Développement de modèles interprétables par des approches d'apprentissage à partir d'images TEP, TDM, et IRM pour la prise en charge de patients atteints de cancer

Thibault Escobar

université
PARIS-SACLAY

Développement de modèles interprétables par des approches d'apprentissage à partir d'images TEP, TDM, et IRM pour la prise en charge de patients atteints de cancer

Development of interpretable models by learning approaches from PET, CT, and MRI images for the management of cancer patients

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°575 : Electrical, Optical, Bio : physics and Engineering (EOBE)
Spécialité de doctorat : Physique et imagerie médicale
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire d'Imagerie Translationnelle en Oncologie**, LITO (Institut Curie, Inserm)
sous la direction d'**Irène Buvat**, PhD, DR,
et le co-encadrement de **Laurence Champion**, MD,
et de **Sébastien Vauclin**, PhD.

Thèse soutenue à Paris-Saclay, le 23 novembre 2023, par

Thibault ESCOBAR

Composition du jury

Membres du jury avec voix délibérative

Jacques DARCOURT MD, PhD, PU-PH, Université Côte d'Azur	Président
Vincent NOBLET PhD, HDR, IR CNRS, Université de Strasbourg	Rapporteur & examinateur
Olivier SAUT PhD, DR CNRS, Université de Bordeaux, INRIA	Rapporteur & examinateur
Stéphanie NOUGARET-JUNG MD, PhD, HDR, PH, Université de Montpellier	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPAST164

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre : Développement de modèles interprétables par des approches d'apprentissage à partir d'images TEP, TDM, et IRM pour la prise en charge de patients atteints de cancer

Mots clés : imagerie médicale, interprétabilité, apprentissage automatique, radiomique, sous-région, voxel

Résumé : Cette thèse en partenariat avec l'Institut Curie et la société DOSIsoft explore l'importance croissante des sciences numériques en santé, en particulier dans le domaine de l'imagerie médicale en oncologie. L'utilisation de différentes techniques d'imagerie médicale, telles que la tomodensitométrie (TDM), la tomographie par émission de positons (TEP) et l'imagerie par résonance magnétique (IRM), joue un rôle essentiel à différents stades du diagnostic, de la planification du traitement, et du suivi des patients atteints de cancer.

L'adoption des approches radiomiques et d'apprentissage automatique dans la pratique clinique semble prometteuse, mais reste limitée aujourd'hui, notamment en raison du manque d'interprétabilité des modèles développés.

Cette thèse vise à relever ce défi en proposant de nouvelles méthodologies intégrant la modélisation, la visualisation, et la cartographie des modèles, avec un accent mis sur la simplicité. L'objectif est de fournir aux cliniciens et aux chercheurs une compréhension approfondie des informations à l'origine des décisions suggérées par les modèles, afin de faciliter leur adoption et leur utilisation en prise de décision clinique.

Cette recherche contribue au développement de modèles et de biomarqueurs robustes, fiables et cliniquement pertinents pour améliorer la prise en charge des patients atteints de cancer.

Title : Development of interpretable models by learning approaches from PET, CT, and MRI images for the management of cancer patients

Keywords : medical imaging, interpretability, machine learning, radiomics, subregion, voxel

Abstract : This thesis in partnership with Institut Curie and the company DOSIsoft explores the growing role of digital sciences in healthcare, particularly in the field of medical imaging for cancer. The use of different medical imaging techniques, such as computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI), plays an essential role for the diagnosis, treatment planning and follow-up of cancer patients.

The adoption of radiomics and machine learning approaches in clinical practice seems promising, but remains limited today, partly due to the lack of interpretability of the associated models.

This thesis addresses this challenge by developing new methods integrating modeling, visualization, and model mapping, with an emphasis on simplicity. The aim is to provide clinicians and researchers with an in-depth understanding of the information underlying the decisions suggested by models, in order to facilitate their adoption and use in clinical decision-making.

This research contributes to the development of robust, reliable and clinically relevant models and biomarkers to improve the management of cancer patients.

Alors que j'écris cette première page en dernier, il me parait nécessaire de prendre le temps de regarder en arrière et de réaliser le chemin parcouru. Ce manuscrit ne reflète, par son contenu, qu'une part de l'aventure. Elle aura été remplie de travail, de défis, mais également de rencontres, d'apprentissage, de rires, de voyages, et d'une multitude de souvenirs et d'anecdotes. Quelques mots pour vous en remercier.

Je tiens tout d'abord à exprimer ma gratitude à ma directrice de thèse, Irène Buvat. Ta confiance, ton soutien, tes idées et ta maîtrise, à la fois larges et précises, nos longues discussions, ainsi que ton enthousiasme et ta capacité à déployer des stratégies efficaces dans toutes les situations, ont été une aide précieuse. Merci de m'avoir laisser essayer, expérimenter, avec une liberté quasi-totale mais toujours justement encadrée. Tu m'as beaucoup appris. Je te remercie infiniment.

L'expérience n'aurait pas du tout été la même si je ne l'avais pas partagée avec Sébastien Vauclin. Je te remercie sincèrement pour ton encadrement. Ton calme, ta patience, ton aide, ton accompagnement dans les moments de réussite tout aussi fidèle dans les périodes plus difficiles, ont été indispensables. Je te dois de nombreuses compétences du domaine industriel. Et merci pour nos multiples rigolades. Tu es une personne formidable.

Laurence Champion. Merci beaucoup pour ton encadrement clinique, permettant constamment de contextualiser mon travail de façon pragmatique et dans son contexte applicatif visé. Ton accueil a toujours été chaleureux à Saint-Cloud et je m'y suis senti à chaque fois à la maison (à l'exception près que je réussissais encore à m'y perdre au bout de trois ans !). Je te remercie également pour ta confiance, ton aide, l'accès aux données, et pour le projet que nous avons initié ensemble grâce à toi.

Je souhaite remercier les membres de mon jury, Stéphanie Nougaret-Jung, Vincent Noblet, Olivier Saut, et Jacques Darcourt. Tout d'abord pour avoir accepté d'en faire partie. Je vous remercie également d'avoir rapporté et examiné en détails mon manuscrit, ainsi que pour la riche discussion que nous avons pu avoir à la suite de ma présentation.

Merci à toute l'équipe DOSIsoft. Marc Uszynski, merci de m'avoir permis de réaliser ce projet à vos côtés. Pascal, ton humour, sérieux sans se prendre au sérieux, le partage de tes expériences et ton aide. Mickaël, grand gourmand au grand cœur. Ronan, etttt ouuuais mon pottttttte ! Le second acolyte. Toujours à deux doigts de rire. Avec tout le reste de l'équipe PLANET, François, Angélique, Delphine, Tangi, Rémi, José, Rahma, Jeremy, Jérôme, Sébastien, je vous remercie d'avoir fait du septième un lieu de travail si agréable. J'utilise encore votre « ET PAF » en toute circonstance !

Merci au LITO et à tous ses membres. Fanny ma marraine du LITO, Christophe notre sauveur à tous, Fahad, Kibrom, Louis, Nicolas, Marie-Judith, Frédérique, Denis, Juliette, Leila, Julie, Arnaud, Narinée, Erwin, Vesna, Laurence, et tous les autres. Merci d'avoir, de près ou d'un peu plus loin, partagé avec moi cette expérience de la thèse, des congrès et voyages, des *abstracts*, des papiers, des challenges, des fameuses réunions radiomiques ! des restos U, du CESFO, des pique-niques, mais surtout des nombreuses discussions et nombreux débats et projets passionnants qui se déroulent au labo.

Merci aux équipes de médecine nucléaire et de radiopharmacie de Saint-Cloud, Romain, Claire, Laurence, Nicolas, Arnaud, tous les autres ! C'était toujours un plaisir de venir même si je devais braver le Transilien et la ligne 13. Tout le monde était à mes petits soins. Merci beaucoup pour votre aide, et surtout vos différentes visions du domaine qui m'ont permis de toujours garder le cap de l'applicatif clinique.

Merci à l'équipe du Centre Antoine Lacassagne, Olivier Humbert, Anne-Capucine Rollet, Jacques Darcourt, Fanny, pour m'avoir permis de travailler avec vous. Votre grand enthousiasme, votre disponibilité, vos avis et vos suggestions, ainsi que votre temps passé pour extraire et expertiser les données, ont été sans faille.

Hervé Brisse, Toulsie, Nayla, de la radiologie de Paris, merci pour l'intérêt que vous avez porté à mes travaux. Merci pour votre bienveillance et votre enthousiasme.

Merci à tous les génies du web, de GitHub à StatQuest, en passant par Toward Data Sciences, Medium, Machine Learning Mastery, et StackOverflow et Cross Validated. Combiné à la recherche scientifique, leur apport à la connaissance est d'une valeur inestimable.

Ma famille. Mes parents, ma sœur et mes frères, la *team* Escobar-Contet grâce à qui j'ai pu aller au bout de mes études, Mamie, Anne, Alain, Thomas, Margaux, I-Chih. Merci de m'avoir soutenu tout le long de cette étape. Merci pour ces week-ends riches en gastronomie et breuvages de qualité. C'était la course à chaque fois, mais je revenais à Paris complètement rechargé. Je vous aime.

Mes amis, mes agriculteurs, mes tractoristes, politiciens et philosophes du dimanche (et du samedi soir), mes sportifs, vous avez trop de casquettes ! Sacrées équipes ! Je rentrais à Paris rechargé certes, mais loin d'être reposé ! Merci pour toutes les escapades, les rigolades, les bêtises en tout genre, votre folie. Les quatre cents coups... on les a largement faits et ce n'est pas fini. Ne changez pas. Certains calmez-vous un peu peut-être, et encore. Mais ne changez pas. Je vous aime comme ça !

Karine, mon amour, ma moitié. Il y aurait tant à écrire. Je suis tombé sur la perle rare. Tu m'apportes tellement. Merci de me soutenir dans tout ce que j'entreprends. Merci pour ta bonté. Merci de m'avoir accompagné corps et âme dans cette aventure.

Mon fiston. C'est tous les jours un bonheur immense de t'admirer, découvrir le monde, progresser, avec tant d'énergie et ton magnifique sourire (sauf quand il faut aller au lit !). Tu es un trésor, une merveille.

Merci à ce qui ne sont plus là. Manou, Papy, Gilles mon parrain. Vous avez été les plus admirables et inspirants à mes yeux, à bien des égards. J'aimerais tant vous ressembler. Je vous dédie ce travail.

Lecteur. Merci à toi d'avoir ouvert ce manuscrit et d'avoir lu jusqu'ici. Allez, tente d'aller un peu plus loin va !

« La vérité est un miroir tombé de la main de Dieu et qui s'est brisé. Chacun en ramasse un fragment et dit que toute la vérité s'y trouve. »

Djalâl ad-Dîn Rûmi

Table des matières

Liste des figures

Liste des tableaux

1 - Introduction

1.1 . Motivations

Tout au long de l'histoire, les avancées technologiques ont toujours suscité une fascination profonde. Parmi tous les domaines, la médecine se distingue comme l'un des plus dynamiques en matière d'innovation. Cette dynamique résonne d'autant plus intensément qu'elle concerne notre vie, aspect fondamental de notre existence. Les avancées dans ce domaine ont permis, au fil du temps, de découvrir des applications cliniques d'une importance capitale, améliorant, par exemple, la détection précoce, le traitement, et le suivi de patients atteints de cancer.

Aujourd'hui, une part majeure de l'innovation a donné naissance à ce que l'on appelle communément « la transition numérique ». Elle englobe l'utilisation des données pour générer de nouvelles connaissances, et développer des outils, dans notre cas bénéfiques pour les patients. Notre ère est profondément marquée par cette transformation, et nous assistons à des changements de plus en plus importants offrant de nombreuses opportunités et défis passionnants pour la recherche scientifique et médicale.

L'imagerie médicale est un domaine intéressant, car il se situe à la croisée de diverses sciences, allant de la physique à la biologie, en passant par les mathématiques, la médecine, et aujourd'hui l'informatique et la science des données. Elle repose sur des principes scientifiques variés, utilisant des techniques sophistiquées pour visualiser l'intérieur du corps. Ces images jouent un rôle essentiel dans le diagnostic, la planification des traitements, et le suivi des patients. L'évolution rapide des technologies associées ouvre de nouvelles perspectives, notamment en matière de précision diagnostique, de personnalisation des traitements, et d'avancées dans la compréhension des maladies. Pour un esprit curieux, la polyvalence de ce domaine et les concepts qui s'articulent entre ces différentes disciplines en font une véritable mine d'or de découvertes et de challenges.

Naturellement, les paragraphes précédents expriment une part de mes opinions personnelles. Ainsi, ma motivation pour entreprendre une thèse de doctorat sur ce sujet était double. D'une part, je suis animé par la curiosité de mieux comprendre le fonctionnement de notre organisme, et d'explorer les outils actuels pour combattre voire éviter ses dysfonctionnements. D'autre part, je souhaitais humblement apporter ma pierre à l'édifice de la science, en espérant que mes travaux puissent offrir de nouvelles perspectives et contribuer à la lutte contre le cancer.

1.2 . Contribution scientifique

La présente thèse contribue au domaine de l'analyse d'images en oncologie, à plusieurs niveaux. Tout d'abord, une méthode de classification interprétable de tumeurs a été développée, permettant de localiser les motifs clés présents dans les lésions tout en maîtrisant la façon dont ils sont formalisés et utilisés. Cette approche offre une compréhension approfondie des caractéristiques des tumeurs, proposant ainsi de nouvelles voies pour l'interprétation diagnostique. De plus, en appliquant cette méthodologie dans différents contextes, cette recherche partage une expérience approfondie, mettant en lumière les défis et les questionnements généraux inhérents à ce domaine complexe, liant apprentissage automatique et sciences de l'image. Enfin, cette thèse propose une vision synthétique qui émerge des résultats obtenus. Elle ouvre sur des perspectives quant aux orientations me semblant pertinentes pour le futur.

1.3 . Résumé des chapitres

Chapitre 2

Ce chapitre introductif sert de base à la compréhension du rôle central de l'imagerie médicale en oncologie. Il souligne l'importance primordiale de l'imagerie médicale dans le domaine de l'oncologie, en se concentrant plus particulièrement sur l'utilisation répandue de la TEP, la TDM, et l'IRM. Le chapitre commence par donner une vue d'ensemble des concepts fondamentaux et des principes physiques et médicaux qui sous-tendent ces modalités d'imagerie. Le chapitre s'articule autour de leur contexte d'utilisation.

Outre l'imagerie elle-même, ce chapitre introduit le concept de biomarqueurs. Les biomarqueurs sont des mesures qui fournissent des informations précieuses sur l'évolution de la maladie, la réponse au traitement, ou encore le pronostic du patient. Ils contribuent à l'évaluation objective et à la quantification des caractéristiques des tumeurs, ce qui permet aux cliniciens d'optimiser les stratégies de soins.

Chapitre 3

Le troisième chapitre de cette thèse propose une exploration approfondie du domaine de la radiomique et de l'apprentissage. Il se penche sur les deux principales approches : la radiomique classique et la radiomique profonde. La radiomique implique l'extraction d'un grand nombre de caractéristiques quantitatives à partir d'images médicales, englobant la forme, l'intensité, la texture, mais également toute mesure conventionnelle telle que le SUVmax en TEP par exemple. Ces caractéristiques constituent une riche source d'informations qui peuvent être analysées pour découvrir de l'information utile dans divers contextes applicatifs. Nous abordons ainsi le concept de représentation dans le contexte des chaînes d'analyse en apprentissage statistique. La

représentation fait référence à la manière dont les images sont transformées et encodées pour faciliter leur analyse ultérieure, transformant les matrices brutes en variables, également appelées caractéristiques ou *features*.

L'approche générale de l'apprentissage automatique pour la classification est présentée, englobant le prétraitement des données, l'extraction et la sélection des caractéristiques, l'entrainement du modèle, et son évaluation. Il donne un aperçu des principaux algorithmes associés.

Enfin, nous abordons les principaux défis, notamment le manque d'interprétabilité, constituant le cœur de notre travail, mais aussi la généralisation des modèles, la variabilité des ensembles de données, et la nécessité d'une validation solide avant toute translation de la recherche vers la pratique de routine.

Chapitre 4

Le premier chapitre lié aux développements originaux de cette thèse présente une nouvelle méthodologie pour aborder la modélisation radiomique en combinant certaines forces de la radiomique classique et de l'apprentissage profond basé sur les réseaux de neurone convolutifs. Elle permet le passage entre l'échelle du voxel de l'image et les prédictions globales à l'échelle du patient, et inversement. En identifiant la localisation de l'information pertinente à l'intérieur de la lésion à la manière des cartes d'activation associées à l'apprentissage profond, notre méthode s'appuie sur des caractéristiques explicitement prédéfinies par des équations et extraites à l'échelle du voxel.

Permettant d'atténuer l'opacité de chacune des deux approches, elle facilite la création d'un modèle de substitution simplifié, dont l'objet est la prédiction de métastases pulmonaires deux ans après le bilan d'extension dans les sarcomes des tissus mous.

Chapitre 5

Dans ce chapitre, nous appliquons la méthode présentée précédemment à une problématique difficile : le diagnostic différentiel entre la récidive et les changements liés au traitement dans les gliomes et les métastases cérébrales. Plus précisément, nous concentrons sur les cas présentant des résultats en IRM et des symptômes ambigus, en utilisant l'imagerie TEP à la 18F-FDOPA. Dans le but d'améliorer la précision du diagnostic, nous introduisons une approche d'imagerie à deux points de temps en incorporant une acquisition tardive en plus de l'acquisition standard (précoce). Cette approche fournit des informations cinétiques supplémentaires, permettant une caractérisation plus complète de la pathologie.

Cependant, cette application spécifique a révélé des limites en termes d'interprétabilité. Bien que la méthode du chapitre précédent nous ait permis d'obtenir des cartes d'explication des modèles, elle n'a pas conduit à des interprétations permettant de formuler des hypothèses claires sur la façon

dont les modèles ont pris des décisions à partir des images. Nous proposons alors une nouvelle représentation qui facilite la reformulation des modèles à l'échelle du voxel, tout en agrégeant simultanément toutes les lésions dans une représentation unifiée pour faciliter l'interprétation globale. En utilisant les modèles de substitution qui en découlent, nos résultats suggèrent une amélioration de la stabilité lors de la simplification des modèles originaux suite à leur interprétation.

La méthode présentée dans ce chapitre, mais surtout son développement progressif en fonction des résultats, souligne l'importance d'adapter les méthodes d'interprétabilité à chaque application clinique spécifiques. Comme tout élément d'une chaine d'analyse en apprentissage automatique (théorème « *no free lunch* »), l'approche adéquate pour interpréter les modèles dépends fortement du contexte.

Chapitre 6

Dans l'avant-dernier chapitre de cette thèse, nous cherchons à évaluer l'impact des contraintes associées aux approches de modélisation radiomique proposées précédemment sur les performances de classification. Notre objectif est d'examiner l'axiome communément accepté qui suggère un compromis entre l'interprétabilité du modèle et sa justesse, selon lequel les modèles plus interprétables ont tendance à être moins précis, et inversement. Pour étudier cette relation, nous nous concentrons sur la prédiction de l'atteinte ganglionnaire lymphatique à partir des caractéristiques de la tumeur primaire en cancer du sein précoce. Ce travail a servi de point de convergence de mon cheminement intellectuel dans mon travail de thèse.

De façon intéressante, les résultats de cette étude suivent une tendance contraire : des approches de modélisation plus simples ont permis d'obtenir de meilleures performances de prédiction. Bien que nous reconnaissons que ces résultats ne remodèlent pas nécessairement l'ensemble du domaine de la modélisation radiomique, ils soulignent l'importance de considérer la simplicité comme une approche viable dans certains contextes. Les résultats de projets annexes en prédiction de survie, mais également en segmentation de lésions, sont mentionnés dans ce chapitre, et appuient ces observations. Ce chapitre représente une contribution au dialogue en cours sur la complexité et la précision des modèles dans le domaine de la radiomique et de l'apprentissage automatique. En remettant en question les hypothèses dominantes et en présentant des preuves, nous invitons à réévaluer les normes établies et encourageons les recherches futures sur des approches de modélisation plus simples.

Chapitre 7

Enfin, nous proposons une conclusion synthétique avec des perspectives ainsi que des considérations éthiques et opinions personnelles.

2 - L'imagerie multimodale en oncologie

Depuis le premier « Röntgenogram » (Figure 2.1), cliché radiographique de la main d'Anna Röntgen en 1895 à la suite de la découverte fortuite des rayons X par son mari Wilhelm Röntgen, l'imagerie médicale a connu un essor considérable [1]. Véritable rupture dans la représentation du corps humain en médecine, elle permet de caractériser macroscopiquement différentes maladies, dans de nombreuses spécialités.

Figure 2.1 – Tirage de la première radiographie de Wilhelm Röntgen de la main de sa femme Anna Röntgen, prise le 22 décembre 1895 [2].

En oncologie, elle constitue un ensemble d'outils de choix utilisés aujourd'hui en conjonction avec un arsenal large d'analyses [3]. Les marqueurs sanguins fournissent une information sur les éléments pathogènes circulants, alors que la réalisation d'une biopsie permet d'investiguer les caractéristiques moléculaires et génomiques du tissu cancéreux. La biopsie est le plus souvent réalisée antérieurement au traitement, en vue du choix de la stratégie thérapeutique la plus appropriée pour le patient. C'est la caractérisation la plus précise possible de la maladie qui permet d'optimiser et d'adapter le traitement à chaque patient, rendant possible une médecine de plus en plus personnalisée. Cependant le cancer est une pathologie complexe, hétérogène, et évolutive. Par exemple, Gerlinger et al. ont montré en 2012 que des biopsies prélevées à différentes positions dans la tumeur pouvaient conduire à des conclusions significativement différentes [4]. De plus, suivre l'évolution de la maladie dans le temps est crucial pour adapter la prise en charge au cours du temps [5]. Or, la biopsie est un acte invasif qui n'est pas envisageable à chaque examen de suivi.

L'imagerie permet la caractérisation à une autre échelle. Elle fournit des informations macroscopiques complémentaires. Surtout, elle permet d'analyser la tumeur dans son ensemble de façon non invasive [6]. Progrès en imagerie médicale et en oncologie sont alors liés, et l'émergence de différentes modalités a permis de visualiser et de mesurer différentes structures et de multiples mécanismes. Du dépistage précoce de certains cancers au suivi du patient, en passant par le diagnostic, le bilan d'extension, et la planification du traitement, l'imagerie multimodale est aujourd'hui incontournable en oncologie.

Ce chapitre décrit brièvement les différentes modalités d'imagerie analysées dans cette thèse. Les exemples de processus biologiques, de mesures, et les types de cancers mentionnés sont ceux qui ont été étudiées au cours de mes travaux.

2.1 . Des phénotypes macroscopiques de différentes caractéristiques biologiques

2.1.1 . La tomodensitométrie

La tomodensitométrie (TDM) (*computed tomography* (CT)) consiste à mesurer l'absorption des rayons X par les tissus. Inventée en 1972 par Godfrey N. Hounsfield et Allan M. Cormack [7], la TDM repose sur les mêmes principes physiques que la radiographie. Le rayonnement électromagnétique X est ionisant, il possède une énergie telle qu'il peut traverser la matière et ainsi être transmis par les tissus biologiques. Lors de leur trajet à travers le patient, les photons X interagissent avec la matière qui le compose. S'il n'est pas absorbé par effet photo-électrique, le photon X subsiste. Cependant, qu'il ait conservé totalement (interaction élastique Rayleigh) ou en partie (interaction inélastique Compton) son énergie, s'il a interagi avec la matière, il aura été diffusé. Il aura alors très probablement changé sa direction de propagation.

Macroscopiquement, si l'on mesure l'intensité I_k du faisceau transmis dans la même k-ième direction x_k que le faisceau incident d'intensité I_0, après qu'il a traversé une épaisseur t_k $[cm]$ de matière hétérogène de la source au détecteur (air, vêtements, différents organes), celle-ci sera atténuée telle que :

$$I_k = I_0 \times e^{-\int_0^{t_k} \mu(x_k) \times dx_k} \tag{2.1}$$

$$\ln \frac{I_0}{I_k} = \int_0^{t_k} \mu(x_k) \times dx_k \tag{2.2}$$

avec $\mu(x_k)$ $[cm^{-1}]$ le coefficient d'atténuation linéique du tissu traversé pour toute valeur de $x_k \in [0, t_k]$.

Les équations (2.1) et (2.2) décrivent les projections selon x_k des coefficients d'atténuation linéique des tissus traversés lors du trajet. Après l'acquisition de nombreuses projections tout autour du patient, différents algorithmes de reconstruction tomographique basés sur la transformée de Radon permettent la reconstruction d'une image numérique de μ en trois dimensions (3D) [8-10].

Le coefficient d'atténuation linéique μ augmente avec la densité et le numéro atomique des atomes composant la matière traversée, et diminue si l'énergie du rayonnement incident augmente. Bien qu'en pratique, il soit impossible d'obtenir un rayonnement parfaitement monochromatique et constant dans le temps, le spectre du rayonnement incident est quasiment identique pour toute projection k. L'imagerie TDM correspond de ce fait à une mesure 3D de la densité tissulaire et atomique du corps. Les matériaux peu denses et ayant des numéros atomiques bas donnent un signal de faible intensité, tandis que ceux qui ont une masse volumique ou des numéros atomiques plus élevés induisent de plus hautes intensités. Le signal est exprimé en unités de Hounsfield (*Hounsfield units* (HU)) correspondant à une mesure relative par rapport à l'eau et l'air tel que :

$$HU_v = \frac{\mu_v - \mu_{eau}}{\mu_{eau} - \mu_{air}} \times 1000 \tag{2.3}$$

$$HU_v \approx \frac{\mu_v - \mu_{eau}}{\mu_{eau}} \times 1000 \tag{2.4}$$

avec μ_v $[cm^{-1}]$ le coefficient d'atténuation linéique moyen des matériaux au voxel v. μ_{eau} et μ_{air}, approximativement égaux à $1cm^{-1}$ et $0cm^{-1}$ respectivement, sont vérifiés lors de la calibration du scanner. La taille des voxels des images TDM actuelles varie d'environ $1mm \times 1mm \times 1mm$ à $2mm \times 2mm \times 2mm$.

Centrées sur $0HU$ pour l'eau et allant de $-1000HU$ pour l'air à plus de $2500HU$ pour l'os cortical, les unités de Hounsfield prennent des valeurs caractéristiques pour de nombreux types de tissus biologiques [10, 11]. En oncologie, les unités de Hounsfield peuvent renseigner sur certains processus tumoraux. Par exemple, bien que similaires à des séquestres osseux, la présence de calcifications (valeurs élevées) dans le volume tumoral de chordomes, des cancers rares affectant la base du crâne et le rachis, est associée au type chondroïde. Elle peut également aider au diagnostic différentiel des chondrosarcomes [12]. La détection de calcifications à l'aide de la TDM est par ailleurs utile dans la caractérisation des sarcomes des tissus mous (*soft tissue sarcomas* (STS)) [13]. Plus globalement, dans différents types de cancers, la détection d'une région centrale hypodense est souvent associée à la présence de tissu nécrotique au sein de la tumeur [11]. Probable conséquence d'une croissance rapide, la nécrose tumorale est aussi fréquemment

associée à des voies de signalisation biologiques entraînant des conséquences fonctionnelles et tumorigènes telles que l'angiogenèse, la libération de promoteurs métastatiques, ou l'augmentation de la résistance à la chimiothérapie [14, 15]. En outre, l'injection d'un produit de contraste en TDM permet d'évaluer la vascularisation de certaines structures [16].

2.1.2 . L'imagerie par résonance magnétique

L'histoire de l'imagerie par résonance magnétique (IRM) (*magnetic resonance imaging* (MRI)) repose sur les travaux de nombreux chercheurs à partir des années 70 [17]. L'IRM exploite le principe de la résonance magnétique nucléaire (RMN). Découverte en 1938 par Isidor I. Rabi, prix Nobel de physique en 1944, et formalisée par Félix Bloch et Edward M. Purcell en 1946, prix Nobel de Physique en 1952, la RMN est une propriété quantique de tout noyau atomique possédant un spin, responsable du moment magnétique intrinsèque $\overrightarrow{\mu}$ $[A \times m^2 = J \times T^{-1}]$ de ce noyau, et placé dans un champ magnétique $\overrightarrow{B_0}$ $[T]$ suivant l'axe z [17]. La RMN concerne différents noyaux chargés mais le proton est le plus utilisé et étudié en IRM du fait de la grande abondance d'hydrogène dans le corps humain. En l'absence de champ magnétique extérieur $\overrightarrow{B_0}$, l'ensemble des moments magnétiques des protons d'un échantillon est orienté de manière aléatoire. En outre, les molécules auxquelles appartiennent ces noyaux sont soumises à l'agitation thermique. La position et l'orientation d'un moment magnétique unitaire varient alors au cours du temps. Pour tout moment magnétique $\overrightarrow{\mu_i}$, il existe toujours un autre moment magnétique proche $\overrightarrow{\mu_j}$ ayant la même orientation mais un sens opposé. De ce fait, l'aimantation nette globale résultante $\overrightarrow{M}$ $[A \times m^{-1}]$ pour l'ensemble de l'échantillon est nulle. En présence de $\overrightarrow{B_0}$, les moments magnétiques entrent en précession autour de z à la fréquence de Larmor telle que :

$$\omega_0 = \gamma \times B_0 \tag{2.5}$$

$$\nu_0 = \frac{\omega_0}{2\pi} = \frac{\gamma}{2\pi} \times B_0 \tag{2.6}$$

avec γ $[rad \times s^{-1} \times T^{-1}]$ le rapport gyromagnétique caractéristique de l'élément étudié, ω_0 $[rad \times s^{-1}]$ la vitesse angulaire de précession autour de z, et ν_0 $[Hz]$ la fréquence associée.

Le moment magnétique du proton ne peut prendre que deux orientations de précession : une telle que $\mu_z = \mu/2$ (état parallèle α) et l'autre telle que $\mu_z = -\mu/2$ (état antiparallèle β). $\overrightarrow{B_0}$ n'a aucun effet sur la phase de précession des moments magnétiques. Soient les axes orthogonaux x_{rot} et y_{rot} dans le plan perpendiculaire à z et en rotation autour de ce même axe à la fréquence ν_0. Pour tout moment magnétique $\overrightarrow{\mu_i}$ en précession autour de z, il existe $\overrightarrow{\mu_j}$ en précession autour de z tel que $\overrightarrow{\mu_{i_{x_{rot},y_{rot}}}} + \overrightarrow{\mu_{j_{x_{rot},y_{rot}}}} = \overrightarrow{0}$. $\overrightarrow{B_0}$ influence donc l'aimantation nette globale selon z uniquement. À l'équilibre,

l'état énergétique α, dans le même sens que $\overrightarrow{B_0}$, est légèrement plus peuplé que l'état excité β plus énergivore. $\overrightarrow{M_0}$, somme vectorielle de tous les moments magnétiques nucléaires, est alors non nulle et dirigée dans la direction de $\overrightarrow{B_0}$.

L'application d'un autre champ magnétique $\overrightarrow{B_1}$ $[T]$ transitoire dans l'axe x_{rot} sous la forme d'une impulsion radio à la fréquence de résonance ν_0 provoque deux phénomènes. L'énergie apportée fait passer les moments magnétiques de l'état α à l'état excité β, entraînant une diminution de l'aimantation longitudinale en z. L'onde radiofréquence a également un effet sur la phase de précession des moments magnétiques. Macroscopiquement, $\overrightarrow{B_1}$ entraîne le changement de direction de l'aimantation nette globale $\overrightarrow{M}$ vers z décroissant en passant par l'axe y_{rot}. Cela s'explique par une cohérence de phase des moments magnétiques en précession autour de z telle que les composantes transversales $\overrightarrow{\mu_{x_{rot},y_{rot}}}$ de certains d'entre eux sont alignées en y_{rot}. C'est la « bascule ». En fonction du temps d'application et de l'intensité de $\overrightarrow{B_1}$, $\overrightarrow{M}$ peut être alignée dans la direction de l'axe y_{rot}.

Le retour à l'équilibre lorsque l'onde radiofréquence disparaît est décrit en fonction du temps t $[s]$ par les équations de Bloch, et régie par les temps de relaxation T1 et T2 tels que :

$$M_z(t) = M_0 \times (1 - e^{-\frac{t}{T1}}) \tag{2.7}$$

$$M_{x,y}(t) = M_{x,y}(t=0) \times e^{-\frac{t}{T2}}. \tag{2.8}$$

$T1$ $[s]$ est la constante de temps nécessaire pour le retour à l'équilibre thermodynamique des moments magnétiques excités par l'impulsion radiofréquence et donc la récupération de la composante $M_z = M_0$ de l'aimantation nette globale. La relaxation longitudinale correspond à un échange d'énergie entre les protons excités par $\overrightarrow{B_1}$ et leur environnement, essentiellement sous forme d'énergie thermique. Due à la perte de cohérence de phase des moments magnétiques au cours du temps, $T2$ $[s]$ est la constante qui régit la décroissance exponentielle de l'aimantation transversale $M_{x,y}$.

Soit x et y les axes orthogonaux du plan transversal fixe perpendiculaire à z. Une antenne fixe en x permet de capter le signal induit par l'aimantation transversale sous la forme d'une onde radio sinusoïdale de fréquence ν_0, amortie selon l'équation (2.8). C'est le « signal de précession libre ». Des gradients de champ magnétique sont appliqués à $\overrightarrow{B_0}$ de sorte que son intensité varie en fonction de x, y, et z. Selon l'équation (2.6), la fréquence $\nu_{0_{x,y,z}}$ varie donc dans l'espace, permettant la localisation du signal.

L'IRM est un domaine de l'imagerie médicale très polyvalent. Il existe une multitude de séquences destinées à mettre en évidence différentes caractéristiques tissulaires, anatomiques, voire des processus fonctionnels. Typiquement, les séquences pondérées en T1 maximisent le contraste entre les tissus

selon leur temps de relaxation longitudinal, tandis que les séquences pondérées en T2 les différencient selon leur relaxation transversale. Tels qu'illustrés en Figure 2.2, les matériaux à T1 longs apparaissent sombres sur les images pondérées en T1. Ceux à T2 longs apparaissent clairs sur les images pondérées en T2 et vice versa.

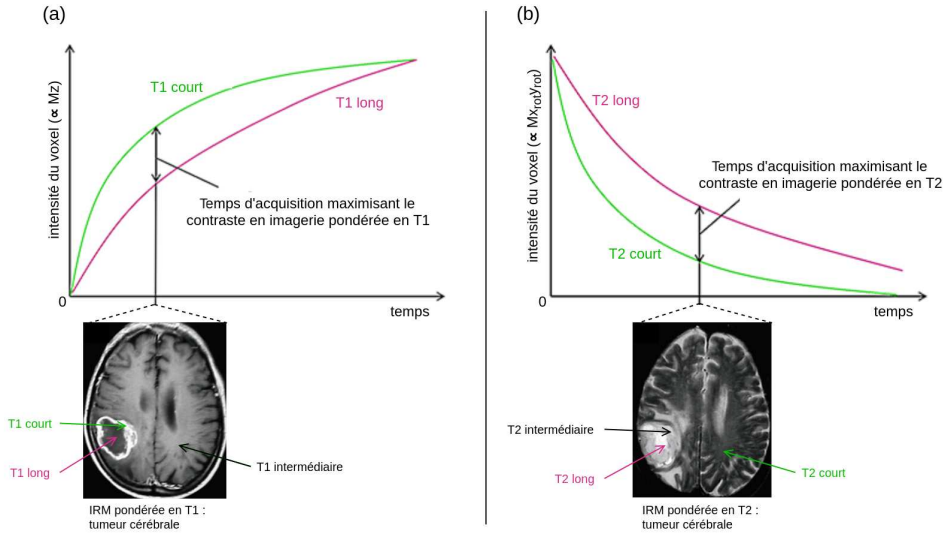


Figure 2.2 – Représentation simplifiée des relaxations longitudinale (a) et transversale (b) au cours du temps pour des tissus ayant des constantes T1 et T2 différentes. Les images associées proviennent de Liu et al. et correspondent aux examens IRM d'un patient atteint d'une tumeur cérébrale [18].

Il existe plusieurs variantes. Par exemple, certaines séquences permettent d'éliminer le signal de la graisse. En pondération T2, les principales sont la séquence pondérée en T2 avec saturation de la graisse (*fat-saturated T2* (T2FS)), et la séquence T2 avec inversion-récupération de tau court (*short tau inversion-recovery T2* (T2-STIR)).

Les facteurs affectant T1 et T2 sont généralement basés sur les mouvements et interactions moléculaires dans les différents tissus biologiques, présentant alors des contrastes caractéristiques [19]. En oncologie par exemple, la mélanine étant une substance paramagnétique favorisant l'état α, elle raccourcit le temps de relaxation longitudinale et fait apparaître le mélanome en hyper intensité sur les images pondérées en T1 [20]. Comme en TDM, l'injection d'un produit de contraste paramagnétique, généralement du gadolinium, permet de mettre en évidence certaines structures [21-23].

L'IRM anatomique a une taille de voxels relativement fine, environ $1mm \times 1mm$ dans le plan, avec une épaisseur de coupe plus importante pour les acquisitions en deux dimensions (2D). Au prix d'un temps d'acquisition plus long, les séquences d'acquisition en 3D permettent une résolution fine en z

également, ainsi qu'une augmentation du rapport signal-sur-bruit [24]. Hormis pour les séquences fonctionnelles (IRMf) et quantitatives (IRMq) visant à fournir des mesures calibrées en unités physiques, la valeur des voxels dans les images IRM ne représente pas une grandeur absolue [25, 26]. L'IRM anatomique est donc une imagerie de contraste nécessitant souvent des traitements de standardisation pour être utilisée dans le cadre d'études statistiques [27-33].

2.1.3 . La tomographie par émission de positons

Bien que la TDM et l'IRM pondérée en T1 ou T2 puissent renseigner sur des processus biologiques, ceux-ci ne sont pas ciblés spécifiquement. On parle d'imagerie anatomique. À l'instar de l'IRMf, la tomographie par émission de positons (TEP) (*positron emission tomography* (PET)) est une modalité d'imagerie fonctionnelle : elle cible une molécule ou un groupement moléculaire impliqué dans des voies de signalisations métaboliques associées à une pathologie.

La TEP est une technique d'imagerie quantitative en médecine nucléaire. Elle repose sur les propriétés radioactives d'un radioisotope avec lequel une molécule vectrice est couplée afin de former un « radiotraceur ». Le vecteur moléculaire du radiotraceur est choisi pour son tropisme, tandis que le radioisotope est responsable du signal capté pour former l'image. Comme son nom l'indique, la TEP implique la désintégration $\beta+$, illustrée en Figure 2.3. De la découverte du positon (ou particule $\beta+$) en 1932 par Carl D. Anderson, aux tomographes actuels à champ de vue axial long voire corps entier, l'histoire de la TEP repose sur de nombreuses avancées physiques, technologiques, pharmacochimiques, et biologiques [34-41].

Injecté au patient avant l'acquisition, le radiotraceur émet des positons. Après un court trajet dans le corps du patient, le positon émis interagit avec un électron, et tous deux s'annihilent produisant une paire de photons γ émis diamétralement tel que :

$$E_\gamma = m_e \times c^2 \approx 511keV \tag{2.9}$$

avec E_γ [$J = 6,24 \times 10^{18} eV$] l'énergie électromagnétique d'un photon émis, m_e [kg] la masse d'un électron, égale à celle d'un positon, et c [$m \times s^{-1}$] la vitesse d'une onde électromagnétique dans le vide [42].

