectif est d'améliorer les classifications existantes avec des prédictions individuelles de la probabilité de rechute à distance grâce aux informations du diagnostic. Pour cela, nous effectuons une comparaison entre les approches statistiques classiques, les modèles mécanistiques et l'apprentissage machine. Nous étudions notamment la classification binaire entre rechute à court et long terme grâce à des méthodes d'apprentissage classiques (régression linéaire, forêts aléatoires, etc.)

Une partie de la méthodologie et des résultats décrits sont présentés dans l'article *Machine learning and mechanistic modelling for prediction of metastatic relapse in breast cancer* ([Nic+19] en cours de publication). Ce chapitre montre notre implication dans les travaux sur l'apprentissage statistique.

6.2 Matériel et méthode

Le jeu de données initial a déjà été analysé dans l'étude de Mascarel et al. avec des outils statistiques standards comme la régression de Cox [Mas+15]. Il inclut 1160 patientes atteintes d'un cancer du sein entre 1988 et 1993 âgées de 32 à 84 ans au diagnostic. Le temps de suivi va de 3 à 18 ans (10,6 ans en moyenne). Une rechute locale ou à distance a été détectée chez 21,85% des patientes et 18,25% sont décédées des suites de leur cancer.

Les caractéristiques rassemblées comprennent 25 variables démographiques et de diagnostic clinique de routine. Les variables cliniques incluent l'âge, la taille de la tumeur au diagnostic, le statut ménopausal, le grade histologique, les classifications T et N, le type histologique et le nombre de ganglions envahis. Sont également mesurés, les récepteurs PR et ER, l'antigène Ki67, les marqueurs membranaires CD24 et CD44, l'enzyme ALDH1, les protéines BCL2, Trio et la cadhérine E. Les tumeurs sont classées HER2 positif ou négatif (variable binaire) en fonction d'une combinaison entre le pourcentage de cellules positives HER2 (variable continue) et l'intensité du HER2 (variable ordinale) [Mas+15].

6.2.1 Sélection et répartition des données

Quatre-vingt patientes sont exclues de l'ensemble initial car leurs données temporelles sont problématiques : date de de rechute antérieure à la date de diagnostic indiquée, date de décès ou de dernier contact antérieure à la date de rechute, etc.

D'autre part, 55% (642) d'entre elles ont subit un traitement locorégional uniquement (chirurgie et/ou radiothérapie), tandis que les autres ont pu être également traitées par chimiothérapie ou hormonothérapie. Comme nous souhaitons modéliser le comportement naturel des métastases, nous restreignons notre étude au premier groupe.

Catégories de classification

On considère le temps de survie globale (***O**verall **S**urvival*) comme le temps écoulé entre la date de diagnostic et celui de décès/perte de vue. Le temps de survie sans progression métastatique (***M**etastatic **F**ree **S**urvival*) est calculé entre la date de progression métastatique ou la date de décès/perte de vue d'une part et la date de diagnostic d'autre part.

La Fig. 6.1 montre la répartition des MFS des 104 patientes concernées. S'il n'est pas vraiment possible de déterminer visuellement un seuil pour les répartir pour la classification, on note que la moyenne et la médiane de la distribution sont respectivement de 5.25 (CI 4.56-5.95) et 4.84 ans. Nous choisissons donc le seuil de 5 ans au delà duquel les rechutes sont considérées à long terme. Ce seuil est également celui considéré en clinique. Il permet de plus de conserver un nombre raisonnable de patientes ayant cette durée de suivi à minima.

Les patientes décédées d'une raison autre que le cancer du sein avant cinq ans ainsi que les patientes décédées avant cinq ans sans métastase sont censurées.

Jeux de données final

En résumé, nous conservons donc une cohorte de **594 patientes** avec traitement **locorégional** qui présente un taux de **rechute métastatique avant 5 ans** de **9.25%**.

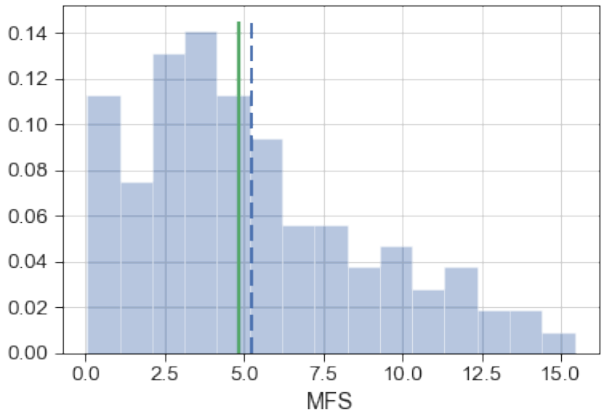


FIGURE 6.1 – Distribution des MFS des 104 patientes qui rechutent (médiane en vert et moyenne en pointillés bleus).

Cette cohorte étant très déséquilibrée, nous en étudions également un sous-ensemble sélectionné de façon à ce que le taux de rechute avant 5 ans y soit de 50%. Pour cela, on conserve l'intégralité des patientes présentant des métastases à 5 ans (55) auxquelles on adjoint le même nombre de patientes sans évènement métastatique durant cette période, tirées aléatoirement sans remise. On obtient alors une seconde cohorte de **110 patientes**.

Dans la suite du document, ces deux cohortes seront désignées par les termes $C_{9.25}$ et C_{50} respectivement.

6.2.2 Statistiques univariées

La différence entre les patientes à rechute rapide et les autres est évaluée pour chaque variable par un test de Student ou de Mann-Whitney (variables continues) et un test du χ^2 ou de Fisher (variables ordinales/catégories).

La corrélation entre les caractéristiques est déterminée par le test de rang de Spearman.

6.2.3 Classification par apprentissage statistique

Dans l'objectif de différencier les patientes présentant une rechute à court terme des autres, nous avons construit des modèles basés sur différents algorithmes d'apprentissage statistique : régression logistique (LR), machines à vecteurs de support (SVM), k-plus proches voisins (kNN), *gradient-boosting* (GB [Fri01]) et forêts aléatoires (RF). Les hyperparamètres de ces algorithmes sont évalués par un *grid-search* systématique à 3 échantillons de validation croisée en évaluant l'aire sous la courbe ROC. Ils sont regroupés en Annexe B.10.

Les modèles sont entraînés et testés par validation croisée. Les échantillons, au nombre de 9 pour une division juste du $C_{9.25}$, sont stratifiés de façon à reproduire la répartition avec/sans évènements de la cohorte (et donc le déséquilibre de $C_{9.25}$).

Nous effectuons 100 fois l'étape de validation croisée en mélangeant systématiquement la composition des échantillons. La graine d'initialisation des RF est également modifiée aléatoirement à chaque fois.

Notes sur le pré-traitement. Contrairement à nos recommandations au chapitre 4, les données cliniques manquantes ont été remplacées par les valeurs calculées par l'algorithme *MissForest* avec 100 arbres *avant* validation croisée, à l'aide du package R *missForest* [SB12]. Les modèles mécanistiques auxquels nous nous comparons n'ont pu être implémentés sous forme de pipeline : la complétion des valeurs manquante n'y est donc pas réitérée à chaque étape de la validation croisée, en raison notamment du temps de calcul nécessaire. Puisque nous souhaitons utiliser des valeurs d'entrée similaires à celles utilisées pour les modèles mécanistiques, nous avons donc également procédé à la complétion des données manquantes une unique fois, avant division du jeu de données. Nous discuterons plus loin de l'impact éventuel de cette concession.

En revanche, la normalisation des données en entrée a bien été faite selon la méthode préconisée. On rappelle que certains modèles ML ne fonctionnent correctement que si les paramètres utilisés sont tous sur la même échelle de valeurs (SVM, régression logistique sans régularisation, cf. section 4.3.) Aussi, les variables sont normalisées par *standard scaling*, à chaque phase de la validation croisée (cf. section 4.4.1).

Métriques. Nous évaluons les résultats de classification grâce à plusieurs métriques : fiabilité de la prédiction (ACC), aire sous la courbe ROC (AUROC), sensibilité (SEN = rappel), spécificité (SPE) ainsi que les PPV (précision), NPV , score F1 (combinaison de la PPV et de la sensibilité) et *average precision* (AP).

6.3 Résultats

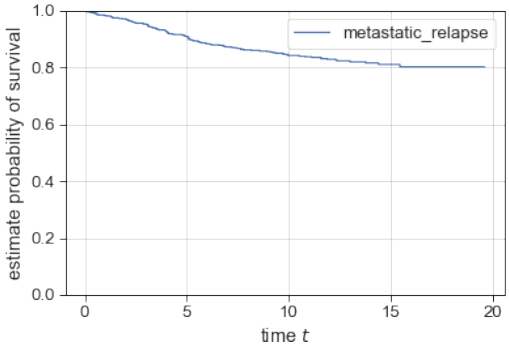
6.3.1 Analyse de survie et univariée

La Fig. 6.2 montre les analyses de survie sans évènement métastatique de la totalité de la cohorte et des exemples suivant la valeur de quelques variables représentatives. Sans surprise la courbe confirme que les patientes ayant une tumeur de grade 3 ou positives au HER2 (cancers à croissance plus rapide) rechutent significativement plus rapidement que les autres grades ou profils génétiques.

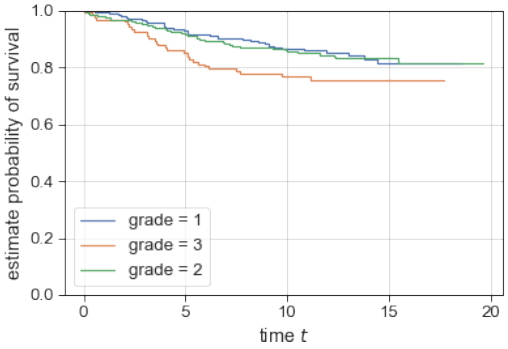
Parmi les variables étudiées, douze s'avèrent significativement associées à la survie sans évènement métastatique : le Ki67 ($p < 10^{-8}$), l'EGFR ($p < 10^{-6}$), l'ER ($p < 10^{-4}$), le PR ($p < 10^{-4}$), le statut HER2 (binaire, $p < 10^{-3}$), le volume clinique ($p < 10^{-3}$), le TNM_T ($p = 0.001$), le CK56 ($p = 0.0012$), l'intensité HER2 (ordinal, $p = 0.0018$), le VIM ($p = 0.01$), le HER2 (continu, $p = 0.03$) et l'absence ou présence de traitement par radiothérapie ($p = 0.04$).

La matrice de corrélation (Fig. 6.3) montre des corrélations fortes pour une partie des variables. Sans surprise, la taille de la pièce clinique est fortement corrélée à la taille histologique et aux classifications TNM [1], les trois descripteurs HER2 sont corrélés entre eux, la ménopause est corrélée à l'âge. Certains marqueurs comme l'ER ou le PR sont inversement corrélés à de nombreux autres (EGFR, VIM, CK56, HER2, KI67).

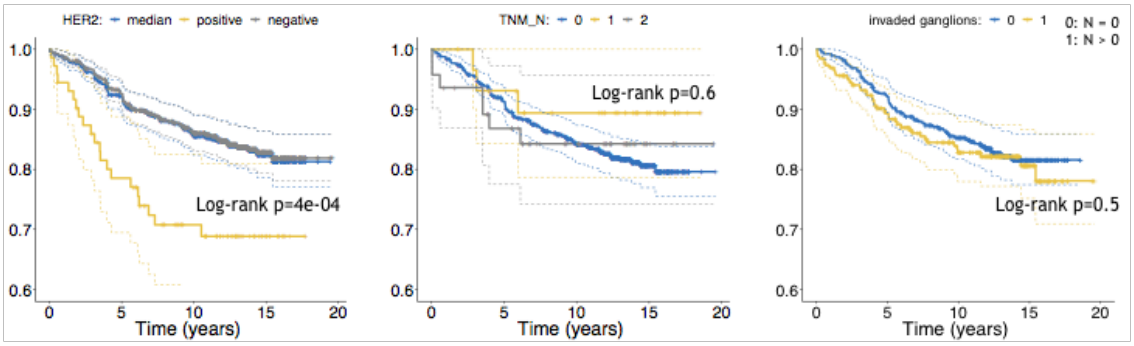
1. Système international de classement des cancers basé sur leur extension anatomique.



(a) MFS total.



(b) MFS en fonction du grade.



(c) MFS en fonction du HER2, du TNM-N et du nombre de ganglions envahis. *source : Chiara Nicolò*

FIGURE 6.2 – Courbes Kaplan-Meier de la survie sans évènement métastatique (MFS).

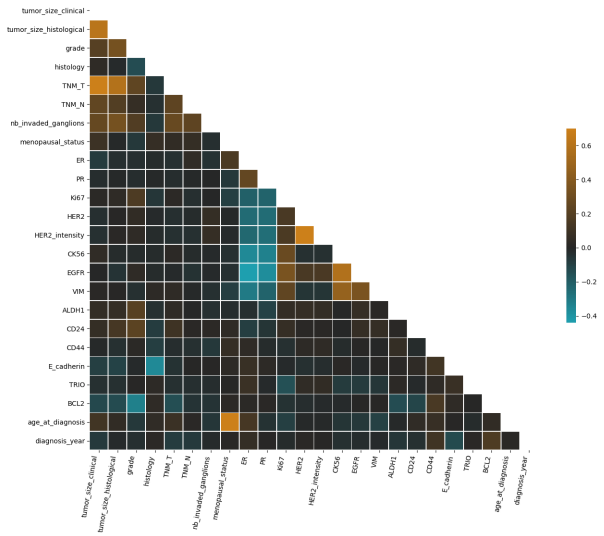


FIGURE 6.3 – Matrice de corrélation (Spearman) des variables cliniques de la cohorte $C_{9.25}$

6.3.2 Apprentissage statistique

Les scores moyens des 100 répétitions de validation croisée des différents algorithmes testés sur la cohorte entière ($C_{9.25}$) sont rassemblés Table 6.1 et représentés Fig. 6.4a. Leurs courbes ROC sont visibles Fig. 6.4b.

Deux algorithmes, les SVM et le GB obtiennent le score maximal à l'entrainement ce qui évoque un risque de sur-apprentissage. Leurs courbes d'apprentissage ainsi que celle des kNN [2] sont visibles Annexe B.12 et confirment cette intuition. Les SVM effectuent une mauvaise performance globale. Malgré une grande fiabilité, le score F1 moyen des kNN est quasi nul et leur AUROC la plus faible. Le GB atteint l'*average precision* la moins mauvaise.

La fiabilité moyenne des RF est de 82.4% sur $C_{9.25}$ (Fig. 6.6a). Avec la LR, ils donnent la meilleure AUROC à 0.75. Toutefois, la fiabilité et l'AUROC ne sont pas les meilleurs métriques à considérer pour un jeu de données non équilibré [SR15]. La NPV moyenne est élevée pour les deux (> 0.90) : la proportion manquée de patientes rechutant avant 5 ans est faible. Cependant, la raison principale est la proportion faible de patientes sans rechute avec 5 ans. La RF a une meilleure spécificité (0.87) et la LR une meilleure sensibilité (0.71). Les scores F1, PPV et AP restent très bas.

$C_{9.25}$	ACC	AUROC	SPE	SEN	NPV	PPV	F1	AP	Train
RF	0.824	0.745	0.866	0.413	0.935	0.239	**0.303**	0.24	0.870
LR	0.664	**0.752**	0.659	**0.712**	**0.957**	0.176	0.282	0.22	0.677
GB	0.898	0.712	0.979	0.106	0.915	0.344	0.161	**0.26**	1.0
kNN	**0.906**	0.623	**0.996**	0.024	0.909	**0.410**	0.046	0.17	-
SVM	0.869	0.639	0.949	0.086	0.911	0.148	0.109	0.14	1.0

TABLE 6.1 – Scores moyens en validation croisée sur $C_{9.25}$

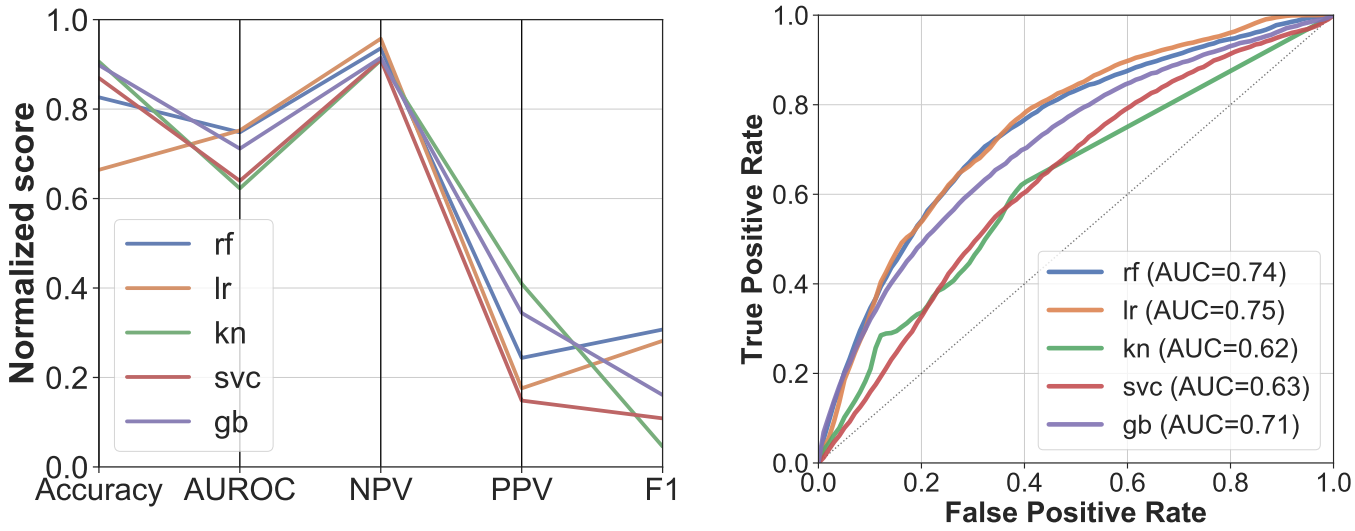
Pour pallier ce problème, nous utilisons un jeu de donnée équilibré sous-échantillonné pour entraîner à nouveau les classifieurs. Les scores obtenus sont regroupés Table 6.2 et Fig. 6.6. Les courbes ROC et graphes précision-rappel sont visibles 6.7 et 6.8 respectivement.

C_{50}	ACC	AUROC	SPE	SEN	NPV	PPV	F1	AP	Train
RF	0.638	0.702	0.699	**0.576**	0.623	0.657	**0.614**	0.67	0.842
LR	**0.659**	**0.706**	**0.781**	0.538	**0.628**	**0.710**	0.612	0.68	0.709
GB	0.605	0.643	0.635	0.575	0.599	0.612	0.592	0.63	1.0
kNN	0.561	0.579	0.662	0.460	0.551	0.577	0.512	0.58	-
SVM	0.567	0.571	0.565	0.569	0.567	0.567	0.567	0.50	1.0

TABLE 6.2 – Scores moyens en validation croisée sur C_{50}

Cette fois, la fiabilité est diminuée par rapport à la cohorte $C_{9.25}$ (66% pour la LR et 65.9% pour les RF), tout comme l'AUROC (0.69 et 0.70) et la NPV (0.63 and 0.62). On observe également une inversion de tendance : la RF acquiert une meilleure sensibilité tandis que la spécificité de la LR augmente. En revanche, les faux positifs sont moins nombreux pour les deux, la PPV et le F1 sont meilleurs pour le C_{50} : 0.66 et 0.61 pour les RF et 0.71 et 0.61

2. Qui, pour rappel, n'ont pas de score d'entrainement.



(a) Scores moyens.

(b) Courbes ROC.

FIGURE 6.5 – Performances des algorithmes testés sur $C_{9.25}$ avec une validation croisée.

pour la LR respectivement. De plus, les courbes précision-rappel montrent une performance bien plus élevée avec des scores AP supérieurs à 0.65.

Les résultats des autres algorithmes sont donnés à titre de comparaison. Leurs performances restent très faibles, les kNN et SVM n'apportent presque aucune information.

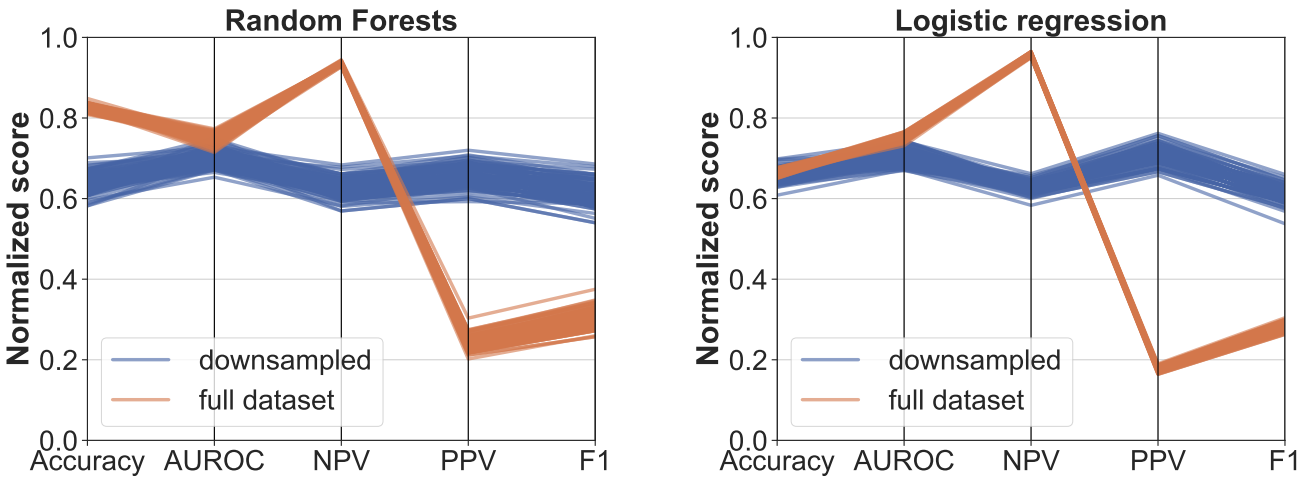
On note également que les modèles entraînés sur le C_{50} semblent bien mieux calibrés (Fig. 6.9), en particulier pour les RF. Les probabilités calculées seraient donc plus fiables.

Les courbes d'apprentissages évaluées sur l'aire sous la ROC sont visibles Fig. 6.10 pour les deux cohortes. Celles des RF (Fig. 6.10a) sont caractérisées par une grande différence de score entre la validation croisée et l'entrainement, ce dernier approchant la valeur maximale même pour un jeu de données de grande taille. À l'inverse, le modèle LR (Fig. 6.10b) a un score d'entrainement bien plus faible et le score de test n'est pas spécialement amélioré passé une certaine quantité de données. Il y a visiblement un problème de sous-apprentissage.

6.3.3 Sélection des variables

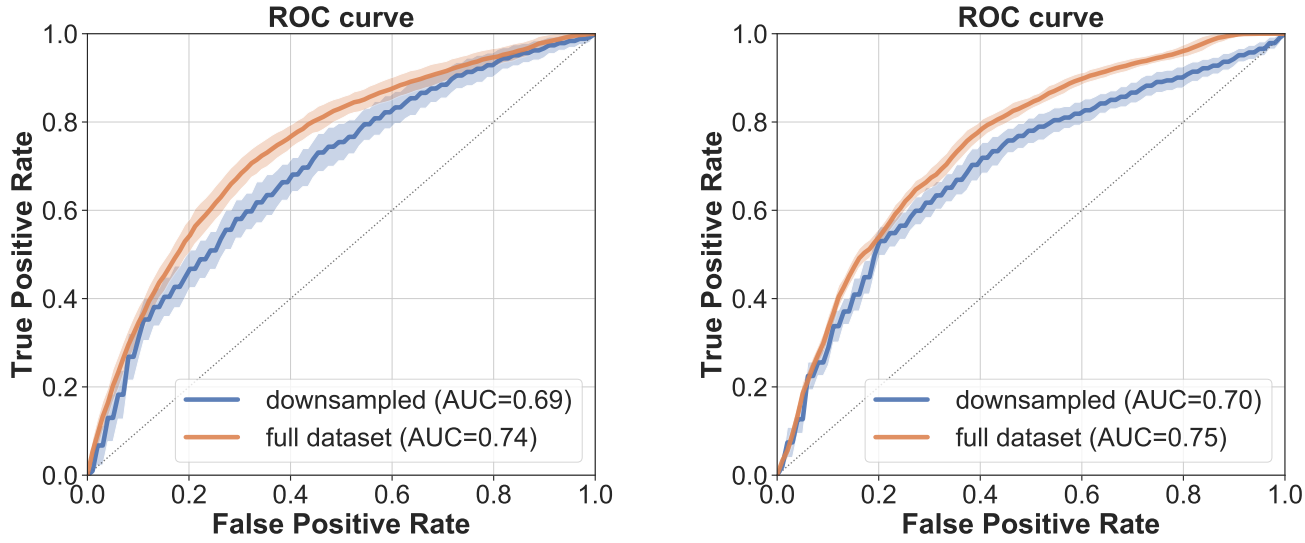
La taille de la cohorte et le sous-apprentissage de certains algorithmes n'encouragent pas à réduire les dimensions du jeux de données en priorité. Nous avons tout de même souhaité comparé les résultats présentés à ceux obtenus avec une sous-sélection de variables, de façon à éliminer les caractéristiques éventuellement inutiles. Nous obtenons deux signatures à utiliser avec les deux cohortes et les deux meilleurs algorithmes, LR et RF :

1. L'ensemble des variables statistiquement significatives à l'analyse univariée (12 caractéristiques, voir 6.3.1).
2. Le même ensemble, à partir duquel on ne conserve qu'une variable par groupe de descripteurs corrélés (la plus significative). Nous obtenons 6 variables : Ki67, EGFR, VIM, le status H2R, la taille et le traitement par radiothérapie.



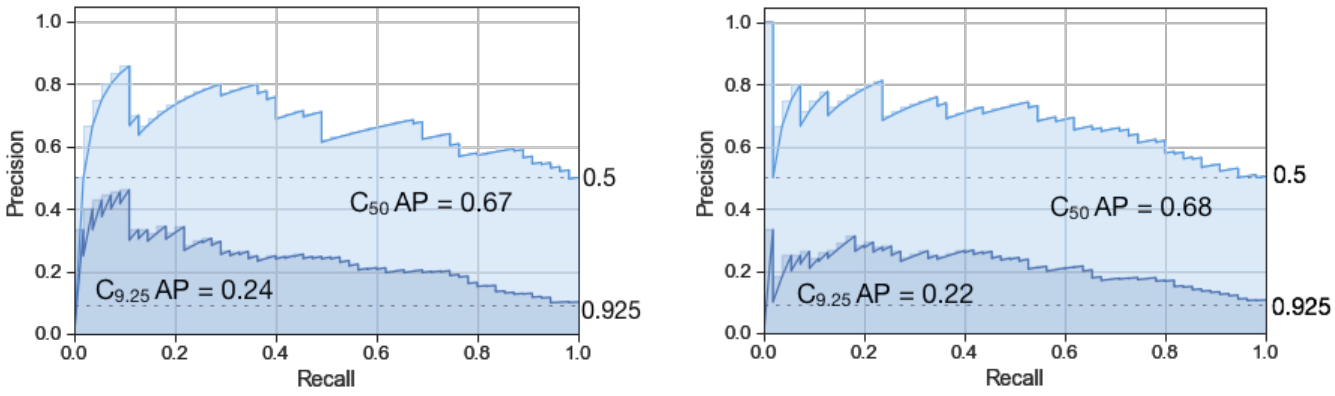
(a) RF (b) LR

FIGURE 6.6 – Scores des modèles RF et LR construits avec l'ensemble des caractéristiques sur les cohortes $C_{9.25}$ et C_{50}.



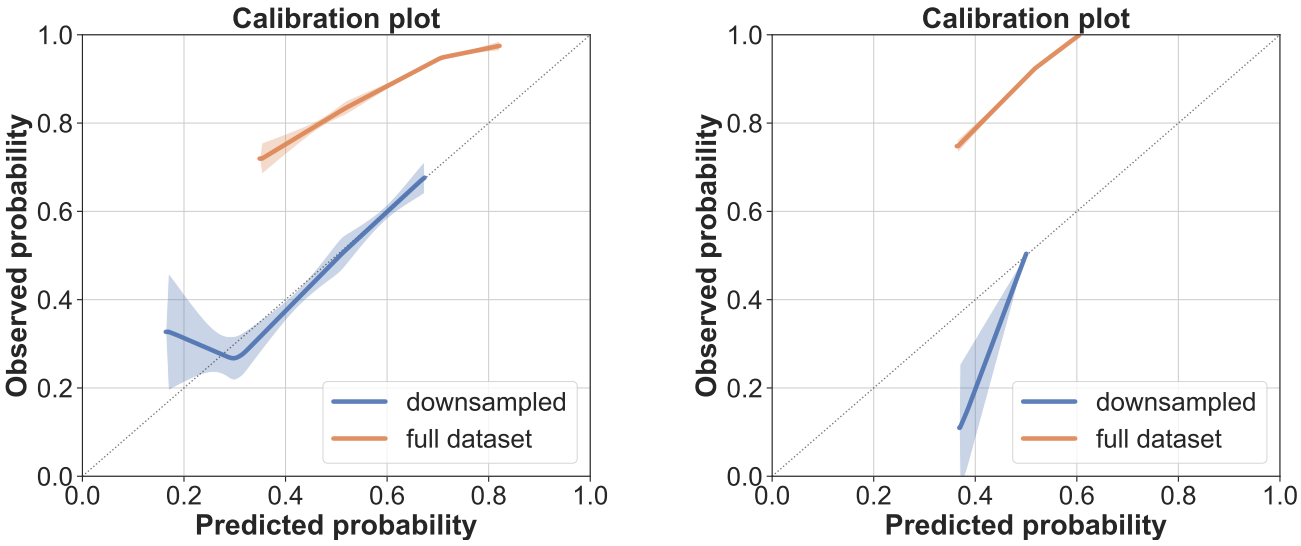
(a) RF (b) LR

FIGURE 6.7 – Courbes ROC des modèles RF et LR construits avec l'ensemble des caractéristiques sur les cohortes $C_{9.25}$ et C_{50}.



(a) RF. (b) LR.

FIGURE 6.8 – Courbes précision-rappel des modèles RF et LR construits avec l'ensemble des caractéristiques sur les cohortes $C_{9.25}$ et C_{50}. Les lignes pointillées représentent la limite informative (= proportion de chaque cohorte).



(a) RF (b) LR

FIGURE 6.9 – Courbes de calibration des modèles RF et LR construits avec l'ensemble des caractéristiques sur les cohortes $C_{9.25}$ et C_{50}. On trace ici la probabilité de survie sans évènement à 5 ans et non celle de rechute.

Les résultats sont visibles en Tables 6.3 pour $C_{9.25}$ et 6.4 pour C_{50}.

$C_{9.25}$	variables	ACC	AUROC	SPE	SEN	NPV	PPV	F1	AP	Train
LR	6	0.704	**0.780**	0.700	**0.739**	**0.963**	0.201	**0.316**	**0.26**	0.708
LR	12	0.720	0.769	0.723	0.690	0.958	0.202	0.313	0.24	0.723
(rappel) LR	25	0.664	0.752	0.659	0.712	0.957	0.176	0.282	0.22	0.677
RF	6	0.765	0.756	0.788	0.540	0.944	0.207	0.299	0.24	0.796
RF	12	0.800	0.762	0.831	0.499	0.942	0.232	**0.316**	0.24	0.835
(rappel) RF	25	**0.824**	0.745	**0.866**	0.413	0.935	**0.239**	0.303	0.24	0.870

TABLE 6.3 – Scores moyens en validation croisée sur $C_{9.25}$ avec trois méthodes de sélection des variables.

C_{50}	variables	ACC	AUROC	SPE	SEN	NPV	PPV	F1	AP	Train
LR	6	0.670	**0.753**	0.770	0.570	**0.642**	0.713	**0.633**	**0.73**	0.693
LR	12	**0.671**	0.740	0.775	0.568	**0.642**	**0.716**	**0.633**	0.70	0.705
(rappel) LR	25	0.659	0.706	**0.781**	0.538	0.628	0.710	0.612	0.68	0.709
RF	6	0.628	0.725	0.669	**0.587**	0.619	0.640	0.612	0.69	0.743
RF	12	0.635	0.709	0.685	0.586	0.623	0.651	0.616	0.68	0.780
(rappel) RF	25	0.638	0.702	0.699	0.576	0.623	0.657	0.614	0.67	0.842

TABLE 6.4 – Scores moyens en validation croisée sur C_{50} avec trois méthodes de sélection des variables.

Pour la LR, les performances sur cohorte entière s'avèrent légèrement meilleures pour les deux signatures courtes. Fiabilité et AUROC augmentent de quelques points, spécificité et sensibilité aussi. Quelques faux positifs semblent avoir disparu puisque la PPV et la spécificité sont (légèrement) améliorées. La signature à 6 variables peu corrélées semble fournir les meilleurs résultats.

En revanche les performances des RF profitent moins de cette réduction de dimensions même si les faux négatifs diminuent sur $C_{9.25}$ (ce qui augmente AUROC et sensibilité).

Il est difficile de tirer une tendance sur C_{50} car la différence entre les résultats est peu significative.

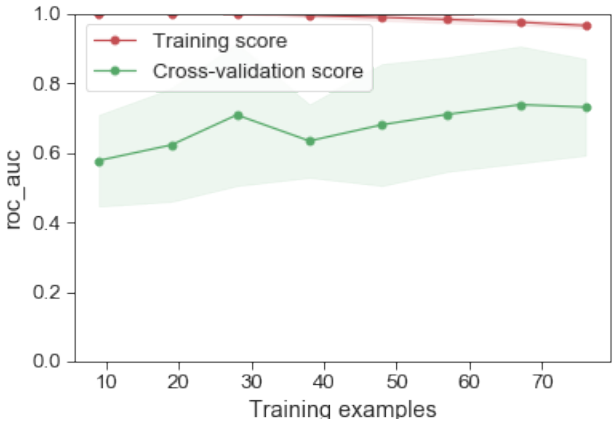
6.4 Discussion

Avec une fiabilité supérieure à 80% et un score AUROC à 0.75, la classification précoce de la rechute métastatique des patientes atteintes d'un cancer du sein semble tout à fait réalisable avec des forêts aléatoires dans la cohorte initiale ($C_{9.25}$). Pourtant l'étude des scores issus de la matrice de confusion (NPV, spécificité, etc.) donne un aperçu plus mitigé et il est au final plus compliqué de conclure sur l'apport des différents classifieurs.

L'objectif principal est en effet d'identifier correctement le maximum de patientes qui rechutent, ce qui revient à minimiser la quantité de faux négatifs. Un modèle qui se trompe rarement sur ce type de patiente a une NPV et une sensibilité élevées. La NPV des cinq modèles s'avère effectivement supérieure à 0.9. Mais puisque le jeu de données est déséquilibré

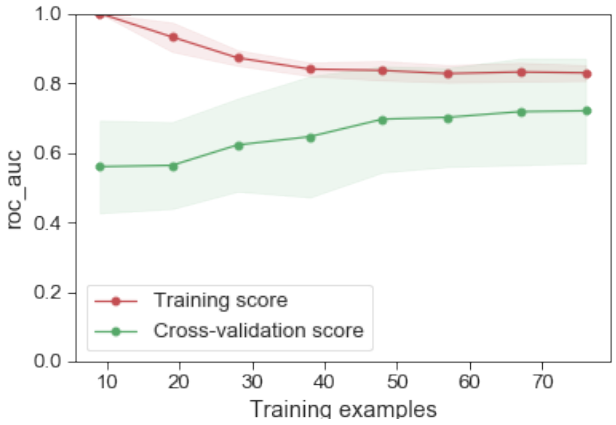
(a1) $C_{9.25}$

(b1) $C_{9.25}$



(a2) C_{50}

(a) RF



(b2) C_{50}

(b) LR

FIGURE 6.10 – Courbes d'apprentissage des modèles RF et LR construits avec l'ensemble des caractéristiques sur les cohortes $C_{9.25}$ et C_{50}. Le score évalué est l'AUROC.