Les photons γ produisent le signal de l'image grâce à un ensemble de détecteurs disposés tout autour du patient en couronnes empilées. Leur détection en coïncidence temporelle permet de les apparier et de déterminer les « lignes de réponse » des désintégrations. Les nombreuses lignes de réponse enregistrées forment des projections autour du patient permettant la reconstruction d'une image localisant la désintégration $\beta+$ en 3D dans le corps [44-47]. La technologie temps de vol produit une estimation probabiliste

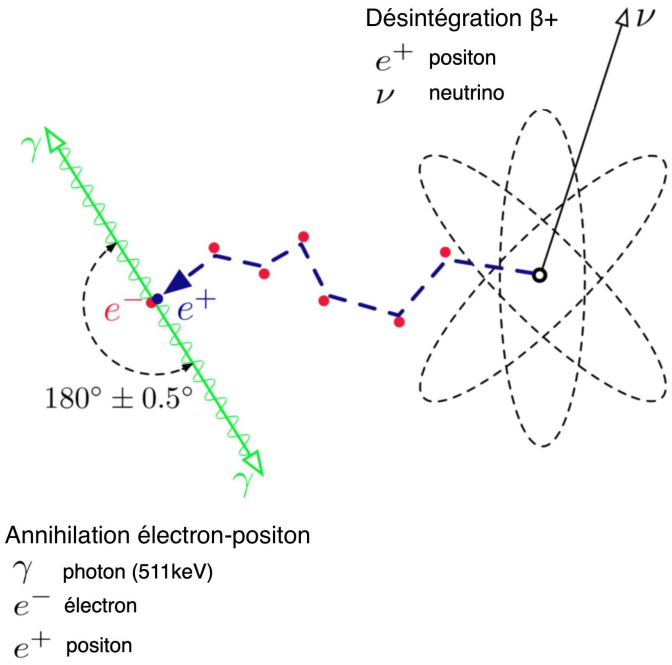


Figure 2.3 – Illustration du mécanisme d'annihilation électron-positon précédé de la désintégration $\beta+$. La distribution des angles d'émission est gaussienne d'espérance égale à $180°$, avec une largeur à mi-hauteur d'environ $0,5°$ (écart-type $\sigma \approx 0,2°$). Cette dispersion est due à l'énergie cinétique non nulle de l'électron et du positon au moment de leur interaction, et à la conservation de la quantité de mouvement [43-45]. Figure adaptée à partir de la thèse de doctorat de Maus [46].

de la position de la désintégration pour chaque ligne de réponse, augmentant ainsi la qualité d'image [48]. L'image est finalement convertie en valeur de fixation normalisée (*standardized uptake value* (SUV)) tenant compte des différences de quantité de radiotraceur injecté et de poids corporel entre les patients telles que, pour tout voxel v :

$$SUV_v(t) = \frac{c_v(t)}{d/m} \tag{2.10}$$

avec d $[MBq]$ la dose de radioactivité injectée, $c_v(t)$ $[MBq \times ml^{-1}]$ la concentration de radioactivité aux voxel v, mesurée au temps t $[s]$ et corrigée de la décroissance radioactive ayant eu lieu entre l'injection et t, et m $[g]$ la masse du patient. En supposant que la masse corporelle humaine est en moyenne égale à $1g \times ml^{-1}$, les valeurs de SUV sont exprimées sans dimension.

Avec une résolution spatiale plus faible qu'en imagerie anatomique et des voxels d'environ $2mm \times 2mm \times 2mm$ dans les machines récentes, la TEP est aujourd'hui couplée à l'imagerie TDM dans les scanners TEP/TDM (PET/CT). Tout comme les rayons X en TDM, les rayons γ sont atténués

par les tissus biologiques traversés lors de leur trajet du point d'émission au capteur. Comme nous l'avons vu dans la Section 2.1.1, l'atténuation dépend du coefficient d'atténuation linéique μ, mesuré en TDM. Les scanners TEP/TDM permettent donc de réaliser la correction d'atténuation du signal TEP, en plus de fournir une information à la fois anatomique et fonctionnelle [36, 49]. Moins répandus, les scanners TEP/IRM (PET/MRI) fournissent également de l'information à la fois anatomique et fonctionnelle [50].

Les cellules cancéreuses ont généralement un métabolisme du glucose accru en raison de leurs besoins énergétiques élevés [38, 39]. De ce fait, le radiotraceur le plus utilisé en oncologie est le fluorodésoxyglucose marqué au fluor 18 (18F-FDG ou plus communément FDG). Analogue du glucose, le FDG est un marqueur de l'activité métabolique, permettant de détecter les tumeurs et de suivre leur évolution [51, 52]. Au-delà du métabolisme glucidique en tant que tel, certains mécanismes physiologiques sous-jacents peuvent être reliés à la fixation du radiotraceur. Par exemple en STS, en plus de la nécrose qui peut être évaluée en identifiant le signal hypométabolique, Rakheja et al. ont relié la fixation de FDG aux caractéristiques histologiques et à l'activité mitotique de la tumeur, et ont montré une corrélation positive significative entre le taux mitotique et le SUVmax, valeur maximale de SUV dans la tumeur [53, 54]. En plus du FDG, il existe de nombreux radiotraceurs permettant la mesure de processus variés. La 18F-fluoroethylL-tyrosine (18F-FET) et la 6-18F-fluoro-L-dopa (18F-FDOPA) sont des exemples d'analogues des acides aminés, utilisés pour l'évaluation des tumeurs cérébrales [40, 41, 55-58]. Moins utilisée en routine clinique, l'acquisition dynamique permet en outre d'évaluer l'évolution temporelle de la distribution du traceur dans l'organisme, rendant possible la mesure de processus fonctionnels additionnels [59].

2.2 . De l'information utile tout au long de la prise en charge et en amont

2.2.1 . L'imagerie du diagnostic à la rémission

L'imagerie médicale est omniprésente en oncologie. La recherche fondamentale et la recherche préclinique permettent d'améliorer les technologies et de découvrir de nouveaux processus physiopathologiques. La recherche translationnelle et la recherche clinique permettent quant à elles d'optimiser la prise en charge des patients.

Les modalités et la façon dont elles sont utilisées dépendent fortement du contexte.

Lorsqu'un patient présente des symptômes ou un résultat de dépistage suggérant un cancer, il est nécessaire de réaliser des examens complémentaires afin de confirmer ou d'infirmer cette suspicion. C'est en conjonction avec des analyses sanguines et un examen clinique que seront prescrits des

examens d'imagerie, voire une biopsie. Dans un contexte diagnostique, la TDM et l'IRM anatomique sont souvent prescrites.

Si le diagnostic d'un cancer est confirmé, l'équipe médicale prescrit d'autres examens pour aider à choisir la stratégie thérapeutique et planifier le traitement. C'est le « bilan d'extension » (« *baseline* »). Par exemple, le médecin détermine le stade du cancer. Pour certains cancers, il est important de connaître le grade de la tumeur ou le groupe de risque auquel le patient appartient. C'est le plus souvent dans ce contexte que sont prescrits les examens TEP/TDM et les séquences IRM correspondant au protocole dédié au type de cancer diagnostiqué. La Figure 2.4 montre les images TEP/TDM et IRM pondérées en T1 et en T2 avec suppression du signal de la graisse, au bilan d'extension, pour deux patients atteints de STS [60, 61].

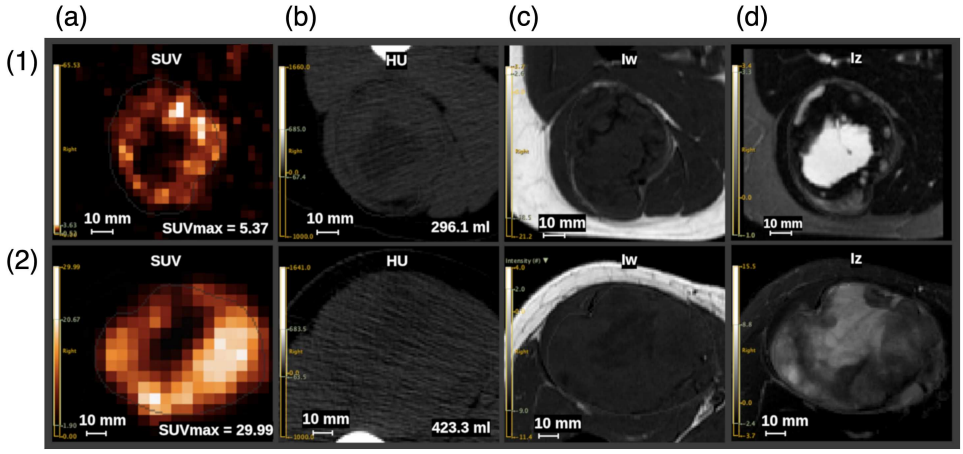


Figure 2.4 – Exemples de coupes transversales de la tumeur de deux patients (1, 2) atteints de STS, imagés au bilan d'extension en TEP (a), TDM (b), IRM T1 (c), et IRM T2 avec suppression du signal de la graisse (d). Figure créée à partir de Vallières et al. [60] et Escobar et al. [61].

Après l'initiation du traitement commence une période de surveillance. Constitué de consultations régulières, le suivi du patient a pour objectif initial l'évaluation de la réponse au traitement et le dépistage d'éventuelles complications. L'équipe médicale s'assure ensuite que le patient n'est pas en situation de récidive. On parle alors de « rémission », et le suivi s'espace progressivement. Cette période de surveillance implique la plupart du temps la prescription d'examens d'imagerie. Parfois, même après un temps de rémission conséquent pouvant se compter en mois voire en années, le patient présente des symptômes cliniques, ou certains éléments de son bilan de suivi nécessitent des investigations plus poussées. Des examens complémentaires peuvent alors être prescrits. Dans le cas de patients souffrant de tumeurs cérébrales traitées par chirurgie et radiothérapie par exemple, l'IRM et la

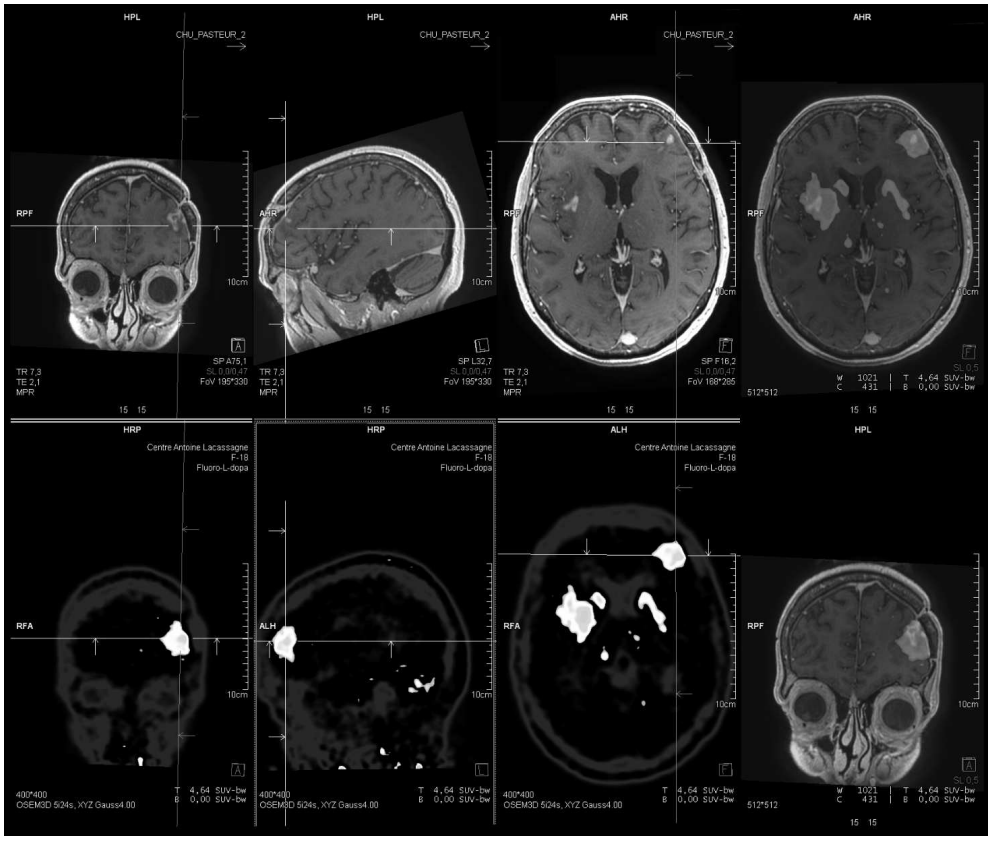


Figure 2.5 – Images de la synthèse du suivi en imagerie IRM pondérée en T1 et TEP à la 18F-FDOPA pour un patient atteint de gliome et traité par chimiothérapie, chirurgie, et radiothérapie. Les examens d'imagerie TEP ont été prescrits dans le cadre d'un diagnostic différentiel entre une récidive et une nécrose radio-induite.

TEP sont utilisées et permettent d'adapter la prise en charge en fonction de la réponse au traitement, mais aussi des potentielles toxicités liées à la radiothérapie [62]. La Figure 2.5 montre les images IRM pondérée en T1 et TEP à la 18F-FDOPA de suivi d'un patient atteint de gliome, en situation de nécrose radio-induite (radionécrose) confirmée. La TEP a été prescrite dans le cadre du diagnostic différentiel entre une récidive et une radionécrose induite par la radiothérapie [63].

Mes travaux se sont concentrés sur les modalités TDM, IRM, et TEP, présentées précédemment. Néanmoins, il existe d'autres modalités d'imagerie utilisées en oncologie [3], comme, par exemple, la radiographie, la mammographie, la tomographie d'émission monophotonique (TEMP) (*single-photon emission computed tomography* (SPECT)), ou encore l'échographie.

2.2.2 . Les biomarqueurs en imagerie oncologique

Pour chaque étape du parcours de soins, la prise de décision repose donc sur le phénotype du patient, évalué grâce à l'expérience de l'équipe médicale, et mesuré via divers « biomarqueurs ». Contraction des termes « marqueurs » et « biologiques », les biomarqueurs sont les mesures « évaluées objectivement comme indicateurs de mécanismes biologiques normaux, de processus pathogènes, ou de réponses pharmacologiques à une intervention thérapeutique » [64]. Certains fournissent une indication sur le risque de développer une complication tandis que d'autres peuvent signaler sa présence. D'autres encore donnent des informations sur la nature et la gravité de la maladie, ou permettent de suivre son évolution sous traitement.

En imagerie anatomique par exemple (TDM, IRM), le critère d'évaluation de la réponse pour les tumeurs solides (*response evaluation criteria in solid tumors* (RECIST)) est utilisé pour évaluer la réponse au traitement et guider la décision de le continuer, de l'arrêter, ou d'en changer [65, 66]. À partir du bilan d'extension, il est basé sur la mesure et l'évolution du plus grand diamètre de lésions dites « cibles » considérées comme représentatives de la maladie. Le critère RECIST est utilisé dans le monde entier en routine clinique, dans la prise en charge de presque tous les patients atteints de tumeurs solides. À l'origine du critère RECIST, le critère de réponse de l'organisation mondiale de la santé (*world health organization* (WHO)) est similaire (Figure 2.6) [67, 68]. Il correspond à la multiplication du plus grand diamètre de la lésion par son diamètre perpendiculaire. Malgré leurs formalismes différents, les critères WHO et RECIST se basent sur le même rationnel biologique : l'évolution de la taille des lésions en tant que marqueur de la charge tumorale.

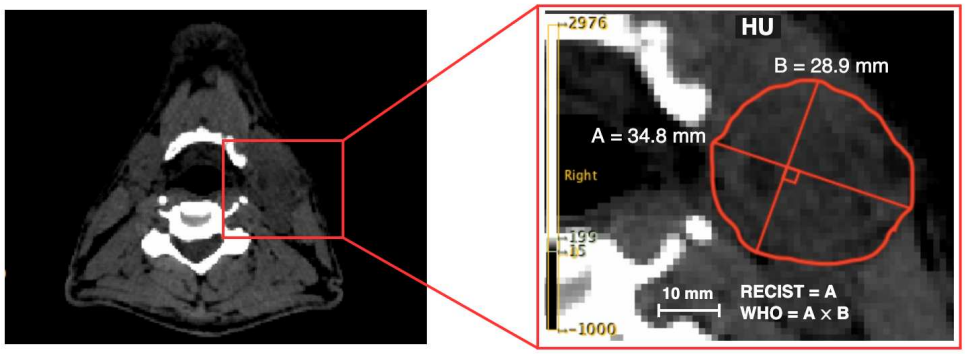


Figure 2.6 – Coupe transversale de l'image TDM d'un patient atteint de cancer de la base du crâne, centrée sur une adénopathie tumorale (lésion ganglionnaire), et illustration des méthodes de mesure associées aux critères RECIST et WHO. Figure créée à partir des données de la compétition « *head and neck tumor segmentation and outcome prediction in PET/CT images, third edition* » (HECKTOR 2022) [69].

En imagerie TEP au FDG, le critère d'évaluation de la réponse des tumeurs solides en tomographie par émission de positons (*positron emission tomography response criteria in solid tumors* (PERCIST)) est également très employé [70]. Il estime la réponse au traitement par la mesure de l'activité métabolique des lésions. Une particularité du critère PERCIST est que le signal TEP utilisé est normalisé par la masse corporelle maigre du patient (*lean body mass* (LBM)) plutôt que par sa masse totale. Le critère PERCIST est basé sur le SUVpeak, correspondant au signal moyen dans la sphère de $1cm^3$ présentant le signal le plus élevé dans la lésion. Il existe différentes définitions du SUVpeak dans la littérature [71]. Défini en Section 2.1.3, le SUVmax correspond quant à lui à la valeur du voxel d'intensité maximale dans une région d'intérêt. De façon similaire, le SUVmean correspond à la valeur moyenne d'intensité calculé en incluant tous les voxels de la région d'intérêt. Conceptuellement, SUV, SUV_{LBM}, SUVmax, SUVpeak, et SUVmean, sont proches, et l'information sémantique qu'ils évaluent en TEP au FDG est commune : le métabolisme glucidique des lésions (Figure 2.7).

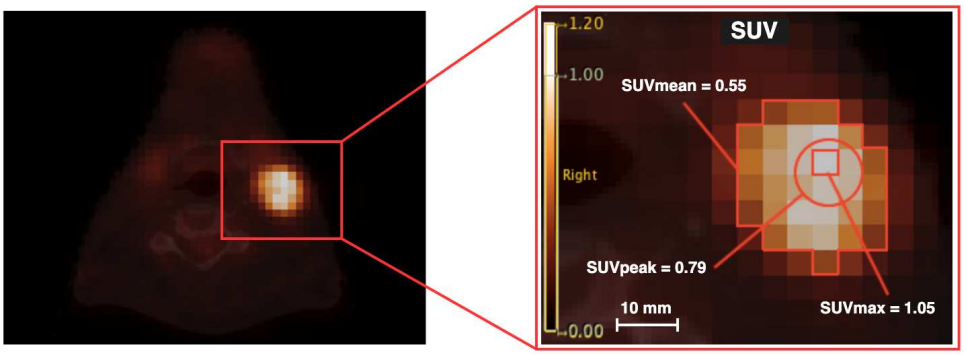


Figure 2.7 – Coupe transversale de l'image TEP au FDG superposée sur l'image TDM d'un patient atteint de cancer de la base du crâne, centrée sur une adénopathie tumorale (lésion ganglionnaire), et illustration du SUVmax, du SUVpeak, et du SUVmean. Figure créée à partir des données de la compétition « *head and neck tumor segmentation and outcome prediction in PET/CT images, third edition* » (HECKTOR 2022) [69].

Il existe de nombreux biomarqueurs en imagerie pertinents dans le contexte de l'oncologie [72]. Le Tableau 2.1 répertorie des exemples fréquemment utilisés en TDM, IRM, et TEP.

Tableau 2.1 – Exemples de biomarqueurs en imagerie TDM, IRM, et TEP, présentés avec leurs définitions et les mécanismes biologiques associés.

Nom	Modalité	Échelle de définition	Information mesurée	Définition
Volume anatomique [73]	Anatomique (TDM, IRM)	Lésion	Charge tumorale	Volume de la région d'intérêt tumorale
Plus grand diamètre	Anatomique (TDM, IRM)	Lésion	Charge tumorale	Distance maximale dans le plan (transversal, coronal, ou sagittal) entre les limites tumorales
Critère WHO [67]	Anatomique (TDM, IRM)	Une ou plusieurs lésions	Évolution de la charge tumorale	Évolution du plus grand diamètre multiplié par son diamètre perpendiculaire au cours du suivi
Critère RECIST [65, 66]	Anatomique (TDM, IRM)	Une ou plusieurs lésions	Évolution de la charge tumorale	Évolution du plus grand diamètre de lésions cibles au cours du suivi
Stade TNM (tumor, nodes, metastases) [74]	Anatomique (TDM, IRM)	Une ou plusieurs lésions	Charge tumorale	Stadification en fonction du nombre, du volume, et de la localisation des tumeurs (T), des ganglions lymphatiques (nodes (N)), et des métastases à distance (M)
SUVmax	TEP	Lésion	Activité métabolique	Valeur maximale dans la région d'intérêt tumorale
SUVmean	TEP	Lésion	Activité métabolique	Valeur moyenne dans la région d'intérêt tumorale
SUVpeak [71]	TEP	Lésion	Activité métabolique	Valeur moyenne dans la sphère de $1cm^3$ contenant l'intensité la plus élevée dans la région d'intérêt tumorale
PERCIST [70]	TEP	Une ou plusieurs lésions	Évolution de l'activité métabolique	Évolution du SUVpeak au cours du suivi
MTV (metabolic tumor volume) [75]	TEP	Lésion	Activité métabolique	Volume de la région d'intérêt tumorale au-dessus d'un certain seuil de fixation
TLG (total lesion glycolysis) [76]	TEP	Lésion	Activité métabolique	$MTV \times SUVmean$ (consommation totale)
Taux de fixation relatif (uptake ratio) [57, 58]	TEP	Lésion, organe, voxel	Activité métabolique	Fixation normalisée par le signal dans un tissu de référence
TMTV (total metabolic tumor volume) [77]	TEP	Corps entier	Activité métabolique	Volume métabolique non physiologique (somme de tous les MTVs)
Dmax (maximum distance) [78]	TEP	Corps entier	Dissémination de la maladie dans le corps	Distance maximale entre deux lésions ou voxels pathologiques
Stade d'Ann Arbor [79]	TEP	Corps entier	Dissémination de la maladie dans le corps et charge tumorale (lymphome)	Stadification en fonction du nombre, du volume, et de la localisation des régions atteintes
TTP (time to peak) [80]	TEP dynamique	Lésion, voxel	Dynamique de l'activité métabolique	Délai entre le début de l'acquisition et le pic d'activité
Pente [80]	TEP dynamique	Lésion, voxel	Dynamique de l'activité métabolique	Pente d'une droite de régression à partir d'un temps t choisi (généralement le pic d'activité)

3 - La radiomique : analyses d'images médicales conduites par les données

Traditionnellement, la recherche de biomarqueurs en imagerie commence par une hypothèse. Les biomarqueurs visent en effet à caractériser un ou plusieurs mécanismes anatomo-fonctionnels d'intérêt définis a priori [72]. C'est la recherche guidée par l'hypothèse (« *hypothesis-driven* » ou « *top-down* » *research*).

Comme nous l'avons vu dans le chapitre précédent, différentes mesures peuvent refléter une même information ou un même mécanisme (Section 2.2.2). L'analyse statistique rétrospective des données est alors nécessaire afin d'identifier les biomarqueurs candidats les plus prometteurs, en vue de validations prospectives voire d'essais cliniques. De plus, bien que facilitant la maîtrise du processus de recherche et la compréhension des résultats, la nécessité de formuler a priori des hypothèses et de définir des métriques associées peut constituer une limite. Même si l'équipe de recherche fait preuve d'intuition et de créativité, elle peut omettre ou ne pas prêter attention à des éléments de l'image pourtant importants. Si tel est le cas, les images seraient sous-exploitées par rapport à l'information qu'elles contiennent.

L'augmentation du nombre de données en santé ainsi que les avancées technologiques ont ouvert la voie à un autre paradigme de recherche pour l'imagerie médicale : la radiomique. Inspirée de domaines médicaux tels que les omiques (eg, génomique, transcriptomique, protéomique, métabolomique), la radiomique repose sur la recherche guidée par les données (« *data-driven* » ou « *bottom-up* » *research*) [81, 82].

La radiomique a été introduite dans les années 2010 par Gillies et al., puis Lambin et al., pour améliorer l'exploitation des images médicales. Elle correspond à l'extraction d'un grand nombre de descripteurs mathématiques à partir de celles-ci [6, 83-89]. En utilisant ces « caractéristiques » (« *features* ») en tant que données d'entrée pour des méthodes d'apprentissage automatique (*machine learning* (ML)), des modèles de classification et de prédiction sont conçus, par exemple pour prédire la réponse à un traitement ou estimer un risque de récidive.

Cette section décrit un état de l'art de l'approche radiomique. La notion de caractéristique est définie, suivie du principe général de l'apprentissage automatique et des algorithmes typiques. L'interprétation des modèles, son importance, et les principales approches sont présentées en fin de section.

3.1 . La représentation des images : les caractéristiques

Pour pouvoir utiliser les images médicales dans des études rétrospectives ou en modélisation statistique, il est nécessaire de les exprimer par des variables. C'est la « représentation » des données [90]. Dans un espace vectoriel, la matrice $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ (« *design matrix* ») représente ces variables pour tout $i \in [1, n]$ et $j \in [1, p]$ tels que :

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix} \tag{3.1}$$

avec $x_j^{(i)}$ la valeur de la j-ième variable (prédicteur, caractéristique, métrique, *feature*) pour le i-ième individu (patient, lésion). n et p sont respectivement le nombre total d'individus et de variables dans l'étude.

Les « caractéristiques radiomiques » peuvent être extraites par des expressions mathématiques bien définies, et remplir $\boldsymbol{X}$ de manière explicite sous la forme d'un tableau. Elles peuvent également être définies directement à partir des données et calculées par des couches de convolution successives dans des réseaux de neurones convolutifs profonds (*convolutional neural networks* (CNN)). Ce sont les « caractéristiques profondes » (« *deep features* ») [91, 92].

3.1.1 . Les caractéristiques radiomiques prédéfinies

Toute métrique extraite d'une image médicale peut être considérée comme une caractéristique radiomique. Celles issues d'un raisonnement sémantique médical, comme le SUVmax par exemple, n'en sont pas exclues. Cependant, c'est l'extraction à haut-débit de caractéristiques provenant du domaine de la vision par ordinateur, non spécifiques du domaine médical, qui a marqué la naissance de la radiomique.

L'extraction des caractéristiques radiomiques prédéfinies suit généralement trois phases : la sélection des régions d'intérêt (*regions of interest* (ROI)), le prétraitement du signal des images, et le calcul des métriques en tant que telles. Plusieurs solutions logicielles existent. Dans cette thèse, la bibliothèque Python PyRadiomics, développée par Van Griethuysin et al., a été utilisée [93, 94]. Les détails des étapes de prétraitement ainsi que la définition mathématique de toutes les caractéristiques sont décrits dans sa documentation. Toutes les visualisations et les manipulations des images ont été faites en utilisant le logiciel LIFEx développé au LITO par Nioche et al. [95, 96].

1 - **La sélection des régions d'intérêt**
Bien que des études récentes s'intéressent au phénotype de régions non tumorales [97, 98], les ROIs correspondent généralement à la segmentation des lésions. Le tracé des contours peut être réalisé manuellement par des experts ou s'appuyer sur des outils d'aide à la segmentation tels que des méthodes de seuillage, de morphologies mathématiques, ou de contours actifs. Des méthodes complètement automatiques existent également. Les approches par apprentissage « profond » (*deep learning* (DL)) montrent des résultats de plus en plus satisfaisants [99-101].

2 - **Le prétraitement du signal**
Dans une étude radiomique, plusieurs traitements sont appliqués au signal des images. L'interpolation permet d'obtenir des voxels isotropes et de même taille. Vient ensuite la discrétisation des valeurs (*binning*) consistant à les regrouper par intervalles (*bins*). Essentiels au calcul de certaines caractéristiques, les intervalles sont définis par leur taille (*fixed bin size* (FBS)) ou par leur nombre dans les ROIs (*fixed bin number* (FBN)) [102, 103]. La discrétisation réduit en outre la dépendance des caractéristiques au bruit dans les images [104]. Une taille fixe est généralement adaptée aux images pour lesquelles la valeur des voxels représente une grandeur absolue (TDM, TEP, IRM quantitative ou fonctionnelle), ou lorsqu'elle est partagée par l'ensemble de la cohorte (IRM anatomique après standardisation du signal). Il n'y a cependant pas de consensus fort quant aux stratégies de rééchantillonage et de discrétisation, ni à propos de la taille ou du nombre d'intervalles [103, 105-110]. Certains filtres peuvent également être appliqués préalablement au calcul des caractéristiques [93, 111].

3 - **Le calcul des caractéristiques**
Les caractéristiques sont ensuite extraites des images prétraitées. Elles peuvent caractériser l'intensité du signal et sa distribution statistique via l'histogramme des valeurs des voxels présents dans les ROIs. Ce sont les caractéristiques de « premier ordre » (moyenne, médiane, maximum, minimum, écart-type, dissymétrie (*skewness*), aplatissement (*kurtosis*), entropie, etc.) [112, 113]. Les caractéristiques de premier ordre ne dépendent pas de la répartition spatiale des valeurs de voxels. Pour différentes régions possédant le même nombre de voxels de mêmes valeurs mais à des positions différentes, l'histogramme est identique et les valeurs des caractéristiques qui en sont extraites sont égales. Les caractéristiques de « second ordre » évaluent quant à elles cet arrangement spatial. Aussi appelées « caractéristiques de texture », elles proviennent de matrices intermédiaires qui en codent

différents aspects. La matrice de co-occurrence des niveaux de gris (*gray level co-occurrence matrix* (GLCM)) renseigne sur la probabilité d'observer des combinaisons de deux valeurs dans une direction de l'espace et à une certaine distance [114]. La matrice des longueurs de niveaux de gris (*gray level run length matrix* (GLRLM)) compte le nombre de voxels consécutifs et de mêmes valeurs alignés dans une direction de l'espace, pour chaque intervalle d'intensité [115]. La matrice des tailles de zones de niveaux de gris (*gray level size zone matrix* (GLSZM)) compte le nombre de voxels connexes de chaque valeur, indépendamment de la direction spatiale [116]. La matrice des différences de tons de gris voisins (*neighbors gray tone difference matrix* (NGTDM)) quantifie la différence entre une valeur de gris et la valeur moyenne de ses voisins à une distance donnée. La somme des différences absolues pour chaque niveau de gris est stockée dans la matrice [117]. Dans une matrice des dépendances de niveaux de gris voisins (*neighboring gray level dependence matrix* (GLDM ou NGLDM)), chaque élément compte le nombre de fois qu'un voxel d'un niveau de gris donné est connexe avec un nombre donné de voxels de même valeur [118]. Enfin, les « caractéristiques de forme » décrivent la géométrie des ROIs, avec des métriques telles que le volume, la sphéricité, l'élongation, ou le rapport surface sur volume [93, 113, 119-121].

Que ce soit pour la sélection des ROIs [122-124], l'interpolation des voxels [125-127], ou la discrétisation du signal [108, 128], de nombreuses études ont montré l'impact du traitement des images sur les valeurs des caractéristiques radiomiques [105-107, 109, 129]. Il est donc crucial d'indiquer les choix effectués. Des efforts ont été entrepris dans le but de standardiser les différentes étapes nécessaires à l'extraction des caractéristiques radiomiques. Dans ce but, l'Initiative de Standardisation des Biomarqueurs en Imagerie (*Image Biomarker Standardization Initiative* (IBSI)) a été créée. IBSI est composée de plus de vingt-cinq équipes de recherche impliquées dans le domaine de la radiomique. Sur la base de consensus, elle fournit des définitions, des directives, ainsi que des données et valeurs de référence [110, 111]. À partir de ce guide devenu incontournable, il peut cependant être difficile de choisir les paramètres optimaux parmi toutes les options disponibles. En fonction de la modalité d'imagerie, des scanners impliqués, des protocoles d'acquisition, des types de cancer, ou des objectifs de l'étude, de nombreux choix méthodologiques demeurent de la responsabilité de l'équipe de recherche. Quelques études ciblant des applications précises ont proposé des chaînes de traitements dédiées, telles que Orlhac et al. en TEP, et Carré et al. et Saint-Martin et al. respectivement en IRM cérébrale et mammaire [31, 89, 130].

Remplissant la matrice $\boldsymbol{X}$, les caractéristiques extraites pour tous les individus de l'étude sont utilisées pour répondre à la question d'intérêt. Différents tests et modélisations statistiques peuvent alors être employés (Section 3.2).

3.1.2 . Les caractéristiques radiomiques profondes

La représentation peut également être « apprise » à partir des données. On parle de DL lorsque l'algorithme, en plus de s'ajuster pour répondre à la question d'intérêt, apprend la représentation des images, les caractéristiques, sans qu'elles soient définies au préalable. Au lieu de traiter les données pour obtenir des caractéristiques de texture ou de forme, un CNN utilise directement les images comme entrée, et apprend à extraire l'information pertinente pour répondre au problème posé. Si les caractéristiques radiomiques prédéfinies le sont indépendamment de la modélisation statistique ultérieure, l'extraction de caractéristiques profondes en fait souvent partie intégrante [91, 92]. Illustrée en Figure 3.1, la structure d'un CNN peut être décrite en deux blocs principaux : l'apprentissage de la représentation (*representation learning*), et la classification (ou la prédiction) [131].

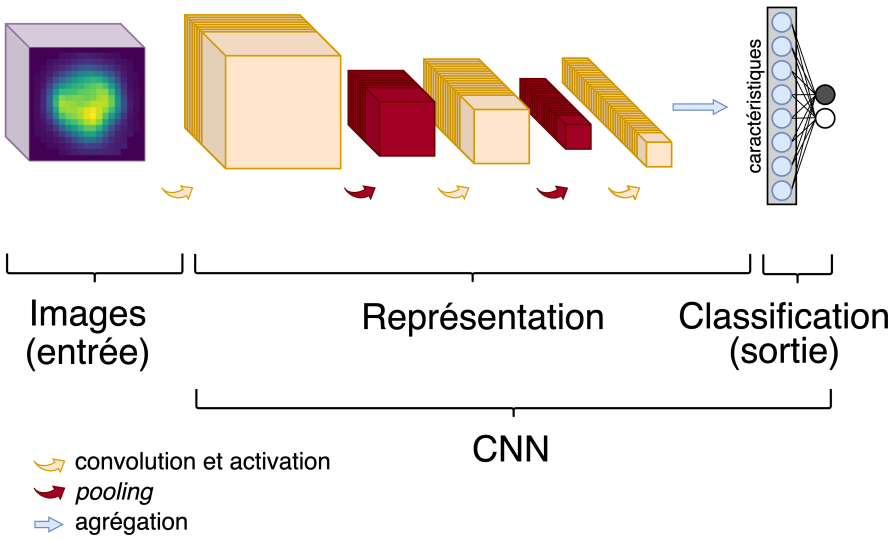


Figure 3.1 – Structure générale d'un CNN.

Un CNN est composé de briques élémentaires connectées entre elles sous forme de couches (*layers*). Les principales sont les couches de convolution, de « *pooling* », et de classification, aussi appelées « couches entièrement connectées ». Le choix dans l'organisation de ces couches dépend fortement de la tâche à effectuer et du contexte, et une infinité d'architectures de CNNs peut être conçue [132]. Dans le cas de la classification, les couches entièrement connectées sont précédées d'une étape « d'agrégation ».

- **Les couches de convolution**

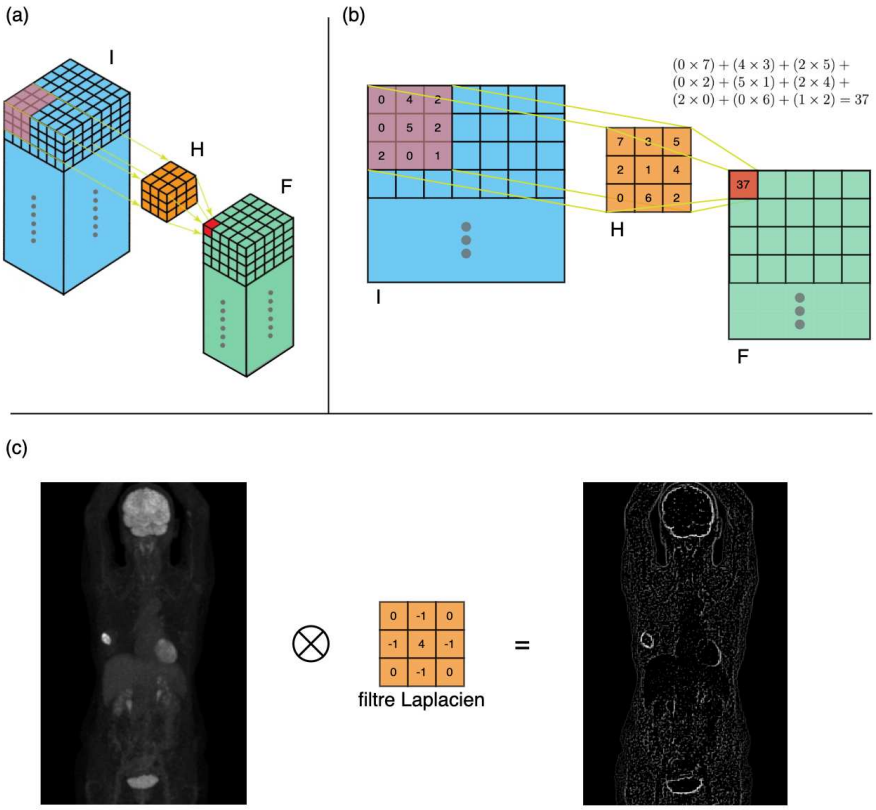


Figure 3.2 – (a) Illustration du principe de la convolution en 3D. (b) Exemple d'application numérique en 2D. Le produit scalaire entre le noyau H et les voxels de l'image source I est calculé localement. Cette opération est répétée pour tout voxel de I pour produire l'image de destination F. Chaque couche de convolution d'un CNN comporte de nombreux filtres. (c) Exemple du produit de convolution 2D de la projection d'intensité maximale (*maximum intensity projection* (MIP)) de l'image TEP d'une patiente atteinte de cancer du sein, par un filtre passe-haut Laplacien faisant apparaître les contours. Figure crée à partir de Bai [133].

Une couche de convolution est constituée d'un ensemble de filtres dont le but est d'extraire des motifs locaux à partir des images. Elle s'appuie sur un ensemble de « noyaux » (« *kernels* ») représentant ces motifs. Chaque filtre est appliqué à l'image d'entrée et produit une image de sortie. C'est la « réponse » de ce filtre, et on parle de « carte de caractéristiques » (« *feature map* »). Une couche de convolution en extrait un nombre égal au nombre de filtres qui la composent. En imagerie numérique 3D, la convolution est définie

pour tout noyau H tel que :

$$F_{x,y,z} = \sum_{h=-k}^{k} \sum_{i=-l}^{l} \sum_{j=-m}^{m} H_{h,i,j} I_{x+h,y+j,z+j} \tag{3.2}$$

avec I et F respectivement les images d'entrée et de sortie, définies dans l'espace par x, y, et z. H est le noyau de convolution d'indices h, i, et j, et correspond à une petite image mobile de dimensions $(2k+1) \times (2l+1) \times (2m+1)$ voxels[1]. Concrètement, cette opération consiste à faire « glisser »[2] le noyau H sur l'image I. Pour chaque position de H centré sur le voxel v de I en x, y, et z, le produit scalaire est calculé : Les voxels qui se superposent sont multipliés entre eux, et la somme de ces produits est affectée à v dans la carte de caractéristiques résultante (Figure 3.2). La convolution est donc une opération linéaire. Dans un CNN, la valeur des poids qui composent les filtres est apprise. Également apprise, une constante est associée au résultat de chaque filtre. C'est le biais. Enfin, l'application d'une « fonction d'activation » donne la possibilité à la couche de convolution d'extraire des caractéristiques liées aux images d'entrée par des relations non-linéaires [134]. La taille de l'image F en sortie est réduite par rapport à l'image I en entrée. Cela est dû au fait que les voxels aux bords de I n'ont pas un voisinage local complet avec lequel faire correspondre le noyau H. Ils ne sont alors pas utilisés en tant que voxels centraux lorsque H parcourt I. Pour contrôler la taille des images en sortie, il est possible d'ajouter artificiellement des voxels aux bords de I. C'est le « *padding* ». Différentes stratégies peuvent être adoptées. Par exemple, la valeur des voxels adjacents aux bords peut être répétée. Il est également possible de reproduire l'image par symétrie à la façon d'un miroir. L'approche la plus simple est généralement celle qui est utilisée. Elle consiste tout simplement à ajouter des voxels de valeur égale à zéro : on parle alors du « *zero-padding* ».