(moins de 10% de rechutes à court terme), c'est la sensibilité qui sera le score le plus révélateur de la qualité du modèle (la NPV croît automatiquement lorsque la prévalence est faible comme dans $C_{9.25}$). Or la sensibilité des classifieurs testés est faible voire très faible, à l'exception notable de celle de la régression logistique, qui indique que plus de 71% des valeurs positives (patientes à rechute précoce) ont été trouvées. Cette sensibilité de la LR est même améliorée à 74% lorsqu'on utilise une sous-sélection de six variables significativement associées à la rechute précoce. Ce score est correct mais toutefois probablement insuffisant pour être valorisé.

L'étude d'une cohorte équilibrée, C_{50}, permet de confirmer que la NPV chute au profit de la PPV pour tous les classifieurs. Malgré une meilleure calibration des modèles, les performances globales sont moyennes avec une fiabilité plafonnée à 67% (LR avec 6 ou 12 variables). Sur le tiers de patients mal prédits en LR, la majorité des erreurs sont des faux négatifs (65%) : ce modèle est donc finalement mal adapté à notre problématique.

Les mécanismes d'erreur de nos classifieurs peuvent être mieux appréhendés avec leurs

courbes d'apprentissage. Celles des RF montrent un écart de score important entre apprentissage et validation qui se réduit légèrement quand on ajoute des données d'entrainement, signe d'une variance conséquente. Il y a donc potentiellement un léger sur-apprentissage qui pourrait bénéficier d'un jeu de donnée de plus grande taille, notamment dans le cas C_{50}. À l'inverse, l'écart entre les courbes de la LR est réduit mais la valeur atteinte est finalement assez basse et dénote d'un modèle biaisé, sûrement trop peu complexe pour modéliser correctement les relations entre les caractéristiques. Réduire le nombre de variables pour simplifier le modèle pourrait ne pas sembler pertinent au vu des résultats faibles à l'entrainement. Pourtant, une sous-sélection de caractéristiques améliore légèrement les résultats en LR, puisque les variables retenues sont moins corrélées (voir section 4.3.1). Nous pourrions donc plutôt envisager de diminuer le terme de régularisation.

En résumé, l'ensemble des résultats obtenus montre qu'une classification de la rechute à court et long terme est envisageable même si encore difficile à mettre en place : les scores reflètent un apport d'information réel mais modeste des classifieurs. Il est nécessaire d'équilibrer le sur-apprentissage des forêts aléatoires et le sous-apprentissage de la régression logistique pour trouver un modèle pertinent.

L'étape de complétion des données manquantes pourrait être un léger facteur de sur-apprentissage dans notre étude. Réalisée en début de traitement sur l'ensemble des données, elle devrait normalement être effectuée de la même manière que la normalisation (c'est à dire à chaque étape de la validation croisée). Cependant, nous souhaitions comparer nos résultats à ceux du modèle mécanistique (voir section suivante) à partir de données strictement identiques. Nous avons donc conservé la même stratégie de complétion que nos collègues, stratégie motivée par le temps de calcul élevé de l'algorithme *MissForest* et des validations croisées des modèles mathématiques.

Nous avons tout de même comparé nos résultats en apprentissage à ceux obtenus avec une complétion réalisée à chaque cycle de validation croisée, avec la médiane des échantillons d'entrainement. Visibles en Annexe B.11, les performances sont quasi identiques en LR et en RF : l'effet de la méthode de complétion s'avère donc finalement négligeable, probablement en raison de la quantité modeste de données incomplètes.

En revanche, il est envisageable que la cible de prédiction, déterminée par un seuil semi-arbitraire à 5 ans, soit non optimale voire erronée. Des recherches plus approfondies sont bien évidemment essentielles sur ce sujet et le passage à une étude de régression est également une étape prochaine nécessaire au regard des erreurs de classification. Quelques premiers résultats avec un seuil positionné sur la médiane ou sur deux ans n'ont à l'heure actuelle pas amélioré les résultats. Cependant, bien qu'arbitraire, le seuil de 5 ans correspond à la référence clinique, car le taux de rechutes métastatiques y est bien plus élevé que dans les années qui suivent. Réussir une classification automatique selon cette référence présente donc un réel intérêt clinique.

Nos modèles pâtissent en outre très probablement du type de caractéristiques rassemblées dans notre jeu de données. Puisque les scores à l'entrainement sont faibles, elles ne sont

probablement pas assez pertinentes ou insuffisantes pour caractériser la vitesse de rechute métastatique. De plus, récoltées fin des années 80 - début 90, les données rassemblées ne correspondent plus toutes à la routine clinique actuelle. L'apprentissage gagnerait donc à être renforcé par des données plus récentes, en terme de patients comme de variables descriptives, tout en veillant à conserver un suivi de 5 ans minimum. Une telle augmentation du volume de donnée serait en outre bénéfique compte tenu de la taille modérée du jeu actuel au regard de la prévalence du cancer de sein. Un second jeu de données indépendant devrait également permettre d'effectuer une validation externe des modèles.

Le besoin d'améliorer l'information d'entrée nous suggère que les modèles ML pourraient probablement bénéficier de l'apport des examens d'imagerie effectués en cours de routine clinique. L'IRM est un élément clé de l'évaluation de la taille tumorale et de la réponse au traitement du cancer du sein. De nombreuses études ont récemment montré l'intérêt des biomarqueurs radiomiques pour décrire les lésions, prédire l'agressivité ou le risque de rechute (voir l'état de l'art par [Cri+18]).

Cependant, le recueil systématique des données d'imagerie, long et fastidieux, n'a pas encore été réalisé sur la cohorte que nous étudions. Les examens ont été effectués dans des centres multiples et sur une période de temps étendue. Les modalités d'acquisition des images sont donc très hétérogènes, ce qui pourrait complexifier l'étude radiomique. Nous avons cependant présenté aux chapitres 2 et 5 une chaîne de traitement apte à améliorer les conditions d'extractions et la fiabilité des descripteurs de texture et de forme.

Cette étude montre également les difficultés qu'il y a à choisir une métrique pour comparer les performances de plusieurs modèles. La problématique et la distribution du jeu de données sont bien sûr les premiers paramètres à prendre en compte. Cependant, il est aussi utile de calculer les scores couramment employées dans la littérature afin de pouvoir s'y mesurer.

6.5 Apports de l'approche mécanistique

La classification de la survie sans évènement est une étape de simplification du problème de l'analyse de la survie des patientes. Nous évoquons ici les deux approches de régression développées par nos coéquipiers.

6.5.1 Modèles d'apprentissage statistique de la survie sans

Forêts aléatoires de survie

Le principal inconvénient des algorithmes de classification utilisés est leur incapacité à gérer les données censurées. L'algorithme des forêts aléatoires de survie (***R****andom* ***S****urvival* ***F****orests*) est une extension des forêts aléatoires pour l'analyse des durées de survie censurées [IK16]. Comme pour les RF, l'objectif est de développer des arbres sur des sous-ensembles de données *bootstrap*, en utilisant une sélection aléatoire de variables à chaque nœud. La différence réside dans le critère de division des nœuds, qui devient un *critère de survie*, impliquant à la fois temps de survie et information de censure. On obtient une estimation

ensembliste de la fonction de risque cumulée pour chaque individu, en moyennant l'estimation cumulée du risque de tous les arbres [IK16].

L'implémentation du module R *randomForestSRC* a été utilisée, avec 100 arbres par forêts et un test de *log-rank* pour diviser les nœuds. Les autres paramètres ont été optimisés de façon à maximiser l'indice de concordance calculé sur les données *out-of-bag*[3].

Comme les RF classiques, les RSF peuvent être utilisées pour évaluer l'importance relative de chaque variable utilisées dans la construction des forêts et effectuer ainsi une sélection de caractéristiques.

Modèle mécanistique de la survie sans évènement

La figure 6.11 résume le principe du modèle de survie développé. Le développement de la tumeur primaire d'un individu i est donnée par un modèle de Gompertz, couramment utilisé pour modéliser la croissance des tumeurs (voir [Nor88] pour un exemple d'application sur le cancer du sein) :

$$V_p^i(t) = e^{\frac{\alpha^i}{\beta^i}(1-e^{-\beta^i t})} \tag{6.1}$$

avec $V_p^i(t)$ le nombre de cellules dans la tumeur au temps t ainsi que α^i et β^i les paramètres de croissance du modèle de Gompertz. β^i est fixé et le volume de la tumeur est converti en nombre de cellules grâce à l'hypothèse $1mm^3 = 10^6$ cellules [SMS95].

Chaque métastase est supposée croître au même taux α^i que la tumeur primaire, à partir du volume V_0 d'une unique cellule. Le nombre total de métastases au temps t dépend d'un paramètre de dissémination μ^i et est donné par l'équation 6.2.

$$N^i(t) = \int_0^t \mu^i V_p^i(s) ds \tag{6.2}$$

Le temps écoulé entre le diagnostic et la détection d'une première métastase ayant atteint le seuil de visibilité est appelé TTR (*time-to-relapse*). Il dépend du volume de la tumeur primaire au diagnostic $V_p^i(t)$, du taux de croissance α^i et du taux de dissémination métastatique μ^i, à condition que le modèle ait prédit la création d'au moins une métastase avant le diagnostic : $N^i(t^i_{diag}) \geq 1$ avec :

$$t^i_{diag} = -\frac{1}{\beta^i} log(1 - \frac{\beta^i}{\alpha^i} log V^i_{diag}) \tag{6.3}$$

où t^i_{diag} est le temps écoulé entre l'apparition d'une première cellule tumorale et la détection d'une tumeur primaire et V^i_{diag} est le volume de la tumeur au diagnostic.

Dans le cas où $N^i(t^i_{diag}) = 0$, le TTR est infini.

La calibration du modèle mécanistique est effectuée en utilisant un modèle à effets mixtes (modèle statistique comportant une part d'effets aléatoires). Plus de détails sur cette étape et le modèle en général sont disponibles dans la publication associée [Nic+19].

3. Non sélectionnées par *bootstrap*.

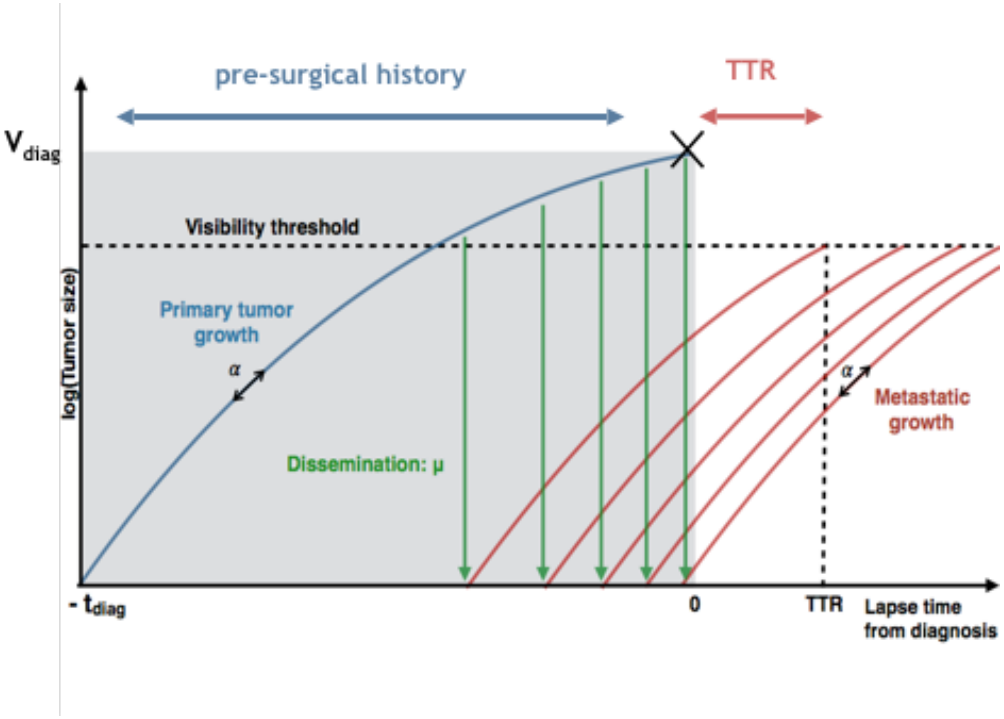


FIGURE 6.11 – La croissance de la tumeur primaire et des métastases est caractérisée par un même paramètre de croissance α. L'émission de métastases à partir de la tumeur primaire se produit à un taux qui dépend de son volume et d'un paramètre de potentiel métastatique μ. Le TTR (*time-to-relapse*) est le temps écoulé entre le diagnostic et le moment où une première métastase atteint une taille suffisante pour être visible. *source : Chiara Nicolò*

Validation des résultats

Les RSF et le modèle mécanistique sont validés par validation croisée à 10 échantillons. Les capacités d'appréciation du temps de survie sont données par l'indice de concordance de Harell ou *c-index*.

Cet indice est utilisé pour évaluer des temps de survie quand des données censurées sont présentes. Pour deux évènements Z_1 et Z_2 arrivant au temps T_1 et T_2, le *c-index* représente la probabilité $P(Z_1 > Z_2 | T_2 > T_1)$ selon le modèle considéré. En d'autres termes, il représente la probabilité qu'un patient 1 qui rechute *avant* un patient 2 obtienne une probabilité de rechute plus que forte que le patient 2. Le c-index est équivalent à l'aire sous la courbe ROC pour un temps t fixé en l'absence de données censurées.

Dans un second temps, les probabilités de rechute à $t = 5$ ans sont calculées pour les patientes de la cohorte $C_{9.25}$ (sans données censurées) pour chaque méthode. La prédiction (binaire) à 5 ans associée est obtenue à partir de ces probabilités en sélectionnant le seuil de binarisation donnant les sensibilité et spécificité les plus élevées.

Ces probabilités et prédictions sont évaluées avec l'aire sous la courbe ROC, la fiabilité, les PPV, NPV, sensibilité et le score F1 ainsi que le score AP.

6.5.2 Résultats et discussion

L'indice de concordance des RSF atteint une valeur maximale de 0.67 (95% CI, 0.66-0.69) pour le sous-ensemble des cinq variables les plus sélectionnées en moyenne par les arbres [4] : le Ki67, la taille de la tumeur, l'âge, l'EGFR et le CD44.

Le modèle mécanistique calibré intègre les variables Ki67 et CD44 dans l'estimation du paramètre α et la variable EGFR dans le paramètre de dissémination μ. Son indice de concordance est de 0.65 (95% CI, 0.60-0.71).

A titre de comparaison, les modèles de Cox ont un indice de concordance compris entre 0.60 et 0.71.

Les résultats obtenus par les modèles de survie à $t = 5$ ans sur la cohorte entière sont visibles Table 6.5. Nous calculons également les prédictions de classification RF sur la signature à cinq variables des RSF.

Étonnamment, les modèles ML de classification à 5 et 25 variables obtiennent des performances similaires voire supérieures à celles du modèle ML de survie (RFS) à 5 ans. Pourtant, ce dernier est capable d'intégrer les données censurées. Les scores obtenus en classification RF sont également légèrement supérieurs à ceux obtenus par le modèle mécanistique.

Toutefois, l'ensemble des modèles de survie présente une sensibilité supérieure à celle des classifieurs. Le classifieur RF à 5 variables améliore un peu cette métrique par rapport à celui à 25 variables (respectivement 0.52 et 0.41), mais le score reste modeste (0.52). L'approche mécanistique obtient le meilleur score (0.75) parmi l'intégralité des modèles étudiés jusqu'ici.

4. On rappelle que cette méthode de sélection présente le risque de voir l'importance relative des variables fortement corrélées diminuer.

Prédiction à 5 ans	AUROC	ACC	PPV	NPV	F1	AP	SEN
Random **S**urvival **F**orests	0.72	0.67	0.17	0.95	0.27	0.22	0.65
Modèle mécanistique	0.73	0.68	0.19	**0.96**	0.30	0.22	**0.75**
Modèle de Cox	0.71	0.71	0.19	0.95	0.29	0.20	0.65
classification RF 5	**0.75**	0.78	0.22	0.94	**0.31**	0.21	0.52
(rappel) classification RF 25	***0.75***	***0.83***	***0.24***	*0.94*	*0.30*	***0.24***	*0.41*

TABLE 6.5 – Scores moyens des modèles de survie sur $C_{9.25}$ à 5 ans. Les modèles de Cox et RSF finaux sont construits avec les cinq variables les plus souvent sélectionnées par la RSF uniquement. Nous les comparons aux RF sur 5 variables et sur 25 variables.

L'objectif de cette étude comparative était d'établir une méthodologie pour utiliser le modèle mécanistique et l'améliorer. Un examen plus fourni des variables cliniques étudiées est prévu, de même qu'une étude détaillée de la puissance prédictive du modèle.

Les classifieurs ML comme la régression logisitique et les forêts aléatoires semblent à même de donner une première estimation de la rechute métastatique du sein à court terme. Les modèles obtenus souffrent toutefois d'un manque de sensibilité, pouvant amener à oublier des patientes susceptibles de rechuter. Ils restent une très bonne première approche, rapide à appliquer, avant de passer aux modèles de survie.

Basé sur les mécanismes biologiques métastatiques, le modèle mécanistique donne un aperçu des processus biologiques à l'œuvre, ce dont les modèles d'apprentissage statistique sont moins capables. Les processus étudiés sont ici simplifiés de façon à conserver deux facteurs, les taux de croissance et de dissémination. Avec un indice de concordance de 0.65, le modèle mécanistique fournit une performance d'évaluation de la survie sans évènements similaire à celle des forêts aléatoires de survie ou du modèle de Cox. Ces scores se rapprochent également de ceux des modèles actuels standard spécifiques au cancer comme *Adjuvant !* (*c-index* = 0.71). Des méthodes avancées d'apprentissage profond n'ont pas spécialement surpassé ces résultats, avec un *c-index* de 0.68 sur le jeu de données BRCA [You+17].

Le modèle proposé du temps de rechute représente un premier essai d'approche prédictive à l'échelle de l'individu et pourra être amélioré de plusieurs façon. D'un côté, de la variance inexpliquée demeure malgré l'inclusion de covariables issues du jeu de données, suggérant que d'autres biomarqueurs pourraient potentiellement améliorer les prédictions.

A cet égard, des modèles incluant des signatures génétiques ont montré un haut potentiel prédictif comparés aux modèles basés sur des variables cliniques et histologiques standards [Vij+02]. D'autre part, le modèle mécanistique pourrait être redéfini en ajoutant le phénomène de dormance des métastases, censé expliquer les rechutes se manifestant de nombreuses années après la chirurgie [UP11]. Nous pensons également aux caractéristiques de texture et de forme à l'imagerie pour améliorer la description des tumeurs primaires.

Les deux approches, mécanistique et apprentissage statistique, pourront toutefois se compléter et s'améliorer l'une et l'autre. Les meilleures résultats pourraient aussi être obtenus par vote conjoint des deux méthodes. En outre, le modèle mécanistique intègre déjà trois des variables sélectionnées par les RSF parmi ses covariables. A l'inverse, les modèles de classification pourront intégrer certains paramètres construits par les modèles mécanistiques.

Discussion générale et perspectives

Sommaire

1 Bilan

Le suivi des patients et la routine clinique en oncologie génèrent des données en nombre et qualité croissants au fur et à mesure des avancées scientifiques et technologiques. Cette richesse d'information couplée à la multiplicité des pathologies et des problématiques est parfois sous-exploitée, par manque de temps, de moyens techniques, de personnel qualifié ou d'outils adaptés.

C'est à ce titre que dans le cadre de ces travaux de thèse, nous nous sommes intéressés à l'exploitation des données de routine, en gardant deux objectifs en tête : automatiser et quantifier la description des lésions d'un patient et utiliser les caractéristiques d'une cohorte pour hâter la prédiction précoce du pronostic clinique.

1.1 Contributions techniques

Nous avons tout d'abord évalué l'apport de l'imagerie IRM, utilisée en oncologie dans de nombreuses étapes de la prise en charge d'un patient. Plus précisément, nous avons choisi d'extraire des **caractéristiques radiomiques IRM en sélectionnant un ensemble de descripteurs simples et bien documentés** de la morphologie, de l'intensité des voxels et de la texture d'une région d'intérêt.

Nous avons notamment estimé l'impact de notre méthode de reconstruction 3D sur les caractéristiques de forme en évaluant leur robustesse suivant la taille du VOI et le positionnement de la grille de discrétisation. Si on confirme qu'une taille minimale est nécessaire pour

des mesures représentatives de la réalité, nous avons également constaté la **reproductibilité forte des marqueurs de forme** aux caractéristiques de reconstruction. Une étude similaire du ré-échantillonnage de la grille a montré qu'une forte augmentation de la résolution de l'image impacte certaines mesures de forme, comme l'aire de la surface. Globalement, **les indicateurs morphologiques les plus robustes sont le volume, l'élongation et le *flatness***.

Des barrières méthodologiques spécifiques à l'IRM en font jusque là une des modalités d'imagerie les moins utilisées pour l'extraction de biomarqueurs [Tra+18]. Nous avons donc implémenté une **correction de la dérive spatiale d'intensité**. Puis, nous avons tenu compte des contraintes liées à l'analyse simultanée de plusieurs examens en incluant une **harmonisation de la signification de l'intensité du signal** par alignement d'histogramme. Ce processus modifie les valeurs des voxels et donc les descripteurs liés (histogramme et texture) de façon à les rendre comparables d'un examen à l'autre. Des indicateurs comme le kurtosis et le skewness varient ainsi nettement après normalisation mais d'autres comme l'**entropie de l'histogramme ou l'homogénéité** y sont moins sensibles. **Les descripteurs des niveaux de gris sont également beaucoup moins stables au ré-échantillonnage** que les caractéristiques morphologiques, car là encore sensibles à l'interpolation des voxels.

Dans un second temps, **nous avons développé des modèles d'apprentissage statistique pour la classification précoce du pronostic des patients**. Comme dans la plupart des jeux de données en oncologie, les cohortes rassemblées contiennent un nombre limité d'observations et la répartition entre les classes est déséquilibrée.

Nous avons adressé le premier problème en limitant les algorithmes testés aux modèles capables de fonctionner avec un ensemble d'entrainement relativement restreint. Nous avons également **réduit les dimensions des jeux de données** afin de ne conserver que les variables les plus informatives. Enfin, nous avons systématiquement opté pour une **validation croisée pour augmenter la puissance statistique** de nos résultats.

Lorsque le déséquilibre de distribution s'est avéré problématique, nous avons ré-échantillonné la cohorte concernée et porté une attention particulière aux mesures de performance de la classification en privilégiant les métriques peu dépendantes de la distribution, comme la sensibilité ou le score F1.

1.2 Contributions cliniques

Grâce à une méthode de traitement simple et dont l'impact sur les mesures a été vérifié, nous avons surmonté les difficultés liées à l'interprétation des niveaux de gris de l'IRM et aux cohortes de taille réduite.

Dans un premier temps, nous avons ainsi pu analyser **le lien entre la texture d'une tumeur au diagnostic et les évènements de rechute** éventuels détectés le long de son suivi (chapitre 3). Le caractère hétérogène, aléatoire sans être forcément très varié des voxels des lésions (représenté par les mesures de l'entropie de l'histogramme et du cluster shade à fine résolution), nous est apparu significativement associé à la rechute des patients atteints d'un carcinome du canal anal.

Par la suite, l'étude de la prédictibilité de la réponse au traitement sur les IRM de STS nous a matériellement permis d'aller encore plus loin (chapitre 5). À l'**information de texture** a été rajoutée l'**étude de la forme** des tumeurs et l'**estimation** des caractéristiques **de sa périphérie** par les spécialistes. Un début d'analyse longitudinale avec le calcul de caractéristiques **delta-radiomiques** sur une courte période s'est avéré plus intéressante que les mesures uniques au diagnostic. De fait, si l'évolution de l'entropie de l'histogramme est encore apparue comme étant une mesure pertinente de la réaction d'une lésion à la chimiothérapie, l'élongation du volume tumoral et la propagation de l'œdème périphérique se sont également révélés d'excellents descripteurs. Les modèles statistiques multivariés ont laissé place aux algorithmes d'apprentissage statistiques pour apporter une information prédictive et détecter la fraction de patients peu sensibles à leur traitement néo-adjuvant. Réalisée **de façon précoce**, cette prédiction permet d'adapter le traitement en route alors qu'il n'était jusque là possible de connaître son impact qu'une fois terminé. Elle dépasse également les performances prédictives de la référence actuelle, le critère RECIST.

Pour aller plus loin, nous avons tenté de classer les voxels appartenant ou non à l'œdème de quelques patients avec de très bons résultats préliminaires.

Même sans tirer partie de l'information radiomique, ces mêmes **modèles d'apprentissage soutiennent la comparaison** de l'analyse multivariée et **des modèles mathématiques** mécanistiques pour prédire les rechutes à distance précoce du cancer du sein (chapitre 6). Cependant, les performances des trois approches s'avèrent acceptables sans être brillantes et profiteraient à coup sûr de l'utilisation de l'imagerie.

2 Limitations

Quelques limitations générales de nos contributions doivent ainsi être discutées.

Tout d'abord, deux choix d'implémentation sont en contradiction avec certaines recommandations obtenues grâce aux études de reproductibilité du chapitre 2.

Le chapitre 3 présente notamment des travaux chronologiquement antérieurs à ceux du chapitre 2, réalisés sur une cohorte IRM n'ayant pas fait l'objet d'une normalisation de l'intensité du signal. Il est donc envisageable que l'absence de traitement ait biaisé le résultat et plus spécifiquement l'importance des descripteurs skewness et cluster shade, considérés potentiellement inconstants par l'étude de reproductibilité. En revanche, le résultat sur l'entropie de l'histogramme a moins de chances d'être modifié puisqu'il s'agit d'une caractéristique stable avec ou sans traitement.

Par ailleurs, les ré-échantillonnages des IRM utilisées dans les chapitres 3 et 5 ont finalement été effectués avec un interpolateur linéaire, quand les conclusions du chapitre 2 suggèrent qu'intensité et texture sont mieux préservées avec une interpolation utilisant le plus proche voisin. Toutefois, des études plus récentes comme celles de Larue et al [**Larue2017**] ont abouti à des conclusions inverses et déconseillent son utilisation. Aussi, l'interpolation linéaire reste un choix convenable et plus de recherches sont nécessaires pour déterminer les raisons de ces résultats opposés.

Nous avons souhaité privilégier l'interprétabilité des descripteurs en nous limitant volontairement à un nombre restreint de caractéristiques radiomiques classiques. Les caractéristiques calculables se comptent en effet par centaines et la probabilité que certaines d'entre elles soient inutiles voire nuisent aux performances de la modélisation est élevée. Limiter le nombre de variables nous a permis de faire ressortir plus facilement les informations d'intérêt et de discuter des phénomènes biologiques sous-jacents. Nous avons ainsi également eu le temps de nous attarder sur les paramètres de l'extraction et de vérifier la stabilité des mesures suivant les manipulations de l'image.

Toutefois l'implémentation actuelle de notre chaîne de traitement autorise l'intégration rapide de nouvelles caractéristiques, comme nous l'avons fait avec les descripteurs des matrices *run length* finalement utilisés dans la segmentation de l'œdème au chapitre 5. D'autres caractéristiques radiomiques pourront donc être incluses dans de futures études.

Nous pensons cependant que les études de prédiction bénéficient essentiellement du développement de nouveaux descripteurs adaptés à chaque pathologie ou problématique, plutôt qu'à l'étude massive d'une grande quantité de caractéristiques. Ainsi, Kritter et al. ont développé un marqueur d'hétérogénéité propre aux gliomes de bas grade en TEP, qui, associés aux variables cliniques, a une capacité diagnostique supérieure aux critères classiques [**Kritter2018**]. Pour notre part, nous avons montré que la description de l'œdème est un facteur pronostic prometteur que nous continuerons à explorer.

Enfin, nous nous sommes concentrés sur les problématiques de classification binaire et n'avons pas creusé le sujet des classifications multiples ou de la régression (quitte à dichotomiser une étude de survie au chapitre 6). En effet, ce type d'apprentissage, plus simple à mettre en place et à évaluer qu'une classification multiple, demande également moins de données qu'une étude de régression. Le résultat binaire qu'il fournit est aussi privilégié en clinique car il simplifie l'analyse et la prise de décision.

3 Discussion

En plus des différentes discussions détaillées dans les chapitres précédents, nous souhaitons développer ici notre point de vue sur les problématiques plus générales de la fouille de données médicales.

Chaque problématique clinique a des besoins différents en terme de performances à atteindre pour qu'un modèle soit validé et transférable. S'ils présentent trop de faux négatifs, les algorithmes de détection d'une rechute ou de segmentation d'une tumeur seront recalés. À l'inverse, des faux positifs fréquents sont à éviter lorsque l'on cherche uniquement à détecter les patients pour qui un traitement lourd s'avère inutile. Le travail conjoint avec une équipe médicale spécialisée est donc indispensable tout le long du développement d'une étude, pour mieux cerner les enjeux et les problématiques, tenir compte des contraintes cliniques, mettre en valeur les résultats positifs et analyser les erreurs.

Il est en effet indispensable de garder en tête l'objectif final de nos travaux et des études radiomiques en général, à savoir l'aide à la décision en clinique. Bien sûr à l'heure actuelle et

même si les résultats sont encourageants, notre méthode est encore loin du transfert technologique effectif et il reste même difficile de conclure sur certains résultats. Notre modèle de prédiction de la réponse au traitement des sarcomes échoue ainsi encore à classer les patients aux tumeurs inhabituelles comme celles présentant une masse nécrotique surabondante. Le modèle de prédiction de la rechute à distance du cancer du sein n'est qu'aux débuts de son développement et profitera nécessairement de l'ajout de l'information des biomarqueurs radiomiques et de la collaboration avec les modèles mécanistiques. Il faudrait en outre valider les résultats sur de grosses cohortes prospectives et multicentriques.

Malgré ces défauts, nous soulignons que les conditions pour valider nos travaux de façon à évoluer vers un prototype utilisable en clinique sont réunies. L'ensemble de nos modèles et analyses a été construit avec les données et recueils issus des protocoles de routine clinique habituels. Ils ne nécessitent pas d'examens médicaux supplémentaires spécifiques, pénibles et contraignants pour le patient et coûteux (en temps, ressources humaines et fonds) pour la clinique.

Notre méthodologie permet de nous adapter aux jeux de données de taille réduite, aux distributions de classes peu équilibrées et aux observations partielles. Les modèles apprennent les paramètres de signatures courtes de caractéristiques interprétables qu'il est possible de relier au contexte biologique. Les temps de construction sont modérés à très réduits et la prédiction de la classe des nouvelles observations est quasi instantanée.

Tous ces facteurs sont indispensables à l'éventuelle création d'un prototype et à son déploiement en clinique. Un modèle qui ne fonctionnerait qu'avec un très grand nombre de données longitudinales ou qui ne serait pas exécutable dans un temps raisonnable (compte tenu des moyens de calcul d'un service radiologique) rendrait en effet tout le travail totalement inutile.

Tout comme l'inévitable manque de données (et le temps à consacrer à leur étiquetage), les éléments que nous venons d'évoquer nous ont en revanche conduit à laisser de côté les méthodes d'apprentissage profond pour les études pronostiques.

Domaine pourtant en vogue grâce à l'augmentation des ressources et de la puissance de calcul, l'apprentissage profond est un champ d'étude prometteur donc les performances surpassent celles des algorithmes d'apprentissage traditionnels, notamment pour analyser des structures complexes dans des jeux de données de hautes dimensions [**LeCun2015**]. Le *deep learning* en imagerie IRM est appliqué avec succès pour la reconstruction d'images, l'analyse radiomique, le recalage, la segmentation, la correction d'artefacts, etc. [**Lundervold2018**].

Les études diagnostiques et pronostiques sont également nombreuses, mais l'analyse des variables obtenues et des phénomènes biologiques liés peut souffrir de l'opacité des structures utilisées (empilement des couches de neurones et complexité des architectures).

Les modèles pré-entrainés sur des bases de données comme ImageNet [**Deng2009**] (*transfert learning*) sont parfois utilisés dans les applications médicales souffrant d'un manque de données ou de moyens de calculs. [**Raghu2019**] montrent cependant que les gains en performance s'avèrent finalement limités et atteignables par des structures beaucoup plus petites.

D'autre part, une étude par [**Geirhos2018**] explique en quoi les CNN sont largement biaisés de façon à privilégier l'information de texture plutôt que l'information sur la mor-

phologie. Or, notre analyse de la réponse au traitement des sarcomes a finalement beaucoup moins bénéficié des caractéristiques de texture que ce à quoi on aurait pu s'attendre compte tenu de leur popularité dans la littérature [5]. Les descripteurs de la morphologie ont largement participé à la prédiction (élongation, volume d'œdème) et s'avèrent également plus robustes aux traitements et à la reconstruction selon nos analyses. Nous pensons donc qu'une approche CNN qui inverse ce biais pour renforcer l'impact de la forme comme celle de [**Geirhos2018**] présente un intérêt tout particulier pour nos travaux.

4 Perspectives

L'étude de la réponse au traitement des STS nous a conduit à en développer deux extensions visant à adresser les limitations liées aux approximations humaines assimilées dans les modèles et à renforcer l'automatisation de l'entièreté de la chaîne de traitement. Ces travaux sont en cours et fournissent des résultats prometteurs.

4.1 Analyse quantitative de l'œdème

La variable prédictive la plus efficace de la réponse au traitement des STS est l'évolution de la quantité d'œdème péri-tumoral (cf. section 5.7). Estimation qualitative, cette variable a nécessité une double vérification par deux radiologues pour s'assurer d'un consensus. Nous avons donc tenté d'en automatiser la segmentation avec une classification des voxels par forêts aléatoires et réseaux de neurones en utilisant plusieurs types de séquences IRM et des cartes de texture.

Les premiers classifieurs développés donnent des résultats très encourageants sur les quelques examens dont le contours de référence est disponible. Les séquences STIR et les cartes associées permettent d'avoir une bonne idée du volume d'œdème, même si quelques sections sont encore omises ou si des vaisseaux sanguins sont encore inclus par erreur dans les contours.

Pour corriger les erreurs de classification restantes, il nous faut encore renforcer l'impact et la qualité de l'information apportée par les autres modalités d'imagerie dans les modèles ou trouver d'autres indicateurs pertinents. L'étude et le recueil de données n'en sont qu'à leur début et des méthodes alternatives comme la segmentation par patches sont encore à tester.

4.2 Segmentation automatique des sarcomes

La seconde étape encore automatisable de notre chaine de traitement est la segmentation des lésions. Nous avons vu un exemple de segmentation manuelle des sarcomes sur plusieurs séquences au chapitre 1, avec un écart entre les volumes obtenus pouvant dépasser 20% du total. Les sarcomes ont à la fois une forme et des bords bien définis qui facilitent la discrimination des limites et une structure complexe de tissus hétérogènes, avec des zones

5. Cette popularité est toutefois peut être justement en partie due à un biais de confirmation provoqué par le nombre conséquent d'études utilisant des CNN.

de nécrose, de fibrose, etc. L'automatisation de la tâche est délicate mais peut bénéficier de l'apprentissage profond et notamment des réseaux de neurones à convolution.

Pour pallier au manque de données de la cohorte de patients atteints d'un STS (IRM T2 de 65 individus), nous construisons un réseau U-Net [**Ronneberger2015**], spécialisé dans la segmentation d'images biomédicales et prévu pour fonctionner avec moins d'images d'entrainement qu'un CNN classique. De plus, nous divisons les volumes 3D par coupe de façon à réaliser des segmentations 2D. Chaque coupe est dupliquée plusieurs fois avec application d'une rotation aléatoire, de façon à multiplier artificiellement les images d'entrainement pour renforcer l'apprentissage.

A l'heure actuelle, cette étude 2D n'est pas encore concluante. Nous projetons donc de rajouter les données d'IRM en T1 car T1 et T2 sont généralement utilisés conjointement par les radiologues pour la délinéation manuelle. Nous souhaitons également tester une méthode de patches 3D ou encore utiliser des auto-encodeurs entrainés sur des coupes sans lésion pour détecter les tumeurs pendant le test. Cette dernière approche présente l'énorme avantage de ne demander qu'une supervision faible consistant uniquement à étiqueter les coupes de façon binaire (absence ou présence d'une tumeur).