1. Bien que cela ne soit pas requis, le noyau de convolution est généralement cubique de dimension $3 \times 3 \times 3$ voxels. Dans ce cas, $k = l = m = 3$ dans l'équation (3.2).

2. L'équation (3.2) ne correspond pas tout à fait au produit de convolution. Dans un CNN, bien que nous appelions cela une convolution, c'est en fait une corrélation croisée qui est utilisée. La seule différence est due au fait qu'ici les indices de H sont ajoutés à ceux de I, alors qu'ils y sont soustraits dans le produit de convolution. L'utilisation de la corrélation croisée permet de préserver une orientation spatiale cohérente entre les images I et F, qui se retrouveraient inversées en x, y, et z si le produit de convolution était utilisé. Hormis cela, les deux opérations sont identiques.

- **Les couches de *pooling***

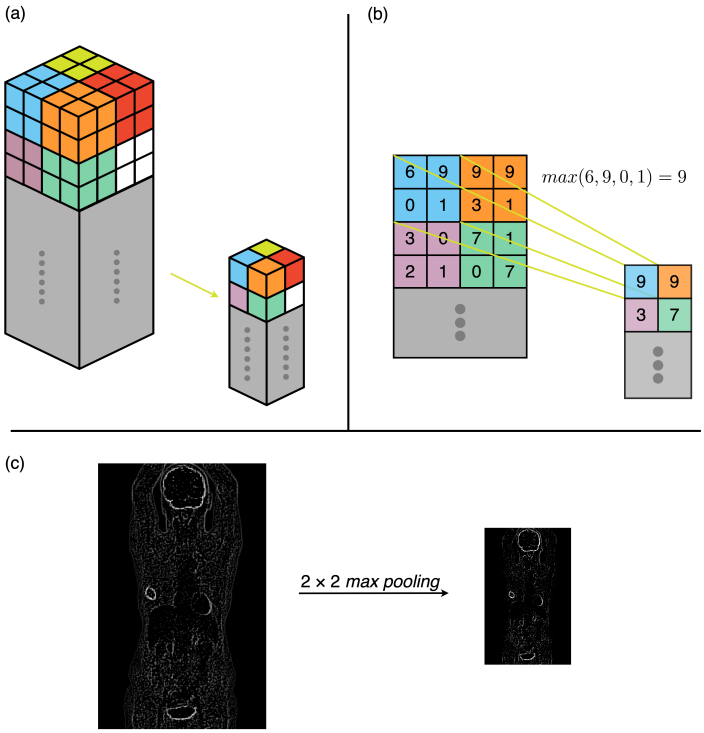


Figure 3.3 – (a) Illustration du *max pooling* en 3D. (b) Exemple d'application numérique en 2D. Dans le cas du *max pooling*, la valeur maximale des voxels superposés au filtre est conservée pour produire l'image en sortie. (c) Exemple d'application du *max pooling* à une carte de caractéristiques provenant de la projection d'intensité maximale (*maximum intensity projection* (MIP)) de l'image TEP d'une patiente atteinte de cancer du sein.

Les opérations de *pooling* sont placées entre les couches de convolution. Leur but est de réduire progressivement la taille des images. Un filtre de sous-échantillonnage parcourt l'image en entrée avec une taille de pas égale à ses dimensions. Par rapport à l'image en entrée, la taille de l'image en sortie est alors diminuée d'un facteur égal à celle du filtre dans chaque dimension (2 dans la Figure 3.3). Bien qu'il existe des variantes [135, 136], les filtres typiques sont le *max pooling*, illustré en Figure 3.3, et l'*average pooling* qui, comme son nom l'indique, calcule la moyenne locale lors du sous-échantillonnage. Un sous-échantillonnage induit nécessairement une perte d'information. Un avantage du *max pooling* est qu'il permet de préserver localement les voxels ayant produit la réponse la plus élevée en sortie de la couche de convolution précédente. Comme nous l'avons vu, les motifs extraits sont appris pour répondre au problème posé. Le *max pooling*

favorise donc les voxels portant une information « importante ». Plus globalement, le *pooling* permet d'extraire des caractéristiques à différentes échelles, tout en préservant une taille de noyaux de convolution fixe et limitée. Sur une grande image, un noyau de dimensions $3 \times 3 \times 3$ voxels extrait des motifs locaux de petite taille, alors que sur une image deux fois plus petite, il extrait des motifs deux fois plus grands. Il n'est alors pas nécessaire d'augmenter le nombre de poids que le CNN doit apprendre pour capturer une information de plus en plus globale (de grande taille et moins localisée). Le champ de réception (« *receptive field* ») d'une couche correspond à l'aire dans l'image initiale qui produit la réponse ponctuelle des filtres de cette couche. En d'autres termes, c'est le champ de vision que cette couche a sur l'image initiale. Les opérations de *pooling* augmentent donc progressivement le champ de réception du CNN.

- **Le bloc de représentation**
  Un réseau de neurones convolutif est principalement un empilement successif de couches de convolution et de *pooling*, auxquelles peuvent venir s'ajouter des modules d'objectifs variés. Par exemple, la normalisation par lots (batch normalization) [137], ou par « étreinte et excitation » (*squeeze-and-excitation*) [138]. Pour davantage de clarté, les couches de convolution et de *pooling* ont été décrites précédemment de façon univariée, c'est-à-dire en ne considérant qu'un seul filtre et donc qu'une seule caractéristique. Il est important de noter qu'un CNN possède une structure multivariée. Chaque couche reçoit plusieurs cartes de caractéristiques en entrée, et en produit plusieurs autres en sortie. Les données traitées dans d'un CNN 3D sont des « tenseurs » à quatre dimensions (4D). Ils sont définis dans l'espace par x, y, et z, mais aussi quantitativement par c (pour *channels*). La dimension c correspond au nombre de cartes de caractéristiques et donc de motifs extraits à chaque étape. Plus nous sommes en profondeur, c'est-à-dire éloignés de l'image initiale, plus les couches extraient des caractéristiques complexes. On dit que l'on augmente « l'abstraction » du signal extrait. En effet, les convolutions des couches profondes extraient des caractéristiques à partir des sorties des couches précédentes, à plus grande échelle, et ainsi de suite. Les premières couches de convolution quantifient des gradients orientés simples, tandis que des couches plus profondes extraient des structures telles que des ellipses par exemple. Comme le noyau est le même pour une carte de caractéristiques donnée, le nombre de poids à apprendre est indépendant de la taille de l'image en entrée. Les paramètres appris ne dépendent que de la taille du noyau et du nombre de cartes de caractéristiques dans les couches de convolution [131].

- **L'agrégation (classification et prédiction)**
  Dans le cas de la classification d'images telle que pour la réalisation d'un diagnostic différentiel, ou lors de la prédiction d'un risque, les données d'entrée et de sortie ne sont pas définies dans le même repère. Prenons le cas de la classification. Pour chaque individu, les images en entrée sont définies dans l'espace par x, y, et z. La sortie est quant à elle un vecteur de probabilité d'appartenance à chacune des classes, qui n'a pas d'information spatiale. En réduisant leur taille, les opérations de *pooling* agrègent progressivement le signal des images, diminuant leur représentation spatiale (dimensions x, y, et z) au profit d'une représentation quantitative (dimension c). Les couches de convolution et de *pooling* ne s'enchaînent cependant pas jusqu'à la perte totale de la dimension spatiale, et la dernière couche fournit un ensemble de petites cartes de caractéristiques. Avant le bloc de classification du CNN, la sortie des convolutions est alors convertie en un vecteur à une dimension (1D), soit par « aplatissement » de la dimension spatiale (« *flatten* »), soit via une opération de « *global pooling* » telle que le « *global average pooling* » (GAP) par exemple (Figure 3.4). L'agrégation par aplatissement (*flatten*) encode une information de localisation implicite associée à chaque motif. En effet, soit i et j deux éléments du vecteur 1D généré, provenant tous deux de la même carte de caractéristiques. Les positions de i et j dans l'espace sont distinctes avant l'agrégation. Par contre, chacune d'entre elles est partagée par l'ensemble des images de l'étude. Si l'algorithme de classification affecte une plus grande une importance à i qu'à j, parce qu'ils proviennent de la même carte et extraient donc le même motif, cette différence d'importance est uniquement due à leur différence de position. En agrégeant le signal dans toutes les directions de l'espace pour chaque caractéristique, le *global pooling* résume l'information spatiale de chaque caractéristique en une seule valeur. Ne préservant que la dimension c, le *global pooling* favorise alors l'invariance par translation du CNN [139, 140]. Bien que cela reste à démontrer, lorsque l'organisation spatiale globale est importante et commune aux images de l'étude, la préservation de l'information de localisation telle que préservée par l'aplatissement pourrait être utile, par exemple en analyse des images du corps entier d'atteintes métastatiques, ou de cancers diffus tels que les lymphomes. L'invariance par translation favoriserait quant à elle l'analyse quantitative de lésions solides, sans organisation spatiale structurée. La localisation dans ce cas pourrait cependant être codée par une caractéristique additionnelle.

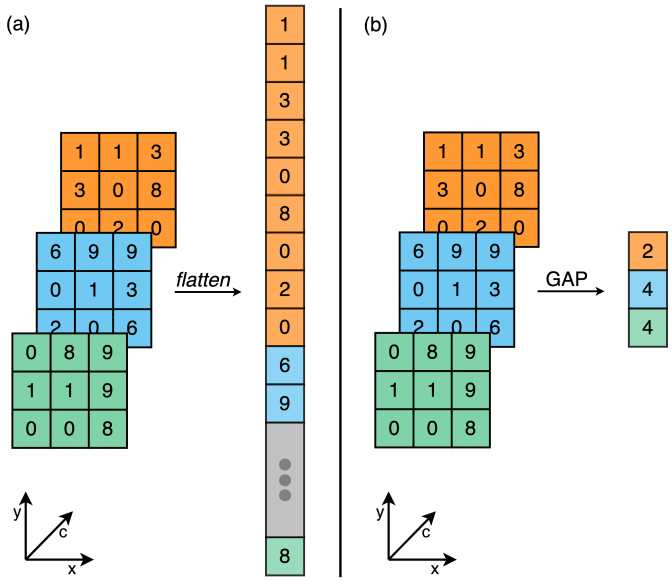


Figure 3.4 – Illustration en 2D du principe d'agrégation par aplatissement (*flatten*) (a), et par *global pooling* moyen (« *global average pooling* » (GAP)) (b).

- **Les couches entièrement connectées**
  Comme nous l'avons vu en Figure 3.1, l'organisation intrinsèque d'un CNN inclut un classifieur en sortie du bloc de représentation, qui s'ajuste pour répondre au problème. On dit que le CNN est entraîné « de bout en bout ». Tout comme la convolution, ce classifieur doit pouvoir être exprimé sous la forme d'un « réseau de neurones artificiel » (« *artificial neural network* » (ANN ou NN)), c'est-à-dire selon le paradigme du perceptron formalisé en 1958 par Rosenblatt [141, 142]. Décrits plus en détails dans la section suivante (3.2), les classifieurs typiques sont les « perceptrons multicouches » (« *multilayer perceptrons* » (MLP)), les « machines à vecteurs de support » (« *support vector machines* » (SVM)) ou « séparateurs à vaste marge », et les « modèles linéaires généralisés » tels que la régression logistique [143, 144].

Le CNN étant entraîné et ses paramètres figés, le vecteur 1D en sortie d'agrégation représente les caractéristiques profondes (Figure 3.1). Pour tout individu, il est possible de l'extraire afin de remplir une matrice $\boldsymbol{X}$. Seules ou en conjonction avec des caractéristiques prédéfinies, les caractéristiques profondes peuvent alors être utilisées en entrée de n'importe quel type de classifieur. Par des méthodes de fouille de données (*data mining*), leur extraction permet également d'étudier l'information qu'elles portent [145].

3.2 . L'apprentissage automatique : le modèle

Le ML est un domaine à l'interface entre la modélisation statistique et l'intelligence artificielle (IA) basé sur le développement d'algorithmes permettant aux ordinateurs d'apprendre à partir de données. Sans être explicitement programmés, ils peuvent améliorer leur performance pour des tâches spécifiques grâce à l'expérience.

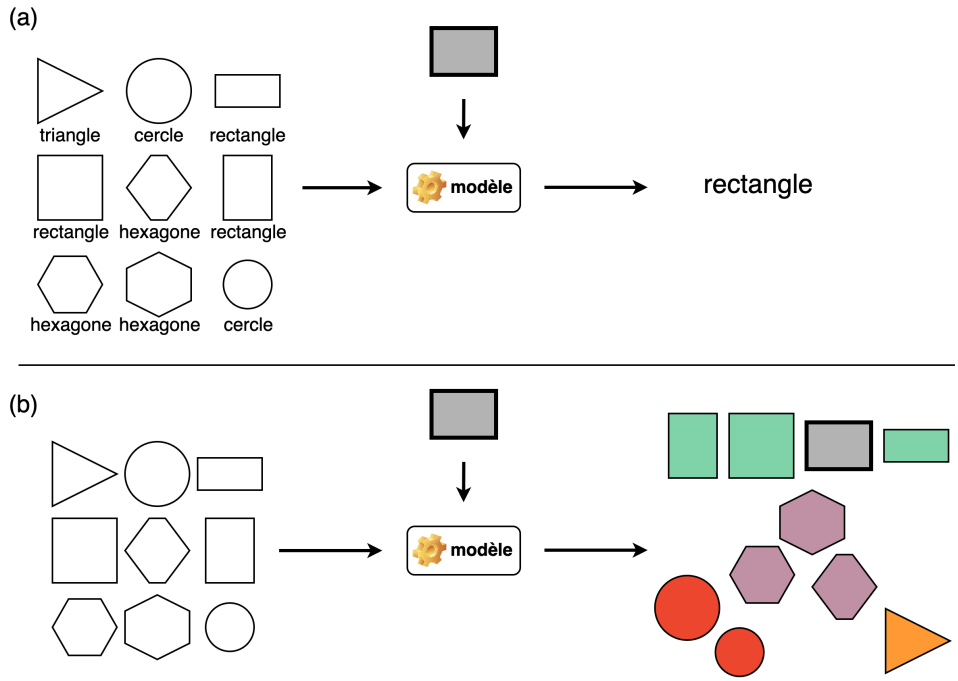


Figure 3.5 – Illustration du principe de la classification en apprentissage supervisé (a), et du *clustering* en apprentissage non supervisé (b).

Il existe de nombreuses approches. Dans l'apprentissage supervisé (*supervised learning*), l'algorithme est entraîné sur un ensemble « d'exemples » étiquetés. La sortie correcte pour chaque entrée est donc connue. C'est le « *label* » ou la « vérité terrain » (« *ground truth* »). L'objectif est que l'algorithme apprenne la correspondance entre l'entrée et la sortie, afin de pouvoir faire des « prédictions [3] », notamment sur de nouvelles données (Figure 3.5 (a)). Dans l'apprentissage non supervisé (*unsupervised learning*), les données d'apprentissage ne sont pas étiquetées. L'algorithme doit découvrir

3. En apprentissage automatique, on dit que le modèle « prédit » la sortie à partir des données d'entrée. Bien que dans certains cas, les modèles d'apprentissage « prédictifs » permettent de modéliser la survenue d'événements dans le futur à partir d'informations à propos du passé, le terme « prédiction » ne porte pas nécessairement une connotation temporelle.

la structure sous-jacente des données grâce à des techniques telles que le « *clustering* » (Figure 3.5 (b)) ou la réduction de la dimensionnalité. L'apprentissage semi-supervisé (*semi-supervised learning*) est à mi-chemin entre les apprentissages supervisé et non supervisé. L'algorithme est entraîné sur un mélange d'exemples étiquetés et non étiquetés. Cela peut être utile lorsqu'il existe une grande quantité de données disponibles, mais que la sortie est connue que pour une petite partie d'entre elles. Dans l'apprentissage par renforcement (*reinforcement learning*), l'algorithme apprend à effectuer des actions dans un environnement afin de maximiser un signal de récompense. L'algorithme apprend par essais et erreurs et ne reçoit pas de commentaire ou d'étiquette explicite.

L'influence des modèles d'apprentissage a considérablement augmenté au cours des dernières décennies et ils sont désormais utilisés dans divers aspects de notre vie [146]. Pour soulager les équipes de soins de tâches répétitives et accélérer certains processus longs et fastidieux, améliorer la reproductibilité des diagnostics, ou fournir des informations exploitables en recherche et en routine clinique, le ML et l'IA présentent de nombreux avantages potentiels en médecine [147-150]. Leurs applications couvrent toute la gamme des spécialités de l'imagerie médicale [82, 151-154]. Dans le premier rapport du groupe de travail sur l'IA de la Société américaine de Médecine Nucléaire et d'Imagerie Moléculaire (*Society of Nuclear Medicine and Molecular Imaging* (SNMMI) AI *task force*) par exemple, Bradshaw et al. ont proposé en 2021 une description des différentes applications du ML en imagerie en médecine nucléaire (Figure 3.6), accompagnée de lignes directrices et de mises en garde [155].

Parmi elles, nous pouvons citer la reconstruction des images [156, 157], leur posttraitement [158] et l'automatisation de certaines tâches d'analyse telles que la segmentation [99, 100], la détection de lésions [159], ou la réalisation de mesures quantitatives [160-162].

Cette thèse se concentre essentiellement sur la radiomique utilisée pour la classification supervisée, en tant qu'outil de découverte de nouveaux biomarqueurs diagnostiques, prédictifs, ou pronostiques [61, 82, 163-165].

3.2.1 . Principe général de l'apprentissage supervisé

En apprentissage supervisé, le vecteur $Y \in \mathbb{R}^n$ représente le résultat connu $y^{(i)}$ pour tout individu i d'un échantillon d'entraînement de n patients tels que :

$$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}. \tag{3.3}$$

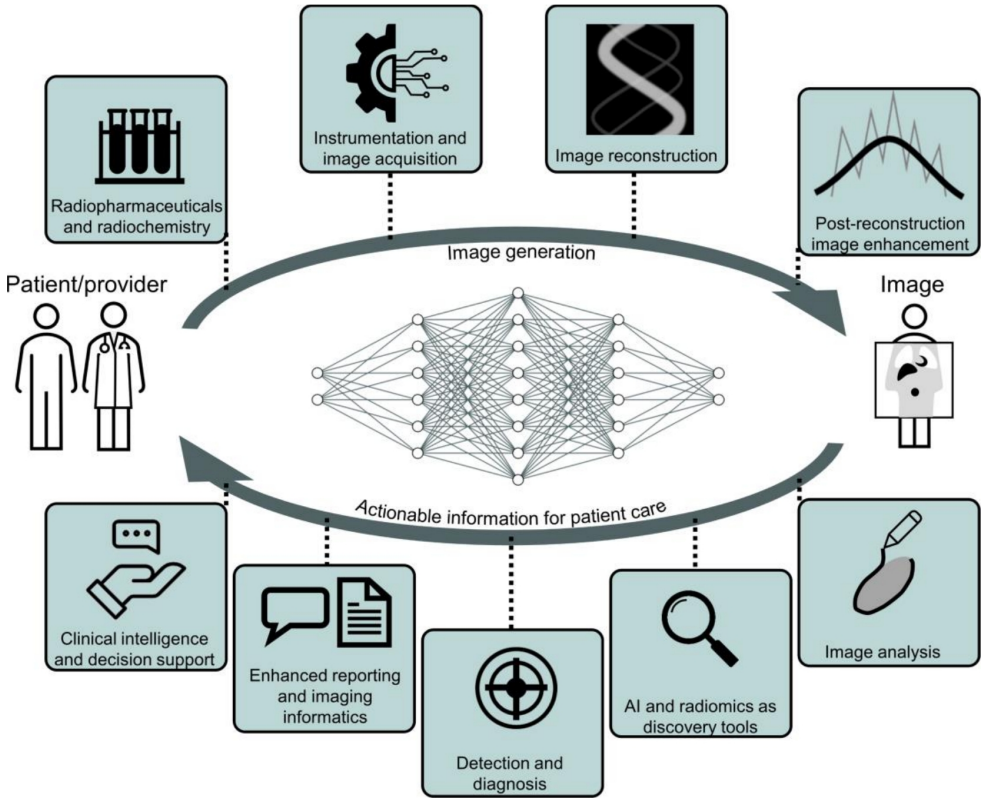


Figure 3.6 – Différents champs d'application de l'apprentissage automatique en imagerie en médecine nucléaire, du patient à l'image et inversement. Figure issue de Bradshaw et al. [155].

Nous voulons utiliser ces informations pour déterminer une fonction f capable d'approcher ce résultat sur la base des p caractéristiques radiomiques extraites des images et contenues dans $\boldsymbol{X}$ définie en (3.1), telle que :

$$f(\boldsymbol{X}) = \hat{Y} \tag{3.4}$$

avec $\hat{Y}$ l'ensemble des prédictions de f pour tout individu $i \in [1, n]$.

Si Y représente une catégorie ou un événement (par exemple un diagnostic différentiel [163]), f est une fonction de « classification », et $\hat{Y}$ est lié à la classe prédite pour chaque individu. Il peut également représenter une probabilité d'appartenance pour chacune des classes. Dans ce cas, on parle alors de modélisation « probabiliste ». Dans d'autres cas, Y est une valeur continue (par exemple le taux d'expression d'un gène [166]), et f est alors une fonction de « régression ».

Un grand nombre de méthodes ont été proposées pour construire des modèles. En 1997, William G. Macready et David Wolpert proposent le théorème du « *no free lunch* ». Ils suggèrent qu'il n'existe pas de « meilleur » algorithme capable de résoudre tous les problèmes avec une efficacité supérieure

aux autres. La performance de tout algorithme dépend de la nature spécifique du problème, des données, des objectifs de l'étude, et des contraintes qui peuvent s'y appliquer [167-169].

De la préparation des données au déploiement, en passant par la sélection des caractéristiques pertinentes et l'ajustement des « hyperparamètres », l'évaluation du modèle, et l'interprétation des résultats, les applications de techniques d'apprentissage supervisé suivent toutefois des chaînes d'analyse s'organisant de façon très similaire.

3.2.2 . L'entraînement : les algorithmes typiques

L'objectif consiste à ajuster les « paramètres » du modèle afin de minimiser l'écart entre les valeurs de sortie prédites, $\hat{Y}$, et les valeurs réelles, Y, sur les données d'apprentissage. Pour ce faire, nous utilisons généralement une « fonction de perte », également appelée « fonction de coût ». Certains des principaux algorithmes d'apprentissage supervisé sont présentés ci-après. Les éléments abordés correspondent à ceux qui ont motivé mes choix de méthodes tout au long de la thèse.

3.2.2.1 . La régression linéaire

Simple, la régression linéaire est un bon exemple pour introduire différents algorithmes d'apprentissage. Utilisée pour prédire la valeur d'une variable continue en fonction des caractéristiques, elle repose sur l'hypothèse selon laquelle $\boldsymbol{X}$ et Y sont reliés par une relation linéaire telle que :

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \beta_0 + \boldsymbol{\epsilon} \tag{3.5}$$

soit

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \beta_0 + \begin{bmatrix} \epsilon^{(1)} \\ \epsilon^{(2)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix} \tag{3.6}$$

avec β_0 le biais, et $\boldsymbol{\beta}$ le vecteur contenant les poids associés à chaque caractéristique $j \in [1, p]$. $\boldsymbol{\epsilon}$ est le vecteur des « résidus ». Pour tout individu $i \in [1, n]$, il correspond à l'écart entre le modèle et les données réelles. Pour un patient i, la prédiction d'un modèle de régression linéaire est donc de la forme :

$$\begin{aligned} \hat{y}^{(i)} &= \beta_0 + \sum_{j=1}^{p} (x_j^{(i)} \times \beta_j) \\ &= \beta_0 + \beta_1 \times x_1^{(i)} + \beta_2 \times x_2^{(i)} + \ldots + \beta_p \times x_p^{(i)}. \end{aligned} \tag{3.7}$$

Différentes fonctions de perte peuvent être utilisées pour apprendre des valeurs de β_0 et $\boldsymbol{\beta}$ afin d'ajuster le modèle, et de nombreuses formulations numériques existent pour minimiser chacune d'entre elles. De même que pour le choix du type de modèle, le choix de la fonction de perte et de la méthode d'optimisation dépend du contexte de l'étude, et des hypothèses statistiques associées [168]. Illustrée en Figure 3.7, la régression linéaire la plus populaire est celle des moindres carrés, qui consiste à choisir les coefficients minimisant la somme résiduelle des carrés (*residual sum of squares*), RSS, telle que :

$$\{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\} = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min}\ RSS(\beta_0, \boldsymbol{\beta}) \tag{3.8}$$

avec

$$\begin{aligned} RSS(\beta_0, \boldsymbol{\beta}) &= \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 \\ &= \sum_{i=1}^{n} (y^{(i)} - \beta_0 - \sum_{j=1}^{p} (x_j^{(i)} \times \beta_j))^2. \end{aligned} \tag{3.9}$$

3.2.2.2 . La régression logistique

Bien que nommée « régression », la régression logistique est un algorithme de classification. C'est un type d'analyse probabiliste utilisé pour modéliser la relation entre les caractéristiques et une ou plusieurs vérités terrain binaires. Elle fait partie des modèles linéaires généralisés. Dans le cas binaire simple à deux classes, elle modélise la probabilité a posteriori pour chaque individu d'appartenir à chacune des deux classes du problème, 1 (positive) et 0 (négative), telles que :

$$\hat{Y} = P(Y = 1|\boldsymbol{X}) = \frac{1}{1 + e^{-D(Y=1|\boldsymbol{X})}} \tag{3.10}$$

et

$$P(Y = 0|\boldsymbol{X}) = 1 - P(Y = 1|\boldsymbol{X}) \tag{3.11}$$

avec $P(Y = k|\boldsymbol{X})$ le vecteur de probabilité d'appartenance à la classe k pour les n individus de l'étude. Dans cette équation, D représente la « fonction de décision » linéaire de P, avec β_0 son biais appris et $\boldsymbol{\beta}$ le vecteur des coefficients appris associés aux p caractéristiques de $\boldsymbol{X}$. Similaire à l'équation (3.5), le vecteur $D(Y = 1|\boldsymbol{X})$ est défini pour tout individu tel que :

$$D(Y = 1|\boldsymbol{X}) = \ln \frac{P(Y = 1|\boldsymbol{X})}{P(Y = 0|\boldsymbol{X})} = \ln \frac{\hat{Y}}{1 - \hat{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \beta_0. \tag{3.12}$$

La probabilité P est reliée à D par une sigmoïde appelée « fonction logistique » $\sigma : \mathbb{R} \to [0\,;1]$, $\sigma(D) = 1/(1 + e^{-D})$. C'est la fonction d'activation, ou « fonction de lien », dont la fonction inverse est la « fonction

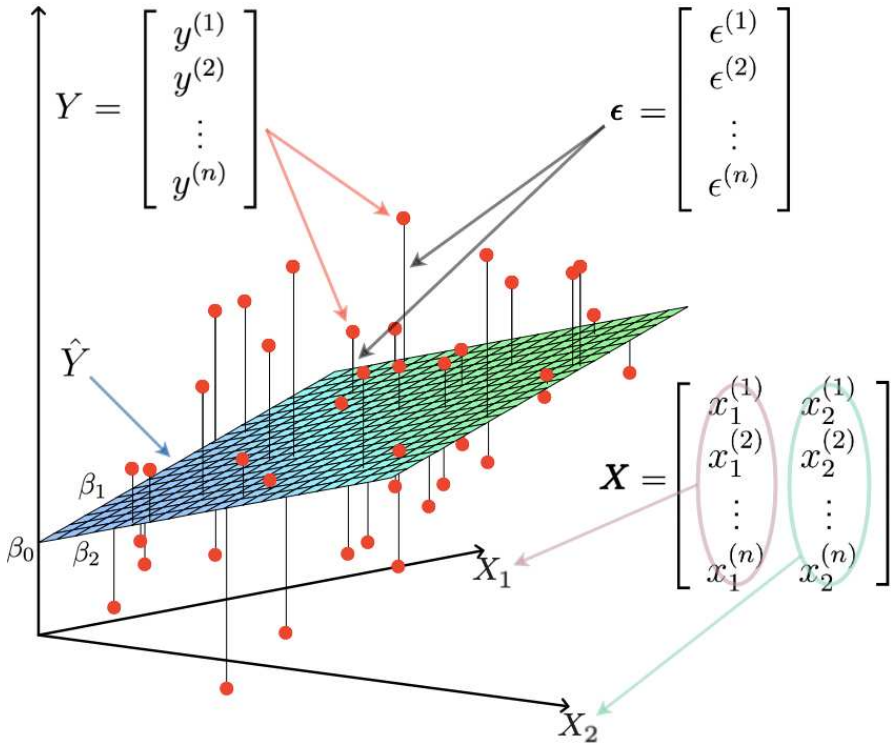


Figure 3.7 – Illustration d'une régression linéaire par moindres carrés avec $\boldsymbol{X} \in \mathbb{R}^{n \times 2}$. $\hat{Y}$ dessine un hyperplan d'ordonnée à l'origine β_0, et de pentes β_1 et β_2 dans les directions respectives X_1 et X_2. C'est la fonction linéaire de $\boldsymbol{X}$ qui minimise la somme des résidus de $\boldsymbol{\epsilon}$ au carré. Figure adaptée de Hastie et al. [168].

logit » $\sigma^{-1} : [0\,;1] \to \mathbb{R}$, $\sigma^{-1}(P) = \ln(P/(1-P))$. Pour un patient i, la prédiction d'un modèle de régression logistique est donc de la forme :

$$\hat{y}^{(i)} = \sigma(\beta_0 + \sum_{j=1}^{p}(x_j^{(i)} \times \beta_j)). \tag{3.13}$$

La fonction de perte associée à la régression logistique est basée sur l'entropie croisée (*cross-entropy* ou *log loss*). Notée L, elle est minimisée dans un problème de classification binaire tel que :

$$\{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\} = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min}\ L(\beta_0, \boldsymbol{\beta}) \tag{3.14}$$

avec

$$L(\beta_0, \boldsymbol{\beta}) = -\sum_{i=1}^{n}(y^{(i)} \times \ln(\hat{y}^{(i)}) + (1 - y^{(i)}) \times \ln(1 - \hat{y}^{(i)})). \tag{3.15}$$

Comme pour la régression linéaire, plusieurs méthodes d'optimisation existent, parmi lesquelles nous pouvons citer la descente de coordonnées et la descente de gradient (Section 3.2.2.5) [170, 171].

Du point de vue de la prédiction, P dessine une « surface de décision » dans l'espace des caractéristiques défini par $\boldsymbol{X}$. Le choix d'un seuil[4] s de probabilité donne une « frontière de décision », correspondant à un hyperplan divisant cet espace en deux parties pour donner $\hat{Y}_s$ dans un contexte de classification binaire stricte. Un exemple en 2D est présenté en Figure 3.8.

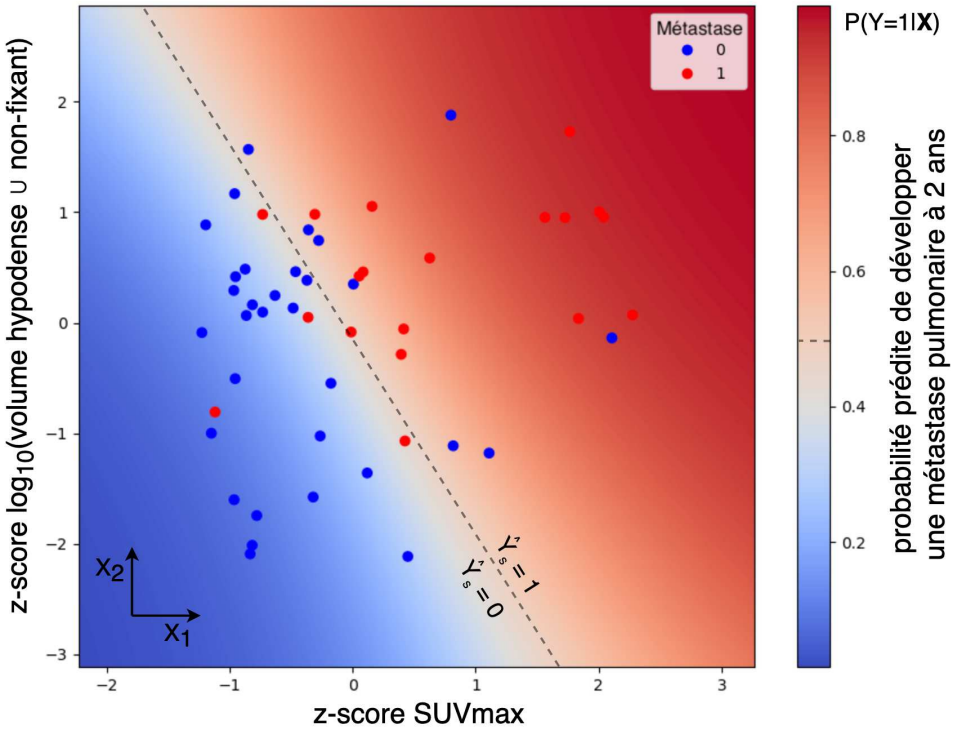


Figure 3.8 – Surface de décision d'un modèle logistique construit à partir de deux caractéristiques (X_1 et X_2) pour prédire la survenue de métastases pulmonaires dans les deux ans suivant le bilan d'extension pour des patients atteints de STS à partir d'images TEP/TDM. Un seuil de probabilité à $0,5$ établit une frontière de décision permettant de classifier automatiquement les patients en deux niveaux de risque. Figure adaptée d'Escobar et al. [174].

3.2.2.3 . La machine à vecteurs de support

Comme la régression logistique, l'objectif d'une SVM est de prédire une classification binaire [168]. Sans entrer dans les détails d'optimisation, elle identifie un hyperplan $X^{\intercal}\boldsymbol{\beta} + \beta_0 = 0$ constituant la frontière de décision

4. Le choix du seuil dépend des objectifs de l'étude [172, 173]. Un seuil bas favorise la sensibilité de détection du modèle, mais entraîne une augmentation du nombre de faux positifs. À l'inverse, un seuil élevé donne un modèle spécifique, au prix d'une augmentation du nombre de faux négatifs. Si la sensibilité et la spécificité du modèle sont aussi importantes, un seuil de 0,5 convient généralement.

entre deux classes, -1 et $+1$, telles que pour $\hat{Y} = \boldsymbol{X}\boldsymbol{\beta} + \beta_0$:

$$\{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\} = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \ (C \times H(\beta_0, \boldsymbol{\beta}) + \frac{1}{2} \times \|\boldsymbol{\beta}\|^2) \tag{3.16}$$

avec

$$H(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \max(0, 1 - \hat{y}^{(i)} \times y^{(i)}). \tag{3.17}$$

H est la « perte de charnière » (« *hinge loss* »). Elle est nulle si $\hat{y}^{(i)}$ et $y^{(i)}$ sont de même signe, et que $|\hat{y}^{(i)}| \geq 1$. Dans tous les autres cas, elle est proportionnelle à $|\hat{y}^{(i)}|$. La valeur C est définie par l'opérateur : c'est un hyperparamètre (Section 3.2.4). Elle contrôle la minimisation de H, par rapport à celle des coefficients de $\boldsymbol{\beta}$ pour maximiser la marge de séparation entre les deux classes (Figure 3.9).

La SVM et la régression logistique sont des modèles linéaires de classification étroitement liés. Elles séparent l'espace des caractéristiques défini par $\boldsymbol{X}$ à l'aide d'un hyperplan. Bien que ces deux méthodes présentent de nombreuses similitudes, nous pouvons citer les différences suivantes [175] :

- Alors que la formulation de la régression logistique est motivée par une modélisation probabiliste, celle de la SVM est une approche géométrique [176].

- Contrairement à la SVM, la régression logistique donne des probabilités calibrées qui peuvent être interprétées comme la confiance dans une décision.

- La SVM ne pénalise pas les exemples pour lesquels une prédiction est fournie du bon côté de l'hyperplan avec suffisamment de distance. Cela peut éviter l'impact des valeurs aberrantes (*outliers*) et constituer un avantage pour la classification binaire stricte de nouvelles données.

- La régression logistique donne une fonction de sortie continue et interprétable (Section 3.3).

- La régression logistique peut simplement être utilisée dans une modélisation bayésienne [177].

- Se basant essentiellement sur des points clés de l'espace des caractéristiques (les vecteurs de support), les SVMs admettent des solutions creuses (*sparse*). Pouvant s'exprimer selon une forme duale, elles sont plus adaptées à l'introduction de non-linéarités, via la « technique du noyau » (« *kernel trick* ») (Section 3.2.2.4) [178].

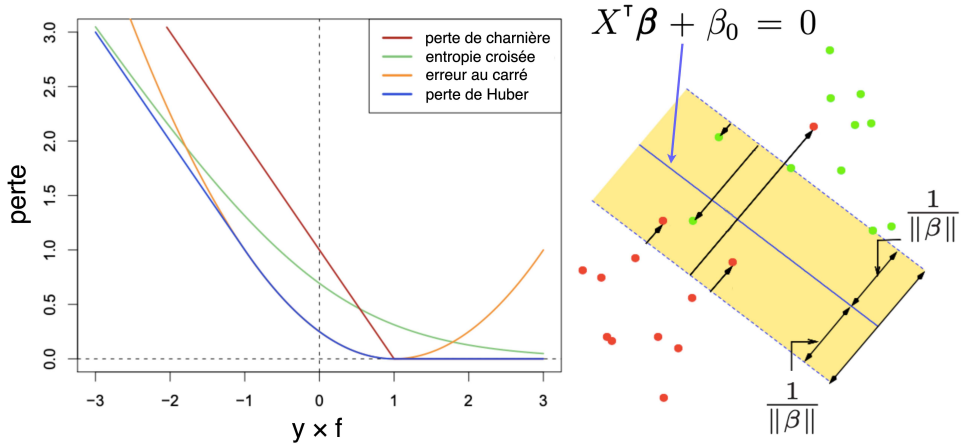


Figure 3.9 – Comparaison de fonctions de perte usuelles et exemple de la frontière de décision d'un modèle à SVM construit à partir de deux caractéristiques. Un seuil à 0 permet de classifier automatiquement les individus. Figure créée et adaptée à partir de Hastie et al. [168].

3.2.2.4 . Introduction à l'apprentissage profond par la technique du noyau

Parfois, le tracé d'un hyperplan n'est pas suffisant pour classer précisément les points. Les données dans leur forme originale ne sont alors pas linéairement séparables. En transformant les données pour créer un espace transitoire de dimension supérieure, il peut cependant devenir possible de trouver un hyperplan satisfaisant (Figure 3.10).

Cette modification de l'espace défini par $\boldsymbol{X}$ peut être réalisée grâce à la technique du noyau. Elle correspond au passage vers un espace de redescription implicite grâce à l'application d'une fonction $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$, avec $q > p$ le nombre de dimensions de cet espace. Implicite également, ϕ est la fonction noyau (*kernel function*). Il existe différents noyaux parmi lesquels nous pouvons citer le noyau polynomial, le noyau de la « fonction de base radiale », et le noyau sigmoïde [178]. La modification de l'espace se traduit par la création de nouvelles variables. Bien que ces variables ne soient pas définies directement à partir des données, elles sont fonctions des caractéristiques initiales. L'attribution de poids à ces variables lors de l'apprentissage constitue alors une forme tacite d'apprentissage de la représentation [131, 178]. Toutefois, l'utilisation de noyaux dans les modèles linéaires rend l'interprétation des résultats plus difficile.