Dans un second temps, l'augmentation artificielle des données (les rotations) est une solution qui doit à terme être complétée par l'ajout de données réelles. Puisque nous n'avons pas besoin cette fois d'une série longitudinale par patient, le recueil de données supplémentaires est facilité. L'ensemble de données publiques "Soft-tissue Sarcoma" par M. Vallières [Val+15] contient par exemples 51 IRM compatibles avec notre étude et d'autres jeux de données destinés à l'étude du grade peuvent être rassemblés à l'Institut Bergonié.

Nous envisageons également d'intégrer dans le réseau les données cliniques de chaque examen pour orienter la segmentation. Les informations comme la localisation de la tumeur dans le corps, sa profondeur sous-cutanée ou encore son grade sont potentiellement reliées à l'aspect de la tumeur et nous supposons que l'ajout de ce type de variables aura une influence sur les résultats de la segmentation automatique. Pour inclure ces méta-données cliniques, il faudrait passer par une approche de réseaux mixtes combinant quelques couches de neurones classiques aux réseaux convolutionnels (voir [**Xu2016** ; **Andrearczyk2016** ; **Cheerla2019**]).

La segmentation de la tumeur et de l'œdème sont les deux dernières étapes conduisant à l'automatisation complète du pipeline des données à la prédiction, dans le cas de l'évaluation précoce de la réponse des STS au traitement. D'autres perspectives à plus ou moins long terme visent à utiliser ce pipeline pour d'autres cas d'application ou à l'enrichir avec d'autres approches.

4.3 Perspectives de recherche générales

Les techniques de traitement et d'analyse de l'image et de classification présentées peuvent bien sûr être adaptées à d'autres pathologies ou problématiques. La prédiction du grade anatomopathologique des sarcomes sur T1-gado par classification ou *clustering* est ainsi à l'état de projet. Nous envisageons également d'inclure l'analyse d'autres séquences ou modalités d'imagerie sur lesquelles les biomarqueurs ont fait leurs preuves comme le TEP-scan [**Baskaran2018** ; Val+15] pour étudier les variations d'hétérogénéités du volume tumoral.

Enfin, les signatures moléculaires issues de l'étude du génome sont une source potentielle d'enrichissement de nos modèles. Nous avons utilisé cette information dans le projet d'évaluation de la rechute métastatique du sein ; nous espérons à terme pouvoir inclure ce type de données pour évaluer le risque de rechute des sarcomes grâce à la signature CINSARC [**Chibon2010**].

L'étude conjointe de l'apprentissage statistique et des modèles mathématiques (équations aux dérivés partielles) est une autre perspective initiée par l'étude évoquée au chapitre 6. Les deux approches ont plusieurs façons de s'enrichir mutuellement.

Le modèle delta-radiomique, en comparant deux points dans le temps, est la première étape vers l'étude longitudinale des lésions. Nous pensons que le développement de modèles de l'évolution de certaines caractéristiques comme le volume des lésions [**Berment2016**] peut fournir des paramètres intéressants à inclure parmi les variables d'apprentissage d'un algorithme de classification. L'importance de l'évolution de l'élongation chez les sarcomes nous incite par exemple à développer un modèle spatial d'évolution de leur forme, sachant qu'un tiers de la cohorte dispose d'un troisième IRM pendant la NAC.

A l'inverse, l'apprentissage statistique peut fournir des intuitions permettant de construire un modèle mathématique, en sélectionnant les variables d'intérêt par exemple [**Rudy2019**]. Il peut également faciliter la personnalisation d'un modèle mécanistique en donnant un a priori sur ses paramètres en fonction de variables non modélisées ou en construisant un modèle réduit pour accélérer son évaluation.

Annexe A

Prédiction de la survie sans progression des cancers du pancréas

Matériel

L'étude est monocentrique et rétrospective. Elle rassemble une cohorte de 89 patients pris en charge au CHU de Bordeaux entre 2010 et 2014 pour un adénocarcinome pancréatique classé initialement *borderline* ou localement avancé. Leur traitement est composé d'une première chimiothérapie puis d'une radiochimiothérapie en l'absence de progression. Les données recueillies regroupent données cliniques et images.

Le recueil clinique est constitué des données dédomgraphique, d'anatomopathologie et de progression/survie : âge, genre, classification de la tumeur, invasion artériel et veineuse, temps de suivi, etc.

Les données images sont issues des scanners de diagnostic (CT_{diag}, d'évaluation post-chimiothérapie (CT_{pCT}, après 4 cures) et post-radiochimiothérapie (CT_{pRCT}) au temps portal.

FIGURE A.1 – Exemple de coupe du CT_{diag} d'un pancréas à temps portal avec sa **ROI**. *source : CHU Bordeaux (Haut l'Évêque).*

Méthode

Des ROI 2D sont contourées à chaque examen : une coupe est sélectionnée de façon à délinéer le plus grand diamètre de tumeur possible sur le CT_{diag}. Elle est ensuite sélectionnée à la même position sur les deux examens suivant. On note que les prothèses biliaires sont exclues pour éviter au mieux les artefacts qu'elles provoquent. Un contourage fidèle de la tumeur étant compliqué à réaliser, nous avons décidé d'exclure toute analyse de forme ou de volume de l'étude.

Les intensité de pixels des CT-scans, mesurées en unités Hounsfield [1], représentent la densité des tissus et sont interprétables quel que soit l'examen considéré. Nous ne cherchons donc pas à normaliser les intensités entre les patients, mais les CT_{pCT} et CT_{pRCT} de chaque patient sont ré-échelonnés en fonction de la moyenne du CT_{diag}. Les coupes et ROI associées sont ré-échantillonnées de façon à obtenir des pixels de $1 \times 1mm^2$ de surface (voir chapitre 2).

Nous procèdons à l'extraction des données d'intensité (histogramme) et de texture (caractéristiques de Haralick à 1, 2, 5 pixels de voisinage dans toutes les directions)(voir chapitre 1).

L'évolution relative entre les examens a été évaluée significative pour la majorité des descripteurs (test de Wilcoxon) ce qui nous a conduit a utiliser les descripteurs delta-radiomiques entre CT_{diag} et CT_{pCT} (Δ_{0-1})et CT_{pCP} et CT_{pRCT} (Δ_{1-2}).

Les valeurs manquantes des données cliniques ont été supplées par des méthodes appropriées au type de variable : médiane pour une variable continue, mode ou valeur neutre pour une variable binaire etc...

La classification a été testée avec une validation croisée à 5 échantillons avec 7 algorithmes aux hyperparamètres choisis empiriquement ; régression logistique (LR), kNN, SVM, SVM linéaire, gradient boosting (GB), forêts aléatoires (RF) et classification naïve bayésienne. Les valeurs des attributs sont normalisées et une analyse en composantes principales est réalisée à chaque cycle de validation pour réduire le nombre de variables (voir chapitre 4).

Résultats

	fréquence	meilleur algo	ACC	AUROC	précision	rappel
Rechute	82%	RF	0.84 ± 0.05	0.63	0.84 ± 0.05	0.98 ± 0.08
Décès	72%	GB	0.8 ± 0.06	0.73	0.83 ± 0.1	0.94 ± 0.07
Rechute précoce	42%	GB	0.62 ± 0.2	0.58	0.60 ± 0.25	0.58 ± 0.14

TABLE A.1 – Meilleur algorithme et scores moyens de la validation croisée pour les trois types de classification de la survie des patients.

Au final plusieurs facteurs ont mené à l'exclusion d'un grand nombre de patients. En raison de la forme allongée du pancréas et du plan de coupe, certaines ROI ont une surface trop réduite (< 50 pixels). D'autre part les images présentent des artefacts nombreux, souvent

1. Échelle quantitative décrivant la radiodensité des tissus[**Feeman2010**].

dus à la présence de prothèses biliaires [2] (voir Fig. A.2). On note également une inflammation post-traitement très visible sur certains examens ainsi que la présence de bile dans les tissus.

En retirant ces cas problématiques, la cohorte a été réduite à **50 individus**.

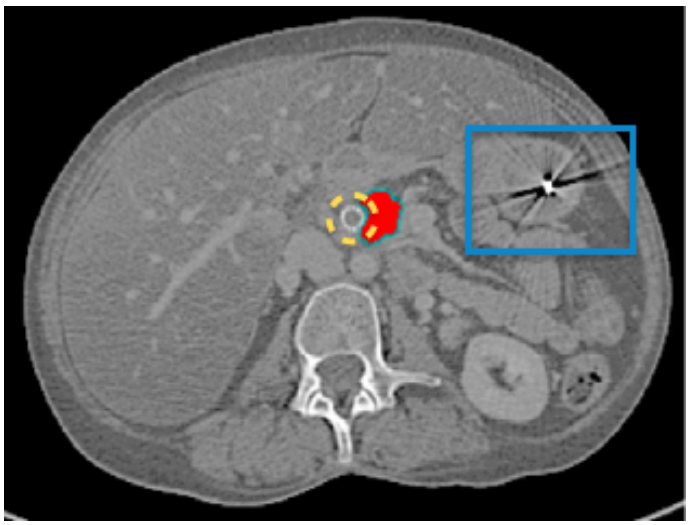


FIGURE A.2 – CT_{pRCT} (contrast enhanced) : **ROI** , **prothèse biliaire** et **artefacts**
source : CHU Bordeaux (Haut l'Évêque)

Les principaux résultats de l'apprentissage statistique sont réunis Table A.1.

• Les modifications de texture se sont avérées plus significatives pour le Δ_{1-2} et ces valeurs ont finalement été les seules conservées.

• Les modèles de prédiction de la survie sans évènement obtiennent les meilleurs résultats. Les forêts aléatoires surclassent les autres algorithmes.

• Les faux négatifs sont peu nombreux, ce qui donne un rappel élevé. L'aire sous la courbe ROC est faible.

• Les modèles de prédiction de la survie globale (binaire) obtiennent de mauvaises performances, sauf avec le Gradient Boosting.

• Les modèles de prédiction de la survie à court ou long terme échouent globalement à classifier les patients selon leur temps médian de rechute.

• Les modèles construits avec les variables cliniques + les variables radiomiques obtiennent de meilleurs résultats que ceux qui n'utilisent qu'un seul type de variable.

Limitations

Les premières limitations viennent du contourage manuel. Les surfaces sont difficiles à délinéer et nous n'avons pas fait vérifier les contours par un second opérateur. En raison du nombre élevé d'examen à délinéer manuellement (89 x trois dates de traitements x jusqu'à trois temps d'injection) et au caractère diffus des lésions, il a été choisi de ne pas procéder à un contourage 3D. Même si la coupe sélectionnée est la plus grande visible au diagnostic,

2. Destinées à améliorer l'écoulement de la bile lorsqu'elle est bloquée par la tumeur.

il est possible que des zones riches en motifs des tumeurs aient été exclues. La propagation de cette coupe sur les deux autres examens est d'ailleurs subjective et donc elle-aussi source d'erreurs.

Malgré une qualité parfois médiocre, on note que le pré-traitement des images est limité. En effet, la présence aléatoire de calcifications et/ou prothèses dans les examens fausse les histogrammes des images entières et diminue la qualité des ré-échelonnages. Toutefois, les CT-scans sont moins dépendant de la normalisation des niveaux de gris que l'IRM. Reste que des artefacts subsistent et faussent probablement le signal.

Du côté des modèles de classification, le principal frein est le déséquilibre de la distribution des classes dans le jeu de données, qui rend les métriques de performance plus difficile à analyser : un score de 84% de bonnes prédictions est difficile à faire valoir quand 82% des patients rechutent ! Ce type de déséquilibre est inhérent aux problématiques cliniques. Il peut être contourné en adaptant les critères d'évaluation des modèles.

Annexe B

Tables et graphiques supplémentaires

1 Impact des traitements de l'IRM sur les caractéristiques radiomiques

points de repère	1	2	10	*moyenne*
moyenne	30.4 (845)	**26.4 (739)**	**34.3 (1110)**	**30.37**
std	8.06 (44.3)	8.45 (45.7)	9.68 (40.2)	**8.73**
H1 entropie	**5.03 (40.8)**	5.26 (42.5)	5.17 (44)	**5.15**
interval	1.06 (7.48)	0.78 (9.78)	0.88 (11.52)	**0.91**
skewness	75.2 (1723)	**80.8 (2180)**	**63.5 (1386)**	**73.17**
kurtosis	**71.5 (1043)**	**86.7 (1108)**	72.6 (1095)	**76.93**
cluster shade	19.1 (203)	**18.66 (200)**	**21.44 (256)**	**19.73**
cluster prominence	11.0 (65.3)	10.9 (115)	**12.2 (124)**	11.37
entropie	8.68 (18.0)	8.55 (19.0)	8.68 (18.4)	**8.64**
homogénéité	**1.86 (7.25)**	1.79 (7.23)	1.81 (7.08)	**1.82**
énergie	**36.2 (146)**	37.6 (154)	**38.2 (155)**	**37.33**
inertie	**9.30 (37.2)**	10.1 (42.5)	**11.0 (45.0)**	10.13
moyenne	23.11	24.67	23.29	23.69

TABLE B.1 – Pourcentages d'erreur absolue moyens (et max) de calcul des paramètres d'intensité et de texture des sarcomes après normalisation.

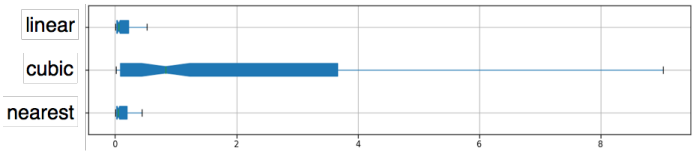


FIGURE B.1 – Diagrammes en boîte : temps de calcul des trois interpolateurs testés. L'interpolation tri-cubique nécessite plus de temps de calcul

2 Prédiction de la réponse au traitement des STS

Characteristics		Patients (n=65)
Gender	Male	38 (58.5)
	Female	27 (41.5)
Age at diagnosis (y), mean ± sd		57.9 ± 12.8
Histotype	Undifferentiated sarcoma[1]	33 (50.8)
	Muscular sarcoma[2]	13 (20)
	M/RC liposarcoma[3]	5 (7.7)
	Other liposarcoma[4]	6 (9.2)
	Synovial sarcoma	7 (10.8)
	MPNST	1 (1.5)
Location	Trunk wall	12 (18.5)
	Pelvic Girdle	2 (3.1)
	Shoulder Girdle	6 (9.2)
	Upper limb	7 (10.8)
	Lower limb	38 (58.5)
Depth	Superficial	4 (6.2)
	Deep	61 (93.8)
LD at baseline (mm), mean ± sd		119 ± 56
Nb cycles	4 cycles	22 (33.8)
	5 or 6 cycles	43 (66.2)

TABLE B.2 – Caractéristiques épidémiologiques au diagnostic.
NOTE.- LD : longest diameter, sd : standard deviation.
Data are numbers of patients with percentages in parentheses, except for age and LD.
1 : myxofibrosarcoma or undifferentiated sarcoma ;
2 : leiomyosarcoma and rhabdomyosarcoma ;
3 : myxoid/round cells liposarcoma ;
4 : pleomorphic or dedifferentiated liposarcoma.
MPNST : malignant peripheral nerve sheath tumor.

3 Prédiction de la rechute métastatique du sein

Variables	Good_HR	Poor_HR	p-value
Baseline clinico-radiological features			
Gender			
Male	11 (68.8)	27 (55.1)	0.393
Female	5 (21.2)	22 (44.9)	
Age at diagnosis (y)	58.8 ± 11.4	57.6 ± 13.3	0.873
Histotype			
Undifferentiated sarcoma[1]	10 (62.5)	23 (46.9)	0.257
Muscular sarcoma[2]	5 (21.2)	8 (16.3)	
M/RC liposarcoma[3]	1 (6.3)	4 (8.2)	
Other liposarcoma[4]	0 (0)	6 (12.2)	
Synovial sarcoma	0 (0)	7 (14.3)	
MPNST	0 (0)	1 (2.1)	
Location			
Trunk wall	2 (12.5)	10 (20.4)	0.146
Pelvic Girdle	2 (12.5)	0 (0)	
Shoulder Girdle	1 (6.3)	5 (10.2)	
Upper limb	2 (12.5)	5 (10.2)	
Lower limb	9 (56.2)	29 (59.2)	
Depth			
Superficial	1 (6.3)	3 (6.1)	1.000
Deep	15 (93.7)	46 (93.9)	
Nb cycles			
4 cycles	11 (68.8)	32 (65.3)	1.000
5 or 6 cycles	5 (21.2)	17 (34.7)	
LD on MRI_0 (mm)	146 (66)	110 (51)	**0.038 ***

Variables	Good_HR	Poor_HR	p-value
MRI_0 to MRI_1			
Change in LD (%)	-11.2 ± 20.8	2.9 ± 19.5	**0.027 ***
RECIST 1.1			
Complete Response	0 (0)	0 (0)	0.112
Partial Response	3 (18.8)	3 (6.1%)	
Stable Disease	13 (81.2)	39 (79.6)	
Progressive Disease	0 (0)	7 (14.3)	
Objective Response			
Yes	3 (18.8)	3 (6.1)	0.154
No	13 (81.2)	46 (93.9)	
Δ_Margin_definition[§]			
Well- or better limited	5 (21.2)	9 (20)	0.490
stable or worst	11 (68.8)	36 (80)	
Δ_Edema			
None or decrease	12 (75)	15 (30.6)	**0.003 ****
Stable or increase	4 (25)	34 (69.4)	
Δ_Peritumoral enhancement[§]			
None or decrease	12 (80)	24 (57.1)	0.134
Stable or increase	3 (20)	18 (42.9)	
Fibro-Necrotic Changes			
No	2 (21.2)	14 (25.6)	0.430
< 50% tumor volume	9 (56.2)	23 (46.9)	
≥ 50% tumor volume	5 (31.3)	12 (24.5)	

TABLE B.3 – Association between demographic and semantic radiological features and histological response.

NOTE.- LD : longest diameter, sd : standard deviation. Data are numbers of patients with percentages in parentheses, except for age, LD and change in LD. 1 : myxofibrosarcoma or undifferentiated sarcoma ; 2 : leiomyosarcoma and rhabdomyosarcoma ; 3 : myxoid/round cells liposarcoma ; 4 : pleomorphic or dedifferentiated liposarcoma. MPNST : malignant peripheral nerve sheath tumor. § : 8 patients had missing values for $\Delta\_Peritumoral_enhancement$ and 4 for $\Delta\_Margin\_definition$ due to defective MR protocol (incomplete acquisition of edema on post contrast T1-WI, different acquisition plan on MRI_0 and MRI_1)
* : $p \leq 0.05$; ** : $p \leq 0.005$.

Variables	Good-HR	Poor-HR	p-value
1st order intensity features			
Δ_H1_Entropy	-0.178 ± 0.546	0.296 ± 0.437	**0.004**
Δ_Std	0.003 ± 0.090	0.095 ± 0.093	**0.005**
Δ_Kurtosis	2.580 ± 10.294	-8.242 ± 24.809	0.061
Δ_Interval	-0.057 ± 0.160	0.042 ± 0.256	0.128
Δ_Mean	-0.113 ± 0.238	-0.086 ± 0.224	0.725
Δ_Skewness	0.103 ± 1.691	0.552 ± 2.249	0.743
2nd order texture features			
Δ_ClusterProminence_1	-86.274 ± 8541	6772.873 ± 14874	0.083
Δ_ClusterProminence_2	-146.031 ± 8191	6456.489 ± 14276	0.092
Δ_ClusterProminence_5	-473.075 ± 6972	5622.833 ± 12714	0.101
Δ_ClusterShade_1	140.667 ± 682	450.330 ± 880	0.361
Δ_ClusterShade_2	142.793 ± 647	416.500 ± 829	0.349
Δ_ClusterShade_5	133.954 ± 551	339.885 ± 708	0.454
Δ_Energy_1	0.046 ± 0.122	0.064 ± 0.139	0.497
Δ_Energy_2	0.046 ± 0.121	0.062 ± 0.138	0.497
Δ_Energy_5	0.044 ± 0.116	0.057 ± 0.132	0.527
Δ_Entropy_1	-0.309 ± 0.889	-0.498 ± 1.159	0.497
Δ_Entropy_2	-0.345 ± 0.955	-0.523 ± 1.229	0.558
Δ_Entropy_5	-0.365 ± 0.991	-0.522 ± 1.279	0.673
Δ_Homogeneity_1	0.016 ± 0.065	0.045 ± 0.096	0.134
Δ_Homogeneity_2	0.024 ± 0.079	0.054 ± 0.112	0.190
Δ_Homogeneity_5	0.029 ± 0.092	0.059 ± 0.127	0.223
Δ_Inertia_1	-0.048 ± 2.253	-0.312 ± 2.978	0.361
Δ_Inertia_2	-0.315 ± 4.054	-0.295 ± 5.462	0.512
Δ_Inertia_5	-0.885 ± 7.219	0.285 ± 11.004	1.000
Shape features			
Δ_Elongation	-0.106 ± 0.207	0.059 ± 0.195	**0.018**
Δ_Equivalent_spherical_radius	-3.248 ± 7.663	1.464 ± 6.273	**0.041**
Δ_Feret_diameter	-5.998 ± 24.753	8.866 ± 20.380	**0.047**
Δ_Flatness	0.189 ± 0.288	0.040 ± 0.240	**0.044**
Δ_Surface_Area	-2919 ± 13395	3147 ± 9852	0.167
Δ_Volume	-48510 ± 199403	89400 ± 252589	0.075
Δ_Roundness	-0.003 ± 0.080	-0.018 ± 0.078	0.575

TABLE B.4 – Association between delta-radiomics features and response in training cohort. Data are givean as mean and standard deviation.

LR	pénalité = 4
SVC	kernel = $\exp(-\gamma\|x - x'\|^2)$ avec $\gamma = 1/nb\_variables$ (rbf) pénalité = 80
kNN	nombre de voisins = 9
RF	arbres = 200 critère de division = entropie nombre minimum d'observations pour diviser une feuille = : 4 nombre minimum d'observations dans une feuille = 2 nombre de variables considérées à la création d'un nœud = $\sqrt{nb\_variables}$

TABLE B.5 – Hyper-paramètres des classifieurs pour la prédiction de la réponse au traitement des STS. Obtenus avec un *grid-search*.

	AUROC	ACC	PPV	NPV	F1	AP	Train
RF	0.55	0.71	0.78	0.24	0.82	0.81	0.99
LR	0.58	0.71	0.79	0.28	0.82	0.87	0.90
(rappel) RECIST	0.66	0.76	0.67	0.21	0.62	0.89	-

TABLE B.6 – Prédiction de la réponse au traitement des STS : scores moyens en validation croisée sur 50 patients avec les caractéristiques sélectionnées par ElasticNet ($\Delta\_Histogram\_Entropy$, $\Delta\_Elongation$, $\Delta\_Equivalent\_radius$, Δ_Edema, $T1\_Kurtosis$, $T1\_Histogram\_Entropy$)

	AUROC	ACC	PPV	NPV	F1	AP	Train
RF	0.54	0.61	0.67	0.36	0.73	0.66	1.00
LR	0.54	0.67	0.73	0.50	0.76	0.77	0.88
(rappel) RECIST	0.72	0.73	0.75	0.67	0.82	0.80	-

TABLE B.7 – Prédiction de la réponse au traitement des STS : scores moyens du test final sur 15 patients avec les caractéristiques sélectionnées par ElasticNet ($\Delta\_Histogram\_Entropy$, $\Delta\_Elongation$, $\Delta\_Equivalent\_radius$, Δ_Edema, $T1\_Kurtosis$, $T1\_Histogram\_Entropy$)

Modèle	**Nb variables**	**ACC**	**AUROC**	**PR**	**RE**	**F1**	**AP**
NN	4	0.88	0.89	0.70	0.37	0.48	0.63
NN	8	0.89	0.92	0.72	0.54	0.59	0.68
(rappel) ***NN***	*7*	*0.91*	*0.93*	*0.75*	0.60	*0.65*	*0.73*

TABLE B.8 – Scores moyens de la détection de l'œdème pour les réseaux de neurones avec différents ensembles de variables.

Patient	% d'oedème	volume oedème (mm³)	volume calculé							
			RF (ElasticNet)		RF (7 variables)		NN (ElasticNet)		NN (7 variables)	
			masque	*probabilité*	*masque*	*probabilité*	*masque*	*probabilité*	*masque*	*probabilité*
A	15.96 %	58194	**-47.5**	-16.1	**42.4**	**35.0**	**-7.1**	**7.7**	19.9	19.4
B	15.3 %	106869	**-49.6**	-16.7	**-33.2**	-12.2	**-66.5**	**-48.7**	**-36.2**	-15.6
C	11.48 %	15409	-26.0	**-2.7**	-19.3	**-5.9**	**-41.0**	**-38.3**	-26.1	-21.1

TABLE B.9 – Écart à la valeur de référence du volume d'œdème calculé sur la prédiction ou sur la carte de probabilités des modèles ML. En vert, écarts $< 10\%$, en rouge $> 30\%$.

LR	pénalité = 0.001
SVC	kernel = $\exp(-\gamma\|x - x'\|^2)$ avec $\gamma = 1/nb\_variables$ (rbf) pénalité = 80
kNN	nombre de voisins = 7
RF	arbres = 50 critère de division = gini nombre minimum d'observations pour diviser une feuille = : 4 nombre minimum d'observations dans une feuille = 2 nombre de variables considérées à la création d'un nœud = $\sqrt{nb\_variables}$
GB	nombre d'estimateurs = 70 taux d'apprentissage = 0.20 profondeur maximale = 6 nombre de variables considérées à la création d'un nœud = $0.7 \times nb\_variables$ dropout = 0

TABLE B.10 – Hyper-paramètres des classifieurs pour la prédiction de la rechute à court terme des cancers du sein. Obtenus avec un *grid-search*.

	complétion	AUROC	ACC	PPV	NPV	F1	AP	Train
LR	médiane	0.75	0.68	0.18	0.96	0.28	0.21	0.69
rappel LR	MissForest	0.75	0.66	0.18	0.96	0.28	0.22	0.68
RF	médiane	0.75	0.83	0.24	0.94	0.31	0.24	0.87
rappel RF	MissForest	0.75	0.83	0.24	0.94	0.31	0.24	0.87

TABLE B.11 – Scores moyens de la validation croisée sur $C_{9.25}$ en fonction de la méthode de complétion des données. MissForest est appliquée une seule fois avant l'apprentissage. La médiane est calculée à chaque cycle de validation croisée avec les valeurs de l'ensemble d'entrainement utilisé.

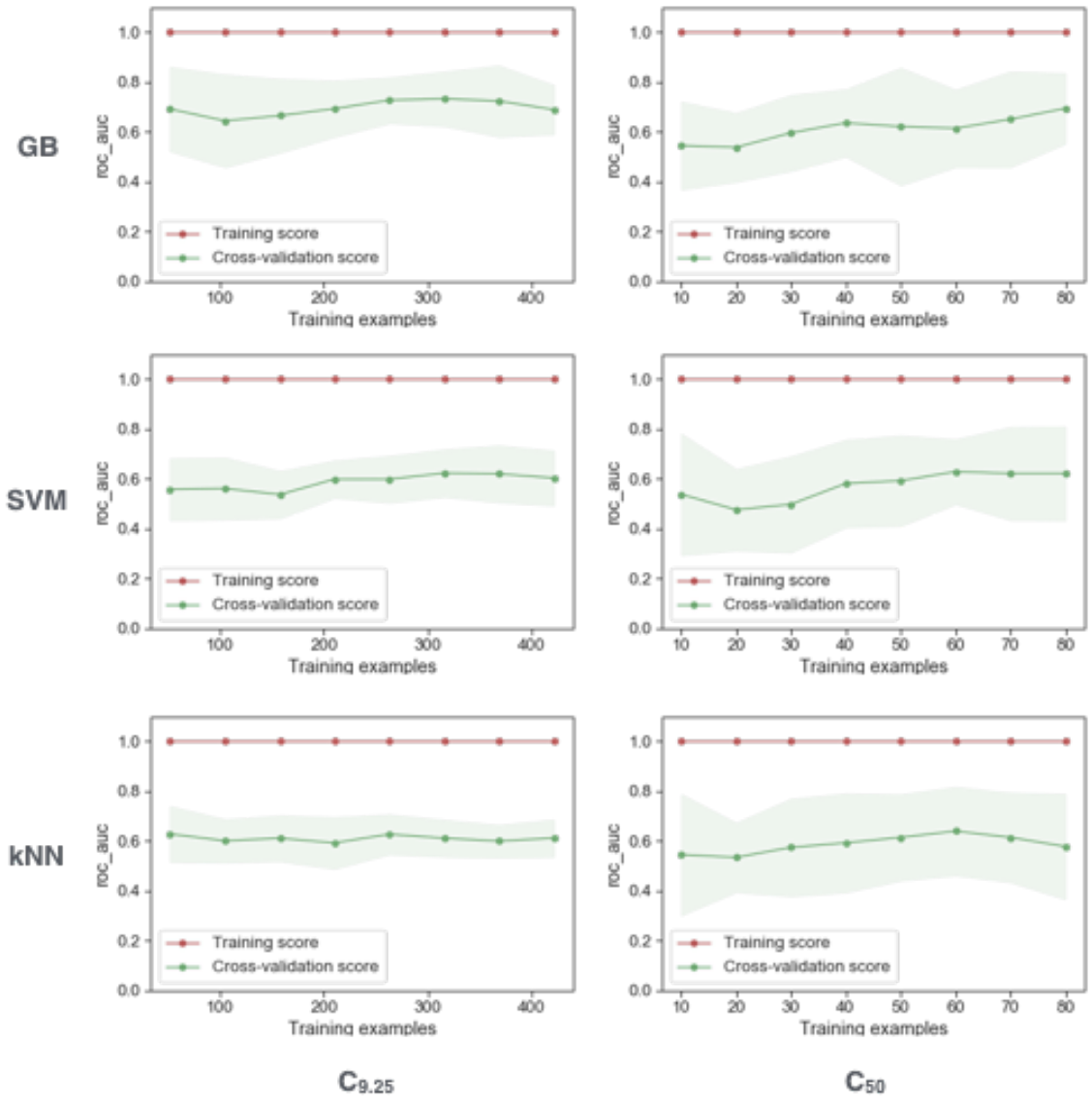


TABLE B.12 – Courbes d'apprentissage des algorithmes GB, kNN et SVM sur les deux cohortes de la rechute métastatique du cancer du sein.

Annexe C

Publications et communications

Publications

Chiara Nicolò, Cynthia Périer, Mélanie Prague, Gaetan MacGrogan, Olivier Saut et Sébastien Benzekry. « Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer ». In : *bioRxiv* (2019) 10.1101/634428

Amandine Crombé*, Cynthia Périer*, Michèle Kind, Baudouin Denis de Senneville, François Le Loarer, Antoine Italiano, Xavier Buy et Olivier Saut. « T2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. » In : *Journal of Magnetic Resonance Imaging* 50 (déc. 2018) 10.1002/jmri.26589 **Ces deux auteurs ont contribué également à ce travail.*

Arnaud Hocquelet, Thibaut Auriac, Cynthia Périer, Clarisse Dromain, Marie Meyer, Jean-Baptiste Pinaquy, Alban Denys, Trillaud Hervé, Baudouin Denis de Senneville et Véronique Vendrely. « Pre-treatment magnetic resonance-based texture features as potential imaging biomarkers for predicting event free survival in anal cancer treated by chemoradiotherapy ». In : *European Radiology* 28 (fév. 2018) 10.1007/s00330-017-5284-z

Communications orales avec acte

PouypoudatPerier PouypoudatPerier

Conférences

- "T2-based MRI-radiomics to improve prediction of histologie response in soft-tissue sarcomas treated by neoadjuvant chemotherapy - preliminary results." *Journées françaises de Radiologie*, Paris, Nov. 2018
- "Textural analysis of pancreatic cancer during radiotherapy and machine learning", *Workshop CGO, Imaging of diagnostic and therapeutic biomarkers in Oncology*, Le Bono, Sept. 2017

Bibliographie

[Aap] *AAPM RT-MAC Challenge 2019*. `http://aapmchallenges.cloudapp.net/competitions/34`. Accessed : 2019-06-03. 2019.

[AC19] Scott A Czepiel. « Maximum Likelihood Estimation of Logistic Regression Models : Theory and Implementation ». In : (sept. 2019).

[AG18] angulakshmi.m Angu et Lakshmipriya GG. « Automated Brain Tumour Segmentation Techniques-A Review ». In : *Imaging systems and technology* (jan. 2018).

[Ahm+13] Arfan Ahmed et al. « Texture analysis in assessment and prediction of chemotherapy response in breast cancer ». In : *Journal of magnetic resonance imaging : JMRI* 38 (juil. 2013).

[AKW08] Omar Al-Kadi et Des Watson. « Texture Analysis of Aggressive and Nonaggressive Lung Tumor CE CT Images ». In : *Biomedical Engineering, IEEE Transactions on* 55 (août 2008), p. 1822 -1830.

[ALM15] Philipp Altrock, Lin Liu et Franziska Michor. « The mathematics of cancer : Integrating quantitative models ». In : *Nature Reviews Cancer* 15 (nov. 2015), p. 730-745.

[Ang+12] Davide Anguita et al. « The 'K' in K-fold Cross Validation ». In : *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. Jan. 2012.

[AP+85] D A Potter et al. « Patterns of recurrence in patients with high grade soft tissue sarcoma ». In : *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 3 (avr. 1985), p. 353-66.

[APKET83] J Anthony Parker, Robert Kenyon et Donald E. Troxel. « Comparison of Interpolating Methods for Image Resampling ». In : *Medical Imaging, IEEE Transactions on* 2 (avr. 1983), p. 31 -39.

[BA+01] James B. Arnold et al. « Qualitative and Quantitative Evaluation of Six Algorithms for Correcting Intensity Nonuniformity Effects ». In : *NeuroImage* 13 (juin 2001), p. 931-43.

[Bak+17] Spyridon Bakas et al. « Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features ». In : *Scientific Data* 4 (sept. 2017).

[Bar+15] Dominique Barbolosi et al. « Computational oncology — mathematical modelling of drug regimens for precision medicine ». In : *Nature Reviews Clinical Oncology* 13 (nov. 2015).

[BE+17] Hassan Bagher-Ebadian et al. « On the Impact of Smoothing and Noise on Robustness of CT and CBCT Radiomics Features for Patients with Head and Neck Cancers ». In : *Medical Physics* 44 (mar. 2017).

[Ben+09] Matthias R. Benz et al. « FDG-PET/CT Imaging Predicts Histopathologic Treatment Responses after the Initial Cycle of Neoadjuvant Chemotherapy in High-Grade Soft-Tissue Sarcomas ». In : *Clinical Cancer Research* 15.8 (2009), p. 2856-2863. eprint : `http://clincancerres.aacrjournals.org/content/15/8/2856.full.pdf`.

[BJ08] Jan-Philip Bergeest et Florian Jäger. « A Comparison of Five Methods for Signal Intensity Standardization in MRI ». In : *Bildverarbeitung für die Medizin 2008*. Sous la dir. de Thomas Tolxdorff et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 36-40.

[Bn+18] Mostefa Ben naceur et al. « Fully Automatic Brain Tumor Segmentation using End-To-End Incremental Deep Neural Networks in MRI images ». In : *Computer Methods and Programs in Biomedicine* 166 (sept. 2018).

[Bog+16] Marta Bogowicz et al. « Stability of radiomic features in CT perfusion maps ». In : *Physics in Medicine and Biology* 61 (déc. 2016), p. 8736-8749.

[Bol+18] Marco Bologna et al. « Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images ». In : *Journal of Digital Imaging* 31.6 (déc. 2018), p. 879-894.

[Bra] *Multimodal Brain Tumor Segmentation Challenge 2018*. `https://www.med.upenn.edu/sbia/brats2018.html`. Accessed : 2019-06-03. 2018.