3.2.2.5 . Le réseau de neurones artificiel

Basé sur le principe du perceptron, un ANN est constitué de plusieurs nœuds, les neurones, organisés en couches successives. La sortie de chaque couche donne l'entrée de la couche suivante (Figure 3.11). L'ANN est l'élément de base du DL (un CNN étant une forme particulière d'ANN). L'idée

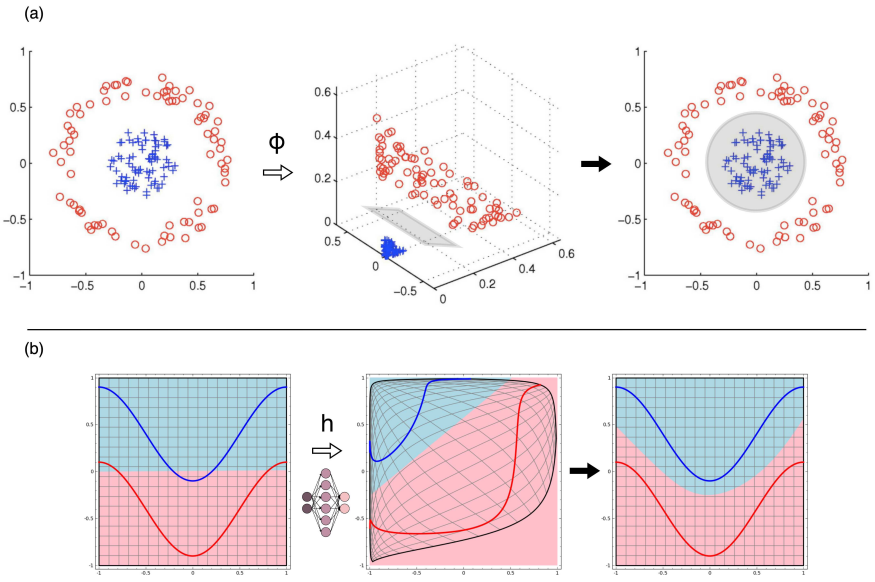


Figure 3.10 – (a) Transformation d'un espace 2D vers un espace 3D via une fonction $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ telle que $\phi(x_1, x_2) = \{z_1, z_2, z_3\} = \{x_1^2, \sqrt{2} \times x_1 \times x_2, x_2^2\}$. Un hyperplan linéaire $Z^\intercal \boldsymbol{\beta_z} + \beta_{0z} = 0$ permet de séparer parfaitement les points en deux classes. Reporté dans l'espace initial, il est défini en fonction de x_1 et x_2 selon une ellipse de la forme $x_1^2 \times \beta_{1z} + \sqrt{2} \times x_1 \times x_2 \times \beta_{2z} + x_2^2 \times \beta_{3z} = 0$. (b) Illustration de la séparation d'un espace caractérisé par deux courbes à l'aide d'un ANN. Dans cet exemple, l'espace 2D initial n'est pas linéairement séparable. Il est transformé en un espace transitoire de dimension supérieure, avant d'être projeté dans un nouvel espace 2D linéairement séparable. La séparatrice apprise dans ce nouvel espace peut alors être reportée dans l'espace initial. Notée h, la transformation peut être vue comme analogue à ϕ. À la différence qu'elle est elle-même apprise par le réseau de neurones, grâce à une ou plusieurs couches intermédiaires. Figure créée et adaptée à partir de Jordan et al. [179] et Olah [180].

centrale des ANNs est d'extraire des combinaisons linéaires des entrées en tant que caractéristiques dérivées, puis de modéliser Y en tant que fonction non-linéaire de ces caractéristiques [181, 182]. Similaire à la régression logistique, chaque neurone prend des valeurs en entrée, les multiplie par un poids, y ajoute un biais, et applique une fonction d'activation pour donner une sortie telle que :

$$a_k^{[l]} = h^{[l]}(\beta_{0_k}^{[l]} + \sum_{j=1}^{p_k}(a_j^{[l-1]} \times \beta_{j_k}^{[l]})) \tag{3.18}$$

avec $a_k^{[l]}$ la sortie du k-ième neurone de la l-ième couche, et $h^{[l]}$ la fonction d'activation de l'ensemble de cette couche. Dans un MLP composé de couches entièrement connectées, le nombre de caractéristiques en entrée, p_k, est égal au nombre de neurones de la couche précédente $(l-1)$.

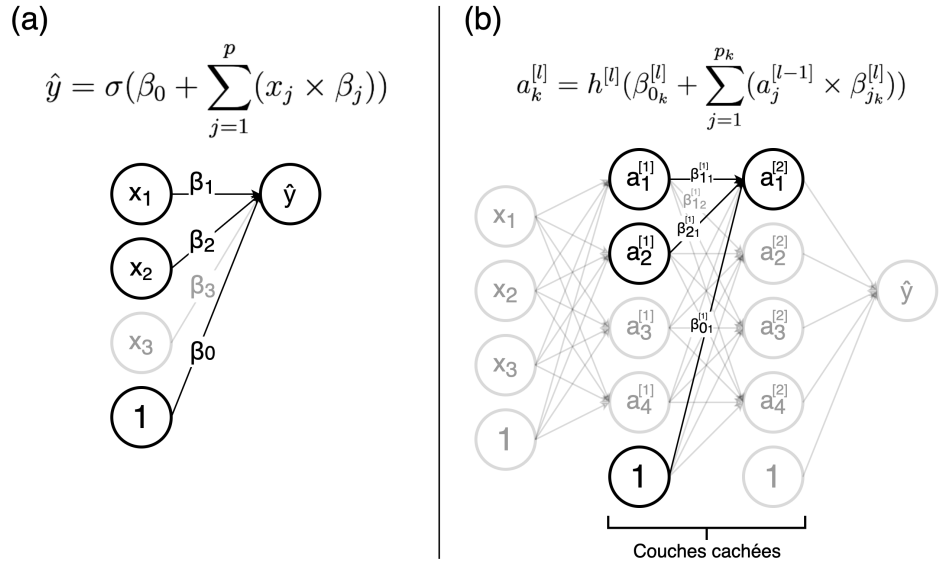


Figure 3.11 – (a) Neurone unique correspondant à une régression logistique exprimée comme un ANN selon le formalisme du perceptron. (b) MLP simple possédant deux couches cachées, contenant quatre neurones chacune.

Les couches intermédiaires qui ne sont pas constituées des nœuds d'entrée ou de sortie sont appelées « couches cachées » (« *hidden layers* »). Comme la technique du noyau, elles permettent de modifier l'espace des caractéristiques de sorte à fournir un espace de redescription linéairement séparable en sortie de l'ANN (Figure 3.10). La Figure 3.12 montre le résultat d'un ANN possédant une couche cachée contenant quatre neurones, entraîné via l'application web Google *A Neural Network Playgroung* [183]. À partir de combinaisons linéaires de deux caractéristiques d'entrée (X_1 et X_2), la couche cachée de l'ANN lui permet de s'ajuster parfaitement aux données, en produisant une surface de décision s'apparentant à une ellipse.

Pour chaque neurone k de l'ANN, l'ensemble des paramètres est constitué du biais et des poids qui doivent être optimisés. Notés $\theta_k = \{\beta_{0_k}, \boldsymbol{\beta}_k\}$ dans l'équation (3.19), ces paramètres sont initialisés de manière aléatoire, puis ajustés itérativement pendant l'apprentissage à l'aide d'algorithmes itératifs d'optimisation basés sur la descente de gradient, tels que la descente de gradient stochastique (*stochastic gradient descent* (SGD)), ou à l'aide de méthodes plus avancées comme l'optimiseur Adam [171]. Une itération t est composée de trois étapes principales :

1 - $\boldsymbol{X}$ est donné en entrée du réseau par lots (*batches*), et l'erreur entre $\hat{Y}^{(batch)}$ et $Y^{(batch)}$ est calculée selon la fonction de perte L à minimiser au cours de l'entraînement. C'est l'étape de « *feedforward* ».

2 - Selon la règle de la dérivation en chaîne, la participation individuelle de chaque neurone k à l'erreur de prédiction est calculée. C'est l'étape de « rétropropagation du gradient ». Elle permet de calculer la dérivée partielle de l'erreur par rapport à tout θ_k.

3 - θ_k est alors mis à jour tel que :

$$\theta_k^{t+1} = \theta_k^t - \eta_k^t \times \frac{\partial L}{\partial \theta_k}(\theta^t) \tag{3.19}$$

avec η_k^t le « taux d'apprentissage » (« *learning rate* »). Il correspond à un facteur de pondération déterminant l'amplitude de la mise à jour à chaque itération t. Il peut être maintenu constant, ou ajusté automatiquement au cours de l'entraînement [184]. Un taux d'apprentissage trop faible entraîne une convergence lente, tandis qu'un taux d'apprentissage trop élevé peut empêcher la convergence voire induire une divergence.

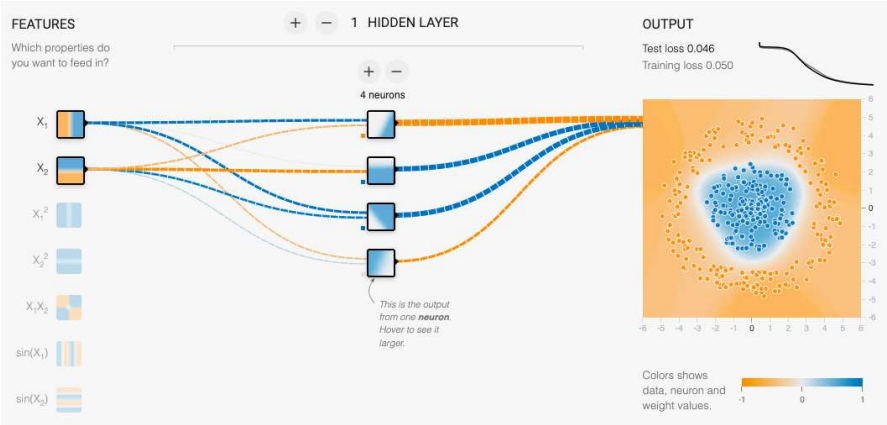


Figure 3.12 – Résultats de l'entraînement d'un MLP possédant une couche cachée contenant quatre neurones, via l'application web Google *A Neural Network Playgroung* [183]. Le MLP parvient à dessiner une surface de décision s'apparentant à une ellipse séparant les points en deux classes à partir de combinaisons linéaires de deux caractéristiques d'entrée (X_1 et X_2).

Au même titre que la fonction de perte [185], que la taille du lot pour une itération (*batch size*) [186], que le nombre de couches et de neurones par couche [187], ou que la fonction d'activation de chaque couche [188-190], le taux d'apprentissage est un hyperparamètre (Section 3.2.4).

Les ANNs sont très versatiles. En 1989, Cybenko et al. suggèrent que toute fonction f définie sur $\mathbb{R}^p$ peut être approchée à l'aide d'un MLP possédant une couche cachée à activation sigmoïde, si elle contient un nombre suffisant de neurones [191]. Ce résultat est affiné en 1991 par Hornik et al.,

montrant que c'est l'organisation en couches qui confère cette propriété à l'ANN, plus que le choix de ses fonctions d'activation [192]. Donnant toute sa puissance à l'ANN, cette flexibilité présente cependant certains inconvénients :

- De très nombreux poids doivent être ajustés, et ce sur un ensemble d'entraînement fini et généralement limité. Cela rend l'ANN plus sujet au « surapprentissage » par rapport à des algorithmes plus simples, tels que la régression logistique ou la SVM linéaire[5]. Cela signifie que par rapport à ces algorithmes, il a plus tendance à apprendre du bruit, des particularités de l'ensemble sur lequel il est entraîné, qui ne sont pas généralisables à des données nouvelles.

- De multiples combinaisons existent pour le choix des différents hyperparamètres d'optimisation. Ces derniers affectant la convergence du modèle [196-198], trouver la combinaison adéquate peut alors s'avérer difficile.

- Comparativement à la régression logistique, la complexité des ANN entraîne la perte de la capacité à fournir des probabilités calibrées. Il est alors plus difficile d'interpréter sa sortie comme une confiance dans sa décision [199].

- Enfin, un ANN ne fournit pas une sortie interprétable en fonction des caractéristiques d'entrée. Comme pour la technique du noyau, l'espace de description d'un MLP est le plus souvent vu comme une « boite noire ». Cela signifie qu'il est difficile de comprendre ou d'expliquer pourquoi il a produit un certain résultat. Discuté en Section 3.3 et tout au long de cette thèse, ce point est important car, dans certains domaines comme la médecine, l'interprétabilité est cruciale [200].

3.2.2.6 . La forêt aléatoire

Les approches d'apprentissage présentées précédemment traitent la non-linéarité en utilisant une stratégie commune : elles modifient transitoirement l'espace d'origine défini par $\boldsymbol{X}$, pour le rendre partitionnable par les hyperplans linéaires qu'elles utilisent finalement. C'est donc la relation entre l'espace original et l'espace de redescription qui permet l'ajustement non-linéaire du modèle.

5. Vrai pour les ANNs mais également pour tout autre algorithme d'apprentissage, plus une méthode est flexible, plus elle fournit des modèles complexes capables de s'ajuster précisément aux données, avec l'inconvénient d'avoir un risque plus élevé de s'ajuster au bruit, et donc d'induire une situation de surapprentissage . C'est le « compromis biais-variance » [168, 193-195].

Les algorithmes basés sur les arbres de décision modélisent quant à eux les données de façon non-linéaire directement dans l'espace d'origine, en utilisant des règles de dichotomie [168, 201, 202]. Les données d'entraînement sont divisées selon les différentes caractéristiques de $\boldsymbol{X}$ de manière séquentielle. L'algorithme mesure la pureté des sous-groupes formés pour déterminer la meilleure division à chaque étape. Plusieurs métriques peuvent être utilisées, incluant « l'impureté de Gini » et l'entropie croisée. Le processus de division est répété jusqu'à la création de sous-ensembles ne contenant qu'une seule classe, ou lorsque certains critères sont satisfaits. Cela peut être une limite sur le nombre de divisions par exemple, ou lorsqu'un embranchement est constitué d'un sous-ensemble ne contenant plus qu'un certain nombre ou qu'une certaine proportion d'individus. En appliquant les règles de division apprises, l'arbre résultant peut alors être utilisé pour faire des prédictions sur de nouvelles données. Grâce à la nature explicite des règles qui le composent, un arbre de décision est facilement interprétable.

Il est cependant enclin au surapprentissage. Les arbres de décision peuvent en effet être profonds, avec de nombreuses branches, et être sensibles à de petits changements correspondant au bruit dans les données. Dans ce cas, ils parviennent à classifier les données d'entraînement, mais sont inefficaces sur de nouvelles données. L'objectif d'une forêt aléatoire est de pallier ce problème en associant de nombreux arbres, entraînés sur différents échantillons provenant des données d'entraînement. Les données sont tirées de façon aléatoire avec replacement. C'est le « *bootstrap* ». La moyenne des prédictions des différents arbres constitue alors $\hat{Y}$, la sortie de la forêt aléatoire. Cette stratégie d'échantillonnage et d'assemblage est appelée « *bagging* » (« *bootstrap aggregating* »)[6]. De cette façon, les erreurs individuelles commises par les arbres s'atténuent mutuellement, de sorte qu'en tant que groupe, ils améliorent la stabilité de la modélisation. Elle est également appliquée aux caractéristiques, afin d'augmenter la variabilité d'information capturée par les différents arbres, et améliorer davantage la robustesse de la forêt aléatoire.

La multiplication du nombre d'arbres composant une forêt aléatoire complexifie cependant son interprétation en tant que modèle. De plus, il est difficile d'estimer la contribution individuelle de chaque arbre aux prédictions globales. Par conséquent, bien que la forêt aléatoire puisse être efficace pour la prédiction, elle ne constitue probablement pas le premier choix dans les applications où l'interprétabilité est requise.

6. Bien que la forêt aléatoire soit considérée comme un algorithme à part entière, les méthodes d'assemblage telles que le *bagging* ne concernent pas uniquement les arbres de décision. La robustesse apportée par ce type de stratégie peut être bénéfique à tout type d'algorithme d'apprentissage.

3.2.2.7 . L'algorithme des K plus proches voisins

Le principe de l'algorithme des k plus proches voisins (*k-nearest neighbors* (k-NN)) est d'identifier les k points dans l'ensemble d'apprentissage qui sont les plus proches d'un nouveau point, afin de produire une prédiction pour ce dernier [168, 203]. Dans le cas d'une classification, nous pouvons alors estimer la classe de ce nouveau point comme l'étiquette majoritaire parmi ses k voisins. Dans un cadre probabiliste, il est possible d'utiliser la fréquence relative de chaque classe parmi les k voisins comme un score pour chacune d'entre elles. Dans un k-NN avec $k = 4$ par exemple, si les voisins du nouveau point en comprennent trois appartenant à la classe 1 et un appartenant à la classe 0, les scores sont de $0,75$ pour la classe 1, et $0,25$ pour la classe 0. Enfin, dans un problème de régression, la valeur prédite pour le nouveau point peut correspondre à la valeur moyenne de ses voisins, de façon analogue à une moyenne mobile.

La valeur de k a un impact sur le modèle. Si elle trop faible, le modèle peut être trop sensible au bruit et aux valeurs aberrantes. D'autre part, si la valeur de k est trop élevée, le modèle peut ne pas être en mesure de capturer la structure sous-jacente des données.

Dans certains cas, il peut être utile de pondérer l'étiquette des voisins en fonction de leur distance par rapport au point à prédire. Généralement, des poids plus élevés sont affectés aux points les plus proches, et vice versa. Cette pondération donne une information supplémentaire sur l'espace des caractéristiques, mais n'est néanmoins pas systématiquement souhaitable. En effet, la position des k voisins peut être affectée par le bruit, ou bien certains d'entre eux peuvent s'avérer être des *outliers* [204-206].

L'un des principaux avantages du k-NN est sa simplicité. Parce qu'il se base sur une simple distance pour identifier les k voisins, il peut être facilement mis en œuvre et appliqué à un large éventail de situations. Non paramétrique, cette méthode ne fait aucune hypothèse sur la structure des données. Elle est alors adaptée aux ensembles de données présentant des structures complexes. Au même titre que l'arbre de décision ou la forêt aléatoire, le k-NN s'ajuste aux données de façon non-linéaire, directement dans l'espace original de $\boldsymbol{X}$.

L'interprétation d'un modèle de k-NN est différente de celle des modèles précédemment présentés. En effet, alors que ces derniers utilisent $\boldsymbol{X}$ pour formuler des règles pouvant être utilisées ensuite indépendamment des données, toute prédiction d'un modèle de k-NN est définie localement, de manière relative à certains points de l'ensemble d'entraînement. Le modèle peut être utilisé pour identifier des groupes d'individus. Cette approche basée sur la similarité peut être utile dans des contextes pour lesquels certains profils typiques d'individus sont facilement identifiables. Une façon d'évaluer localement l'importance d'une caractéristique peut alors être de mesurer l'ef-

fet de chaque caractéristique sur la distance entre le nouveau point et ses voisins. Par exemple, si une caractéristique a un effet significatif sur cette distance, elle peut être considérée comme importante. Parce qu'il ne modélise pas explicitement leur relation avec la variable cible, le modèle de k-NN ne renseigne toutefois pas sur l'impact de chacune des caractéristiques dans le processus de décision en général. Cela peut rendre difficile la compréhension du modèle, limitant alors l'utilité des k-NNs pour certaines applications.

3.2.2.8 . Les modèles de survie

Les modèles de survie constituent un ensemble de méthodes statistiques utilisées pour analyser et prédire le temps entre une situation initiale (en oncologie, généralement le diagnostic ou le bilan d'extension), et l'occurrence d'un événement d'intérêt. Ils sont particulièrement utiles dans les domaines tels que la médecine, pour laquelle le moment où les événements se produisent peut être d'une importance capitale. Comme pour les modèles de régression et de classification, il en existe plusieurs types [207]. Une approche courante de l'analyse de survie consiste à utiliser les estimateurs empiriques de Kaplan-Meier, représentant la probabilité de survenue d'un événement dans le temps pour différents groupes définis a priori [208]. D'autres techniques basées sur les algorithmes d'apprentissage présentés précédement incluent la régression de Cox ou « modèle de risque proportionnel de Cox » [209, 210], proche de la régression logistique [211, 212], la forêt de survie aléatoire, basée sur la forêt aléatoire [213], et l'estimateur de Kaplan-Meirer pondéré par k-NN [214].

3.2.3 . L'évaluation du modèle

Le ML ne consiste évidemment pas seulement à ajuster le modèle aux données d'apprentissage. Une fois que le modèle est entraîné, il est important d'évaluer sa performance, pour s'assurer qu'il est capable de « généraliser », en faisant des prédictions correctes sur de nouvelles données. Cette étape d'évaluation est cruciale, que ce soit dans un contexte de déploiement en vue de l'automatisation d'une tâche, ou bien lors d'une étude exploratoire. En effet, même dans un contexte de découverte guidée par les données, l'efficacité d'un modèle sur des données avec lesquelles il n'a pas été entraîné permet de savoir si le signal capturé est robuste et réellement informatif.

Comme nous l'avons vu dans la section précédente (3.2.2.5), si un modèle produit une bonne performance sur les données d'apprentissage, mais que cette dernière diminue fortement lors de son application à de nouvelles données, nous sommes dans une situation de suapprentissage. D'autre part, s'il n'est pas en mesure de capturer des tendances clés dans les données, nous sommes en situation de « sousapprentissage » : soit le modèle est mal ajusté, soit les données ne contiennent tout simplement pas d'information substantielle sur le processus qui nous intéresse.

Nous pouvons imaginer un processus caché parmi l'ensemble de ceux qui produisent les données. Il peut par exemple s'agir d'un processus qui induit une augmentation du risque de métastases, et qui se reflète (partiellement) dans les images. Nous ne connaissons pas tous les détails à propos de ce processus. Nous avons simplement un ensemble d'exemples concrets (les images), étiquetés en rapport avec le résultat de ce processus (par exemple le statut métastatique deux ans après le bilan d'extension). Admettons que les données d'entraînement portent de l'information à propos de ce processus. Parce que cette information est partagée par l'ensemble de la population étudiée, elle est généralisable à de nouvelles données provenant de cette même population mais n'ayant pas servi à l'entraînement. Cependant, les données d'entraînement contiennent aussi de l'information qui leur est propre et qui n'a pas de lien avec le processus d'intérêt. Cela peut être du bruit, ou un signal spécifique à l'échantillon d'entraînement non lié à ce que l'on cherche. L'objectif est alors d'apprendre du mieux possible l'information généralisable, tout en évitant au maximum d'ajuster le modèle au bruit.

3.2.3.1 . L'échantillonnage des données

Une stratégie empirique pour répondre à cet objectif consiste à conserver des données hors de l'ensemble d'entraînement, afin de les utiliser pour évaluer le modèle une fois entraîné. C'est l'ensemble de validation. Plus l'ensemble d'apprentissage est grand, meilleur est le modèle. Plus l'ensemble de validation est grand, plus l'évaluation est fiable.

Pour les ensembles de données de petite taille souvent rencontrés en médecine, une stratégie de validation croisée peut être utilisée. Elle consiste à diviser les données en k « plis » (« *k-fold* »), chaque pli constituant un ensemble de validation. Le modèle est alors entraîné et évalué k fois sur des ensembles d'entraînement et de validation légèrement différents. La performance globale du modèle est ainsi estimée via la moyenne estimations sur l'ensemble des plis, associée à son écart-type, son coefficient de variation, et à l'estimation d'un intervalle de confiance (*confidence interval* (CI)).

Il existe plusieurs types de validation croisée [215]. Illustrées en Figure 3.13, nous pouvons citer la « validation croisée classique à k plis » et la « validation croisée de Monte Carlo ». Le cas extrême correspond à un nombre de plis égal au nombre d'individus. Dans ce cas, on parle de « *leave-one-out cross-validation* », et un seul individu est utilisé à chaque pli pour évaluer le modèle. Facultative, la « stratification » garantit que la distribution des classes est maintenue dans chaque pli.

Une autre stratégie consiste à utiliser le *bootstrap*. Parce que chaque individu peut être tiré plusieurs fois du fait du replacement, certains ne sont pas tirés. L'ensemble des individus qui n'ont pas été tirés lors d'un échantillonnage *bootstrap* constitue l'ensemble *out of bag* (OOB), qui peut être utilisé pour la validation.

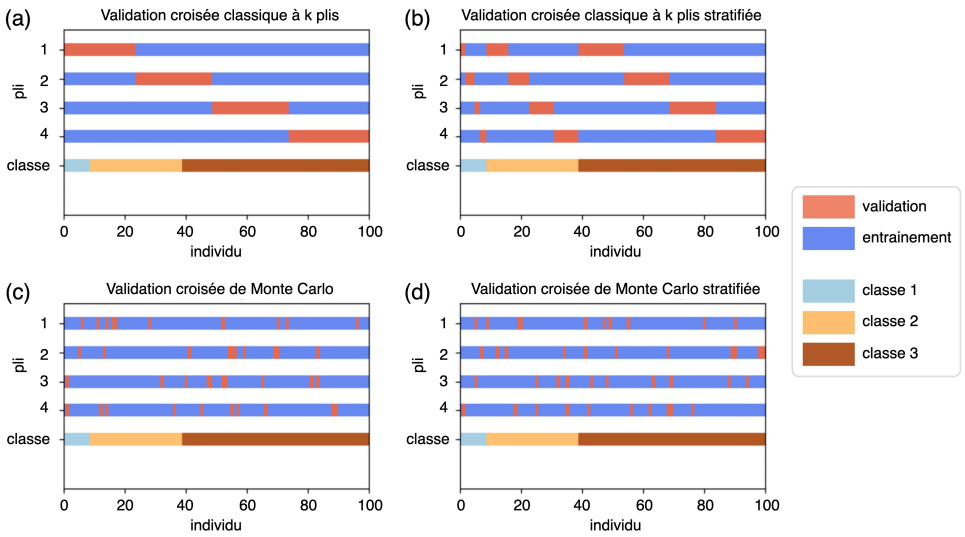


Figure 3.13 – (a, b) Validation croisée classique à k plis et variante stratifiée. (c, d) Validation croisée de Monte Carlo et variante stratifiée. Dans cet exemple de classification à trois classes, le nombre de plis est de $k = 4$. Figure créée et adaptée à partir de la documentation de la bibliothèque Python Scikit-Learn. [215-217]

Une seule exécution de la procédure de validation croisée peut tout de même donner lieu à une estimation bruitée, et différents fractionnements des données peuvent donner différents résultats. La validation croisée répétée permet d'améliorer l'estimation de la performance d'un modèle. Il suffit de répéter la procédure de validation croisée en modifiant l'ordre des individus à chaque fois [218].

3.2.3.2 . Les métriques d'évaluation

Bien que la fonction de perte minimisée durant l'entraînement puisse être utilisée lors de l'évaluation, plusieurs métriques existent pour évaluer un modèle d'apprentissage. Il est nécessaire de choisir des métriques adaptées aux objectifs de l'étude. Les métriques de classification les plus courantes sont présentées ci-après.

- **La matrice de confusion**
  Dans le cas d'une classification binaire stricte, les métriques sont généralement basées sur la « matrice de confusion », rassemblant le nombre de vrais positifs (*true positive* (TP)), faux positifs (*false positive* (FP)), vrais négatifs (*true negative* (TN)), et faux négatifs (*false negative* (FN)) (Tableau 3.1).

Tableau 3.1 – Matrice de confusion.

	prédiction positive	prédiction négative
vérité positive	TP	FN
vérité négative	FP	TN

- **L'*accuracy* (Acc)**
  C'est l'exactitude du modèle. Elle donne la proportion de prédictions correctes :
  $$Acc = \frac{TP + TN}{TP + FP + TN + FN}. \tag{3.20}$$

- **La sensibilité (Se)**
  C'est la capacité de détection du modèle. Également appelée « taux de vrais positifs » ou « rappel », la sensibilité correspond à la proportion de cas positifs correctement classifiés tels que :
  $$Se = \frac{TP}{TP + FN}. \tag{3.21}$$

- **La spécificité (Sp)**
  Il s'agit de l'aptitude du classifieur à identifier correctement les individus négatifs :
  $$Sp = \frac{TN}{FP + TN}. \tag{3.22}$$
  La capacité de détection du modèle est cruciale, mais il est tout aussi important de s'assurer que seuls les individus réellement positifs sont classés dans la catégorie correspondante, de façon « spécifique ». La spécificité est aussi appelée « taux de vrais négatifs ».

- **La *balanced accuracy* (Bacc)**
  L'*accuracy* ne tient pas compte de l'équilibre des classes dans l'ensemble de données. Elle peut alors s'avérer trompeuse en cas de fort déséquilibre. Par exemple, si un modèle prédit systématiquement la classe positive, dans un ensemble de données contenant 90% d'exemples positifs et 10% d'exemples négatifs, il sera considéré comme exact à 90% ($Acc = 0,9$) même s'il ne capture aucune information utile dans les données. C'est pourquoi d'autres métriques sont souvent utilisées, comme la *balanced accuracy*. Elle correspond à la moyenne de l'exactitude du modèle calculée pour chacune des classes, c'est à dire à la moyenne de la sensibilité et de la spécificité :
  $$Bacc = \frac{Se + Sp}{2}. \tag{3.23}$$

Dans l'exemple présenté plus haut, $Se = 1$, $Sp = 0$, et $Bacc = 0,5$, ce qui montre que malgré une exactitude élevée due à la proportion de chaque classe, le modèle est aléatoire lorsqu'il s'agit de discriminer deux individus appartenant à des classes différentes.

- **La statistique J de Youden (J)**
  Équivalente à la *balanced accuracy*, la statistique J de Youden s'exprime telle que :
  $$J = Se + Sp - 1. \tag{3.24}$$

- **La courbe ROC**
  Les métriques précédentes nécessitent une prédiction binaire impliquant le plus souvent le choix d'un seuil sur la sortie du modèle. Si nous abaissons le seuil pour qu'un résultat soit classé comme positif, nous augmentons la sensibilité (car nous avons plus de chances d'identifier les cas positifs) mais nous diminuons la spécificité (car nous avons désormais plus de faux positifs). La courbe ROC (*receiver operating characteristic*) est une représentation graphique qui montre le compromis entre la sensibilité et la spécificité pour tous les seuils de classification faisant changer leur valeur. Elle consiste à représenter Se en fonction de $1 - Sp$.

- **L'aire sous la courbe ROC (AUC ou AUROC)**
  L'aire sous la courbe ROC (*area under the ROC curve* (AUC ou AUROC)) est une mesure de la performance indépendante du seuil. Elle peut être interprétée comme sa performance de rang. Soit deux individus, A appartenant à la classe positive et B à la classe négative. Pour tous individus A et B, l'AUC représente la probabilité que la probabilité prédite pour A soit supérieure à celle prédite pour B. Une AUC de 0,5 représente une classification aléatoire, tandis qu'une AUC de 1 indique une classification parfaite.

- **L'indice de concordance (c-index)**
  De façon semblable à l'AUC, l'indice de concordance (*concordance index* (c-index)) est utilisée le plus souvent dans les modèles de survie. Il mesure la probabilité que la prédiction du temps d'occurrence pour l'événement d'intérêt soit dans le bon ordre pour un ensemble de paires de patients, A et B, de l'ensemble de données.

- **Le score de Brier (B)**
  Dans un contexte probabiliste, il est possible d'évaluer quantitativement la prédiction pour chaque individu, en comparant la probabilité prédite au véritable *label*. Le score de Brier mesure l'erreur quadratique moyenne dans l'ensemble de données. Analogue à l'*accuracy*, il

est utilisé pour évaluer l'exactitude des prédictions probabilistes avec :

$$B = \frac{1}{n} \times \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2. \tag{3.25}$$

- **Le score de Brier stratifié moyen (ASB)**
  Comme l'*accuracy*, le score de Brier peut s'avérer trompeur en cas de fort déséquilibre de classes. C'est dans ce contexte que Wallace et al. ont introduit la notion de « score de Brier stratifié » (« *stratified Brier score* ») [219]. Il décompose le score de Brier pour chacune des classes, de façon analogue à Se et Sp. Le score de Brier stratifié moyen (*average stratified Brier score*), ASB, peut alors être défini par :

$$ASB = 1 - \frac{SB_1 + SB_2}{2} \tag{3.26}$$

avec

$$SB_1 = \frac{1}{n_1} \times \sum_{i=1}^{n} ((y^{(i)} - \hat{y}^{(i)})^2 \times [y^{(i)} = 1]) \tag{3.27}$$

et

$$SB_0 = \frac{1}{n_0} \times \sum_{i=1}^{n} ((y^{(i)} - \hat{y}^{(i)})^2 \times [y^{(i)} = 0]). \tag{3.28}$$

avec n_1 et n_0 les nombres de patients appartenant respectivement aux classes 1 et 0. SB_1 et SB_0 correspondent au score de Brier stratifié associé. $[...]$ est une parenthèse d'Iverson, qui vaut 1 lorsque la condition entre les parenthèses est vraie et 0 sinon [7]. Contrairement à B que l'on cherche à minimiser, ASB doit être maximisé. Un modèle parfait donne un score de 1 alors qu'un modèle qui prédit toujours la mauvaise classe donne un score de 0. Un modèle totalement sous-ajusté qui prédit toujours une probabilité de 0,5 donne un score de 0,75, et un modèle prédisant systématiquement la classe majoritaire donne un score de 0,5, comme la *balanced accuracy* [8].

7. La formulation de ASB avec SB_1 et SB_2 est équivalente au calcul de $1-B$ avec une pondération pour chaque individu, inversement proportionnelle à la fréquence de sa classe dans l'ensemble de données. Cette stratégie peut également être appliquée à toute fonction de perte lors de l'entraînement du modèle. L'entropie croisée équilibrée (*balanced cross-entropy*), par exemple, est une fonction de perte couramment utilisée lors de l'ajustement d'un modèle logistique [220].

8. En cas de classification stricte, un modèle prédisant toujours la classe majoritaire est tout autant inutile qu'un modèle prédisant systématiquement une probabilité égale à 0,5. Par contre, ce n'est pas le cas dans un cadre probabiliste dans lequel la probabilité prédite par le modèle est calibrée. En effet, si nous interprétons la probabilité comme la confiance que le modèle accorde dans sa décision, il est préférable d'obtenir des prédictions totalement incertaines, que des prédictions fausses avec un taux élevé de certitude. Par analogie avec notre prise de décision, il vaut mieux préserver le doute que faire une erreur associée à un excès de confiance.

3.2.4 . La sélection des hyperparamètres et des caractéristiques

Comme nous l'avons vu avec les différents algorithmes d'apprentissage (Section 3.2.2), les hyperparamètres correspondent à l'ensemble des paramètres qui doivent être ajustés par l'opérateur, afin d'obtenir le modèle le plus efficace. Il peut s'agir de la méthode d'optimisation, plus spécifiquement de l'architecture d'un ANN, du nombre d'arbres d'une forêt aléatoire, ou de tout autre paramètre n'étant pas directement appris lors de l'entraînement. Leur ajustement consiste à trouver la combinaison optimale pour un problème donné.

Il existe différentes approches pour le réglage des hyperparamètres. Il est possible de les définir a priori selon certains objectifs ou du fait de contraintes s'appliquant à l'étude. Il existe également des méthodes automatiques. Nous pouvons citer la grille de recherche (*grid search*), la recherche aléatoire (*random search*), et l'optimisation par modélisation bayésienne (*bayesian model-based optimization*) [196-198, 221].

3.2.4.1 . La fuite de données

Le choix de la méthode d'ajustement des hyperparamètres dépend de la situation. Il convient néanmoins d'évaluer l'impact des hyperparamètres sur le modèle. Dans ce contexte, nous pouvons utiliser la validation croisée pour sélectionner la combinaison qui donne la meilleure performance. Le partitionnement en deux ensembles (entraînement et validation) peut s'avérer insuffisant dans ce cas. Bien qu'il ait été montré que cela n'est la plupart du temps pas critique lors de l'utilisation de modèles simples [222], l'évaluation de la performance peut conduire à un biais optimiste lorsqu'elle est effectuée simultanément avec l'optimisation des hyperparamètres. En effet, en ajustant les hyperparamètres en fonction de la performance évaluée sur l'ensemble de validation, il s'opère une forme d'apprentissage, et il est possible que la combinaison choisie soit surajustée aux données de validation. Le cas échéant, nous sommes dans une situation de surapprentissage [223]. On parle alors de « fuite de données » [224]. Pour éviter ce problème, un troisième sous-ensemble de données peut être mis de côté : l'ensemble de test. Après l'apprentissage des paramètres lors de l'entraînement, et l'ajustement des hyperparamètres lors de la validation, c'est sur cet ensemble que l'évaluation du modèle sera finalement réalisée. Une stratégie de validation croisée imbriquée peut alors être adoptée (Figure 3.14). Elle permet d'obtenir une estimation non biaisée de la performance du modèle, mais aussi d'évaluer la stabilité des hyperparamètres en fonction de l'échantillonnage des données.

Toutefois, le nombre de patients disponibles est souvent insuffisant dans les études radiomiques pour réaliser une approche imbriquée. En effet, l'augmentation des subdivisions lors de l'échantillonnage implique une diminution du nombre d'individus pour l'entraînement (réduisant les capacités d'appren-

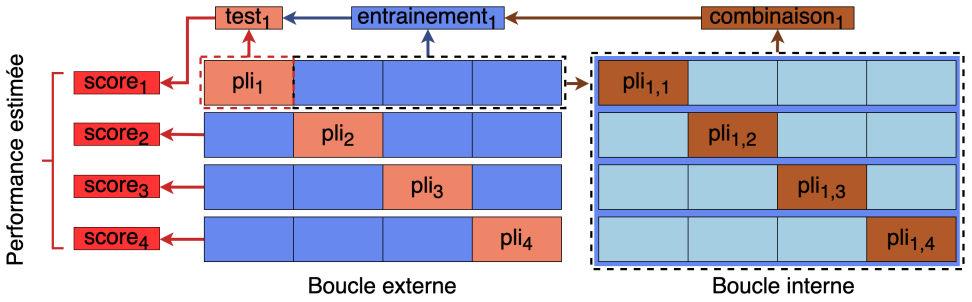


Figure 3.14 – Validation croisée imbriquée à 4×4 plis. Chaque pli de la boucle externe définit un ensemble de test. Il définit également un ensemble avec lequel une combinaison d'hyperparamètres est estimée via une validation croisée. C'est la boucle interne. Pour chaque pli de la boucle externe, la combinaison estimée dans la boucle interne est utilisée, le modèle est entraîné, puis évalué sur l'ensemble de test. La performance estimée du modèle correspond alors à la moyenne sur l'ensemble des plis de la boucle externe.

tissage), la validation, et le test (réduisant la fiabilité des estimations). De toute façon, même si l'estimation des paramètres du modèle, de ses hyperparamètres, et de sa performance n'est pas biaisée méthodologiquement, elle a un niveau de fiabilité très faible dans le cas d'un petit nombre d'individus, en raison du biais d'échantillonnage intrinsèquement causé par la petite taille de l'échantillon. Dans ce cas, le modèle n'est probablement pas exportable pour faire des prédictions sur de nouvelles données. Si l'objectif est d'obtenir des informations à partir des données plutôt que de déployer un modèle prédictif avec une estimation précise de la performance, un test de permutations peut être utilisé. Il permet de s'assurer que le modèle ne s'ajuste pas au bruit uniquement, et que l'information capturée permet de prédire la classe des patients dans cet ensemble de données avec une performance significativement différente du hasard [225]. L'ensemble de la chaîne d'analyse est répétée t fois en effectuant des permutations aléatoires des *labels* à chaque fois. Pour chaque itération, le score estimé par validation croisée est sauvegardé, ce qui conduit à une distribution nulle des t meilleurs scores. Illustrée en Figure 3.15, cette distribution montre la performance estimée lorsqu'il n'y a pas de relation réelle entre les caractéristiques et les *labels*. Sur la base de cette distribution nulle, la p-*value* empirique associée au score observé pour le modèle obtenu avec les *labels* corrects peut être calculée par :

$$p\text{-}value = \frac{C+1}{t+1} \tag{3.29}$$

avec C le nombre d'itérations pour lesquelles le score est supérieur à celui estimé avec les vrais *labels*. La meilleure valeur p possible est $1/(t+1)$. La pire est 1.