[Bra+19] Nathaniel Braman et al. « Association of Peritumoral Radiomics With Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2 (ERBB2) –Positive Breast Cancer ». In : *JAMA Network Open* 2 (avr. 2019), e192561.

[Bre01] Leo Breiman. « Random Forests ». In : *Mach. Learn.* 45.1 (oct. 2001), p. 5-32.

[Bre05] M.F. Brennan. « Soft tissue sarcoma : Advances in understanding and management ». In : *The surgeon : journal of the Royal Colleges of Surgeons of Edinburgh and Ireland* 3 (juil. 2005), p. 216-23.

[BW61] William Bennett et George Weiss. « Electrical Noise ». In : *Physics Today* 14 (jan. 1961), p. 54.

[Cel08] Alain Celisse. « Optimal cross-validation in density estimation with the L^2-loss ». In : *The Annals of Statistics* 42 (déc. 2008).

[CH67] T.M. Cover et P.E. Hart. « Nearest Neighbor Pattern Classification ». In : *IEEE Transactions on Information Theory* 13 (jan. 1967), p. 21-27.

[Cha+02] Nitesh Chawla et al. « SMOTE : Synthetic Minority Over-sampling Technique ». In : *J. Artif. Intell. Res. (JAIR)* 16 (jan. 2002), p. 321-357.

[Cha+11] Marie Chavent et al. « Clustering of variables via the PCAMIX method ». In : *International Classication Conference*. Saint Andrews, United Kingdom, juil. 2011.

[Che+18] Dmitry Cherezov et al. « Delta radiomic features improve prediction for lung cancer incidence : A nested case–control analysis of the National Lung Screening Trial ». In : *Cancer Medicine* 7 (déc. 2018).

[CJG15] Daniel Commenges et Hélène Jacqmin-Gadda. *Dynamical Biostatistical Models*. CRC Press, oct. 2015.

[com16] pyradiomics community. *Frequently Asked Questions*. 2016. url : `https://pyradiomics.readthedocs.io/en/latest/faq.html`.

[Cor+15] Thibaud Coroller et al. « CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma ». In : *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* 114 (mar. 2015).

[Cor+17a] Valentina Corino et al. « Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions ». In : *Journal of magnetic resonance imaging : JMRI* 47 (juin 2017).

[Cor+17b] Francois Cornelis et al. « Precision of manual two-dimensional segmentations of lung and liver metastases and its impact on tumour response assessment using RECIST 1.1 ». In : *European Radiology Experimental* 1 (déc. 2017).

[Cou+17] Sophie Cousin et al. « Clinical, radiological and genetic features, associated with the histopathologic response to neoadjuvant chemotherapy (NAC) and outcomes in locally advanced soft tissue sarcoma (STS) patients (pts). » In : *Journal of Clinical Oncology* 35 (mai 2017), p. 11014-11014.

[Cri+18] Paola Crivelli et al. « A New Challenge for Radiologists : Radiomics in Breast Cancer ». In : *BioMed Research International* 2018 (oct. 2018), p. 1-10.

[Cro+18a] Amandine Crombé et al. « High-grade soft-tissue sarcoma : optimizing injection improves MRI evaluation of tumor response ». In : *European Radiology* 29 (juil. 2018).

[Cro+18b] Amandine Crombé* et al. « T2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. » In : *Journal of Magnetic Resonance Imaging* 50 (déc. 2018).

[Cro17] Amandine Crombé. « Optimal combination of conventional MRI and DCE-MRI parameters to early predict pathologic response to anthracycline-based neoadjuvant chemotherapy for locally advanced high-grade soft-tissue sarcomas ». In : 2017.

[CSM04] Guylaine Collewet, Michal Strzelecki et François Mariette. « Influence of MRI acquisition protocols and image intensity normalization methods on texture classification ». In : *Magnetic resonance imaging* 22 (fév. 2004), p. 81-91.

[CTH84] Richard W. Conners, Mohan M. Trivedi et Charles A. Harlow. « Segmentation of a high-resolution urban scene using texture operators ». In : *Computer Vision, Graphics, and Image Processing* 25 (1984), p. 273-310.

[Cui+11] Chun-Yan Cui et al. « Quantitative analysis and prediction of regional lymph node status in rectal cancer based on computed tomography imaging ». In : *European radiology* 21 (juin 2011), p. 2318-25.

[DAKM17] Adrien Depeursinge, Omar Al-Kadi et Joseph Mitchell. *Biomedical Texture Analysis : Fundamentals, Tools and Challenges*. London : Academic Press, août 2017.

[Dep+19] Adrien Depeursinge et al. « Multiscale lung texture signature learning using the Riesz transform ». In : *Med. Image Comput. Comput.-Assist. Interv.-MICCAI* (août 2019), p. 517-524.

[DG06] Jesse Davis et Mark Goadrich. « The Relationship Between Precision-Recall and ROC Curves ». In : *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York, NY, USA : ACM, 2006, p. 233-240.

[Dia+11] Idanis Diaz et al. « A critical review of the effects of de-noising algorithms on MRI brain tumor segmentation ». In : *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2011 (août 2011), p. 3934-7.

[Dif+17] Sarah Differding et al. « Radiation dose escalation based on FDG-PET driven dose painting by numbers in oropharyngeal squamous cell carcinoma : a dosimetric comparison between TomoTherapy-HA and RapidArc ». In : *Radiation Oncology* 12 (mar. 2017), p. 59.

[DJ+13] Roger D James et al. « Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II) : A randomised, phase 3, open-label, 2x2 factorial trial ». In : *The lancet oncology* 14 (avr. 2013).

[Don+17] Yuhao Dong et al. « Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI ». In : *European Radiology* 28 (août 2017).

[Duf+17] Michael Duffy et al. « Clinical use of biomarkers in breast cancer : Updated guidelines from the European Group on Tumor Markers (EGTM) ». In : *European Journal of Cancer* 75 (avr. 2017), p. 284-298.

[Dur+19] L Duron et al. « Gray-level discretization impacts reproducible MRI radiomics texture features ». In : *PLOS ONE* 14 (mar. 2019), e0213459.

[DWK05] Dursun Delen, Glenn Walker et Amit Kadam. « Predicting breast cancer survivability : A comparison of three data mining methods ». In : *Artificial intelligence in medicine* 34 (juil. 2005), p. 113-27.

[Fav+16] Xenia Fave et al. « Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer ». In : *Translational Cancer Research* 5 (août 2016), p. 349-363.

[Fav+17] Xenia Fave et al. « Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer ». In : *Scientific Reports* 7 (avr. 2017), p. 588.

[FD+14] Manuel Fernandez-Delgado et al. « Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ? » In : *Journal of Machine Learning Research* 15 (oct. 2014), p. 3133-3181.

[FIC19] Ihab F. Ilyas et Xu Chu. *Data Cleaning*. ACM Books, juil. 2019.

[Fri01] Jerome Friedman. « Greedy Function Approximation : A Gradient Boosting Machine ». In : *Annals of Statistics* 29 (oct. 2001), p. 1189-1232.

[FS10] George Forman et Martin Scholz. « Apples-to-apples in Cross-validation Studies : Pitfalls in Classifier Performance Measurement ». In : *SIGKDD Explor. Newsl.* 12.1 (nov. 2010), p. 49-57.

[Gan+10] Balaji Ganeshan et al. « Texture analysis of non-small cell lung cancer on unenhanced computed tomography : Initial evidence for a relationship with tumour glucose metabolism and stage ». In : *Cancer imaging : the official publication of the International Cancer Imaging Society* 10 (juil. 2010), p. 137-43.

[Gan+11] Balaji Ganeshan et al. « Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis : Preliminary evidence of an association with tumour metabolism, stage, and survival ». In : *Clinical radiology* 67 (sept. 2011), p. 157-64.

[GBG01] Valeriy Gavrishchaka et Supriya B. Ganguli. « Support vector machine as an efficient tool for high-dimensional data processing : Application to substorm forecasting ». In : *Journal of Geophysical Research* 106 (déc. 2001), p. 29911-29914.

[GBZB02] Miguel Ángel González Ballester, Andrew P. Zisserman et Michael Brady. « Estimation of the partial volume effect in MRI ». In : *Medical Image Analysis* 6.4 (2002), p. 389 -405.

[Gei+19] Robert Geirhos et al. « ImageNet-trained CNNs are biased towards texture ; increasing shape bias improves accuracy and robustness. » In : *ICLR*. OpenReview.net, 2019.

[Gil+10] Robert Gillies et al. « The Biology Underlying Molecular Imaging in Oncology : From Genome to Anatome and Back Again ». In : *Clinical radiology* 65 (juil. 2010), p. 517-21.

[Gin21] Corrado Gini. « Measurement of Inequality of Income ». In : *Economic Journal* 31 (jan. 1921).

[GKH15] Robert Gillies, Paul Kinahan et Hedvig Hricak. « Radiomics : Images Are More than Pictures, They Are Data ». In : *Radiology* 278 (nov. 2015), p. 151169.

[GMSP13] Baptiste Gregorutti, Bertrand Michel et Philippe Saint-Pierre. « Correlation and variable importance in random forests ». In : *Statistics and Computing* 27 (oct. 2013).

[Gne+16] Khemara Gnep et al. « Haralick textural features on T2-weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer ». In : *Journal of magnetic resonance imaging : JMRI* 45 (juin 2016).

[GO+18] Jessica Goya-Outi et al. « Computation of reliable textural indices from multimodal brain MRI : Suggestions based on a study of patients with diffuse intrinsic pontine glioma ». In : *Physics in Medicine and Biology* 63 (avr. 2018).

[Goh+10] Vicky Goh et al. « Magnetic Resonance Imaging Assessment of Squamous Cell Carcinoma of the Anal Canal Before and After Chemoradiation : Can MRI Predict for Eventual Clinical Outcome ? » In : *International journal of radiation oncology, biology, physics* 78 (fév. 2010), p. 715-21.

[Goh+11] Vicky Goh et al. « Assessment of Response to Tyrosine Kinase Inhibitors in Metastatic Renal Cell Cancer : CT Texture as a Predictive Biomarker ». In : *Radiology* 261 (août 2011), p. 165-71.

[Gra98] GG Grabenbauer. « Concomitant radiotherapy and chemotherapy is superior to radiotherapy alone in the treatment of locally advanced anal cancer. Results of a phase III randomized trial ». In : *Strahlentherapie und Onkologie* 174 (fév. 1998), p. 108-109.

[Gri+17] Joost van Griethuysen et al. « Computational Radiomics System to Decode the Radiographic Phenotype ». In : *Cancer Research* 77 (nov. 2017), e104-e107.

[Gro+17] Alessandro Gronchi et al. « Histotype-tailored neoadjuvant chemotherapy versus standard chemotherapy in patients with high-risk soft-tissue sarcomas (ISG-STS 1001) : An international, open-label, randomised, controlled, phase 3, multicentre trial ». In : *The Lancet Oncology* 18 (mai 2017).

[Gru15] Joel Grus. *Data Science from Scratch : First Principles with Python*. 1st. O'Reilly Media, Inc., 2015.

[GRWW17] David Grimes, Daniel R Warren et Samantha Warren. « Hypoxia imaging and radiotherapy : Bridging the resolution gap ». In : *The British Journal of Radiology* 90 (mai 2017), p. 20160939.

[Gua+16] Yue Guan et al. « Whole-Lesion Apparent Diffusion Coefficient-Based Entropy-Related Parameters for Characterizing Cervical Cancers ». In : *Academic Radiology* 23 (sept. 2016).

[Gun+12] Leonard Gunderson et al. « Long-Term Update of US GI Intergroup RTOG 98-11 Phase III Trial for Anal Carcinoma : Survival, Relapse, and Colostomy Failure With Concurrent Chemoradiation Involving Fluorouracil/Mitomycin Versus Fluorouracil/Cisplatin ». In : *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30 (nov. 2012).

[GW50] Brier GW. « Verification of forecasts expressed of probability ». In : *Monthly Weather Review* 78 (jan. 1950), p. 1-3.

[HA04] Victoria Hodge et Jim Austin. « A Survey of Outlier Detection Methodologies ». In : *Artificial Intelligence Review* 22 (oct. 2004), p. 85-126.

[Hai+19] Jinjin Hai et al. « Fully Convolutional DenseNet with Multiscale Context for Automated Breast Tumor Segmentation ». In : *Journal of Healthcare Engineering* 2019 (jan. 2019), p. 1-11.

[Has09] Trevor Hastie. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, jan. 2009.

[Hat+14] Mathieu Hatt et al. « 18F-FDG PET Uptake Characterization Through Texture Analysis : Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort ». In : *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 56 (déc. 2014).

[Hay+15] Koichi Hayano et al. « Texture Analysis of Non–Contrast-Enhanced Computed Tomography for Assessing Angiogenesis and Survival of Soft Tissue Sarcoma ». In : *Journal of computer assisted tomography* Publish Ahead of Print (mar. 2015).

[Hen+17] Shelley Henderson et al. « Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer ». In : *European radiology* 27 (mai 2017).

[HJ+19] Seung Hyuck Jeon et al. « Delta-radiomics signature predicts treatment outcomes after preoperative chemoradiotherapy and surgery in rectal cancer ». In : *Radiation Oncology* 14 (déc. 2019).

[HL13] Feng Hu et Hang Li. « A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model : NRSBoundary-SMOTE ». In : *Mathematical Problems in Engineering* 2013 (nov. 2013).

[Hoc+16a] Arnaud Hocquelet et al. « Magnetic resonance texture parameters are associated with ablation efficiency in MR-guided high-intensity focussed ultrasound treatment of uterine fibroids ». In : *International Journal of Hyperthermia* (oct. 2016), p. 1-8.

[Hoc+16b] Arnaud Hocquelet et al. « Three-Dimensional Measurement of Hepatocellular Carcinoma Ablation Zones and Margins for Predicting Local Tumor Progression ». In : *Journal of Vascular and Interventional Radiology* 27 (mai 2016).

[Hoc+18] Arnaud Hocquelet et al. « Pre-treatment magnetic resonance-based texture features as potential imaging biomarkers for predicting event free survival in anal cancer treated by chemoradiotherapy ». In : *European Radiology* 28 (fév. 2018).

[Hon+15] Charles Honoré et al. « Soft tissue sarcoma in France in 2014 : Epidemiology, classification and organization of clinical care ». In : *Journal of visceral surgery* 152 (juin 2015).

[HOU+17] Clémence HOUARD et al. « Role of 18 F-fluorodeoxyglucose Positron Emission Tomography-Computed Tomography in post-treatment evaluation of anal carcinoma ». In : *Journal of Nuclear Medicine* 58 (mar. 2017), jnumed.116.185280.

[HP96] Gudbjartsson H et Samuel PATZ. « The Rician distribution of noisy MRI data ». In : *Magnetic Resonance in Medicine* 36 (jan. 1996), p. 331-333.

[HSD73] R. M. HARALICK, K. SHANMUGAM et I. DINSTEIN. « Textural Features for Image Classification ». In : *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (nov. 1973), p. 610-621.

[Höc+96] Michael HÖCKEL et al. « Hypoxia and Radiation Response in Human Tumors ». In : *Seminars in radiation oncology* 6 (fév. 1996), p. 3-9.

[IK16] H. ISHWARAN et Udaya KOGALUR. « randomForestSRC : Random Forests for Survival, Regression and Classification (RF-SRC) ». In : (jan. 2016).

[Iss+10] Rolf ISSELS et al. « Neo-adjuvant chemotherapy alone or with regional hyperthermia for localised high-risk soft-tissue sarcoma : A randomised phase 3 multicentre study ». In : *The lancet oncology* 11 (juin 2010), p. 561-70.

[JH05] Eric J HALL. « Dose-painting by numbers : A feasible approach ? » In : *The lancet oncology* 6 (mar. 2005), p. 66.

[Joh+13] Hans J. JOHNSON et al. *The ITK Software Guide*. Third. *In press*. Kitware, Inc. 2013.

[Jun+05] Jaber JUNTU et al. « Bias Field Correction for MRI Images ». In : *Computer Recognition Systems*. Sous la dir. de Marek KURZYŃSKI et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, p. 543-551.

[Jun+11] Jaber JUNTU et al. « Classification of Soft Tissue Tumors by Machine Learning Algorithms ». In : *Soft Tissue Tumors*. Sous la dir. de Fethi DERBEL. Rijeka : IntechOpen, 2011. Chap. 3.

[Kim+12] Woojae KIM et al. « Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine ». In : *Journal of breast cancer* 15 (juin 2012), p. 230-8.

[KLP11] John KALBFLEISCH et Ross L. PRENTICE. « The Statistical Analysis of Failure Time Data, Second Edition ». In : (jan. 2011), p. 247 -277.

[KM58] EL KAPLAN et P MEIER. « Nonparametrics estimates for incomplete observations ». In : *Journal of the American Statistical Association* 53 (jan. 1958), p. 457-480.

[Koc+16] Rohit KOCHHAR et al. « The assessment of local response using magnetic resonance imaging at 3- and 6-month post chemoradiotherapy in patients with anal cancer ». In : *European Radiology* 27 (avr. 2016).

[KTV85] Serge KOSCIELNY, M TUBIANA et Alain-Jacques VALLERON. « A simulation model of the natural history of human breast cancer ». In : *British journal of cancer* 52 (nov. 1985), p. 515-24.

[Lam+12] Philippe Lambin et al. « Radiomics : Extracting more information from medical images using advanced feature analysis ». In : *European journal of cancer (Oxford, England : 1990)* 48 (mar. 2012), p. 441-6.

[Lei+15] Ralph T. H. Leijenaar et al. « The effect of SUV discretization in quantitative FDG-PET Radiomics : the need for standardized methodology in tumor texture analysis ». In : *Scientific reports*. 2015.

[Lel+14] Benoît Lelandais et al. « Fusion of multi-tracer PET images for Dose Painting ». In : *Medical Image Analysis* 18 (oct. 2014).

[Li+03] Xueqing Li et al. « Using low-discrepancy sequences and the Crofton formula to compute surface areas of geometric models ». In : *Computer-Aided Design* 35 (août 2003), p. 771-782.

[Li+17] Qihua Li et al. « A Fully-Automatic Multiparametric Radiomics Model : Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme ». In : *Scientific Reports* 7 (déc. 2017).