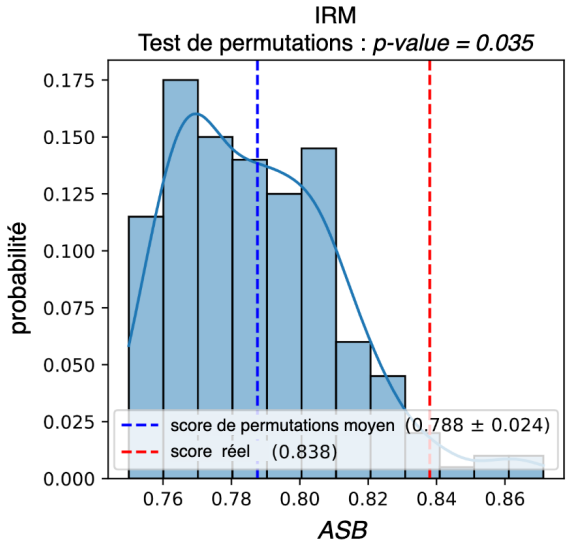


Figure 3.15 – Représentation graphique de la distribution nulle d'un test de permutations effectué dans le cadre de la prédiction du risque de métastases pulmonaires deux ans après le bilan d'extension en STS à partir d'images IRM. Le score utilisé ici est l'ASB. Figure adaptée d'Escobar et al. [61].

3.2.4.2 . La sélection des caractéristiques

En identifiant et en sélectionnant les caractéristiques les plus pertinentes, elle constitue une étape clé de la chaîne d'analyse. Elle réduit la complexité du modèle et permet d'améliorer sa performance, son interprétabilité, ainsi que sa capacité de généralisation. Il existe de nombreuses méthodes, généralement classées en trois catégories : les méthodes de filtrage (*filter methods*), les méthodes d'enveloppement (*wrapper methods*), et les méthodes intégrées (*embedded methods*) [193, 218, 226-231]. Quelques exemples sont présentés ci-après.

Les méthodes de filtrage fonctionnent indépendamment de l'entraînement du modèle. De façon supervisée ou non, elles sélectionnent les caractéristiques en fonction de mesures statistiques.

- **La sélection univariée**
  La sélection univariée des caractéristiques en fonction de leur corrélation avec la cible Y est un exemple d'approche supervisée.

- **La suppression des caractéristiques redondantes**
  Également basée sur la mesure de corrélations, une approche non supervisée peut consister à minimiser la redondance dans l'ensemble des caractéristiques de $\boldsymbol{X}$. En effet, de nombreuses caractéristiques radio-

miques peuvent être fortement corrélées entre-elles voire colinéaires [9]. La colinéarité est une association linéaire entre deux caractéristiques, tandis que la multicollinéarité désigne une situation dans laquelle plus de deux caractéristiques sont linéairement liées. La colinéarité et la multicollinéarité peuvent nuire à la stabilité du modèle [218]. Une solution couramment employée consiste à calculer la corrélation deux à deux pour toutes les paires de caractéristiques, et ne préserver qu'une caractéristique pour chaque paire corrélée au-delà d'un certain seuil. Une autre façon de procéder consiste à calculer le « facteur d'inflation de la variance » (*variance inflation factor* (VIF)) pour chaque caractéristique. Il estime l'augmentation de la variance d'un coefficient de régression causée par la multicollinéarité. Contrairement aux corrélations par paires, il quantifie dans quelle mesure chaque caractéristique introduit une redondance dans les données tout en considérant toutes les caractéristiques ensemble. Pour chaque j-ième caractéristique de $\boldsymbol{X}$, il est calculé en effectuant une régression linéaire par rapport aux autres caractéristiques avec :

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.30}$$

avec R_j^2 le coefficient de détermination de la régression sur la j-ième caractéristique. Un VIF élevé signifie que la caractéristique j peut être approchée par la combinaison linéaire d'autres caractéristiques de $\boldsymbol{X}$. Les caractéristiques hautement redondantes peuvent alors être supprimées de façon itérative en identifiant celles qui ont le VIF le plus élevé, jusqu'à ce que le VIF maximum dans l'ensemble de données soit inférieur à un certain seuil [193].

Les méthodes enveloppantes consistent à construire un modèle à l'aide d'un sous-ensemble de caractéristiques et à évaluer sa performance. Le processus est répété pour différents sous-ensembles, et celui qui donne la meilleure performance est sélectionné.

9. Le choix d'utiliser une mesure de corrélation linéaire, comme la corrélation de Pearson, ou une mesure de corrélation de rang (non-linéaire), comme la corrélation de Spearman, dépend du contexte. Le concept de redondance dépend de la manière dont le modèle utilise les caractéristiques. Par exemple, dans le cas d'un modèle linéaire tel que la régression logistique, deux caractéristiques peuvent avoir une forte corrélation de Spearman sans pour autant être redondantes, notamment si elles ont une faible corrélation de Pearson et créent un nuage de points courbé caractérisé par une classe aux extrémités et une autre dans l'angle, de façon semblable aux SVMs à noyaux et aux ANNs.

- **La sélection séquentielle en avant**
  La sélection séquentielle en avant (*sequential forward selection* (SFS)) consiste à ajouter de façon itérative une caractéristique au modèle et à le réentraîner. À chaque étape, la caractéristique ajoutée est celle associée à la meilleure performance. Une fois que toutes les caractéristiques sont ajoutées au modèle, ou qu'un certain nombre est atteint, le sous-ensemble final est choisi.

- **La sélection séquentielle en arrière**
  La sélection séquentielle en arrière (*sequential backward selection* (SBS)) adopte une approche inverse. Le modèle est d'abord entraîné avec toutes les caractéristiques, puis elles sont itérativement supprimées en fonction de sa performance.

- **L'élimination récursive de caractéristiques**
  L'élimination récursive de caractéristiques (*recursive feature elimination* (RFE)) est semblable à la SBS, à la différence qu'au lieu de supprimer les caractéristiques en fonction de la performance du modèle, elle utilise le rang des poids (ou de l'importance) que les caractéristiques ont dans la prédiction.

Les méthodes intégrées font partie intégrante de l'entraînement du modèle. Elles consistent à ajuster le modèle de façon parcimonieuse, c'est-à-dire en le contraignant de sorte qu'il n'utilise qu'un sous-ensemble des caractéristiques de $\boldsymbol{X}$.

- **Le LASSO**
  Le LASSO (*least absolute shrinkage and selection operator*) sélectionne des caractéristiques en ajoutant un terme de « régularisation » à la fonction de perte minimisée durant l'entraînement d'un modèle linéaire généralisé, d'une SVM, ou d'un ANN. Ainsi, elle pénalise la somme des valeurs absolues des coefficients (régularisation « L1 »). La force de la régularisation est contrôlée par un hyperparamètre α tel que :

$$\{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\} = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min}\ L_{L1}(\beta_0, \boldsymbol{\beta}) \tag{3.31}$$

  avec

$$L_{L1}(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \alpha \times \sum_{j=0}^{p} |\beta_j| \tag{3.32}$$

  avec L la fonction de perte initiale[10]. En pénalisant la valeur des

10. Également noté λ dans de nombreuses formulations, la force de régularisation est exprimée par son inverse noté C dans la bibliothèque Python Scikit-Learn, de façon cohérente avec la formulation de la SVM (équation (3.16)).

coefficients, la régularisation contraint le modèle à conserver et à attribuer des poids élevés uniquement aux caractéristiques pertinentes, et à abaisser les autres vers zéro, les supprimant ainsi du modèle.

Le modèle final correspond à celui issu de l'entraînement sur toute la base de données une fois les hyperparamètres ajustés et les caractéristiques sélectionnées. Une stratégie d'assemblage peut également être adoptée à cette étape [232].

3.2.5 . La chaîne d'analyse radiomique typique et les principaux défis actuels

Le processus général d'analyse radiomique est illustré en Figure 3.16, partant des images (Chapitre 2) et aboutissant au modèle, en passant par l'extraction des caractéristiques. Comme nous l'avons vu, la radiomique se présente actuellement sous deux formes. Les caractéristiques extraites obtenues à l'aide de définitions mathématiques précises sont dites « *hand-crafted* », ou « *engineered* ». Leur définition ne dépend pas des données, seules leurs valeurs en dépendent (Section 3.1.1). On parle dans ce cas de la radiomique « classique ». Dans le cas des caractéristiques profondes, leur définition dépend des données, ainsi que leurs valeurs (Section 3.1.2). Quelle que soit la nature des caractéristiques, elles sont utilisées pour alimenter un modèle de classification ou de prédiction (Section 3.2).

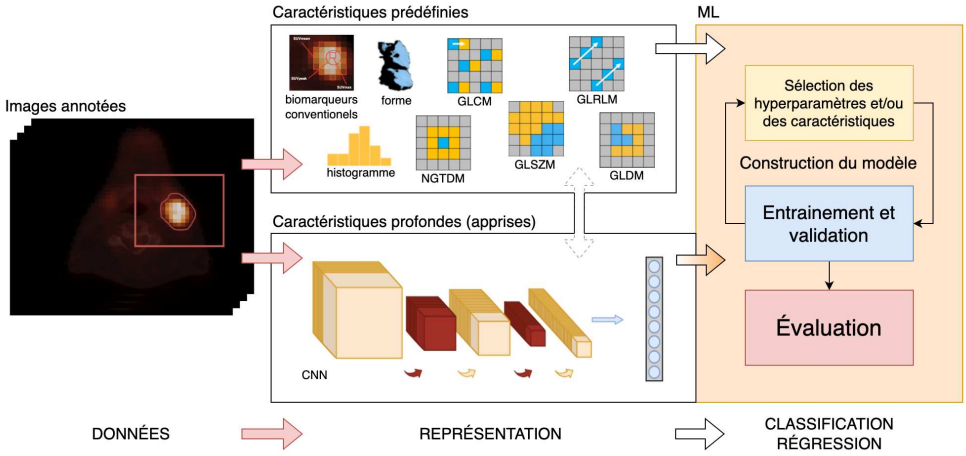


Figure 3.16 – Processus général de l'analyse radiomique, des images au modèle. Dans le cas des caractéristiques profondes provenant d'un CNN, l'étape de classification ou de régression peut être réalisée à la volée : c'est l'entraînement de bout en bout (flèche orange dégradé). Elles peuvent également être extraites afin d'être utilisées comme des variables descriptives classiques (flèche en pointillés ascendante). Inversement, les caractéristiques prédéfinies peuvent être intégrées dans le CNN (flèche en pointillés descendante).

Les modèles radiomiques peuvent être développés pour la classification des tumeurs et des patients, la prédiction de la réponse au traitement, et celle du temps d'occurrence d'événements d'intérêt et de la survie. Dans différents domaines de l'oncologie, la radiomique se développe rapidement. Cependant, elle doit encore relever des défis, car aucun modèle n'est encore utilisé en routine clinique. Ces défis concernent toutes les étapes du processus, notamment la standardisation des définitions [233, 234] et des approches [235, 236], l'accès sécurisé aux données et l'entraînement de modèles robustes sur des échantillons représentatifs [237], ainsi que le manque de reproductibilité des études [233, 238].

Afin de relever ces défis, de nombreuses équipes ont proposé des lignes directrices. Parmi elles, nous pouvons citer IBSI [110], les listes de vérification TRUE (« *is it true, repeatable, useful, and explainable ?* ») [239], TRIPOD (*transparent reporting of a multivariable prediction model for individual prognosis or diagnosis*) [235, 240], et MI-CLAIM (*minimum information about clinical artificial intelligence modeling*) [241], les recommandations des rapports de la SNMMI AI *task force* à propos du développement [155], de l'évaluation [242], et du déploiement des modèles [243], ou bien encore le score de qualité radiomique (*radiomic quality score* (RQS)) [236].

3.3 . L'interprétation des résultats : l'information

« Dans la vie, rien n'est à craindre, tout est à comprendre. »

Marie Curie

Une étude proposant un modèle d'aide à la décision a été publié en 2014 par Aerts et al. [244]. Le RQS a été proposé par plusieurs co-auteurs en 2017 (Lambin et al. [236]), et une revue systématique par Sanduleanu et al. est parue en 2018 et a classé cette étude en première position parmi 41 autres selon ce score [245]. Pourtant, une élégante réanalyse des données par Welch et al. en 2019 a démontré la mauvaise interprétation des résultats initiaux, mettant en évidence que le modèle proposé reflétait en fait le volume de la tumeur [246]. Cet exemple illustre qu'au-delà de suivre des recommandations d'experts, il est crucial de maitriser le contexte applicatif de chaque étude, mais surtout de comprendre l'information capturée par le modèle ainsi que son comportement [200]. Cela aiderait à éviter les biais de confirmation, correspondant à la tendance à interpréter des informations d'une manière qui confirme ou soutient ses croyances a priori [247]. Cela aiderait également à éviter les bais de publication, désignant en science le fait que les chercheurs et journaux scientifiques ont plus tendance à publier des expériences ayant obtenu un résultat positif que des expériences ayant obtenu un résultat négatif [248-250]. Ces biais donnent aux lecteurs une perception biaisée de l'état de la recherche et donc de la connaissance.

En médecine, la fiabilité est un principe fondamental. Pour qu'un processus soit fiable, il doit être transparent, vérifiable, robuste, répétable, équitable, non biaisé, et sécurisé. Plus qu'un élément parmi la liste de ces prérequis, l'interprétabilité peut constituer une pierre angulaire de la fiabilité d'un modèle, car la maîtrise de tout système commence par sa compréhension.

L'interprétabilité désigne dans quelle mesure une analyse peut être comprise, ou le degré avec lequel un humain peut prédire de manière cohérente le résultat d'un modèle [251-254]. Plus la décision d'une machine affecte la vie d'une personne, plus il apparait important pour la machine d'expliquer son comportement. Cependant, les algorithmes de ML ne fournissent pas toujours des prédictions compréhensibles facilement. Cela peut constituer un obstacle à leur adoption dans des contextes à forts enjeux tels que l'imagerie médicale, l'oncologie, et plus généralement la médecine [200].

3.3.1 . L'importance de l'interprétabilité en médecine

3.3.1.1 . L'interprétabilité aide le médecin et le patient

L'interprétabilité est importante en médecine car elle peut contribuer à améliorer la confiance dans les résultats. Lorsque le fonctionnement interne d'un modèle est difficile à comprendre ou à interpréter, il est plus difficile d'avoir confiance dans ses prédictions ou ses recommandations. Dans ce contexte, Vinuesa et al. ont réalisé une enquête en 2020 dans laquelle ils mentionnent que les professionnels de santé ont régulièrement une forte méfiance envers les systèmes d'IA [255]. En utilisant des modèles interprétables ou des techniques d'explication, il est possible de mieux comprendre les facteurs qui déterminent les résultats produits par le modèle. Cela permettrait d'exploiter ces résultats pour prendre des décisions éclairées, en conjonction avec l'ensemble des analyses réalisées lors de la prise en charge.

Il est également possible que la situation inverse se produise. Des résultats automatiques globalement satisfaisants à l'échelle d'une population d'individus (eg, un ensemble de patients) pourraient conduire une équipe de soins à accorder trop de confiance au modèle. Si tel est le cas, une baisse d'attention dans la prise en charge des patients pourrait apparaître, et avoir des conséquences au niveau individuel (eg, pour un patient particulier) [256, 257]. C'est le « biais d'automatisation ». Il correspond à la tendance à surévaluer les observations et les analyses statistiques ou informatiques par rapport à celles des êtres humains [258-260]. L'interprétabilité aiderait à l'éviter, en permettant aux praticiens de comprendre le modèle et donc de l'intégrer dans leur pratique de manière contrôlée. Par exemple, si un modèle fait des prédictions incohérentes, ou si leurs explications sont inattendues, l'équipe de soins pourra les confronter aux autres informations en leur possession. Il est important de toujours replacer la prédiction dans le contexte général de la prise en charge du patient, qui souvent dépasse largement le modèle [200].

3.3.1.2 . L'interprétabilité aide au contrôle et à la maintenance du modèle

Les modèles de ML ne peuvent être audités que lorsqu'ils peuvent être interprétés. L'interprétabilité peut par exemple aider les entités de contrôle à comprendre les limites et les risques du modèle, leur permettant ainsi de mettre en place des mesures de protection appropriées pour les patients [261]. Elle est utile afin de garantir un déploiement stable, responsable, mais aussi éthique [243, 262]. Il est important de s'assurer que le modèle est juste et impartial, par exemple, qu'il n'est pas biaisé contre certains groupes sociaux. L'interprétabilité peut aider à l'identification de tels biais.

Il est aussi essentiel de surveiller et d'évaluer le modèle au fil du temps pour s'assurer qu'il continue d'être performant, et ne change pas de manière inattendue. Pour ce faire, on peut utiliser des tests statistiques et des mesures de performance, mais aussi son interprétation pour mieux comprendre l'évolution potentielle de sa prise de décision. Ce processus d'évaluation et de contrôle continu peut aider à identifier tout changement du comportement du modèle, pour prendre des mesures correctives si nécessaire.

3.3.1.3 . L'interprétabilité aide le développeur

Lorsqu'un modèle n'est pas performant, son interprétation peut aider à comprendre pourquoi, et comment y remédier (Figure 3.17) [165, 200, 263].

La radiomique nécessite de grandes bases de données d'images pour construire et valider des modèles. Dans un contexte multicentrique par exemple, en combinant des images provenant de sources multiples telles que des scanners différents, les bases de données peuvent être augmentées. Cependant, lorsque des méthodes d'analyse ou des modèles radiomiques sont développés dans un centre, ils ne sont pas toujours aussi performants dans un autre. Cela peut être dû à des biais dans l'échantillon de patients constituant la base de données d'entraînement, agissant comme des facteurs confondants [264, 265], ou bien parce que l'information capturée par le modèle et les caractéristiques qui le composent s'ajustent partiellement aux particularités des images de cette base de données : leur bruit, leur aspect plutôt flou ou non, leur résolution, etc. [266]. Certaines caractéristiques sont sensibles aux variations de scanners et de protocoles d'imagerie. Comprendre comment le modèle utilise les images, via quelles caractéristiques et lesquelles sont les plus importantes, à quoi correspondent-elles en termes de signal, et comment elles se comportent à travers les différentes configurations d'acquisition, peut aider à mitiger ce problème. Par exemple via « l'adaptation de domaine », ou « l'harmonisation». Une technique simple peut consister à aligner les distributions des caractéristiques en fonction du domaine de provenance des images (centre, protocole, scanner, etc.). Dans cet objectif, la méthode *ComBat* a été proposée initialement en génomique pour minimiser « l'effet de lot » (« *batch effect* ») [267, 268]. Elle a été introduite en

imagerie médicale par Fortin et al. [269, 270], puis plus spécifiquement dans le domaine de la radiomique par Orlhac et al. [271-274].

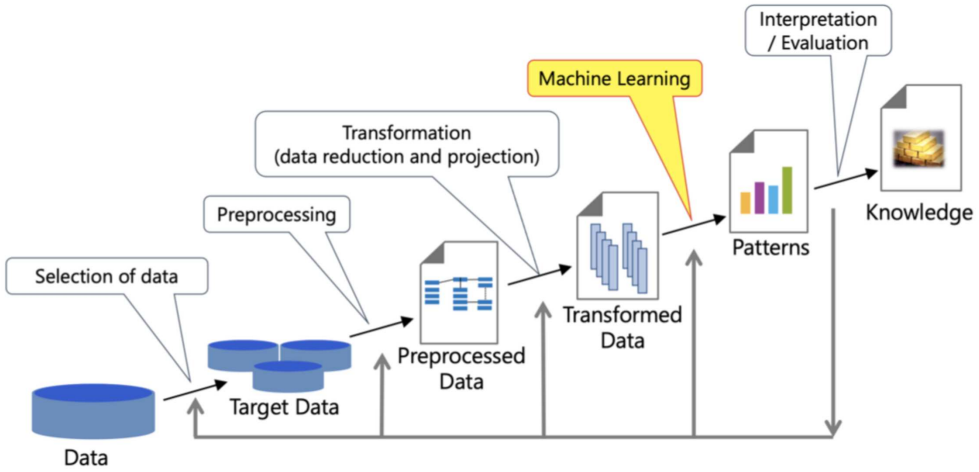


Figure 3.17 – Illustration d'un processus cyclique d'utilisation du ML en sciences. Figure issue de Rudin et al. [263], adaptée de Fayyad et al. [165].

3.3.1.4 . L'interprétabilité aide le chercheur

Les modèles radiomiques sont souvent considérés comme des outils permettant d'automatiser et d'aider les médecins dans la prise en charge des patients. Le déploiement de tels modèles nécessite une forte capacité de généralisation, c'est-à-dire la capacité d'être applicables dans un contexte multicentrique, donc construits avec une quantité suffisante de données représentatives de la population. Au-delà de l'objectif de déploiement d'un modèle prédictif en pratique, nous pouvons utiliser les modèles pour générer des nouvelles hypothèses grâce à leur interprétation sémantique. Cela permettrait de tirer profit des informations présentes dans les images, et d'améliorer notre compréhension de la relation entre leur contenu et ce que nous voulons prédire, même avec des ensembles de données petits et hétérogènes incompatibles avec le déploiement de modèles prédictifs (Section 3.2.4.1) [61, 82, 163-165]. Cette approche conduite par les données pourrait être utilisée lorsque l'on connaît peu les caractéristiques tumorales associées à un résultat, afin de mettre en évidence les motifs liés à la décision du modèle, faciliter l'émergence de nouvelles hypothèses biologiques et médicales, ou guider la recherche de nouveaux biomarqueurs (Figure 3.17).

3.3.2 . Taxonomie de l'interprétabilité

L'interprétailité en ML et en IA est un domaine très vaste. De nombreux éléments sont à considérer et de nombreuses techniques existent pour améliorer la compréhension des modèles [254, 275, 276]. Cette section en présente un aperçu non exhaustif.

3.3.2.1 . La transparence algorithmique

Elle concerne la manière dont l'algorithme s'ajuste aux données pour fournir un modèle, le type de relations qu'il peut apprendre, et dans quelle mesure son optimisation est maîtrisée. Par exemple, nous avons une certaine compréhension du fonctionnement d'un CNN, quel que soit le problème auquel il est appliqué. Il utilise des gradients dans un réseau de millions de neurones, pour apprendre des filtres à plusieurs échelles et des poids de classification. Cependant, la façon dont il s'ajuste aux données n'est pas connue précisément, et de nombreuses recherches sont encore en cours, notamment à propos de l'impact de l'initialisation [277], de l'ordre des données en entrée [278], ou de la taille des lots [186]. La transparence algorithmique d'un CNN peut donc être considérée comme inférieure à celle d'un modèle tel que la régression logistique, bien étudiée et comprise.

3.3.2.2 . L'interprétabilité globale

Elle fait référence à la compréhension de la structure du modèle dans son ensemble une fois entraîné, et vise à comprendre comment il fait des prédictions : quelles informations sont importantes, quelles relations ont-elles entre elles et avec la cible, et comment pouvons-nous expliquer le lien entre les entrées et les sorties du modèle. Pour les arbres, ce seraient les valeurs de coupure associées aux caractéristiques et où elles se situent dans l'arborescence. Cela peut également être l'ensemble des poids d'une régression linéaire ou logistique.

Ce niveau d'interprétabilité peut donc être lié à la notion d'importance des caractéristiques. De nombreuses techniques existent pour déterminer cette importance. Dans le cas des ANNs par exemple, la technique du « *drop-out feature ranking* » permet d'ordonner les caractéristiques selon leur importance à partir de l'effet de l'abandon de chaque neurone sur la performance du modèle [279]. Applicable à tout type de modèle, l'importance par permutations consiste à estimer l'importance d'une caractéristique en évaluant la performance du modèle lorsque les valeurs de cette caractéristique sont mélangées aléatoirement entre les individus de l'ensemble de données [280]. L'importance d'une caractéristique est toujours définie dans le contexte des autres caractéristiques contenues dans le modèle.

La notion d'importance pour une caractéristique n'a pas de définition précise. Estimée de différentes façons, elle peut prendre différentes formes, et correspondre à différentes informations. Par exemple, elle peut être liée à l'influence qu'a la caractéristique sur la valeur de sortie du modèle (eg, poids de régression), ou bien faire directement référence à son impact sur la performance selon une ou plusieurs métriques (eg, *drop-out feature ranking*, importance par permutations).

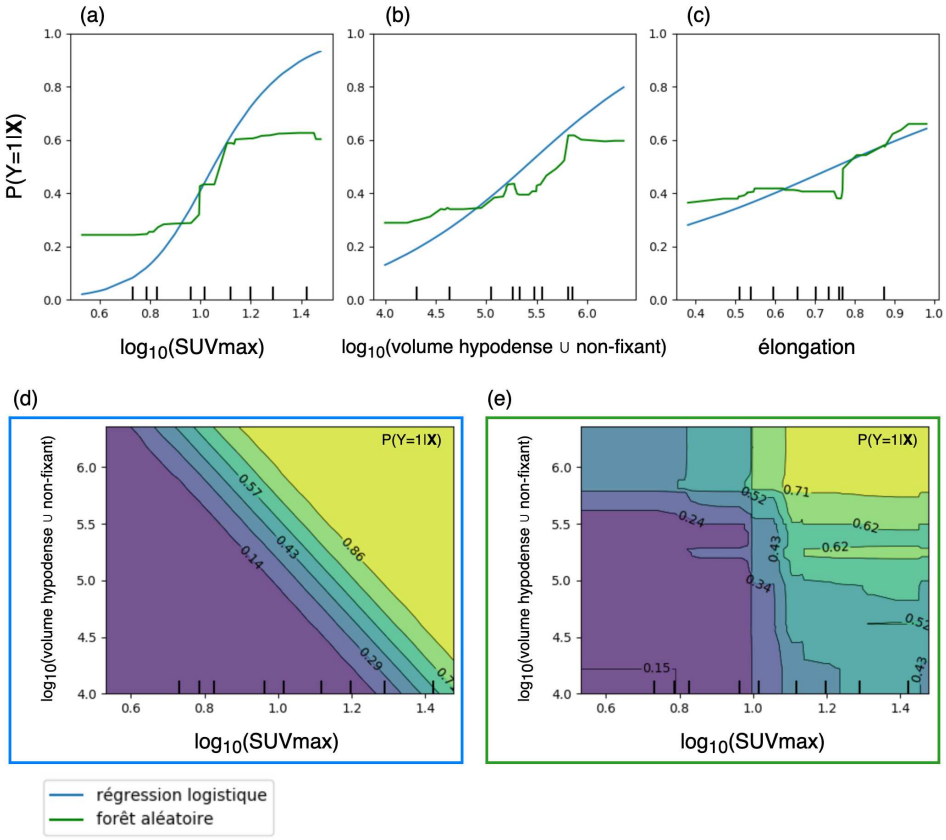


Figure 3.18 – (a, b, c) Graphiques de dépendance partielle 1D pour trois caractéristiques utilisées dans une régression logistique et une forêt aléatoire, pour prédire la survenue de métastases pulmonaires dans les deux ans suivant le bilan d'extension pour des patients atteints de STS à partir d'images TEP/TDM. (d, e) Graphiques de dépendance partielle 2D pour les deux premières caractéristiques. En 2D, le graphique de dépendance partielle est semblable à la surface de décision d'un modèle à deux caractéristiques (Figure 3.8).

Toutefois, il est difficile de comprendre un processus dans son ensemble s'il a un trop haut niveau de complexité. Son interprétabilité diminue donc à mesure que le nombre de facteurs qui le régissent augmente. Pourtant, nous pouvons facilement comprendre un seul ou un nombre limité de poids. Ainsi lorsque nous comprenons un modèle, nous ne considérons en fait que des parties de celui-ci.

Par exemple, en tant que généralisation de la notion de poids, les « graphiques de dépendance partielle » permettent d'évaluer empiriquement l'évolution de la prédiction du modèle en fonction de celle d'une ou de plusieurs caractéristiques (généralement deux au maximum), en les marginalisant par rapport à toutes les autres. Pour tout individu i, la prédiction $\hat{y}_S^{(i)}$ est calculée en faisant varier la valeur des caractéristiques étudiées, $\boldsymbol{X}_S$, tout en préser-

vant la valeur des autres caractéristiques, $\boldsymbol{X}_C$, fixe[11]. La valeur moyenne de $\hat{y}_S^{(i)}$ sur l'ensemble des i pour toutes les combinaisons de valeurs pour $\boldsymbol{X}_S$ correspond à la dépendance partielle de la prédiction par rapport à ces caractéristiques [168, 254] (Figure 3.18).

Il demeure que l'interprétabilité dépend de la façon avec laquelle le modèle utilise les caractéristiques. Un modèle contraignant des relations monotones garantit que la relation entre une caractéristique et la prédiction va toujours dans le même sens, ce qui facilite sa compréhension globale. Une augmentation de la valeur de x_j entraîne soit toujours une augmentation, soit toujours une diminution de $\hat{y}$. Contrainte encore plus forte aidant davantage à la compréhension, la linéarité garantit que l'effet d'une caractéristique sur la sortie du modèle est constant, en termes de sens, mais aussi quantitativement. Une augmentation d'une unité de x_j fait varier $\hat{y}$ de β_j quelle que soit leur valeur. Cela garantit que l'importance relative globale de chaque caractéristique par rapport aux autres est vraie en tout point de l'espace défini par $\boldsymbol{X}$ (Section 3.3.2.3). De façon intermédiaire, certaines fonctions simples, monotones mais non-linéaires, préservent la possibilité d'une interprétation globale du modèle. C'est le cas pour les modèles linéaires généralisés tels que la régression de Cox (Section 3.2.2.8) ou la régression logistique (Section 3.2.2.2).

La régression logistique transforme la somme pondérée D en probabilités via la fonction logistique $\sigma(D)$ (équations (3.10), (3.11), et (3.12)). Soit $odds = \frac{P(y=1|X)}{P(y=0|X)}$ la « cote » (« *odds* ») de la classe positive (combien de fois l'individu à plus ou moins de probabilité d'appartenir à la classe positive qu'à la classe négative), les caractéristiques de $\boldsymbol{X}$ sont alors proportionnelles à son logarithme $\ln(odds)$, et une augmentation d'une unité de x_j le fait varier de β_j. Comme la fonction $\ln$ peut parfois apparaître délicate à interpréter, une exponentielle permet d'exprimer la cote directement avec :

$$odds = e^{\beta_0+\beta_1\times x_1+\ldots+\beta_j\times x_j+\ldots+\beta_p\times x_p}. \tag{3.33}$$

Nous pouvons ensuite exprimer ce qu'il se passe lorsque l'on augmente x_j d'une unité :

$$OR_{x_j} = \frac{odds_{x_j+1}}{odds_{x_j}} = \frac{e^{\beta_0+\beta_1\times x_1+\ldots+\beta_j\times(x_j+1)+\ldots+\beta_p\times x_p}}{e^{\beta_0+\beta_1\times x_1+\ldots+\beta_j\times x_j+\ldots+\beta_p\times x_p}} = e^{\beta_j} \tag{3.34}$$

avec OR_{x_j} le « rapport de cote » (« *odds ratio* ») associé à x_j. Une augmentation de x_j d'une unité fait varier la cote, $odds$, de façon multiplicative selon un facteur égal à e^{β_j}. Plus globalement, pour toute variation u de x_j, $odds_{x_j+u} = odds_{x_j} \times e^{\beta_j\times u}$.

11. Il est à noter que dans la plupart des cas d'application, la présence de corrélations entre les caractéristiques de $\boldsymbol{X}$ fait qu'il n'est pas réaliste de faire varier la valeur de certaines d'entre elles tout en fixant strictement toutes les autres.

Prenons un individu i dont la cote initiale est de $odds^{(i)} = 3$. Cela signifie que sa probabilité prédite pour $y = 1$ est trois fois plus élevée que pour $y = 0$. Si le poids β_j appris par le modèle et associé à x_j est de $0,69$, alors augmenter x_j d'une unité multiplie la cote par $e^{0,69} \approx 2$, la faisant passer de 3 à environ 6. Si au contraire, x_j perd une unité ($u = -1$), la cote est divisée par 2 (multipliée par $0,5$) et devient $odds^{(i)}_{x_j-1} = 3 \times e^{-0,69} \approx 1,5$. Enfin, si ce sont deux unités qui sont ajoutées, la cote passe de 3 à $odds^{(i)}_{x_j+2} = 3 \times e^{0,69\times 2} \approx 12$, et l'individu i a désormais 12 fois plus de chance d'appartenir à la classe positive qu'à la classe négative.

L'interprétation additive d'une fonction linéaire semble être la plus naturelle pour l'humain. L'interprétation multiplicative du rapport de cote demande en effet déjà un peu d'habitude. Par contre, à cause de la transformation logarithmique globale, elle peut être plus intuitive que l'interprétation additive directe de D. En outre, elle est toujours définie de façon relative à une cote de base. Cela peut être la cote moyenne d'un groupe d'individus de référence par exemple, ou bien celle d'un objectif à atteindre. En fonction des cas d'application, interpréter une prédiction relativement à une autre plutôt que de façon absolue peut tout autant constituer un avantage qu'un inconvénient. L'interprétation globale et numériquement précise qui caractérise les modèles linéaires généralisés est rare voire absente dans les autres types de modèles. Par exemple, il n'est pas possible de relier précisément l'évolution de la prédiction aux caractéristiques incluses dans une forêt aléatoire, de façon générale, avec une équation composée d'éléments compréhensibles par l'humain (Figure 3.18). L'importance globale des caractéristiques dans ce cas n'est qu'une approximation de la façon dont le modèle les utilise (Section 3.3.2.4).

3.3.2.3 . L'interprétabilité locale

L'interprétabilité locale correspond à l'examen de la prédiction du modèle à l'échelle individuelle : quel est le résultat pour un individu donné et pourquoi (Figure 3.19).

Dans un modèle multivarié, deux individus peuvent en effet avoir la même valeur de prédiction, mais pour des raisons différentes. Tel qu'illustré dans la Figure 3.18 (d, e) par exemple, un patient peut avoir une probabilité prédite élevée à cause d'un SUVmax élevé, ou bien si le volume de sa lésion hypodense ou ne fixant pas le FDG est important. Elle peut aussi être le résultat d'une infinité de combinaisons de valeurs intermédiaires pour ces deux caractéristiques.

Dans les modèles non-linéaires, l'impact d'une caractéristique varie dans l'espace. La méthode LIME (*local interpretable model-agnostic explanations*), par exemple, consiste à créer un modèle naturellement interprétable (eg, linéaire) basé sur un petit voisinage d'un individu d'intérêt, et à l'utiliser pour

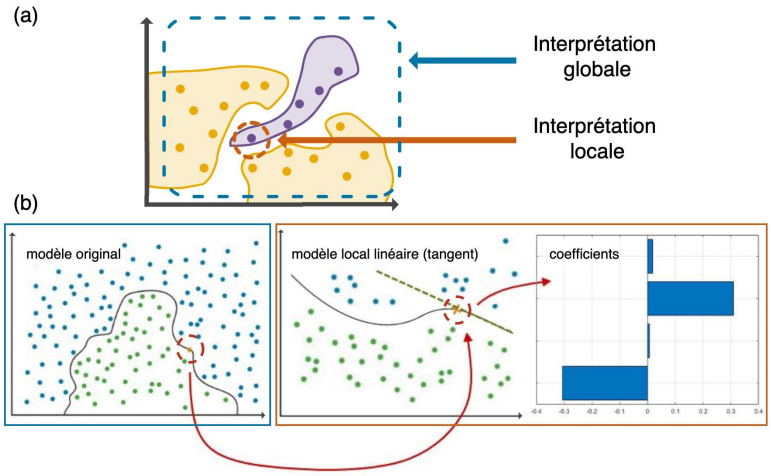


Figure 3.19 – (a) Illustration générale des notions d'interprétation globale et locale. (b) Illustration du principe de la méthode d'explication locale LIME. Figure créée et adaptée à partir de la rubrique sur l'interprétabilité du site web mathworks.com [281].

expliquer les prédictions du modèle pour cet individu en identifiant l'importance des caractéristiques (eg, les poids de régression) à cet endroit précis de l'espace (Figure 3.19 (b)) [282].

Nous pouvons également citer la méthode SHAP (*Shapley additive explanations*) [283], basée sur les « valeurs de Shapley » provenant de la théorie des jeux [284]. Elle calcule l'importance des caractéristiques pour chaque individu en estimant la participation marginale de chacune d'entre elles à la prédiction pour ce dernier grâce à un système de « coalitions de joueurs ». Les caractéristiques représentent les joueurs, et les différentes coalitions correspondent à différentes situations dans lesquelles certaines caractéristiques ont été « masquées ». Pour chaque individu i, pour chaque caractéristique $x_j^{(i)}$, il est possible de calculer la différence de prédiction pour les c_j paires de coalitions, A_{k_j} et B_{k_j}, telles que X_j est masquée dans A_{k_j} mais pas dans B_{k_j}. Cela correspond aux « contributions » $d_{k_j}^{(i)}$ de la caractéristique $x_j^{(i)}$ aux différentes coalitions dans lesquelles elle est présente. La valeur de Shapley $\varphi_{x_j^{(i)}}$ est alors calculée en prenant la moyenne sur c_j, normalisée par le nombre total p de caractéristiques dans le modèle tel que :

$$\varphi_{x_j^{(i)}} = \frac{1}{p} \times \frac{1}{c_j} \times \sum_{k_j=1}^{c_j} d_{k_j}^{(i)}. \tag{3.35}$$

En fonction du contexte d'utilisation, des hypothèses statistiques du problème, du type de modèle expliqué, et de comment est réalisé le masquage, la méthode SHAP converge avec d'autres méthodes d'explicabilité.

Une importance globale des caractéristiques peut être exprimée en agrégeant les importances locales (eg, avec la moyenne sur i en valeur absolue).

En imagerie, la cartographie de la contribution de chaque voxel à la sortie du modèle permet de mettre en évidence la localisation des sous-régions qui sont importantes pour prendre une décision ou faire une prédiction, contribuant ainsi à augmenter la transparence et l'interprétabilité. Ce sont les méthodes de « saillance ». Plusieurs approches ont été proposées dans le contexte de l'apprentissage profond, plus particulièrement lors de l'utilisation de CNNs [285-290]. Génériques, les méthodes LIME et SHAP peuvent également estimer l'importance des voxels, et produire des cartes.

3.3.2.4 . Interprétabilité intrinsèque ou explicabilité *post-hoc*

Alors que certains modèles sont intrinsèquement interprétables, d'autres nécessitent d'utiliser des techniques additionnelles pour comprendre leur processus de décision. L'explication la plus fidèle d'un modèle est le modèle lui-même puisqu'il se représente parfaitement. La régression linéaire, la régression logistique, et l'arbre de décision sont considérés comme interprétables. La simplicité de leur structure ainsi que des relations avec lesquelles ils relient $\boldsymbol{X}$ à Y les rendent directement et facilement compréhensibles pour l'être humain.

D'autre part, les modèles complexes, comme les forêts aléatoires, les SVMs à noyau, ou les ANNs et CNNs profonds, ont une complexité telle qu'il est impossible de les comprendre directement. Ils peuvent cependant devenir explicables dans certains cas, grâce à une méthode *post-hoc*, correspondant à une approximation interprétable du modèle original, par exemple au moyen des méthodes LIME, SHAP, des graphiques de dépendance partielle, du *drop-out feature ranking*, ou encore de l'importance par permutations.

3.3.2.5 . Interprétabilités et explicabilités agnostiques ou spécifiques aux modèles

Certaines approches (LIME, SHAP, dépendance partielle, importance par permutations) sont applicables à tous les modèles. Ce sont les techniques agnostiques dites « *model-agnostic* ».

D'autres sont spécifiques à certains types de modèles (coefficients de régression, règles de dichotomie d'un arbre, cartes de saillance, *drop-out feature ranking*). Dans les modèles basés sur des arbres par exemple, l'importance des caractéristiques peut être estimée via la diminution moyenne de l'impureté. Cette méthode consiste à mesurer la diminution de l'impureté (eg, Gini) des groupes formés par dichotomie, pour chaque caractéristique, dans tous les arbres, et à considérer la moyenne comme importance globale de cette caractéristique. La caractéristique qui entraîne la plus forte diminution moyenne de l'impureté est alors considérée comme la plus importante globalement et ainsi de suite.

4 - Analyse multimodale supervisée de tumeurs avec des caractéristiques radiomiques définies à l'échelle du voxel mettant en évidence des motifs biologiquement interprétables

Ce chapitre propose une approche méthodologique d'interprétation des modèles radiomiques, ainsi que les résultats obtenus sur une base de données publique [60]. Après la présentation de résultats préliminaires sous forme d'un e-poster au congrès annuel de 2021 de la SNMMI [174], ces travaux ont fait l'objet d'une publication dans le journal Medical Physics [61].

4.1 . Introduction

La compréhension précise du fonctionnement des modèles de classification et de prédiction est la plupart du temps incluse dans les recommandations citées en Section 3.2.5. Dans la pratique, cette question ne constitue cependant pas le cœur du domaine de la radiomique, et l'emphase est généralement mise sur la performance et la capacité de généralisation. Néanmoins, elle est de plus en plus considérée [243, 291-294], et des études se prétendant interprétables sont publiées en radiomique classique [295-300], en DL [301-304], ou via des approches hybrides [305-307]. Souvent, les auteurs emploient des modèles simples et parcimonieux, ou des approches *post-hoc* pour « retrouver » l'importance des caractéristiques, non intrinsèquement disponible lors de l'utilisation de certains modèles. Ce sont ces méthodes de ML qui semblent donner le caractère « interprétable » aux analyses proposées. Le sont-elles vraiment ? Cela suffit-il pour comprendre quelles informations le modèle capture dans les images et relier ses prédictions à un rationnel médical ? Si l'image comme donnée d'entrée peut être comprise selon les principes physiques qui la régissent, et si le modèle renseigne sur son processus de décision, il demeure que le pont entre les deux, la représentation des images par des variables, n'est pas toujours aussi interprétable qu'il n'y paraît.

4.1.1 . Le cas de la radiomique classique

Utilisant la régression de Cox admettant des coefficients clairement définis, le modèle de Aerts et al. confondant le volume tumoral aurait pu être considéré comme interprétable [244]. Les auteurs l'ont d'ailleurs expliqué, probablement à tort [246], comme capturant « l'hétérogénéité tumorale ». Même si les caractéristiques radiomiques *handcrafted* sont mathématique-

ment bien définies, l'interprétation des modèles basés sur celles-ci reste souvent difficile. Connaître l'importance des variables dans le modèle et comment il les utilise est crucial, mais l'interprétation des caractéristiques peut elle-meme être un défi. Elles sont parfois trop complexes pour qu'on puisse les interpréter intuitivement et les raccrocher de façon fiable au signal qu'elles capturent [93, 291, 308]. Cela rend difficile la compréhension de la biologie qui les sous-tend, et donc leur lien avec l'état du patient. Leur combinaison dans un modèle multivarié ne fait que compliquer encore plus cette tâche d'interprétation, même lorsqu'il s'agit de modèles « explicables » (eg, utilisant LIME ou SHAP), ou simples et « interprétables » (eg, régression linéaire ou logistique, arbre de décision).

De plus, les caractéristiques prédéfinies sont généralement calculées à partir du contenu d'une ROI entière avec des mesures agrégatives. Cela signifie que lors de l'extraction d'une caractéristique radiomique pour une ROI donnée, les voxels qui composent cette ROI sont agrégés pour donner la valeur scalaire de la caractéristique. Conceptuellement, cela limite grandement la décomposition de l'information qu'elle porte selon sa part descriptive, qui capture un motif d'intérêt du signal, et sa part agrégative, résumant ce motif dans l'espace. Il est donc difficile de relier les sorties du modèle à une caractérisation distribuée au niveau du voxel ou de sous-régions. Déchiffrer un système ou un processus complexe implique pourtant souvent de le décomposer.

> *« Le second [principe de la méthode], de diviser chacune des difficultés que j'examinerais, en autant de parcelles qu'il se pourrait, et qu'il serait requis pour les mieux résoudre. »*
>
> René Descartes, Discours de la méthode, 1637

4.1.2 . Le cas de la radiomique profonde

Tirant parti de la structure des CNNs naturellement décomposée en blocs de filtrage (extraction de motifs locaux) et de *pooling* ou d'agrégation (Section 3.1.2), de nombreuses cartographies de saillance ont été proposées dans le contexte de l'apprentissage profond (Section 3.3.2.3).

Pourtant, bien que ces méthodes soient prometteuses notamment pour détecter certains biais et facteurs confondants [264], elles présentent des limites, telles qu'une résolution grossière ou une représentation éparse. Plus important encore, les CNNs sont des modèles boites noires. Les caractéristiques qu'ils extraient ne sont pas clairement définies mathématiquement. La complexité des modèles et l'abstraction de l'information extraite limitent donc leur transparence. En effet, savoir où se trouve l'information pertinente dans l'image peut être utile, mais n'indique pas comment cette information est utilisée [200]. Kumar et al. proposent par exemple une chaîne d'analyse

radiomique par DL entraînée de bout en bout pour prédire la malignité de nodules pulmonaires à partir d'images TDM [302]. En employant une technique basée sur la déconvolution [290], des cartes de saillance sont associées à chaque prédiction. La Figure 4.1 montre les résultats pour quatre nodules, deux bénins et deux malins. Chaque image est associée à la carte d'explication pour sa prédiction. Les zones importantes mises en évidence se situent au niveau des nodules. Dans un cas sur deux pour chaque classe, elles pointent également la zone pleurale ou parenchymateuse la plus proche, de façon cohérente avec les résultats de Hosny et al. utilisant une autre méthode de cartographie [303]. Bien que cela puisse donner une certaine confiance quant au fait que le modèle capture une information pertinente, il reste difficile voire impossible d'expliquer pourquoi chaque nodule a été classifié dans une classe et pas dans l'autre.

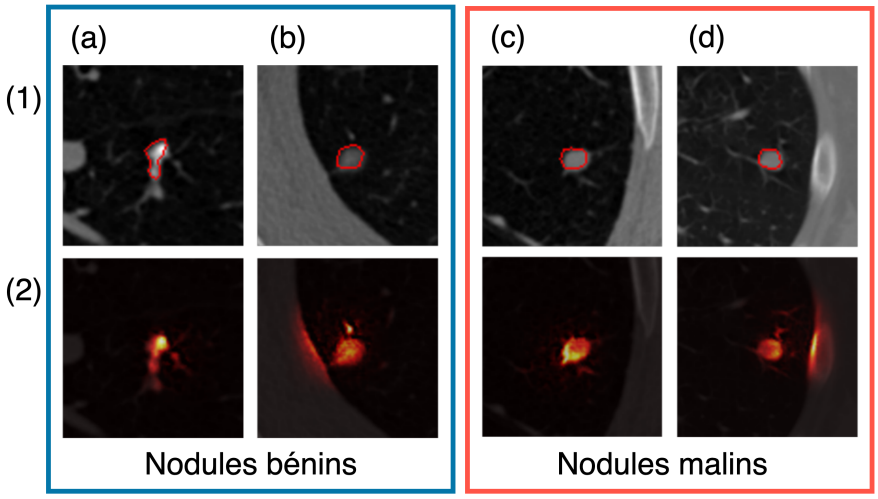


Figure 4.1 – Nodules pulmonaires bénins (a, b) et malins (c, d) imagés et délinéés en TDM (1), associés aux cartes de saillance expliquant les prédictions (2). Figure créée et adaptée à partir de Kumar el al. [302].

4.1.3 . La limite de l'explicabilité *post-hoc*

Il est impossible que les méthodes *post-hoc* représentent exactement ce que le modèle original calcule. Si elles le faisaient, elles seraient inutiles, puisque l'explication serait le modèle lui-même (Section 3.3.2.4). Cela pose la question du risque, pour toute méthode d'explication, de ne pas être fidèle à l'information réellement capturée, au moins dans certaines zones de l'espace de représentation (localement pour certains individus). Cela peut conduire à des explications potentiellement trompeuses, que ce soit dans le cas spécifique des méthodes de saillance [309-312], ou plus globalement pour toutes les approches *post-hoc* [313, 314].

Par exemple, Adebayo et al. ont publié une étude en 2018 proposant de mesurer le lien réel entre certaines méthodes de saillance, les modèles, et leurs données d'entrée [315]. Ils ont observé que certaines méthodes peuvent s'avérer très peu sensibles au modèle expliqué ainsi qu'aux données d'entraînement, et être finalement très similaires à de simples filtres déterministes détecteurs de contours (« *edge detectors* »). Cette étude a elle-même été critiquée par Tomsett et al. en 2019, observant plusieurs limites dans la formulation des évaluations [316].

Si les méthodes d'explication sont difficiles à évaluer, il devient difficile de développer des modèles pour lesquels nous aurons confiance dans l'ensemble de ses décisions, même si elles sont « expliquées ». Si nous ne pouvons pas savoir avec certitude si l'explication est correcte, nous ne pouvons pas savoir si nous devons lui faire confiance. Cela nous ramène au point de départ de l'interprétabilité des modèles complexes (eg, CNN), et par extension de leur fiabilité [243].

4.2 . La cartographie de décision radiomique : vers une méthode interprétable, spatialement distribuée, et clairement définie

Des méthodes combinant les informations spatiales et quantitatives liées aux sorties d'une manière simple et fidèle au modèle original sont nécessaires. La cartographie des caractéristiques prédéfinies pourrait atténuer la conséquence de la complexité mathématique de leur définition sur leur manque d'interprétabilité. En tant que méthode intermédiaire entre la définition spatiale des CNNs et la définition quantitative des caractéristiques prédéfinies, c'est ce que nous proposons dans cette étude : une cartographie des sorties d'un modèle de ML intrinsèquement interprétable basée sur des caractéristiques prédéfinies, dans l'objectif de faciliter son interprétation biologique.

Peu d'études utilisent la cartographie des caractéristiques radiomiques prédéfinies. Wu et al. [317-319], Xu et al. [320], et Even et al. [321] ont utilisé des méthodes de *clustering* non supervisé pour identifier des sous-régions tumorales, afin de les associer a posteriori à la survie des patients. Beaumont et al. ont utilisé une approche de forêt aléatoire pour prédire la localisation de la récidive grâce à des caractéristiques et des vérités terrain définies à l'échelle du voxel [322]. Vuong et al. ont étudié la radiomique par patchs avec activation binaire pour identifier l'emplacement des régions responsables d'une classification donnée [323]. À notre connaissance, bien que la radiomique *handcrafted* soit largement utilisée, notamment lorsque les ensembles de données sont petits et ne se prêtent pas à la radiomique profonde, aucune approche n'a été proposée pour cartographier quantitativement, au niveau du voxel, la sortie de modèles basés sur de telles caractéristiques.

4.2.1 . Matériels et méthodes

Cette section décrit comment la régression logistique peut être utilisée pour faire le lien entre une classification binaire probabiliste et une carte de caractérisation quantitative et interprétable définie à l'échelle du voxel.

4.2.1.1 . Formulation

Pour permettre une cartographie de la sortie du modèle dans le cadre d'une analyse utilisant des caractéristiques mathématiquement bien définies, ces dernières sont initialement extraites au niveau du voxel. Une fenêtre glissante 3D cubique est utilisée. Pour chaque position du cube centré sur le voxel v à l'intérieur de la ROI, les caractéristiques radiomiques sont calculées dans ce cube et les valeurs résultantes sont attribuées à v dans les cartes de caractéristiques résultantes (Figure 4.2).

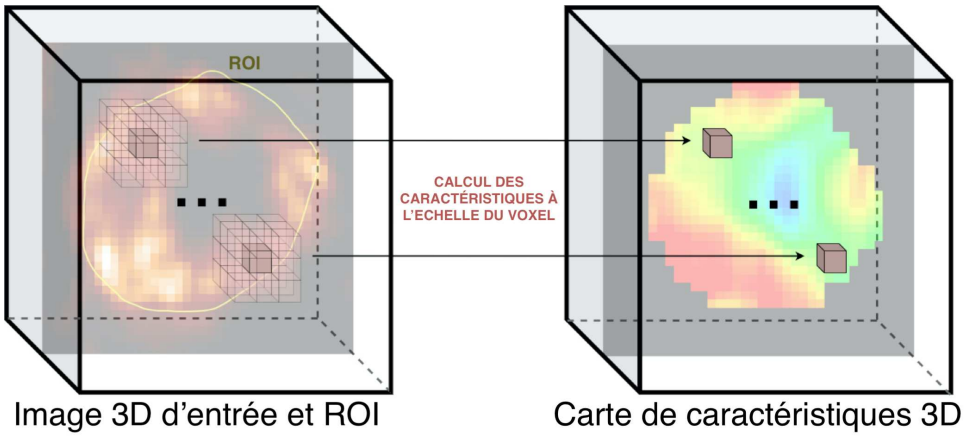


Figure 4.2 – Extraction des caractéristiques à l'échelle du voxel à l'aide d'un noyau 3D glissant de dimensions choisies. Dans cet exemple, une carte de caractéristiques est calculée à l'aide d'un noyau de $3\times3\times3$ voxels et le résultat est attribué au voxel central de cette fenêtre dans la carte de caractéristiques 3D. Ce processus est répété pour toutes les caractéristiques et tous les voxels à l'intérieur de la ROI. Figure adaptée d'Escobar et al. [61]

La valeur de la j-ième caractéristique attribuée au voxel v est notée $x_j^{(i,v)}$, composante de $X^{(i,v)}$ l'ensemble des caractéristiques définies à l'échelle du voxel pour la tumeur i. La j-ième caractéristique radiomique $g_j^{(i)}$ pour i est obtenue en faisant la moyenne sur tous les voxels de la ROI telle que :

$$g_j^{(i)} = \frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}} x_j^{(i,v)} \tag{4.1}$$

avec $m^{(i)}$ le nombre total de voxels dans la ROI.

Chaque tumeur est ainsi décrite par un vecteur de caractéristiques $G^{(i)}$ composé de p caractéristiques. Si chaque patient ne possède qu'une seule

tumeur, les données sont représentées par la matrice $\boldsymbol{G} \in \mathbb{R}^{n \times p}$ contenant l'ensemble des caractéristiques agrégées de tous les patients. Définie en Section 3.2.2.2, l'équation (3.10) modélisant la probabilité d'appartenance à la classe positive admet une fonction de décision linéaire, D, pouvant ici s'exprimer telle que :

$$D(G^{(i)}) = \ln \frac{\hat{y}^{(i)}}{1 - \hat{y}^{(i)}} = G^{(i)\intercal}\boldsymbol{\beta} + \beta_0 = \beta_0 + \sum_{j=1}^{p}(g_j^{(i)} \times \beta_j). \tag{4.2}$$

L'application de β_0 et $\boldsymbol{\beta}$ à $X^{(i,v)}$ à l'échelle du voxel donne $DV^{(i)}$ la carte de décision radiomique (*radiomic decision map* (RDM)), quantitative, qui distribue spatialement la décision pour i dans la ROI telle que :

$$DV^{(i)} = D(X^{(i,v)}) = X^{(i,v)\intercal}\boldsymbol{\beta} + \beta_0 = \beta_0 + \sum_{j=1}^{p}(x_j^{(i,v)} \times \beta_j). \tag{4.3}$$

La rétroprojection de β_0 et $\boldsymbol{\beta}$ à $X^{(i,v)}$ préserve la quantification probabiliste. En effet, en raison de la nature linéaire de la moyenne ainsi que de D, la valeur moyenne $\overline{DV}^{(i)}$ de $DV^{(i)}$ sur tous les voxels de la ROI est égale à $D(G^{(i)})$.

$$\begin{aligned}
\overline{DV}^{(i)} &= \frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}}(\beta_0 + \sum_{j=1}^{p}(x_j^{(i,v)} \times \beta_j)) \\
&= \beta_0 + \sum_{j=1}^{p}(\frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}} x_j^{(i,v)} \times \beta_j) \\
&= \beta_0 + \sum_{j=1}^{p}(g_j^{(i)} \times \beta_j) \\
\overline{DV} &= D.
\end{aligned} \tag{4.4}$$

La fonction de décision D, exprimée en fonction de la moyenne des caractéristiques locales donnant les caractéristiques agrégées G, est égale à la moyenne de la fonction de décision locale, exprimée en fonction des caractéristiques locales X sur tous les voxels de la ROI pour chaque patient.

Ainsi, nous pouvons exprimer la probabilité pour un patient i d'appartenir à la classe positive directement au niveau du voxel en utilisant :

$$\hat{y}^{(i)} = \frac{1}{1 + e^{-\left(\frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}}(X^{(i,v)\intercal}\boldsymbol{\beta}) + \beta_0\right)}}. \tag{4.5}$$

La méthode proposée produit donc des RDMs qui quantifient la contribution marginale de chaque voxel à la classification des patients, mettant en évidence les sous-régions les plus contributives au sein de la ROI (Figure 4.3).

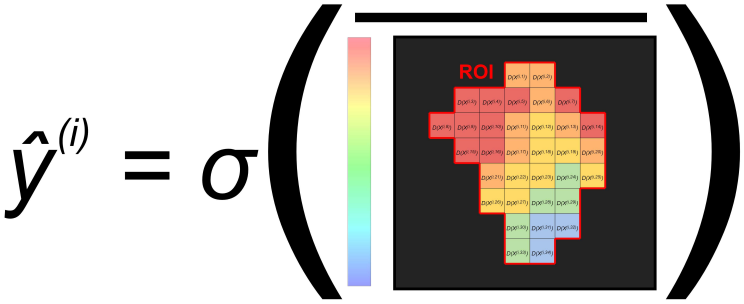


Figure 4.3 – Illustration de l'expression à l'échelle du voxel de $\hat{y}^{(i)}$ la probabilité prédite pour un patient i d'appartenir à la classe positive en fonction de l'imagerie de sa tumeur. Les voxels dont la valeur est fortement positive (rouge) font augmenter la probabilité d'appartenance à la classe positive, tandis que ceux ayant un signal fortement négatif (bleu) participent à abaisser la probabilité, et sont plutôt associés à une classification dans la classe négative.

Certaines caractéristiques, telles que les caractéristiques de forme, ne peuvent pas s'exprimer au niveau du voxel en employant une boite glissante, mais elles peuvent néanmoins contribuer à une fonction de décision :

$$\begin{aligned} D([G^{(i)}; G'^{(i)}]) &= G^{(i)\intercal}\boldsymbol{\beta} + G'^{(i)\intercal}\boldsymbol{\beta'} + \beta_0 \\ &= \frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}} (X^{(i,v)\intercal}\boldsymbol{\beta}) + G'^{(i)\intercal}\boldsymbol{\beta'} + \beta_0 \end{aligned} \tag{4.6}$$

où $G'^{(i)}$ représente le vecteur composé de p' caractéristiques $g_{j'}^{\prime(i)}$ sans formulation locale qui est concaténé à $G^{(i)}$, et $\beta'_{j'}$ les coefficients appris associés à ces caractéristiques.

Comme dans l'équation (4.5), la probabilité qu'un patient i donné appartienne à la classe positive peut être exprimée par :

$$\hat{y}^{(i)} = \frac{1}{1 + e^{-\left(\frac{1}{m^{(i)}} \times \sum_{v=1}^{m^{(i)}} (X^{(i,v)\intercal}\boldsymbol{\beta}) + G'^{(i)\intercal}\boldsymbol{\beta'} + \beta_0\right)}}. \tag{4.7}$$

Lorsque le modèle inclut des caractéristiques sans formulation à l'échelle du voxel, seule une partie du modèle est expliquée par les RDMs. Les coefficients associés aux caractéristiques qui peuvent être cartographiées indiquent leur importance par rapport à celles qui ne peuvent pas l'être.

4.2.1.2 . Patients et données

Nous avons utilisé un ensemble de données public de 51 patients atteints de STS pour lesquels les images TEP au FDG, TDM, et IRM pondérées

en T1 et T2 avec suppression du signal de la graisse (T2-SG) sont disponibles avec la ROI tumorale [60]. Cette base de données inclut également des informations cliniques et de suivi.

Au cours de la période de suivi, 19 patients ont développé des métastases pulmonaires et 32 n'en ont pas développé. La tâche consistait à prédire l'apparition de ces métastases dans les deux ans suivant le bilan d'extension. Les images TEP/TDM ont toutes été acquises au Centre Universitaire de Santé McGill (Montréal, Canada) à l'aide du même scanner (Discovery ST, GE Healthcare) pour les 51 patients. Les examens IRM ont été réalisés dans le cadre des soins de routine, avec des protocoles hétérogènes selon les patients. Les images IRM pondérées en T1 étaient disponibles pour les 51 patients. Deux types de séquences pondérées en T2 avec suppression du signal de la graisse ont été acquises, à savoir des séquences T2FS (26 patients), et des séquences T2-STIR (25 patients) (Section 2.1.2). Chaque tumeur a été détourée manuellement par un radio-oncologue sur les images T2-SG, et les ROI ont été propagées aux images TEP, TDM, et T1 après un recalage rigide. Des informations détaillées sur la base de données sont fournies par Vallières et al. [60].

4.2.1.3 . Traitement et représentation des images

Prétraitement des images IRM

Les images T1 ont été corrigées du « champ de biais » (« *bias field* ») avec l'algorithme N4ITK [324], avec des paramètres par défaut et le masque du corps comme région pour l'estimation du champ de biais. Aucune correction n'a été possible sur les images T2-SG car il n'y avait pas de signal significatif en dehors de la ROI tumorale pour estimer le champ de biais. L'effet de la correction sur le signal des images T1 a été évalué par inspection visuelle et de manière quantitative. Pour chaque patient, des sphères de $239 \pm 52mm^3$ (volume moyen $\pm$ 1 écart-type) ont été dessinées manuellement dans le tissu adipeux (23 ± 9 sphères par patient). Des sphères de $243 \pm 45mm^3$ ont également été tracées dans le tissu musculaire (16 ± 6 sphères par patient). Avant et après l'application de l'algorithme N4ITK, le coefficient de variation intratissus CV_t a été calculé dans chaque image et chaque tissu tel que :

$$CV_t = \frac{\sigma_t}{\mu_t} \tag{4.8}$$

avec σ_t et μ_t l'écart-type et l'intensité moyenne dans le tissu t. En outre, le contraste intertissus CIT a été calculé pour s'assurer que la correction a atténué les inhomogénéités intratissus tout en préservant le contraste intertissus tel que :

$$CIT = \frac{|\mu_{graisse} - \mu_{muscle}|}{\mu_{graisse} + \mu_{muscle}}. \tag{4.9}$$

Enfin, le coefficient de variation conjointe CJV a été utilisé comme mesure combinée des variations intratissus et de la séparation intertissus tels que :

$$CJV = \frac{\sigma_{graisse} + \sigma_{muscle}}{|\mu_{graisse} - \mu_{muscle}|}. \tag{4.10}$$

Pour ces trois mesures, l'effet global de la correction du champ de biais par l'algorithme N4ITK sur l'ensemble des images a été évalué à l'aide du test de rang signé par paires de Wilcoxon [325].

Dans les images IRM T1 et T2, la valeur d'un voxel ne peut pas être facilement interprétée en termes de quantité physique, et le même type de tissu peut donner des valeurs de voxels différentes entre différentes acquisitions, même si les images sont acquises chez le même patient en suivant le même protocole. Une version adaptée de la méthode *White Stripe* a été utilisée pour standardiser de façon linéaire les images en se basant sur la graisse comme tissu de référence pour toutes les images T1 [28]. Les sphères dessinées manuellement dans le tissu adipeux ont été utilisées. Pour chaque patient, la valeur de chaque voxel de l'image a été modifiée linéairement de sorte que le signal moyen dans toutes les sphères soit égale à 0 avec un écart-type de 1 telles que :

$$IW_v = \frac{I_v - \mu_{graisse}}{\sigma_{graisse}} \tag{4.11}$$

avec I_v l'intensité de chaque voxel v dans l'image T1 après application de la méthode N4ITK, $\sigma_{graisse}$ et $\mu_{graisse}$ l'écart-type et la moyenne d'intensité dans le tissu graisseux de référence, et IW_v la valeur standardisée au voxel v. Comme aucun tissu de référence ne pouvait être utilisé pour standardiser les images T2-SG, un *z-score* basé sur la ROI a été utilisé, de sorte que ce soit la valeur moyenne dans chaque tumeur qui soit de 0 avec un écart-type de 1 telle que :

$$IZ_v = \frac{I_v - \mu_{tum}}{\sigma_{tum}} \tag{4.12}$$

avec I_v l'intensité dans l'image T2-SG, σ_{tum} et μ_{tum} l'écart-type et la moyenne d'intensité dans la ROI tumorale, et IZ_v le *z-score* (valeur standardisée) au voxel v. La différence fondamentale entre les équations (4.11) et (4.12) est que l'équation (4.11) préserve les variabilités interpatients de l'intensité du signal entre les tumeurs alors que ce n'est pas le cas pour l'équation (4.12).

Calcul des caractéristiques

Les images ont été rééchantillonnées en utilisant l'interpolation *B-spline* de troisième ordre afin qu'elles aient des voxels isotropes. Les images TEP ont été exprimées en SUV, rééchantillonnées en voxels de $3mm \times 3mm \times 3mm$, et une discrétisation de taille de *bins* fixe de $0,3125SUV$ a été utilisée [326]. Les images TDM exprimées en HU ont été rééchantillonnées à des voxels de $1mm \times 1mm \times 1mm$ et une taille de bin fixe de $10HU$ a été utilisée. Les images T1 et T2-SG prétraitées ont été rééchantillonnées en voxels de $1mm \times 1mm \times 1mm$ et la taille des *bins* a été fixée de manière à définir 128 *bins* entre la valeur minimale et la valeur maximale dans l'ensemble de la cohorte, ce qui correspond à $0,1668$ pour les images T1 et à $0,05611$ pour les images T2-SG. Les voxels dont les valeurs sont inférieures à $-230HU$ ou supérieures à $600HU$ ont été exclus de la ROI en TDM afin de limiter la présence de voxels d'air et d'os dans la ROI, tout en conservant les valeurs éventuellement associées aux hypodensités tumorales et aux calcifications.

Des caractéristiques radiomiques de premier ordre, de la GLCM, la GLDM, la GLRLM, et la NGTDM ont été extraites localement dans toutes les ROIs à l'aide d'une fenêtre glissante de $9 \times 9 \times 9$ voxels, ce qui a permis d'obtenir 308 cartes de caractéristiques radiomiques par patient (77 cartes de caractéristiques par modalité).

Comme définie dans les équations (4.1) et (4.2), la valeur moyenne dans les ROIs a été calculée pour chaque caractéristique afin d'obtenir deux vecteurs de caractéristiques agrégées de 154 composantes chacun par patient, un vecteur provenant des cartes de caractéristiques TEP/TDM (composé de 77 caractéristiques TEP et de 77 caractéristiques TDM), et un autre provenant des cartes de caractéristiques IRM (composé de 77 caractéristiques T1 et de 77 caractéristiques T2-SG). En outre, des caractéristiques de volume et de forme ont été calculées à partir du masque de segmentation rééchantillonné de la TDM (voxels de $1mm \times 1mm \times 1mm$), produisant 14 caractéristiques supplémentaires qui ont été ajoutées aux caractéristiques TEP/TDM et IRM pour produire deux vecteurs de 168 caractéristiques.

En utilisant une fenêtre glissante plus petite que la totalité de la ROI, il est probable de manquer certaines informations impliquant des voxels qui sont à une distance supérieure à la distance maximale dans la fenêtre. Pour ces caractéristiques, la fenêtre glissante forme des bornes limitant la mesure. Cela peut toutefois être vu comme un avantage plutôt qu'un inconvénient dans la pratique. En effet, le calcul des caractéristiques directement à partir de la ROI peut conduire à des corrélations élevées avec le volume ou la forme de la tumeur [107, 246, 327-329], alors que leur calcul à l'échelle du voxel permet de l'éviter [329]. Afin d'estimer la potentielle perte d'information liée à l'extraction locale, nous avons calculé la valeur absolue maximale du coefficient de corrélation de Pearson, $|r_{max_c}|$, pour chaque caractéristique

classique c par rapport aux caractéristiques calculées à l'échelle du voxel agrégées et aux caractéristiques de volume et de forme. Inversement, nous avons également calculé $\left|r_{max_j}\right|$ pour toute caractéristique j de $\boldsymbol{G}$, afin d'évaluer l'apport potentiel d'information lors de l'extraction à l'échelle du voxel.

Toutes les caractéristiques utilisées dans ce travail sont répertoriées avec leur définition à l'adresse suivante :
https://pyradiomics.readthedocs.io/ [93].

4.2.1.4 . Classification probabiliste

Dans cette section, les caractéristiques TEP/TDM et IRM ont été utilisées séparément. Nous avons entraîné deux modèles distincts afin de tester notre méthode dans deux contextes différents. L'objectif était de déterminer si les modèles étaient basés sur des zones communes, des zones spécifiques aux informations portées par chaque modalité, ou une combinaison des deux.

Réduction de la multicolinéarité

Les caractéristiques ont d'abord été sélectionnées de façon non supervisée (Section 3.2.4.2). De nombreuses caractéristiques radiomiques peuvent être fortement corrélées, voire colinéaires. En utilisant la corrélation de Pearson par paires sur les ensembles de caractéristiques TEP/TDM et IRM, un seuil sur la valeur absolue du coefficient de corrélation $|r|$ a été initialisé à 1. Tant qu'il y avait une multicollinéarité parfaite dans les données (déterminant de la matrice de corrélation de Pearson égal à 0), ce seuil a été diminué itérativement par pas de $0,001$. Au cours de ce processus, si deux caractéristiques étaient corrélées de telle sorte que leur valeur $|r|$ dépassait le seuil, la caractéristique présentant la valeur de $|r|$ moyenne la plus élevée avec les autres caractéristiques était supprimée. Ensuite, la sélection des caractéristiques a été effectuée en calculant le VIF. Les caractéristiques hautement redondantes (multicolinéaires) ont été éliminées en supprimant celles ayant le VIF le plus élevé de manière itérative, jusqu'à ce que le VIF maximum soit inférieur à 10 dans l'ensemble des caractéristiques [193].

Modélisation multivariée

Dans la suite de ce chapitre, les modèles basés sur les images TEP/TDM et IRM sont respectivement désignés par M1 et M2.

Les caractéristiques résultantes de l'étape de réduction de la multicollinéarité ont ensuite été sélectionnées via la méthode supervisée SFS (Section 3.2.4.2) en optimisant l'ASB (Section 3.2.3.2). Une régression logistique avec l'entropie croisée équilibrée comme fonction de perte et une régularisation LASSO (équation (3.32)) a été utilisée pour modéliser la probabilité d'apparition de métastases pulmonaires. L'algorithme de descente de coordonnées déterministe Liblinear a été utilisé [330], avec une tolérance de 10^{-4} ou un nombre maximal d'itérations égal à 100 comme critères d'arrêt.

Une grille de recherche (Section 3.2.4) a été utilisée pour déterminer le terme de régularisation optimal C du LASSO et le nombre de caractéristiques à conserver lors de la sélection SFS. Pour rappel, $C = 1/\alpha$ correspond à l'inverse de la force de régularisation. Ainsi, une valeur plus faible de C signifie une régularisation plus importante. Dix valeurs ont été définies pour C, de $0,1$ à 100 sur une échelle logarithmique de base 10. Pour chaque valeur de C, la procédure SFS a été effectuée par validation croisée stratifiée à 5 plis répétée 200 fois. Les données d'apprentissage ont été utilisées pour standardiser [1] les caractéristiques par un *z-score* à chaque itération de la procédure de validation croisée. La moyenne et l'écart-type du score ASB ont été enregistrés avec le sous-ensemble de caractéristiques associé. Le paramètre C et le sous-ensemble de caractéristiques retenus ont été sélectionnés manuellement sur la base d'un compromis entre la maximisation du score ASB moyen, et la minimisation de son écart-type et de son coefficient de variation, tout en favorisant les modèles les plus parcimonieux (peu de caractéristiques) et les plus régularisés (faible C).

Pour évaluer si notre approche donnait des résultats trop optimistes en adaptant les données au bruit, un test de permutations a été effectué (Section 3.2.4.1). L'ensemble de la chaîne d'apprentissage automatique, y compris la sélection SFS et l'optimisation de C via la grille de recherche, a été répété 200 fois en effectuant des permutations aléatoires des *labels* des patients à chaque itération. Pour chaque itération, le meilleur score ASB moyen a été enregistré, ce qui a conduit à une distribution nulle des 200 meilleurs scores de validation croisée. Pour rappel, cette distribution montre la performance estimée lorsqu'il n'y a pas de relation réelle entre les caractéristiques et les *labels*. Sur la base de cette distribution nulle, la p-*value* empirique associée au score ASB observé pour les modèles obtenus avec les *labels* corrects a pu être calculée (équation (3.29)).

En plus du score ASB, la perte standard du score de Brier, la courbe ROC moyenne, et son aire sous la courbe (AUC) associée à son écart-type ont été calculées comme figures de mérite.

Bagging et comparaison avec des biomarqueurs habituels

Pour construire les modèles, 1000 échantillons *bootstrap* ont été tirés. Les coefficients des fonctions de décision des 1000 modèles ont été moyennés pour obtenir les fonctions de décision linéaires finales de M1 et M2, désignées par D_{M1} et D_{M2}.

1. La régularisation LASSO impose une contrainte sur la valeur des coefficients. Cette valeur est dépendante de l'échelle et de l'amplitude de chaque caractéristique dans l'ensemble de données. Il est donc nécessaire de les centrer et les réduire, afin que la régularisation ait la même force pour toutes les caractéristiques. Pour les mêmes raisons, nous devons exprimer les caractéristiques dans des échelles comparables pour les ordonner selon leur importance grâce à leurs coefficients appris.

Les modèles M1 et M2 ont été comparés à des biomarqueurs habituels (Tableau 2.1). Pendant le rééchantillonnage *bootstrap* de la procédure de *bagging*, les données d'entraînement ont été utilisées à chaque itération pour standardiser les caractéristiques en utilisant un *z-score*, et les AUC ont été calculées sur la base des échantillons OOB pour les prédictions des modèles ainsi que pour le volume tumoral anatomique (*anatomical tumor volume* (ATV)), le SUVmax, le volume tumoral métabolique (*metabolic tumor volume* (MTV)) et la glycolyse totale de la lésion (*total lesion glycolysis* (TLG)).

Cartes de décision

Pour fournir la décision des modèles, les coefficients $\boldsymbol{\beta}$ appris s'appliquent à la valeur standardisée des caractéristiques agrégées auxquelles ils sont associés. L'écart-type σ_{G_j} et la moyenne μ_{G_j} sur tous les patients pour chaque caractéristique agrégée j de $\boldsymbol{G}$ impliquée dans les modèles finaux M1 et M2 ont donc été utilisés pour standardiser les cartes de caractéristiques correspondantes pour chaque patient i dans l'ensemble de données tel que :

$$z_j^{(i,v)} = \frac{x_j^{(i,v)} - \mu_{G_j}}{\sigma_{G_j}} \tag{4.13}$$

avec $z_j^{(i,v)}$ la valeur standardisée de la j-ième caractéristique au voxel v pour le patient i, et $x_j^{(i,v)}$ sa valeur originale.

Après avoir rééchantillonné toutes les cartes de caractéristiques sur une grille commune de voxels de $1mm \times 1mm \times 1mm$ grâce à une interpolation B-spline de troisième ordre, les RDMs $DV_{M1}^{(i)}$ et $DV_{M2}^{(i)}$ ont été obtenues pour chaque patient i en rétroprojetant au niveau du voxel les coefficients appris.

4.2.2 . Résultats

4.2.2.1 . Traitement et représentation des images

Prétraitement des images IRM

La Figure 4.4 montre un exemple de coupe pour un patient, avec le champ de biais estimé par l'algorithme N4ITK superposé à l'image T1 brute correspondante (a), l'image T1 brute seule (b), et l'image T1 corrigée (c). D'un point de vue qualitatif, la correction N4ITK a amélioré les images T1 par rapport à leur version brute. Ce résultat est plus visible dans le tissu adipeux, avec un signal visuellement plus homogène après la correction. L'image T2-SG pour ce patient au même emplacement de coupe (d) montre qu'il n'y a pas de signal significatif en dehors de la ROI tumorale pour estimer le champ de biais pour cette séquence.

L'impact quantitatif de la correction sur les coefficients de variation intratissus $CV_{graisse}$ et CV_{muscle}, le contraste intertissus CIT, et le coefficient

de variation conjointe CJV pour l'ensemble des patients est montré en Figure 4.4 (e). La valeur médiane sur l'ensemble des patients a été significativement réduite après la correction N4ITK pour $CV_{graisse}$ (p-*value* $< 0,0001$), CV_{muscle} (p-*value* $= 0,0058$), et CJV (p-*value* $< 0,0001$), tandis que le contraste CIT n'a pas significativement baissé (p-*value* $= 0,0788$). Ces résultats suggèrent que l'algorithme N4ITK a réduit les inhomogénéités tout en préservant la séparation intertissus et le contraste global.

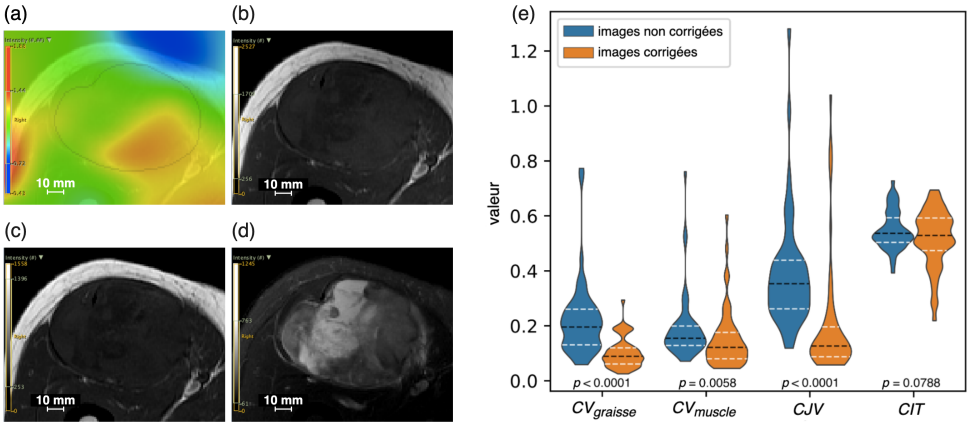


Figure 4.4 – Résultats de la correction du champ de biais en T1 par l'algorithme N4ITK. (a) Champ de biais estimé superposé à l'image T1 brute correspondante pour un patient. (b) Image T1 corrigée correspondante. (c) Image T1 brute correspondante. (d) Image T2-SG au même emplacement de coupe. (e) Diagrammes en violon de l'impact quantitatif de la correction du champ de biais par l'algorithme N4ITK sur l'ensemble des patients selon $CV_{graisse}$, CV_{muscle}, CJV, et CIT.

Calcul des caractéristiques

Quatre exemples de cartes de caractéristiques pour un patient sont montrées dans la Figure 4.5 avec leur valeur moyenne dans la ROI, mettant en évidence plusieurs motifs différents.