[Lin03] Joakim Lindblad. « Surface Area Estimation of Digitized Planes Using Weighted Local Configurations ». In : *Discrete Geometry for Computer Imagery*. Sous la dir. d'Ingela Nyström, Gabriella Sanniti di Baja et Stina Svensson. Berlin, Heidelberg : Springer Berlin Heidelberg, 2003, p. 348-357.

[LJ10] Jun-Tao LI et Ying-Min JIA. « An Improved Elastic Net for Cancer Classification and Gene Selection ». In : *Acta Automatica Sinica* 36 (juil. 2010), p. 976-981.

[LK00] Lance Liotta et E.C. Kohn. « Invasion and metastases ». In : *Cancer Medicine* (jan. 2000), p. 121-131.

[LL12] Gaetan Lehmann et David Legland. « Efficient N-Dimensional surface estimation using Crofton formula and run-length encoding, Kitware INC(2012) ». In : *Insight Journal* (2012).

[Lon12] Dan Longo. « Longo DLTumor heterogeneity and personalized medicine. N Engl J Med 366 :956-957 ». In : *The New England journal of medicine* 366 (mar. 2012), p. 956-7.

[LPXP00] Dzung L. Pham, Chenyang Xu et Jerry Prince. « A Survey of Current Methods in Medical Image Segmentation ». In : *Annual review of biomedical engineering* 2 (fév. 2000), p. 315-37.

[LSDMJ19] Rebecca L. Siegel, Kimberly D. Miller et Ahmedin Jemal. « Cancer statistics, 2019 ». In : *CA : A Cancer Journal for Clinicians* 69 (jan. 2019).

[Mas+15] Isabelle de Mascarel et al. « Comprehensive prognostic analysis in breast cancer integrating clinical, tumoral, micro-environmental and immunohistochemical criteria ». In : *SpringerPlus* 4 (sept. 2015), p. 528.

[MB05] Ch M. Bishop. *Neural Networks For Pattern Recognition*. T. 227. Jan. 2005.

[MC+96] J M Coindre et al. « Prognostic factors in adult patients with locally controlled soft tissue sarcoma. A study of 546 patients from the French Federation of Cancer Centers Sarcoma Group ». In : *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 14 (mar. 1996), p. 869-77.

[Men+14] Bjoern Menze et al. « The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) ». In : *IEEE Transactions on Medical Imaging* 99 (déc. 2014).

[MN+12] M Mizuki Nishino et al. « Personalized Tumor Response Assessment in the Era of Molecular Medicine : Cancer-Specific and Therapy-Specific Response Criteria to Complement Pitfalls of RECIST ». In : *AJR. American journal of roentgenology* 198 (avr. 2012), p. 737-45.

[Moo+09] Stella Mook et al. « Mook S, Schmidt MK, Rutgers EJ, van de Velde AO, Visser O, Rutgers SM, Armstrong N, van't Veer LJ, Ravdin PMCalibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant ! program : a hospital-based retrospective cohort study. Lancet Oncol 10 : 1070-1076 ». In : *The lancet oncology* 10 (oct. 2009), p. 1070-6.

[Mor+18] Takayasu Moriya et al. « Unsupervised Segmentation of 3D Medical Images Based on Clustering and Deep Representation Learning ». In : (avr. 2018).

[MPH14] Rebecca Muirhead, Mike Partridge et Maria Hawkins. « A Tumor Control Probability Model for Anal Squamous Cell Carcinoma ». In : *International Journal of Radiation Oncology*Biology*Physics* 90 (sept. 2014), S399-S400.

[MV93] James C Mullikin et Piet W Verbeek. « Surface area estimation of digitized planes ». In : *Bioimaging* 1.1 (1993), p. 6-16. eprint : `https://onlinelibrary.wiley.com/doi/pdf/10.1002/1361-6374%28199303%291%3A1%3C6%3A%3AAID-BIO3%3E3.0.CO%3B2-3`.

[New+12] Paul Newton et al. « A Stochastic Markov Chain Model to Describe Lung Cancer Growth and Metastasis ». In : *PloS one* 7 (avr. 2012), e34637.

[Ng+12] Francesca Ng et al. « Assessment of Primary Colorectal Cancer Heterogeneity by Using Whole-Tumor Texture Analysis : Contrast-enhanced CT Texture as a Biomarker of 5-year Survival ». In : *Radiology* 266 (nov. 2012).

[Nic+19] Chiara Nicolò et al. « Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer ». In : *bioRxiv* (2019).

[Nio+18] Christophe Nioche et al. « LIFEx : A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity ». In : *Cancer Research* 78 (juin 2018), canres.0125.2018.

[Nke+16] Gabriel Nketiah et al. « T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness : preliminary results ». In : *European Radiology* 27 (déc. 2016).

[NKUZ00] László Nyúl, Jayaram K. Udupa et Xuan Zhang. « New Variants of a Method of MRI Scale Standardization. » In : *IEEE transactions on medical imaging* 19 (mar. 2000), p. 143-50.

[Noa+11] Simon Noah et al. « Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent ». In : *Journal of Statistical Software* 39 (mar. 2011).

[Noo+18] AM Noone et al. « SEER cancer statistics review, 1975-2015 ». In : *Bethesda, MD : National Cancer Institute* (2018).

[Nor+10] J Northover et al. « Chemoradiation for the treatment of epidermoid anal cancer : 13-year follow-up of the first randomised UKCCCR Anal Cancer Trial (ACT I) ». In : *British journal of cancer* 102 (mar. 2010), p. 1123-8.

[Nor88] Larry Norton. « A Gompertzian Model of Human Breast Cancer Growth ». In : *Cancer Research* 48.24 Part 1 (1988), p. 7067-7071. eprint : `https://cancerres.aacrjournals.org/content/48/24_Part_1/7067.full.pdf`.

[OPM02] Timo Ojala, Matti Pietikäinen et T Maenpaa. « Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns ». In : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24 (août 2002), p. 971-987.

[Orl+14] Fanny Orlhac et al. « Tumor Texture Analysis in F-18-FDG PET : Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis ». In : *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 55 (fév. 2014).

[Orl+18] Fanny Orlhac et al. « Validation of a method to compensate multicenter effects affecting CT radiomic features ». In : *Radiology* 291 (jan. 2018).

[Orl15] Fanny Orlhac. « Beyond the measurement of SUV in PET imaging : Properties and potential of the parameters of texture to characterize tumors ». Theses. Université Paris Sud - Paris XI, sept. 2015.

[Par+14] Chintan Parmar et al. « Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation ». In : *PloS one* 9 (juil. 2014), e102107.

[Par+19] Ji Eun Park et al. « Reproducibility and Generalizability in Radiomics Modeling : Possible Strategies in Radiologic and Statistical Perspectives ». In : *Korean Journal of Radiology* 20 (juil. 2019), p. 1124-1137.

[Par95] Charles Parisot. « The DICOM standard ». In : *The International Journal of Cardiac Imaging* 11.3 (sept. 1995), p. 171-177.

[Ped+11] F. Pedregosa et al. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12 (2011), p. 2825-2830.

[Pin88] Juan Pineda. « A Parallel Algorithm for Polygon Rasterization ». In : *In Proceedings of Siggraph '88*. 1988, p. 17-20.

[PPP17] Kedar Potdar, Taher Pardawala et Chinmay Pai. « A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers ». In : *International Journal of Computer Applications* 175 (oct. 2017), p. 7-9.

[QR03] Maurer, Rensheng Qi et Vijay Raghavan. « A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary images ». In : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25 (mar. 2003), p. 265-270.

[RAM87] S.J Roan, J.K Aggarwal et W.N Martin. « Multiple resolution imagery and texture analysis ». In : *Pattern Recognition* 20 (déc. 1987), p. 17-31.

[Rat+18] Saima Rathore et al. « Deriving stable multi-parametric MRI radiomic signatures in the presence of inter-scanner variations : survival prediction of glioblastoma via imaging pattern analysis and machine learning techniques ». In : fév. 2018, p. 8.

[RB+08] Matthias R Benz et al. « Combined Assessment of Metabolic and Volumetric Changes for Assessment of Tumor Response in Patients with Soft-Tissue Sarcomas ». In : *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 49 (oct. 2008), p. 1579-84.

[Ret+97] Michael Retsky et al. « Computer simulation of a breast cancer metastasis model ». In : *Breast cancer research and treatment* 45 (oct. 1997), p. 193-202.

[Rim14] Lorenzo Rimoldini. « Weighted skewness and kurtosis unbiased by sample size and Gaussian uncertainties ». In : *Astronomy and Computing* 5 (juil. 2014).

[RNZ18] Muhammad Imran Razzak, Saeeda Naz et Ahmad Zaib. « Deep Learning for Medical Image Processing : Overview, Challenges and the Future ». In : *Classification in BioApps : Automation of Decision Making*. Sous la dir. de Nilanjan Dey, Amira S. Ashour et Surekha Borra. Cham : Springer International Publishing, 2018, p. 323-350.

[RP14] François Rousseau et Nicolas Passat. « L'analyse et le traitement d'images IRM cérébrales. Techniques de l'Ingénieur ». In : *Techniques de l'ingénieur* (2014).

[RSR04] Antoine Rosset, Luca Spadola et Osman Ratib. « OsiriX : An Open-Source Software for Navigating in Multidimensional DICOM Images. » In : *J. Digital Imaging* 17.3 (2004), p. 205-216.

[RTA09] Michael B. Richman, Theodore B. Trafalis et Indra Adrianto. « Missing Data Imputation Through Machine Learning Algorithms ». In : *Artificial Intelligence Methods in the Environmental Sciences*. Sous la dir. de Sue Ellen Haupt, Antonello Pasini et Caren Marzban. Dordrecht : Springer Netherlands, 2009, p. 153-169.

[RV+12] Emmanuel Rios Velazquez et al. « A semiautomatic CT-based ensemble segmentation of lung tumors : Comparison with oncologists' delineations and with the surgical specimen ». In : *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* 105 (nov. 2012).

[Sam63] A.L. Samuel. « Some studies in machine learning using the game of checkers ». In : *Computers and Thought* (jan. 1963), p. 71-105.

[Sap+17] Maristella SAPONARA et al. « (Neo)adjuvant treatment in localised soft tissue sarcoma : The unsolved affair ». In : *European Journal of Cancer* 70 (jan. 2017), p. 1-11.

[SB12] Daniel STEKHOVEN et Peter BÜHLMANN. « MissForest ?Non-parametric missing value imputation for mixed-type data ». In : *Bioinformatics (Oxford, England)* 28 (jan. 2012), p. 112-8.

[Sen+13] E SENKUS et al. « Primary breast cancer : ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up on behalf of the ESMO Guidelines Working Group ». In : *Annals of Oncology* 24 (oct. 2013).

[Sha+11] Mohak SHAH et al. « Evaluating intensity normalization on MRIs of human brain with multiple sclerosis ». In : *Medical image analysis* 15 (avr. 2011), p. 267-82.

[SL91] S.Rasoul SAFAVIAN et David LANDGREBE. « "A Survey of Decision Tree Classifier Methodology." IEEE Transactions on Systems ». In : *Systems, Man and Cybernetics, IEEE Transactions on* 21 (juin 1991), p. 660 -674.

[SMS95] John SPRATT, John MEYER et John SPRATT. « Rates of growth of human solid neoplasms : Part II ». In : *Journal of surgical oncology* 60 (oct. 1995), p. 137-46.

[Sol+15] Theodoros SOLDATOS et al. « Multiparametric Assessment of Treatment Response in High-Grade Soft-Tissue Sarcomas with Anatomic and Functional MR Imaging Sequences ». In : *Radiology* 278 (sept. 2015), p. 142463.

[Son+16] Jiangdian SONG et al. « Association between tumor heterogeneity and progression-free survival in non-small cell lung cancer patients with EGFR mutations undergoing tyrosine kinase inhibitors therapy ». In : t. 2016. Août 2016, p. 1268-1271.

[SR15] Takaya SAITO et Marc REHMSMEIER. « The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets ». In : *PLOS ONE* 10.3 (mar. 2015), p. 1-21.

[Sta+09] Silvia STACCHIOTTI et al. « High-Grade Soft-Tissue Sarcomas : Tumor Response Assessment—Pilot Study to Assess the Correlation between Radiologic and Pathologic Response by Using RECIST and Choi Criteria 1 ». In : *Radiology* 251 (avr. 2009), p. 447-56.

[Sta+12] Silvia STACCHIOTTI et al. « Tumor response assessment by modified Choi criteria in localized high-risk soft tissue sarcoma treated with chemotherapy ». In : *Cancer* 118 (déc. 2012).

[Sub+13] Nagesh SUBBANNA et al. « Hierarchical Probabilistic Gabor and MRF Segmentation of Brain Tumours in MRI Volumes ». In : t. 16. Sept. 2013, p. 751-8.

[SUH+18] M SHAFIQ UL HASSAN et al. « Voxel size and gray level normalization of CT radiomic features in lung cancer ». In : *Scientific Reports* (juil. 2018).

[SWS17] Dinggang SHEN, Guorong WU et Heung-Il SUK. « Deep Learning in Medical Image Analysis ». In : *Annual review of biomedical engineering* 19 (mar. 2017).

[Szc+08] Piotr Szczypiaski et al. « MaZda-A software package for image texture analysis ». In : *Computer methods and programs in biomedicine* 94 (nov. 2008), p. 66-76.

[SZE02] J.G. Sled, Alex Zijdenbos et Alan Evans. « A nonparametric method for automatic correction of intensity nonuniformity in MRI data ». In : *Medical Imaging* 17 (jan. 2002), p. 87-97.

[SZH17] Shuang Song, Yuanjie Zheng et Yunlong He. « A review of Methods for Bias Correction in Medical Images ». In : *Biomedical Engineering Review* 3 (jan. 2017).

[Thi+09] Guillaume Thibault et al. « Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification ». In : nov. 2009.

[Tia+14] Fang Tian et al. « Response assessment to neoadjuvant therapy in soft tissue sarcomas : using CT texture analysis in comparison to tumor size, density, and perfusion ». In : *Abdominal imaging* 40 (déc. 2014).

[Tix+11] Florent Tixier et al. « Intratumor Heterogeneity Characterized by Textural Features on Baseline F-18-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer ». In : *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 52 (fév. 2011), p. 369-78.

[Tom76] Ivan Tomek. « Two modifications of CNN ». In : *IEEE Transactions on Systems, Man, and Cybernetics* 6 (nov. 1976).

[Tra+18] Alberto Traverso et al. « Repeatability and Reproducibility of Radiomic Features : A Systematic Review ». In : *International Journal of Radiation Oncology*Biology*Physics* 102 (juin 2018).

[TS+14] Russell T Shinohara et al. « Statistical normalization techniques for magnetic resonance imaging ». In : *NeuroImage. Clinical* 6 (déc. 2014), p. 9-19.

[UP11] Jonathan Uhr et Klaus Pantel. « Controversies in clinical cancer dormancy ». In : *Proceedings of the National Academy of Sciences of the United States of America* 108 (juil. 2011), p. 12396-400.

[Val+15] Martin Vallières et al. « A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities ». In : *Physics in medicine and biology* 60 (juin 2015), p. 5471-5496.

[VAPM14] Francisco J. Valverde-Albacete et Carmen Peláez-Moreno. « 100Normalized Information Transfer Factor Explains the Accuracy Paradox ». In : *PloS one* 9 (jan. 2014), e84217.

[Vij+02] Marc Vijver et al. « A Gene-Expression Signature as a Predictor of Survival in Breast Cancer ». In : *The New England journal of medicine* 347 (déc. 2002), p. 1999-2009.

[VL63] V.N. Vapnik et A Lerner. « Pattern Recognition Using Generalized Portrait Method ». In : *Automation and Remote Control* 24 (jan. 1963), p. 774-780.

[Wan17] Lidong Wang. « Heterogeneous Data and Big Data Analytics ». In : *Automatic Control and Information Sciences* 3 (août 2017), p. 8-15.

[Why+19] Phil Whybra et al. « Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging ». In : *Scientific Reports* 9 (déc. 2019).

[Wis+10] Gordon Wishart et al. « PREDICT : A new UK prognostic model that predicts survival following surgery for invasive breast cancer ». In : *Breast cancer research : BCR* 12 (jan. 2010), R1.

[YA16] Stephen Yip et Hugo Aerts. « Applications and limitations of radiomics ». In : *Physics in Medicine and Biology* 61 (juil. 2016), R150-R166.

[Yan+18] Fei Yang et al. « Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction : A simulation study utilizing ground truth ». In : *Physica Medica* 50 (juin 2018), p. 26-36.

[You+17] Safoora Yousefi et al. « Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models ». In : *Scientific Reports* 7 (déc. 2017).

[Yus+06] Paul A. Yushkevich et al. « User-Guided 3D Active Contour Segmentation of Anatomical Structures : Significantly Improved Efficiency and Reliability ». In : *Neuroimage* 31.3 (2006), p. 1116-1128.

[Zac+09] Evangelia Zacharaki et al. « Classification of brain tumor type and grade using MRI texture in a Machine Learning technique ». In : *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 62 (déc. 2009), p. 1609-18.

[ZH15] H Zou et T Hastie. « Regularization and Variable Selection via the Elastic Nets ». In : *J. Royal Stat. Soc. B* 67 (jan. 2015), p. 301-320.

[Zha+14] Binsheng Zhao et al. « Exploring Variability in CT Characterization of Tumors : A Preliminary Phantom Study ». In : *Translational oncology* 7 (fév. 2014), p. 88-93.

[Zhu+12] Ying Zhu et al. « Semi-Automatic Segmentation Software for Quantitative Clinical Brain Glioblastoma Evaluation ». In : *Academic radiology* 19 (mai 2012), p. 977-85.

[Zie03] Eric R Ziegel. « The Elements of Statistical Learning ». In : *Technometrics* 45.3 (2003), p. 267-268. eprint : `https://doi.org/10.1198/tech.2003.s770`.

[Zwa+16] Alex Zwanenburg et al. « Image biomarker standardisation initiative - feature definitions ». In : *CoRR* abs/1612.07003 (2016). arXiv : `1612.07003`.

[Gal74] M. M. Galloway. « Texture analysis using grey level run lengths ». In : *NASA STI/Recon Technical Report N* 75 (juil. 1974).