La figure 4.6 montre les coefficients $|r_{max_c}|$ et $\left|r_{max_j}\right|$, pour toute caractéristique classique c calculée directement au niveau de la ROI par rapport à toute caractéristique j de $\boldsymbol{G}$ utilisées dans cette étude et inversement. La plupart des caractéristiques d'une approche avaient au moins une caractéristique équivalente dans l'autre méthode, avec un $|r_{max}| \geq 0,75$. Bien que la quantité d'information potentiellement perdue ($|r_{max_c}| < 0,75$) lors de l'extraction à l'échelle du voxel soit plus importante que l'information potentiellement gagnée ($\left|r_{max_j}\right| < 0,75$), les corrélogrammes associés aux variables concernées montrent leur redondance, formant des groupes de corrélations et réduisant de ce fait la quantité d'information réellement omise. Indépendamment de leur importance pour la décision des modèles entraînés, ces résultats suggèrent que l'approche à l'échelle du voxel n'a pas perdu d'informations substantielles par rapport à l'approche radiomique traditionnelle.

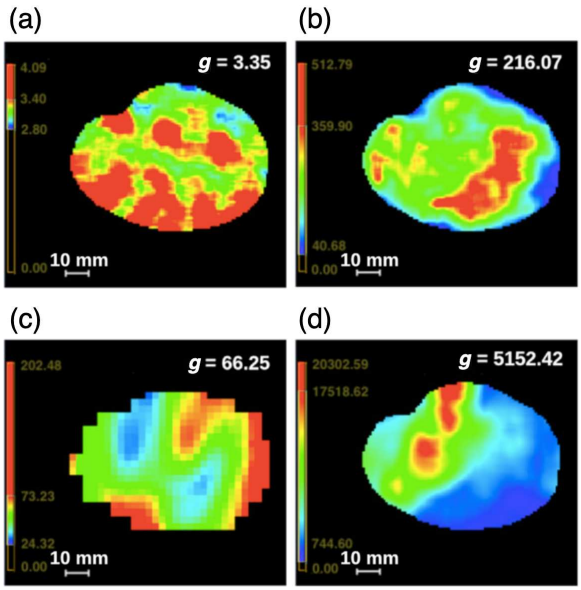


Figure 4.5 – Exemple de cartes de caractéristiques radiomiques. Comme les modèles ont été entraînés en prenant comme entrées la valeur moyenne à l'intérieur de la ROI, il n'était pas nécessaire de rééchantillonner toutes les cartes de caractéristiques sur une grille commune. Les cartes de caractéristiques avaient donc des résolutions spatiales différentes ($3mm \times 3mm \times 3mm$ pour la TEP, et $1mm \times 1mm \times 1mm$ pour la TDM et l'IRM). (a) Entropie de premier ordre en TDM. (b) Non-uniformité du niveau de gris (*gray level non-uniformity* (GLNU)) de la GLDM en IRM T1. (c) Contraste de la GLCM en TEP. (d) Grandes longueurs de niveau de gris élevé (long run high gray level emphasis (LRHGLE)) de la GLRLM en IRM T2-SG. Figure adaptée d'Escobar et al. [61].

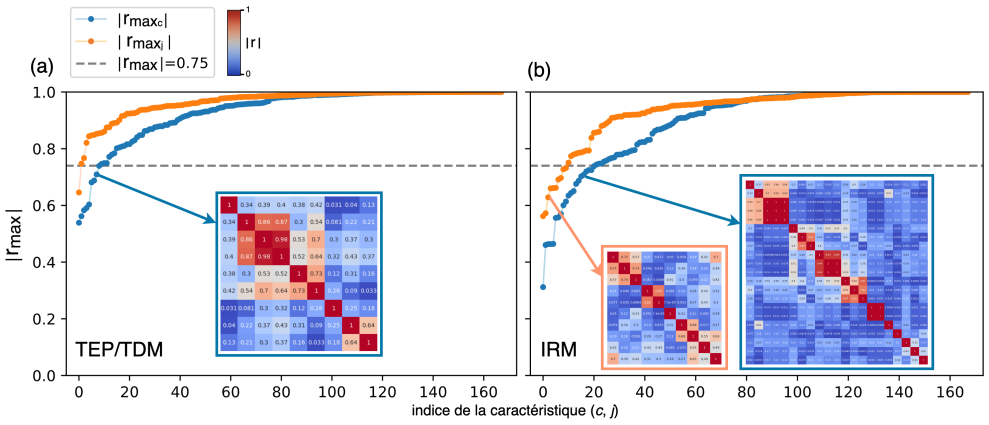


Figure 4.6 – Évaluation de la perte et du gain potentiels d'information lors de l'extraction des caractéristiques à l'échelle du voxel par rapport à l'extraction classique à partir de la ROI en TEP/TDM (a) et IRM (b). Les corrélogrammes associés montrent la valeur absolue des coefficients de corrélation de Pearson comparant deux à deux les caractéristiques ayant un $|r_{max}| < 0,75$.

4.2.2.2 . Classification probabiliste

Réduction de la multicolinéarité

Un total de 25 caractéristiques (dont 4 caractéristiques de forme) et 26 (dont 4 caractéristiques de forme) sur 168 ont été sélectionnées respectivement à partir de la TEP/TDM et de l'IRM après réduction de la multicollinéarité. La valeur du VIF des caractéristiques sélectionnées est indiquée dans le Tableau S1 (Annexe I). La Figure S1 (Annexe II) représente les matrices de corrélation de Pearson de ces caractéristiques pour la TEP/TDM (a) et l'IRM (b). Comme attendu, plusieurs caractéristiques sont redondantes au niveau de la ROI, et seule une fraction d'entre elles a été retenue après réduction de la multicollinéarité.

Modélisation multivariée

Les distributions nulles des 200 modèles aléatoires issus des tests de permutations sont présentées en Figure 4.7 pour les modèles M1 (a) et M2 (b), avec les performances réelles en validation croisée. Via les grilles de recherche, 5 caractéristiques ont été retenues pour la TEP/TDM et l'IRM avec $C = 2,2$. Les scores ASB moyens associés ($\pm$ 1 écart-type) étaient de $0,872 \pm 0,056$ (p-*value* $= 0,005$) pour la TEP/TDM et de $0,838 \pm 0,065$ (p-*value* $= 0,035$) pour l'IRM, significativement plus élevés que ceux des modèles aléatoires dans les deux cas. Les résultats de la construction des modèles M1 et M2 sont résumés dans le Tableau 4.1.

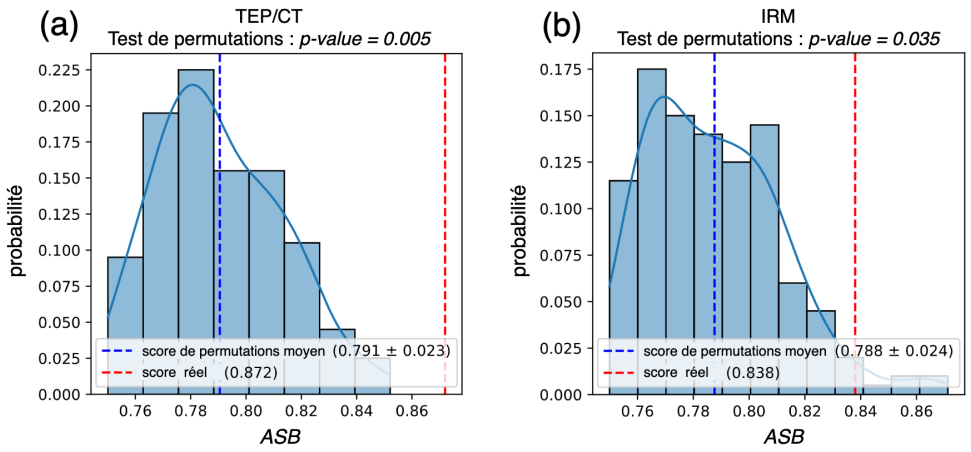


Figure 4.7 – (a) TEP/TDM. (b) IRM. Distributions des scores ASB des tests de permutations pour les paramètres de construction des modèles M1 et M2. Figure adaptée d'Escobar et al. [61].

Tableau 4.1 – Performances en validation croisée pour l'optimisation par grille de recherche des paramètres de régularisation LASSO et de sélection des caractéristiques. Tableau adapté d'Escobar et al. [61].

Paramètres de construction des modèles	**M1**	**M2**
C	$2,2$	$2,2$
Nombre de caractéristiques sélectionnées	5 (1 caractéristique de forme)	5 (1 caractéristique de forme)
ASB ($\pm$ 1 écart-type)	$0,872 \pm 0,056$	$0,838 \pm 0,065$
Brier *score loss* ($\pm$ 1 écart-type)	$0,133 \pm 0,057$	$0,167 \pm 0,068$
ROC AUC ($\pm$ 1 écart-type)	$0,910 \pm 0,094$	$0,853 \pm 0,115$

Bagging et comparaison avec des biomarqueurs habituels

Les fonctions de décision linéaires D_{M1} et D_{M2} sont reportées dans les équations (4.14) et (4.15) avec l'écart-type associé à chaque caractéristique sur les 1000 échantillons *bootstrap*.

$$\begin{aligned} D_{M1} = &- 0,653(\pm 0,623) \times TDM_{GLDM_{LDLGLE^*}} \\ &+ 1,711(\pm 0,745) \times TEP_{1^{er}ordre_{minimum}} \\ &+ 2,655(\pm 0,907) \times TEP_{1^{er}ordre_{skewness}} \\ &+ 1,496(\pm 0,600) \times TEP_{GLCM_{corrélation}} \\ &+ 0,953(\pm 0,710) \times FORME_{élongation} \\ &+ 0,673(\pm 0,428) \end{aligned} \tag{4.14}$$

$$\begin{aligned} D_{M2} = &- 1,325(\pm 0,735) \times T1_{1^{er}ordre_{énergie}} \\ &- 1,729(\pm 0,698) \times T1_{GLDM_{SDLGLE^{**}}} \\ &+ 1,032(\pm 0,470) \times T2\text{-}SG_{1^{er}ordre_{RMS^{***}}} \\ &+ 1,895(\pm 0,731) \times T2\text{-}SG_{1^{er}ordre_{énergie}} \\ &+ 1,197(\pm 0,577) \times FORME_{sphéricité} \\ &+ 0,857(\pm 0,444) \end{aligned} \tag{4.15}$$

Les fonctions de densité de probabilité (*probability density function* (PDF)) des distributions OOB pour l'AUC pour les prédictions de M1 et M2, pour l'ATV, le SUVmax, le MTV, et le TLG sont résumées dans le Tableau 4.2 et montrées dans la Figure 4.8.

*Large dépendance de faible niveau de gris (*large dependence low gray level emphasis* (LDLGLE)).

**Petite dépendance de faible niveau de gris (*small dependence low gray level emphasis* (SDLGLE)).

***Moyenne quadratique (*root mean square* (RMS)).

Tableau 4.2 – Résumé des AUC OOB pour les prédictions de M1 et M2, pour l'ATV, le SUVmax, le MTV, et le TLG. Tableau adapté d'Escobar et al. [61].

OOB AUC	**M1**	**M2**	**ATV**
Moyenne ($\pm$ 1 écart-type)	$0,883 \pm 0,086$	$0,840 \pm 0,090$	$0,691 \pm 0,107$
95% CI	$[0,660\,;1,000]$	$[0,622\,;0,974]$	$[0,472\,;0,890]$
Maximum PDF (mode)	$0,908$	$0,858$	$0,703$

OOB AUC	**SUVmax**	**MTV**	**TLG**
Moyenne ($\pm$ 1 écart-type)	$0,806 \pm 0,094$	$0,594 \pm 0,117$	$0,728 \pm 0,110$
95% CI	$[0,612\,;0,971]$	$[0,361\,;0,818]$	$[0,501\,;0,931]$
Maximum PDF (mode)	$0,789$	$0,569$	$0,749$

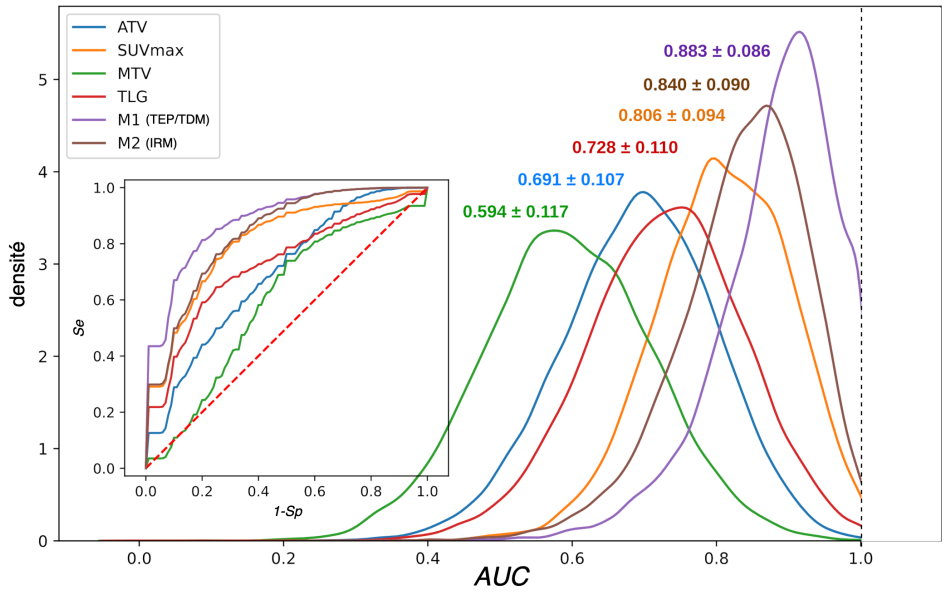


Figure 4.8 – PDFs des distributions OOB pour l'AUC pour les prédictions de M1 et M2, pour l'ATV, le SUVmax, le MTV, et le TLG. Les courbes ROC moyennes associées à ces distributions sont montrées dans la sous-figure de gauche. Figure adaptée d'Escobar et al. [61].

Cartes de décision

Des exemples de coupes représentatives de RDMs DV_{M1} (a) et DV_{M2} (b), d'images TEP (c), TDM (d), T1 (e), et T2-SG (f) sont présentés en Figure 4.9. Confortées par les équations (4.14) et (4.15), les RDMs ont révélé des motifs prédictifs interprétables et cohérents entre les patients. En particulier, les cartes DV_{M1} ont mis en évidence des sous-régions tumorales localisées de forte fixation de FDG ainsi que de larges régions homogènes présentant un faible métabolisme. Elles ont également fait ressortir des sous-

régions hypodenses. Les sous-régions mises en évidence par les cartes DV_{M2} étaient globalement bien colocalisées avec celles hypodenses et non avides en FDG dans les cartes DV_{M1}. Ce sont des sous-régions de faible intensité sur les images T1 et de de haute intensité sur les images T2-SG. Ces sous-régions correspondent à de la nécrose suspectée. Dans les cartes DV_{M1}, les sous-régions caractérisées par une forte fixation de FDG localisée et hétérogène incluaient la plupart du temps le voxel associé au SUVmax.

Le risque pour le patient 1 a été bien prédit avec une probabilité élevée pour M1 $(0,85)$ et M2 $(0,88)$, ce qui est cohérent avec le SUVmax élevé $(29,99)$ et la grande sous-région nécrotique observée dans la tumeur. Avec un risque bien prédit pour M1 $(0,82)$ et M2 $(0,58)$, l'image TEP du patient 2 a montré une sous-région nécrotique plus petite mais un SUVmax élevé $(27,80)$, ce qui est cohérent avec le fait que la probabilité prédite est plus faible pour M2 que pour M1. Le volume nécrotique des patients 2 et 3 étaient comparables, avec un SUVmax plus faible $(5,37)$ pour le patient 3. Cela pourrait expliquer la probabilité prédite plus faible pour M1 $(0,02)$, que pour M2 $(0,67)$ donnant un faux positif. M1 $(0,27)$ et M2 $(0,20)$ ont bien prédit de faibles probabilités, proches, pour le patient 4. La probabilité prédite pour le patient 5 a conduit à un faux négatif pour M1 $(0,08)$ et à un vrai positif pour M2 $(0,57)$, toujours en accord avec le SUVmax relativement faible $(4,21)$ de ce patient. Enfin, les probabilités prédites pour le patient 6 avec un SUVmax de $7,15$ et un grand volume nécrotique ont conduit à un vrai positif pour M1 $(0,67)$, illustrant la supériorité de M1 sur le SUVmax dans ce cas malgré leur cohérence. Résumés dans le Tableau 4.3, ces résultats suggèrent que deux informations biologiques capturées par des motifs locaux dans les images étaient associées au risque d'apparition de métastases pulmonaires dans cet ensemble de données : le développement de la nécrose dans la tumeur et son métabolisme élevé en glucose.

En outre, pour des valeurs ou combinaisons intermédiaires du SUVmax et du volume de nécrose suspectée, ou lorsque ce dernier n'est pas franc à l'IRM, la forme sphérique [2] de la tumeur est apparue comme un motif global prédictif supplémentaire.

2. Que ce soit M1 qui utilise une caractéristique nommée « élongation », ou M2 employant la « sphéricité », tous deux pénalisent les tumeurs sphériques. En effet, bien que cela paraisse contre-intuitif, $FORME_{élongation} = \sqrt{\lambda_2[mm]/\lambda_1[mm]}$ correspond à la racine carrée de la deuxième valeur propre spatiale sur la première, traduisant une forme sphérique pour les valeurs élevées (proches de 1), et une forme plus allongée pour les valeurs plus faibles (proches de 0). Cette caractéristique doit donc être interprétée plutôt comme la « non-élongation ». D'autre part, $FORME_{sphéricité} = \sqrt[3]{36\pi \times V[mm^3]^2}/A[mm^2]$ est une mesure de rondeur relative à une sphère, avec des valeurs intuitives, plus élevées (proches de 1) pour les formes sphériques et plus faibles (proches de 0) dans tous les autres cas.

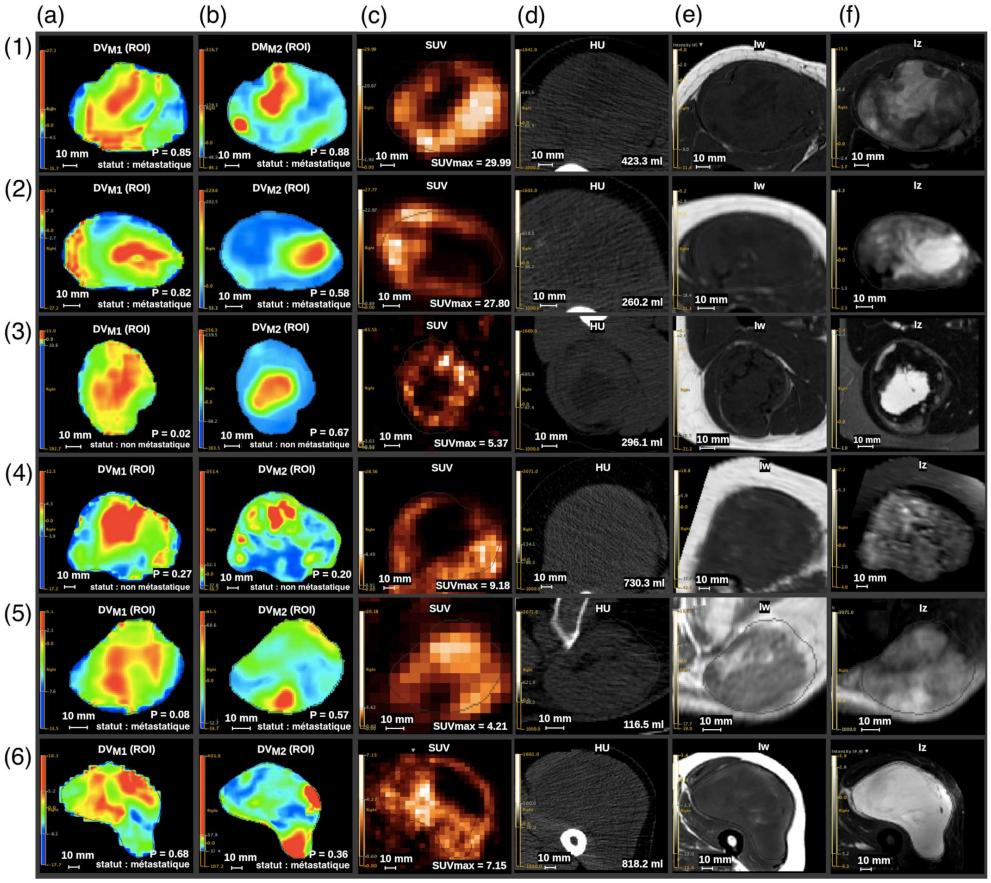


Figure 4.9 – Exemples de coupes de RDMs DV_{M1} (a) et DV_{M2} (b), d'images TEP (c), TDM (d), T1 (e), et T2-SG (f) pour six patients (1-6). Figure adaptée d'Escobar et al. [61].

Tableau 4.3 – Résumé des observations des RDMs D_{M1} et D_{M2} analysées conjointement avec les images TEP/TDM et IRM.

Patient	P_{M1}	P_{M2}	SUVmax	Volume nécrotique (grand/intermédiare/petit)	Forme en 3D (allongée/intermédiaire/isotrope)	Métastase (oui/non)
1	0,85 (TP)	0,88 (TP)	29,99	grand	allongée	oui
2	0,82 (TP)	0,58 (TP)	27,80	petit	allongée	oui
3	0,02 (TN)	0,67 (FP)	5,37	intermédiaire	allongée	non
4	0,27 (TN)	0,20 (TN)	9,18	grand (peu visible en IRM)	intermédiaie	non
5	0,08 (FN)	0,57 (TP)	4,21	petit	isotrope	oui
6	0,67 (TP)	0,36 (FN)	7,15	grand (peu visible en IRM)	isotrope	oui

4.3 . Le modèle substitut global : reformuler un modèle simple

Le fait qu'un événement ou un phénomène puisse être expliqué par diverses causes est appelé « l'effet Rashomon ». Il tire son nom du film « Rashōmon » d'Akira Kurosawa, dans lequel un meurtre est décrit par quatre témoins, de quatre manières contradictoires mais toutes aussi réalistes, donc probables [331]. En ML, il désigne l'existence de différents modèles équivalents en performances [164, 332-334]. C'est « l'ensemble Rashomon ». Deux modèles d'un tel ensemble peuvent différer par leur forma-

lisme (eg, régression logistique, forêt aléatoire, ANN), ou bien parce qu'ils approchent le problème en utilisant des informations différentes mais toutes aussi prédictives (eg, charge tumorale ou métabolisme du glucose, phénotype en imagerie ou caractérisation génomique).

En outre, comme nous l'avons vu en Section 2.2.2 du Chapitre 2, une même information sémantique peut être mesurée de différentes façons. En imagerie TEP au FDG, nous pouvons citer le SUVmax, le SUVmean, le SUVpeak, ou encore le MTV et le TLG, partageant un objectif commun, l'estimation du métabolisme du glucose de la tumeur, mais le mesurant sous des aspects différents. Enfin, de la même manière que plusieurs modèles formulés différemment peuvent produire des espaces de décision similaires (Figure 3.18), différentes formulations mathématiques pour une variable peuvent être plus ou moins équivalentes en pratique lorsqu'elles sont appliquées aux données. Vanderhoek et al. ont étudié l'effet de différentes formulations pour le SUVpeak et ont observé que bien qu'impactant l'estimation de la réponse au traitement au niveau individuel, elles étaient équivalentes à l'échelle d'une population de patients [71]. Ce phénomène se matérialise également par le niveau élevé de redondance dans un ensemble de différentes caractéristiques radiomiques (Section 4.2.2.2). Cela souligne l'importance de prendre conscience des différentes façons dont un concept peut être évalué.

Une particularité des caractéristiques radiomiques classiques et profondes est que malgré leur opacité, utilisées en groupe, elles confèrent une certaine flexibilité à la modélisation. Laissant « parler les images », une chaîne d'analyse générique peut alors être appliquée à de nombreux problèmes différents de manière efficace si de l'information utile est présente dans les images. Dans le cas de la recherche de biomarqueurs en imagerie oncologique, si une forte performance prédictive sur la distribution qui a généré les données est généralement une condition nécessaire pour qu'un modèle soit utile, elle n'est souvent pas suffisante. Il est nécessaire que le modèle soit robuste aux variations courantes des données ou aux différences dans les instruments de mesure, et invariant aux caractéristiques sensibles. Il doit en outre s'accorder avec les connaissances et l'usage des utilisateurs finaux, de l'équipe de soins, ou des experts médicaux [72, 334].

Reformuler spécifiquement des variables qui codent une information proche de celle capturée par le modèle est une façon de tirer parti de l'ensemble Rashomon. L'interprétation peut en effet guider la formulation de biomarqueurs candidats ou aider à les combiner pour produire une signature. La construction d'autres modèles, plus simples, pourrait alors favoriser leur compréhension, leur utilisation en routine clinique, voire améliorer leur robustesse.

> *« Essentiellement, tous les modèles sont faux,*
> *mais certains sont utiles. »*
>
> George Box, 1987 [335]

4.3.1 . Matériels et méthodes

À partir de notre lecture des RDMs et des équations des modèles, et pour évaluer la validité de nos interprétations, nous avons construit un modèle de substitution simplifié à partir de M1, nommé M1'.

L'objectif était d'approcher la fonction de prédiction de M1 aussi fidèlement que possible avec la fonction de prédiction du modèle de substitution M1', avec la contrainte que M1' soit interprétable non seulement dans sa formulation comme M1 (régression logistique), mais aussi selon ses caractéristiques d'entrée.

Nous avons donc formulé des caractéristiques plus simples et plus facilement interprétables reflétant le développement nécrotique à l'intérieur du volume anatomique de la tumeur avec des images TEP/TDM. Nous avons calculé le volume absolu (V) et le volume relatif ($rV = V/ATV$) caractérisés soit par un faible métabolisme ($< 40\% SUVmax$ en TEP), soit par un signal hypodense ($< 20HU$ ou $< 30HU$ en TDM), soit par une mesure combinée de ces deux motifs en utilisant les opérateurs union ($\cup$) ou intersection ($\cap$). La transformation logarithmique de base 10 a également été appliquée à ces nouvelles caractéristiques ainsi qu'à l'ATV, au SUVmax, au MTV, au TLG, et aux caractéristiques de forme, afin de permettre une plus grande flexibilité pour la modélisation, et tenir compte des distributions asymétriques.

Nous avons finalement construit M1' en entraînant un modèle logistique avec la sortie prédite de M1 comme cible, en suivant la même procédure d'apprentissage automatique mais en utilisant uniquement ces caractéristiques. Toutes les nouvelles caractéristiques sont répertoriées dans le Tableau S2 (Annexe III) avec leur définition.

Les prédictions OOB de M1 et M1' ont été comparées via les corrélations de Pearson (r_P) et de Spearman (r_S). La performance de M1' sur les vrais *labels* a été estimée grâce à l'AUC OOB.

4.3.2 . Résultats

De façon conforme à nos interprétations, trois caractéristiques ont été automatiquement sélectionnées pour approximer les prédictions de M1 : $\log_{10}(SUVmax)$, $\log_{10}(V_{<20HU \cup <40\% SUVmax})$, et $FORME_{élongation}$. La fonction de décision linéaire de *bagging* de M1', $D_{M1'}$, est rapportée dans l'équation (4.16) avec l'écart-type associé à ces caractéristiques sur les 1000 échantillons *bootstrap*.

$$\begin{aligned} D_{M1'} = &+ 3,243(\pm 1,251) \times \log_{10}(SUVmax) \\ &+ 2,070(\pm 0,745) \times \log_{10}(V_{<20HU \cup <40\% SUVmax}) \\ &+ 0,940(\pm 0,466) \times FORME_{élongation} \\ &- 0,468(\pm 0,482). \end{aligned} \tag{4.16}$$

Les corrélations de Pearson et Spearman entre les prédictions OOB de M1 et M1' était de $r_P = 0,874$ et $r_S = 0,800$ respectivement. L'AUC OOB moyenne pour M1' était de $0,830 \pm 0,089$ (95% CI $[0,620 ; 0,972]$), légèrement inférieure aux résultats pour le modèle original donnant $0,883 \pm 0,086$ (95% CI $[0,660 ; 1,000]$). Une comparaison des sorties des modèles est présentée dans la Figure 4.10.

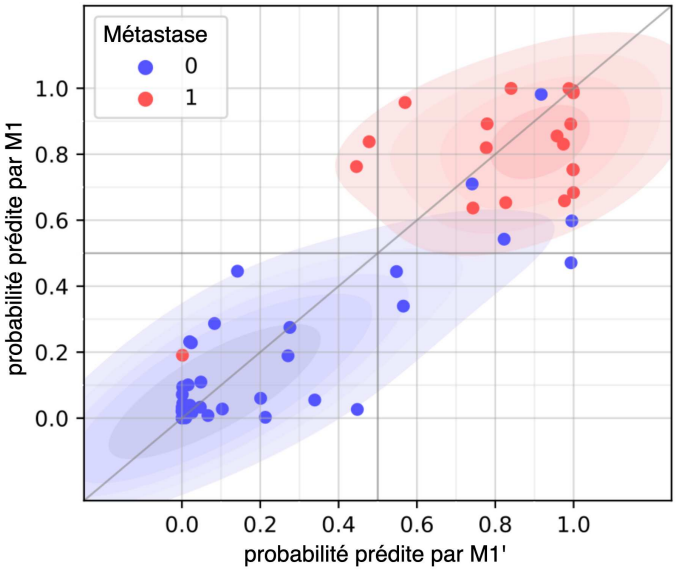


Figure 4.10 – Graphique conjoint de dispersion et d'estimation de densité comparant les sorties probabilistes de M1' et M1 sur l'ensemble de données. La couleur des points représente le *label* des patients correspondants (bleu : pas d'occurrence de métastase pulmonaire, rouge : occurrence de métastase pulmonaire). Figure adaptée d'Escobar et al. [61].

4.4 . Discussion

Dans cette étude, nous avons proposé une méthode pour identifier et caractériser les sous-régions tumorales qui déterminent les prédictions de modèles construits à l'aide de caractéristiques radiomiques prédéfinies. Notre approche est basée sur des caractéristiques calculées à l'échelle du voxel à l'aide d'une fenêtre glissante, moyennées ensuite dans la ROI pour la modélisation probabiliste ultérieure. La rétroprojection des coefficients du modèle dans l'espace distribué au niveau du voxel produit une carte de décision pour chaque patient.

Lors de l'utilisation de modèles linéaires généralisés tels que la régression logistique, ces cartes préservent la quantification probabiliste modélisée à l'échelle de la ROI. L'activation logistique σ de la valeur de décision moyenne des voxels dans la ROI, ajoutée à la combinaison linéaire des caractéristiques

qui ne peuvent pas être cartographiées, est égale à la probabilité modélisée d'appartenir à la classe positive pour chaque patient. Ainsi, les cartes de décision résultantes sont directement liées aux modèles qu'elles cartographient, sans approximation, et montrent partiellement la contribution marginale de chaque voxel dans la ROI du patient au risque modélisé d'apparition de métastases. Dans le cas de la combinaison de caractéristiques à l'échelle du voxel avec des caractéristiques sans formulation locale, comme les caractéristiques de forme dans notre cas, les cartes de décision n'expliquent qu'une partie du modèle. Néanmoins, en montrant les voxels qui contribuent le plus localement, et en les analysant conjointement avec les images d'entrée, les RDMs proposés augmentent l'interprétabilité des modèles.

Techniquement, notre approche est comparable à la carte d'activation de classe (*class activation map* (CAM)) associée aux modèles d'apprentissage profond [287]. En effet, le principe de la CAM est d'utiliser le *global average pooling* dans une architecture de CNN pour calculer la moyenne de tous les voxels dans les dernières cartes de caractéristiques afin de produire la sortie basée sur une couche entièrement connectée unique. Une fois le CNN entraîné, la rétroprojection des coefficients linéaires de cette couche avant la fonction d'activation permet d'obtenir les CAMs. Dans le cas de classifications binaires, une couche entièrement connectée unique avec une activation sigmoïde correspond à une régression logistique, ce qui rend notre méthode proche de l'approche CAM en apprentissage profond. L'utilisation de la valeur moyenne des caractéristiques sur l'ensemble des voxels dans les (dernières) cartes de caractéristiques permet d'entraîner les modèles sur des images ou des ROI de différentes tailles. Comme nous l'avons vu en Section 3.1.2, elle favorise également l'invariance des modèles par translation. Il s'agit de différences essentielles par rapport aux autres approches de saillance qui n'utilisent pas le *global average pooling*, et relient une ou plusieurs couches entièrement connectées à tous les voxels des dernières cartes de caractéristiques (*flatten*). En outre, lors de l'utilisation de plusieurs couches de classification, l'utilisation des gradients rétropropagés n'est pas sans risque lorsqu'on les interprète comme l'importance des voxels (Section 4.1.3) [310, 312].

Notre approche utilise une fenêtre glissante de dimensions choisies pour calculer les caractéristiques locales de chaque voxel des images d'entrée, donnant des cartes de décision de résolution relativement fine, directement comparables aux images d'entrée pour l'analyse conjointe et l'identification de sous-régions tumorales d'intérêt. Cette approche contraste avec la plupart des stratégies de sous-échantillonage utilisées dans l'apprentissage profond, qui produisent des cartes de saillance de résolution grossière. En termes d'interprétabilité, la sélection du champ de réception du modèle par le choix de la taille de la fenêtre glissante définit l'environnement local dans lequel le modèle capture les informations autour de chaque voxel.

Par rapport à l'extraction classique des caractéristiques radiomiques, l'impact du bruit aléatoire ou des biais sur l'extraction à l'échelle du voxel a été étudié par Bernatowicz et al. [336]. Les auteurs ont conclu que l'extraction des caractéristiques à l'échelle du voxel est plus affectée que le calcul des caractéristiques directement au niveau de la ROI, comme attendu. En effet, lors de l'extraction d'une caractéristique radiomique au niveau de la ROI pour un patient donné, tous les voxels de la ROI sont pris en compte pour donner la valeur scalaire de la caractéristique. Comme la fenêtre de notre approche est plus petite que la ROI, le nombre de voxels contribuant au calcul est plus faible, et donc le résultat est plus affecté par le bruit et les biais. Pour une comparaison équitable entre les deux approches, il faudrait comparer la moyenne des valeurs des caractéristiques calculées au niveau du voxel sur l'ensemble de la ROI (comme nous le faisons pour notre étape de modélisation) aux valeurs des caractéristiques directement calculées sur toute la ROI. On s'attend à ce qu'une telle agrégation lisse l'impact du bruit aléatoire et des biais. Par contre, lorsque nous rétroprojetons les fonctions de décision au niveau du voxel après la modélisation, nous revenons à un espace local plus sujet aux biais et au bruit. Cependant, cela peut être un atout car en mettant en évidence les motifs pouvant être dus au bruit ou à un biais, l'approche les rendra détectables, évitant ainsi une interprétation trompeuse, alors qu'ils pourraient demeurer non détectés dans une approche uniquement définie à l'échelle des ROIs.

Notre méthode présente également des similitudes avec les approches d'apprentissage par instances multiples *(multiple instance learning*), dans lesquelles chaque individu est représenté par un « sac d'instances » (« *bag of instances* ») [337]. Ici, le sac correspond à la ROI de la tumeur du patient, et les voxels à l'intérieur de celle-ci représentent les multiples instances.

Les RDMs utilisent des caractéristiques prédéfinies. Bien que cela puisse être vu comme un manque d'optimisation par rapport aux approches de DL, cela rend notre méthode plus adaptée aux petites cohortes difficilement compatibles avec l'entraînement de modèles profonds [145]. En imagerie médicale, les données sont très différentes de celles, telles qu'ImageNet, qui sont généralement utilisées pour évaluer les modèles de vision par ordinateur, dont les CNNs dominent actuellement le domaine [338]. Les formes irrégulières et le signal 3D des tumeurs, souvent avec des frontières floues et un faible niveau d'abstraction, ne ressemblent pas aux images 2D basées sur des objets et des concepts précis (eg, chien, chat, voiture, vélo, travail, repos, nage, course). De plus, les tailles des ensembles de données sont complètement différentes. Par exemple, alors qu'ImageNet est composé de 1,4 million d'images, les bases de données d'images médicales en comptent généralement quelques dizaines à quelques centaines. Alors qu'avec l'augmentation des données, un CNN peut améliorer et optimiser les caractéristiques qu'il apprend, les ca-

ractéristiques prédéfinies restent les mêmes. En radiomique, aucun avantage clair de l'utilisation de caractéristiques profondes apprises par un CNN par rapport aux caractéristiques prédéfinies ou inversement n'a encore été démontré dans la littérature, et les résultats dépendent ici aussi du contexte applicatif [339-346]. Plus important encore, les caractéristiques profondes, bien qu'optimisées pour un problème et définies localement, n'ont pas de définition mathématique explicite. Les méthodes de cartographie associées permettent donc de localiser l'information pertinente mais elles n'expliquent pas comment le signal est capturé. Grâce à la nature explicite des caractéristiques dont elles sont constitués, les RDMs sont mathématiquement bien définies pour chaque voxel à l'intérieur de la ROI, spatialement mais aussi quantitativement, ce qui facilite leur interprétation.

Notre méthode gère également les modèles reposant sur des caractéristiques sans signification locale. Elle est donc compatible avec des modèles impliquant même des caractéristiques ne relevant pas de l'imagerie, comme des caractéristiques cliniques ou génomiques. De tels modèles holistiques pourraient tout de même bénéficier de RDMs caractérisant les motifs locaux des images contribuant partiellement à leur décision.

L'objectif de la présente étude était de mettre en évidence des motifs interprétables, par opposition à la construction du modèle le plus performant compte tenu de toutes les informations disponibles, auquel cas nous aurions inclus les quatre modalités d'imagerie dans un seul modèle. La transparence globale, locale, et algorithmique est respectée par notre approche simple utilisant la régression logistique.

Nous pouvons nous demander pourquoi nous avons besoin de modèles linéaires généralisés pour les RDMs. Au-delà de leur interprétabilité, la raison principale est qu'à l'échelle du voxel, de nombreuses valeurs $x_j^{(i,v)}$ sont hors distribution par rapport à $g_j^{(i)}$ utilisée pour construire un modèle. Cela est dû à l'agrégation utilisant la moyenne dans la ROI, lissant les valeurs extrêmes de x_j en donnant la distribution de g_j (équation (4.1)). Par conséquent, lorsque nous le rétroprojetons à l'échelle des voxels, pour beaucoup d'entre eux nous extrapolons le modèle. Par exemple, pour une fonction non-monotone sur $\mathbb{R}$ mais avec une sortie strictement croissante en fonction de $g_j \in [a, b]$ (eg, dépendance partielle croissante, valeurs de Shapley et coefficients tangents positifs), mais décroissante pour toute valeur hors distribution de $g_j > b$ (eg, dépendance partielle décroissante, valeurs de Shapley et coefficients tangents négatifs), un voxel tel que $x_j^{(i,v)} > b$ pourrait apparaître comme diminuant la probabilité prédite pour le patient i. Cela induirait en erreur, car il l'aurait en fait augmentée dans l'espace agrégé, en augmentant la valeur de $g_j^{(i)}$ (Figure 4.11). Plus globalement, bien que préservant le rang de contribution des voxels dans la ROI s'il est monotone, un modèle non-linéaire ne quantifierait pas de façon fidèle la décision modélisée

au niveau du patient une fois distribuée à l'échelle du voxel. Ceci est dû au fait que d'après l'inégalité de Jensen, l'égalité $\overline{f(X)} = f(\overline{X})$ est vraie pour tout vecteur $X \in \mathbb{R}$ si et seulement si f est une fonction ni convexe, ni concave sur $\mathbb{R}$, donc affine[3] (équation (4.4)) [347]. Cela dit, comme nous l'avons vu en Section 3.2.2 du Chapitre 3, la plupart des approches de ML relient la sortie à l'entrée avec une relation non-linéaire en modifiant l'espace original avant de le séparer linéairement. En analyse d'images, nous pouvons considérer que l'espace original correspond au signal (eg, les valeurs de SUV des voxels à l'intérieur des ROIs), tandis que l'extraction et l'ingénierie des caractéristiques peuvent être considérées comme une forme de modification de cet espace original. De nombreuses caractéristiques radiomiques sont définies à l'aide d'équations non-linéaires, fonctions du signal des images. La non-linéarité est donc introduite à ce niveau dans notre travail, bien que nous utilisions ensuite un modèle linéaire généralisé. La sélection des caractéristiques est supervisée et dépend des images, de la tâche, ainsi que du classifieur utilisé ensuite. Elle correspond donc à une optimisation de la représentation pour une séparation par un hyperplan, même si ce n'est pas un apprentissage de celle-ci. De cette façon, l'opacité de la modélisation est isolée au niveau de la définition des caractéristiques finalement sélectionnées.

En termes de performance de prédiction, le modèle M1 (TEP/TDM) a donné une AUC plus élevée et une perte de score de Brier plus faible que le modèle M2 (IRM). Néanmoins, M2 a donné une AUC plus élevée que le SUVmax, qui était le biomarqueur « conventionnel » avec la meilleure performance. Bien que les performances des modèles radiomiques puissent être en partie expliqués par un surapprentissage probable, les résultats de M2 ainsi que les sous-régions nécrotiques mises en évidence par les RDMs associées soulignent l'importance de l'évaluation de la nécrose pour établir le pronostic de patients atteints de STS. Les sous-régions nécrotiques ont été observées de manière cohérente dans les RDMs du modèle M1 (TEP/TDM), qui a également affiché des valeurs de décision élevées dans les sous-régions présentant une forte fixation de FDG. Cela suggère que la combinaison de la nécrose et de régions tumorales hautement métaboliques au bilan d'extension est hautement prédictive du risque d'apparition de métastases au cours du suivi. Cette interprétation a été renforcée par la conception du modèle substitut plus simple M1', dans lequel le SUVmax et le volume tumoral hypodense ou non métaboliquement actif ont été automatiquement sélectionnés pour produire des résultats proches de ceux obtenus avec M1 (avec une caractéristique de forme commune mesurant l'isotropie de la tumeur). En raison de la

3. Dans le contexte du ML, nous utilisons le terme « linéaire » pour faire référence aux modèles s'ajustant aux données en utilisant des hyperplans. En tant que tel, un hyperplan peut avoir un biais β_0 non nul. Tel que définie en algèbre linéaire, la fonction associée à un « modèle linéaire » est en fait une fonction affine.

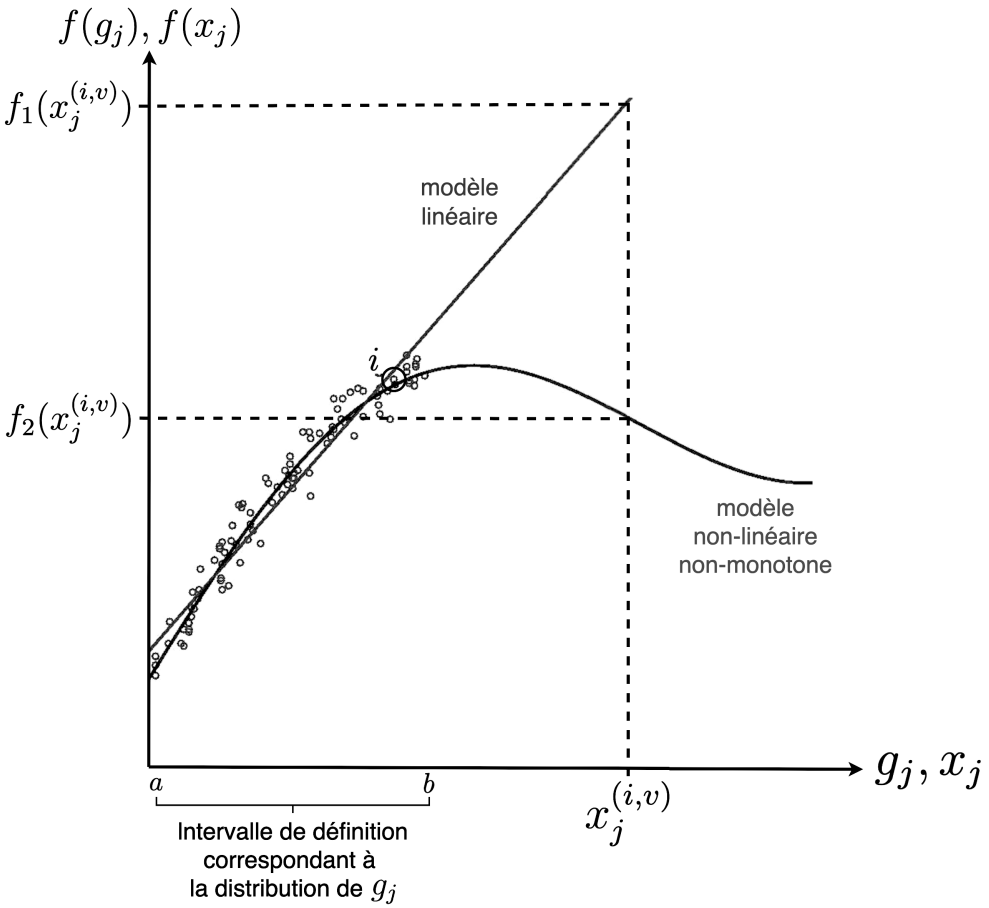


Figure 4.11 – Illustration d'un impact potentiel de l'utilisation d'un modèle non-linéaire (f_2) entraîné à partir d'une variable x_j moyennée donnant g_j. L'estimation de la participation du voxel v à la décision prédite pour le patient i dépend de $x_j^{(i,v)}$, hors distribution par rapport à g_j. Le résultat fourni par f_2 est trompeur, et largement sous-estimé par rapport à celui fourni par f_1.

petite taille de l'ensemble de données, les distributions d'AUC OOB ont des intervalles de confiance importants. Les différences entre les modèles et les biomarqueurs ont donc une puissance et une signification statistique faibles.

Vallières et al. ont rapporté une AUC OOB globalement meilleure que la nôtre ($0,976 \pm 0,002$ et $0,984 \pm 0,002$ en utilisant la technique du $0,632+$ *bootstrap* [348, 349]) à partir du même ensemble de données [60]. Néanmoins, la complexité et le très grand nombre de paramètres d'extraction de caractéristiques qui ont été testés augmentent le risque d'adapter les données au bruit plutôt qu'au signal réel en raison de multiples comparaisons. Nous pensons que ces différences peuvent alors être dues à un certain surapprentissage. Le résultat de leur test de permutations semble également le montrer, avec une AUC élevée de $0,895 \pm 0,04$ (intervalle minimum-maximum : $[0,745\ ; 0,988]$) pour les modèles aléatoires de la distribution

nulle, bien que systématiquement inférieure à la véritable AUC ($997/1000$ permutations) et donnant une p-*value* $= 0,004$. En outre, ce test de permutation ne couvrait pas l'ensemble de leur chaine d'optimisation. Il est ainsi possible que la performance des modèles aléatoires ait été sous-estimée et qu'une fuite de données ait conduit à une surestimation des performances réelles et de leur signification statistique. Nos résultats sont toutefois cohérents avec leurs interprétations. D'après leurs résultats de corrélation univariée et leurs modèles multivariés, ils ont suggéré que la présence d'une sous-région nécrotique à l'intérieur de la tumeur serait associée à un risque plus élevé de métastases. Ils ont par ailleurs suggéré que la présence de sous-régions de fixation de FDG élevée pourrait jouer un rôle important dans la caractérisation des tumeurs à haut risque. Cependant, les images fusionnées à partir desquelles les caractéristiques ont été extraites rendent difficile l'identification précise de l'information capturée par chaque modalité. De plus, l'interprétation biologique de leurs résultats n'était soutenue par aucune carte d'importance et donc limitée à une interprétation basées sur la définition mathématique des caractéristiques.

Nos résultats sont cohérents avec plusieurs autres études en imagerie TEP et IRM [53, 54, 350-354], ainsi qu'avec les systèmes de stadification des STS basés sur la biopsie et montrant une capacité à prédire le développement de métastases et la mortalité [355]. En effet, le système de classification des STS du *National Cancer Institute* repose sur l'histologie, la localisation, et la nécrose tumorale. Le système de classification de la *Fédération Nationale des Centres de Lutte Contre le Cancer* est également basé sur la différenciation tumorale, l'activité mitotique, et la nécrose de la tumeur. Mentionné en Section 2.1.3 du Chapitre 2, au-delà de la nécrose identifiée en TEP/TDM grâce au signal hypométabolique dans la tumeur, Rakheja et al. ont rapporté une corrélation positive entre le SUVmax et le taux mitotique, inclus dans les caractéristiques des systèmes de stadification.

Nos résultats sont ainsi conformes aux connaissances actuelles sur les STS, et la méthode proposée n'a pas permis de découvrir de nouveaux motifs prédictifs dans ce contexte médical. En revanche, cette cohérence suggère que cette méthode entièrement basée sur les données pourrait être utilisée lorsque les caractéristiques de la tumeur associées à un résultat sont peu connues, afin de mettre en évidence les motifs régionaux qui déterminent la décision du modèle, ce qui pourrait faciliter l'émergence de nouvelles hypothèses.

Cette étude présente des limites. Certaines d'entre elles sont liées à la chaine de modélisation par apprentissage. Tout d'abord, malgré leur grande efficacité pour trouver un bon sous-ensemble de caractéristiques, les approches séquentielles de sélection sont sujettes au surapprentissage en raison de leur mode de fonctionnement intrinsèque de comparaisons multiples. De plus, l'évaluation de la performance des modèles M1 et M2 est peut-être op-

timiste car elle a été effectuée en même temps que l'optimisation des hyperparamètres, sans effectuer une validation croisée imbriquée (Section 3.2.4). Le nombre de patients disponibles est souvent insuffisant dans les études radiomiques pour suivre une telle approche, notamment dans le cas de pathologies rares, comme dans la cohorte de 51 patients analysée ici. Pourtant, la découverte de nouvelles caractéristiques tumorales prédictives serait particulièrement intéressante dans le cas de cancers rares pour lesquels la faible quantité de données disponibles limite la connaissance médicale à leur sujet. Notre objectif était de démontrer comment obtenir des cartes d'importance informatives dans un tel contexte, plutôt que de déployer un modèle à visée d'automatisation. Nous avons donc utilisé un test de permutations pour nous assurer que les informations capturées par les modèles n'étaient pas du bruit.

Une autre limite est que les valeurs locales moyennées dans la ROI ne sont pas nécessairement égales ou même corrélées aux valeurs des caractéristiques directement calculées à partir du ROI. Cela rend notre cartographie incompatible avec les signatures radiomiques déjà publiées, qui sont presque toujours calculées directement à partir de la ROI. De plus, certaines caractéristiques radiomiques demeurent difficiles à interpréter malgré leur définition mathématique précise, et cette complexité n'est compensée ici que par l'identification locale de l'information pertinente sans perte de résolution spatiale par rapport aux images originales. Il pourrait encore être utile de développer une méthodologie pour convertir facilement une signature radiomique compliquée en une signature plus simple et plus robuste qui pourrait même mieux se généraliser. Bien que les similitudes dans les résultats du modèle simplifié M1' par rapport à M1 suggèrent qu'il est possible de reformuler un modèle compliqué en conservant une représentation fidèle, nous n'avons malheureusement pas eu accès à un ensemble de données externes pour comparer la robustesse et l'exportabilité des deux modèles.

Une limitation potentielle est également que l'utilisation d'une fenêtre glissante pourrait manquer certaines informations globales dans la ROI. Néanmoins, nous avons observé empiriquement que la plupart des caractéristiques classiques avaient des valeurs fortement corrélées avec le volume, la forme de la tumeur, ou même certaines caractéristiques locales moyennées. Par conséquent, nous ne pensons pas manquer d'informations cruciales pour la caractérisation des tumeurs.

4.5 . Conclusion

Nous avons décrit une méthode générique basée sur des caractéristiques radiomiques prédéfinies calculées localement pour caractériser spatialement et quantitativement les sous-régions et le signal à l'origine des prédictions de modèles.

Lorsque le nombre de données est limité, nous avons démontré comment

cette méthode permet une interprétation cohérente des modèles et identifie des biomarqueurs potentiellement utiles pour la classification ou la stratification des patients. Étant techniquement applicable à tout problème approché à l'aide de la radiomique, cette méthode pourrait contribuer à accroître notre compréhension des informations pertinentes apportées par les images médicales lorsque l'on connaît peu les informations qu'elles portent associées à la question d'intérêt.

À l'avenir, si l'identification et l'interprétation de sous-régions peuvent être associées à des relations causales par les experts médicaux, on pourrait être en mesure d'adapter localement et de personnaliser le traitement de chaque patient en fonction de l'expression phénotypique de sa maladie, comme Reuzé et al. l'ont proposé dans le contexte de la radiothérapie [112].

5 - Cartes de décision radiomiques discriminant progression tumorale et nécrose radio-induite chez des patients atteints de tumeurs cérébrales

En collaboration avec le Centre Antoine Lacassagne (CAL, Nice, France), nous présentons ici l'application de la méthode proposée dans le chapitre précédent à une base de données de patients atteints de tumeurs cérébrales, pour le diagnostic différentiel entre progression tumorale et nécrose radio-induite après traitement par chirurgie et radiochimiothérapie. Les résultats de cette étude ont fait l'objet d'une présentation orale au congrès annuel de 2022 de la SNMMI [163], et ont été récompensés par le premier prix des jeunes investigateurs du Conseil de Physique, Instrumentation, et *Data Sciences* (PIDSC *Young Investigator Award* [356]).

5.1 . Introduction

Les gliomes sont, avec les métastases cérébrales, les cancers affectant le cerveau les plus fréquents [357]. La thérapie standard pour les tumeurs cérébrales comprend généralement la combinaison d'une chirurgie et de séances de radiothérapie et de chimiothérapie [358-360]. L'IRM est la modalité d'imagerie de première intention pour le diagnostic et la surveillance [361, 362].

La surveillance post-thérapeutique des patients atteints de gliomes suit généralement les recommandations du groupe de travail sur l'évaluation de la réponse en neuro-oncologie (*working group on response assessment in neuro-oncology* (RANO)) [359, 363], et repose sur des examens cliniques et l'imagerie IRM. Les métastases cérébrales sont hétérogènes par nature, la variabilité du type et du nombre de métastases représente un défi. Leur prise en charge est ainsi moins codifiée que l'approche plus systématique appliquée à la gestion des gliomes. Elle dépend des caractéristiques de la maladie et des manifestations cliniques [358, 364-366].

Dans tous les cas, l'apparition possible d'une nécrose tissulaire induite par le traitement est le principal effet secondaire à moyen terme de la radiothérapie [367-370]. La biopsie est considérée comme la technique la plus fiable pour distinguer la récidive tumorale d'une radionécrose induite par le traitement [367-369]. Cependant, il s'agit d'une procédure invasive qui comporte les risques généralement associés aux procédures chirurgicales. De plus, une lésion au cours du suivi peut inclure tissus nécrotiques et tissus tumoraux en croissance [371]. Cela complexifie la caractérisation de la lésion, surtout lors

de la réalisation d'une biopsie avec un unique ou un nombre réduit de points de prélèvement.

Souvent, les symptômes cliniques ne permettent pas de distinguer radionécrose et récidive. En outre, sur l'IRM, la radionécrose peut être très similaire à la progression de la tumeur lors d'une récidive. Ainsi, il est parfois difficile de différencier les deux situations lors du suivi [56, 368, 369, 371, 372]. Il est pourtant crucial pour l'équipe de soins de déterminer l'étiologie d'une lésion observée sur l'imagerie de suivi, car les stratégies de prise en charge de la récidive tumorale et de la nécrose liée à la toxicité du traitement sont différentes [364, 369]. Par conséquent, des méthodes de diagnostic différentiel, notamment par imagerie, sont nécessaires. De telles méthodes pourraient réduire le nombre d'interventions chirurgicales, ce qui augmenterait la survie des patients et améliorerait leur qualité de vie. Dans les situations nécessitant une intervention, ces techniques pourraient être utilisées pour guider la biopsie ou le prélèvement de tissus dans la zone concernée.

L'imagerie TEP a ainsi été proposée pour réaliser ce diagnostic différentiel [41, 372-374]. Plusieurs radiotraceurs ont été étudiés dans ce contexte, incluant le FDG [375-377], les radiotraceurs impliquant la choline tels que la 18F-fluorocholine et la 11C-choline [378-380], les analogues de nucléosides comme la 18F-fluorothymidine (18F-FLT) [381-383], ainsi que les analogues des acides aminées, avec la 11C-méthionine (11C-MET), la 18F-FET, et la 18F-FDOPA (Section 2.1.3) [56, 375, 384-387].

5.2 . Utilisation de la radiomique pour le diagnostic différentiel en TEP statique et double temps à la 18F-FDOPA

Du fait d'une incorporation amplifiée des acides aminés lors de la synthèse protéique associée au développement des tumeurs cérébrales, l'imagerie TEP utilisant les analogues des acides aminés montre un bon contraste entre la fixation physiologique du cerveau et la fixation lésionnelle, particulièrement en comparaison avec le FDG (Figure 5.1). En effet, leur faible captation par le parenchyme cérébral limite la présence du bruit de fond gênant l'interprétation des images [384, 388, 389]. Ce contraste est lié, au moins en partie, à la surexpression du transporteur d'acides aminés LAT-1 dans les cellules des tumeurs et leur système vasculaire [384, 390-393].

Parmi les trois principaux radiotraceurs analogues des acides aminés cités précédemment, la 11C-MET nécessite un cyclotron dans l'hôpital où les images TEP sont acquises en raison de la courte demi-vie du 11C ($t_{1/2} \approx 20min$), ce qui limite ses applications cliniques. Les radiotraceurs marqués au 18F ($t_{1/2} \approx 110min$) tels que la 18F-FET et la 18F-FDOPA ne souffrent pas de cette limite et correspondent également mieux au processus relativement lent de la synthèse protéique [384, 394, 395]. En outre, la 18F-FDOPA dispose d'une autorisation de mise sur le marché en France pour diverses indica-

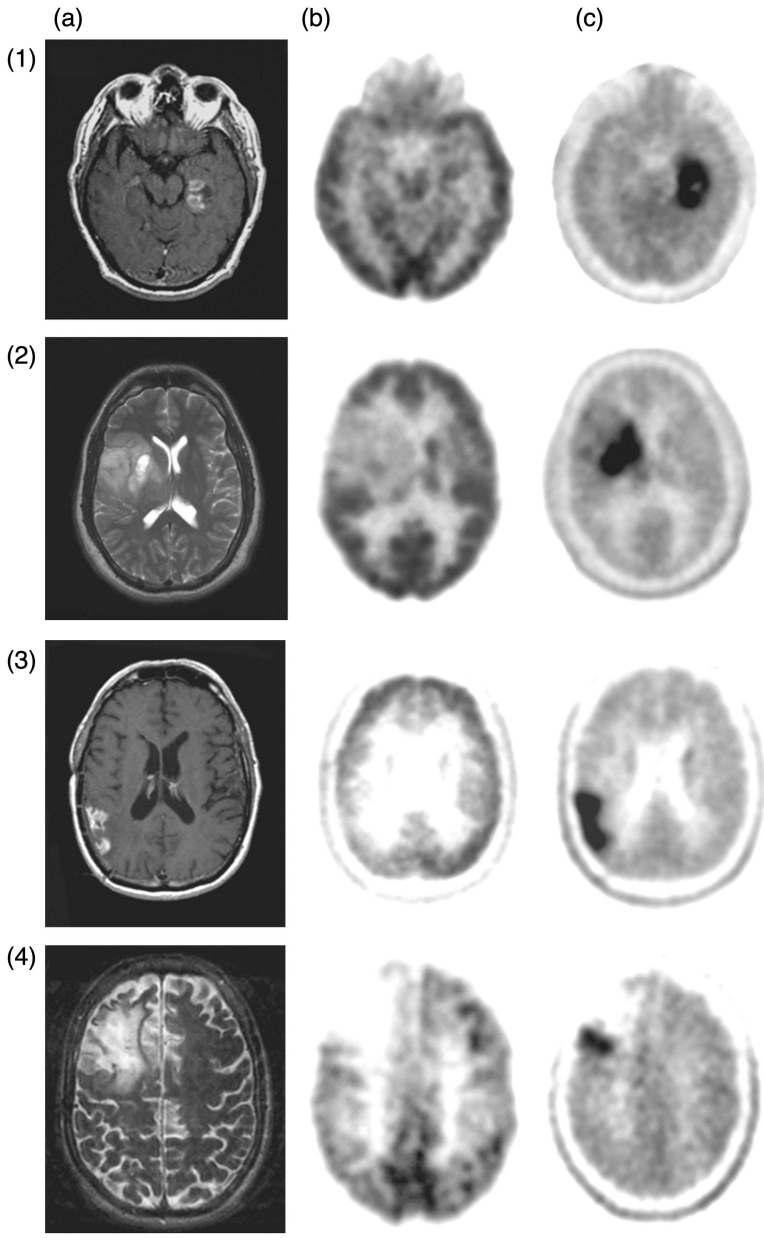


Figure 5.1 – Exemples de coupes cérébrales d'images IRM (a) et TEP au FDG (b) et à la 18F-FDOPA (c) pour quatre patients atteints de gliomes. (1, 2) Tumeurs nouvellement diagnostiquées. (3, 4) Tumeurs récidivantes. Figure adaptée de Chen et al. [388].

tions, dont le diagnostic différentiel entre récidive et radionécrose de tumeurs cérébrales [396-398]. D'autre part, la 18F-FET, présentant des contrastes similaires [387], n'est pas aisément disponible dans tous les pays [384, 399]. Ces éléments ont amené l'équipe du CAL à étudier la 18F-FDOPA dans le contexte du diagnostic différentiel entre radionécrose et récidive tumorale.

En plus de la captation tumorale spécifique, bien que des études suggèrent une implication importante de la nature inflammatoire de certains tissus fixant la 18F-FDOPA [385, 400], on ne connaît pas précisément le degré ni les mécanismes de fixation des acides aminés dans les lésions tumorales et nécrotiques [384, 385, 389, 400]. Parce que les cellules tumorales fixent la 18F-FDOPA, une lésion litigieuse au suivi en IRM qui fixerait le radiotraceur serait alors plutôt associée à une progression, tandis qu'une lésion négative en TEP suggérerait une toxicité du traitement. Néanmoins, parce que la nécrose peut être associée à une inflammation adjacente, les acides aminés et donc la 18F-FDOPA peuvent avoir un tropisme pour certaines lésions présentant une nécrose radio-induite, augmentant le risque de faux positifs et limitant ainsi sa spécificité pour la détection des progressions [385].

En 2014, Herrmann et al. dans le contexte des gliomes, et Lizarraga et al. dans le contexte des métastases cérébrales, ont proposé chacun une échelle de caractérisation visuelle [57, 58]. Toutes deux sont basées sur l'intensité du signal dans la lésion par rapport au striatum controlatéral et varient d'un score bas pour les lésions ne présentant pas de signal à un score élevé pour les lésions présentant un signal supérieur au signal striatal. L'échelle de Herrmann utilise un score à cinq valeurs, $\{-2\ ;-1\ ;0\ ;1\ ;2\}$, tandis que celle de Lizarraga en comporte quatre, $\{0\ ;1\ ;2\ ;3\}$. Herrmann et al. ont rapporté une $Se = 0,852$ et une $Sp = 0,724$ ($Bacc = 0,788$), semblables à Lizarraga et al. rapportant une $Se = 0,813$ et une $Sp = 0,843$ ($Bacc = 0,828$). Dans les deux cas, le score seuil permettant de discriminer au mieux les patients était celui correspondant à un signal de même intensité dans la lésion et le striatum controlatéral (0 pour Herrmann et al. et 2 pour Lizarraga et al.). En comparaison, leurs analyses semi-quantitatives univariées, impliquant le rapport d'intensité maximale (max) ou moyenne (mean) de la lésion (L) par rapport au striatum (S) controlatéral ou au cortex cérébral « normal » (N) (L/Smax, L/Smean, L/Nmax, L/Nmean) ainsi que le SUVmax et le SUVmean, ont donné des performances similaires mais plus faibles que l'analyse visuelle dans les deux cas. Ces résultats suggèrent qu'il n'est pas facile de définir objectivement des variables capables d'automatiser avec une exactitude équivalente ou supérieure le diagnostic réalisé par les experts médicaux lorsqu'ils sont aidés par une échelle de caractérisation visuelle. De plus, leurs résultats montrent qu'aucun des seuils optimaux appliqués aux variables impliquant les striata n'est égal ou proche de 1, alors que cette valeur représenterait une intensité identique entre les striata et les lésions, comme c'est le cas pour les échelles visuelles, et tel que rapporté par Chen et al. [388]. Il semble que l'appréciation globale des experts médicaux porte une information, implicite ou non, qui n'est pas codée par les seuls rapports d'intensité entre lésion et striatum ou cortex cérébral.

Dans le cadre d'une étude s'appuyant sur l'échelle de Lizarraga et visant à évaluer l'impact de la 18F-FDOPA sur le diagnostic et la décision thérapeutique d'un concertation multidisciplinaire de neuro-oncologie par son ajout aux critères habituels de l'IRM, l'équipe du CAL a constitué une base de données incluant des patients atteints de glioblastomes (un type de gliome) et de métastases cérébrales, imagés à la 18F-FDOPA [56]. Leurs résultats confortent l'utilité de ce radiotraceur, avec un changement de diagnostic et de stratégie thérapeutique lié à l'interprétation de la TEP chez 4/12 (33%) patients atteints de glioblastomes. Pour les métastases cérébrales, le changement de diagnostic concernait 16/41 (39%) patients, et celui de la stratégie de traitement en concernait 7/41 (17%). Tel que montré en Figure 5.2, que ce soit pour les glioblastomes ou les métastases cérébrales, l'utilisation de la 18F-FDOPA a globalement amélioré le diagnostic différentiel. Un élément intéressant de cette étude est que même s'il y a eu une amélioration du diagnostic et de la prise en charge, l'accord entre les membres des réunions de concertation multidisciplinaire de neuro-oncologie a évolué lorsque la 18F-FDOPA a été incluse, dans certains cas à la hausse, mais dans d'autres à la baisse. Cela suggère une dépendance des interprétations et donc du diagnostic et de la prise en charge thérapeutique, à la composition et l'expérience des membres de l'équipe de soins, donc de probables disparités entre différents centres hospitaliers.

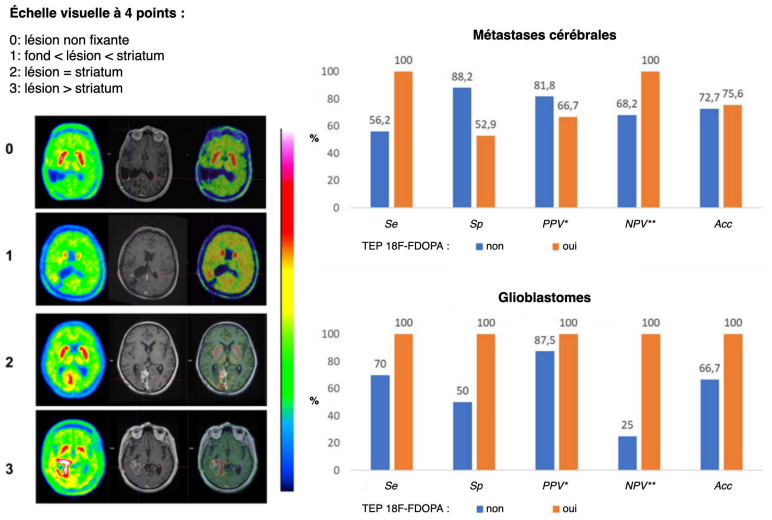


Figure 5.2 – Exemples d'images IRM et TEP à la 18F-FDOPA, et diagnostic avant (bleu) et après (orange) la TEP. *$PPV = TP/(TP+FP)$: *positive predictive value*. **$NPV = TN/(TN+FN)$: *negative predictive value*. Figure adaptée de Humbert et al. [56].

Les études mentionnées précédemment ont utilisé l'imagerie TEP statique, c'est-à-dire avec l'acquisition standard d'une seule image, lors d'un seul examen. Les informations fournies par des paramètres issus de la TEP

dynamique ont été étudiées dans l'évaluation de la récidive tumorale avec la 18F-FET, connue pour présenter un pouvoir discriminant équivalent à la 18F-FDOPA [80, 387, 401-403]. La TEP dynamique à la 18F-FDOPA a également été utilisée pour distinguer les gliomes récidivants de haut grade et de bas grade, ainsi que pour leur caractérisation moléculaire [404, 405]. Récemment, Zaragori et al. ont étudié la TEP dynamique à la 18F-FDOPA pour distinguer la récidive de la radionécrose [406]. Pour les paramètres statiques, ils ont rapporté des L/Smax, L/Smean, L/Nmax, et L/Nmean plus élevés dans les tumeurs récidivantes, en cohérence avec la littérature. En ce qui concerne les caractéristiques cinétiques, ils ont observé des associations entre la progression tumorale et un temps au pic de la courbe activité-temps (*time-to-peak* (TTP)) précoce, ainsi qu'une pente fortement négative entre 10 et 30 minutes (non significative après ajustement de la p-*value* de Mann-Whitney avec la correction de Benjamini-Hochberg [407, 408]). Cependant, aucun avantage supplémentaire n'a été observé en termes de performances de classification lors de l'utilisation de la TEP dynamique. Dans ce contexte, pour améliorer le confort des patients, et éviter la difficulté logistique et les coûts d'utilisation de la TEP dynamique tout en continuant à étudier la pertinence de motifs cinétiques dans les images, l'équipe du CAL a proposé une acquisition statique « tardive », en plus de l'acquisition standard « précoce » [384]. Également adoptée par Lohmann et al. dans une étude utilisant la 18F-FET [409], cette approche est ici nommée « double temps ».

En ce qui concerne la radiomique (18F-FDOPA et 18F-FET), plusieurs études ont montré un potentiel intéressant pour le diagnostic différentiel entre progression et nécrose radio-induite, ou plus globalement pseudoprogression [410-414]. Plus précisément, Ahrari et al. ont combiné les approches dynamique et radiomique pour différencier les gliomes de haut grade des changements liés au traitement [415]. Leurs performances étaient similaires au niveau multicentrique en utilisant différents modèles radiomiques de ML, et les auteurs ont rapporté une plus-value marginale pour la radiomique comparativement à l'analyse basée sur des variables simples telles qu'utilisées par la même équipe dans l'étude de Zaragori et al. [406]. L'ensemble de ces études suivait une approche radiomique « classique », avec une extraction de caractéristiques à l'échelle de la ROI. Ainsi, les informations de l'image sur lesquelles les modèles radiomiques étaient basés demeurent peu claires. Elles ne permettent pas non plus d'identifier explicitement des sous-régions pertinentes dans le volume lésionnel. Dans cette étude, nous avons donc appliqué notre méthode de cartographie de décision radiomique avec l'objectif de révéler des motifs quantitatifs interprétables capturés par les modèles entraînés pour réaliser le diagnostic différentiel. Cette approche a été appliquée en combinaison avec l'utilisation de l'imagerie double temps.

5.2.1 . Matériels et méthodes

5.2.1.1 . Patients et données

Notre étude rétrospective a porté sur une base de données composée de patients atteints de glioblastomes, et ayant bénéficié d'une TEP/TDM à la 18F-FDOPA au service de médecine nucléaire du CAL entre 2012 et 2019. Des informations médicales détaillées sur les données sont fournies par Rollet dans sa thèse de médecine [384].

Critères d'inclusion et d'exclusion

Les critères d'inclusion de l'étude comprenaient une histologie confirmée de glioblastome, un traitement antérieur de radiochimiothérapie, une IRM cérébrale de suivi disponible avec une incertitude entre progression tumorale et lésion radio-induite, une TEP/TDM réalisée dans les trois mois suivant l'IRM, et un suivi clinique minimum de trois mois. Les patients étaient également soumis à des critères d'exclusion. Les patients qui suivaient un traitement par anti-VEGF [1] au moment de l'acquisition de la TEP/TDM ont été exclus. Ont également été exclus les patients dont le suivi clinique et radiologique ne permettait pas le diagnostic différentiel en l'absence de preuve histologique.

Enfin, les lésions présentant des examens IRM suspects mais sans fixation en TEP n'ont pas été prises en compte dans cette étude, et ont d'emblée été considérées comme radionécrotiques. Ainsi, seules les images des lésions présentant une fixation suspecte (positive) de la 18F-FDOPA ont été utilisées. Ce choix a été fait pour correspondre au réel défi de discriminer la toxicité de la progression dans le cas d'une TEP positive à la 18F-FDOPA. En effet, aucun paramètre radiomique ou (semi-)quantitatif n'est nécessaire pour caractériser un signal nul (ou un bruit de fond).

Plusieurs examens TEP/TDM à la 18F-FDOPA acquis chez un même patient ont été inclus, mais seulement en cas d'épisodes indépendants de récidive ou de radionécrose, ou au moins avec six mois d'intervalle. Ainsi, 96 études (lésions) provenant de 87 patients [2] ont été incluses.

Caractéristiques de l'échantillon

Tous les patients ($n = 87$) ont bénéficié d'une première intervention chirurgicale pour obtenir la preuve histologique d'un glioblastome, avec 44 (50,6%) d'entre eux ayant eu une résection complète, 25 (28,7%) ayant eu une résection partielle, et 18 (20,7%) ayant seulement eu une biopsie. L'échantillon étudié se composait de 41 hommes et 46 femmes, donnant un $ratio_{sexe} \approx 0,89$. L'âge médian des patients au moment de l'étude de chaque lésion était de 60 ans ($n = 96$). En comparaison, la littérature sur le

1. Les traitements anti-VEGF modifient la structure des vaisseaux, entrainant une modification de l'aspect en imagerie. Ce facteur confondant potentiel a été éliminé en excluant les patients ainsi traités de l'étude.

2. 1 examen : 79 patients, 2 examens : 7 patients, 3 examens : 1 patient.

sujet suggère que l'incidence du glioblastome en Europe et aux États-Unis d'Amérique est $1,6$ fois plus élevée chez les hommes que chez les femmes, et qu'elle est maximale vers 65 à 70 ans [416-418]. Un biais d'échantillonnage sous-estimant l'incidence relative des hommes par rapport aux femmes ainsi que l'âge au diagnostic semble donc être présent dans notre base de données.

Diagnostic final

Le diagnostic final utilisé comme vérité terrain était basé sur un suivi de trois à six mois. Des données pathologiques étaient également disponibles pour 27 lésions (28,1%). Ainsi, les progressions étaient au nombre de 69 (71,9%), et les radionécroses de 27 (28,1%). Il est important de rappeler qu'une lésion cérébrale au suivi peut inclure du tissu tumoral en progression et des zones nécrotiques et inflammées (radionécrose). Par conséquent, la vérité terrain utilisée étant binaire (récidive ou radionécrose), elle peut être imparfaite pour certains patients.

Protocole d'imagerie

Avant l'imagerie TEP/TDM à la 18F-FDOPA, les patients devaient être à jeun de protéines depuis au moins quatre heures. Les patients sans contre-indication ont reçu $100mg$ de Carbidopa en prémédication une heure avant l'injection [419]. Une injection intraveineuse de $2MBq \times kg^{-1}$ de radiotraceur a ensuite été réalisée. Chaque étude comprenait deux acquisitions TEP/TDM statiques de 10 minutes, réalisées 20 minutes (TEP20) et 90 minutes (TEP90) après l'injection. Sur un même scanner pour tous les patients (Biograph mCT, Siemens Healthineers), la correction d'atténuation a été effectuée à partir de la TDM acquise avec une énergie de $120kV$ et une intensité de $80mA$. Les images ont été reconstruites en utilisant la technique OSEM (*ordered subset expectation-maximization*) avec 24 sous-ensembles et 5 itérations, sans modélisation de la fonction de réponse du détecteur. Les images reconstruites ont été corrigées de l'atténuation, la diffusion, et des détections fortuites.

5.2.1.2 . Traitement et représentation des images

Prétraitement des images et obtention des images double temps

Les images TEP90 ont été recalées automatiquement sur les images TEP20 grâce à l'algorithme BRAINFit implémenté dans 3D Slicer en mode rigide avec six degrés de liberté (*degree of freedom* (6 DOF)) et tous les autres paramètres par défaut [420, 421]. En utilisant LIFEx, la lésion et le striatum controlatéral ont été délinéés pour chaque étude via un seuil à 50% du SUVmax striatal sur l'image TEP20. Raffinées manuellement, les segmentations ont ensuite été propagées à l'image TEP90.

Pour chaque patient, le rapport entre la fixation à l'échelle du voxel v et le SUVmax ou le SUVmean du striatum controlatéral à la lésion d'intérêt a été calculé pour obtenir des images normalisées par le striatum (L/S), désignées

ici par L/Smax20, L/Smax90, L/Smean20, et L/Smean90, et définies telles que :

$$L/Smax20_v = \frac{TEP20_v}{TEP20_{SUVmax_{striatum}}} \tag{5.1}$$

$$L/Smax90_v = \frac{TEP90_v}{TEP90_{SUVmax_{striatum}}} \tag{5.2}$$

$$L/Smean20_v = \frac{TEP20_v}{TEP20_{SUVmean_{striatum}}} \tag{5.3}$$

$$L/Smean90_v = \frac{TEP90_v}{TEP90_{SUVmean_{striatum}}}. \tag{5.4}$$

Après recalage, les images double temps ont été obtenues en soustrayant les images à 90 minutes des images à 20 minutes, donnant les images TEP20-90, L/Smax20-90, et L/Smean20-90 définies telles que :

$$TEP20\text{-}90_v = TEP20_v - TEP90_v \tag{5.5}$$

$$L/Smax20\text{-}90_v = L/Smax20_v - L/Smax90_v \tag{5.6}$$

$$L/Smean20\text{-}90_v = L/Smean20_v - L/Smean90_v. \tag{5.7}$$

Calcul des caractéristiques

Le calcul des caractéristiques radiomiques suit les mêmes étapes que pour la prédiction des métastases pulmonaires en STS (Section 4.2.1.3 du Chapitre 4). Les paramètres d'extraction sont les suivants :

- **Interpolation** : *B-spline* de troisième ordre : $1mm \times 1mm \times 1mm$ (taille d'origine : $1,02mm \times 1,02mm \times 2,03mm$)
- **Discrétisation** : taille de *bins* fixe :
  - TEP : $0,1SUV$ (100 *bins* de 0 à 10) [384]
  - L/S : $0,02$ (100 *bins* de 0 à 2)
- **Caractéristiques** :
  - premier ordre ($p_{po} = 18$)
  - GLCM ($p_{GLCM} = 24$)
  - GLDM ($p_{GLDM} = 14$)
  - GLRLM ($p_{GLRLM} = 16$)
  - NGTDM ($p_{NGTDM} = 5$)
  - total par image ($p_{tot} = 77$)
- **Fenêtre glissante** : $5 \times 5 \times 5$ voxels
- **Agrégation** : moyenne
- **Modèle final** : *bagging*

Chaque patient possédait 9 images : TEP20, TEP90, L/Smax20, L/Smax90, L/Smean20, L/Smean90, TEP20-90, L/Smax20-90, et L/Smean20-90. Avec 77 cartes de caractéristiques par image, le nombre de cartes de caractéristiques extraites des images pour chaque lésion était de $77 \times 9 = 693$. Pour la stratégie double temps, en plus des cartes calculées sur les images issues de la soustraction entre les deux temps, nous avons calculé les soustractions entre les cartes elles-même, pour les images TEP, L/Smax, et L/Smean, produisant $3 \times 77 = 231$ cartes supplémentaires. Le nombre total de cartes de caractéristiques pour chaque lésion était donc de $p = 693 + 231 = 924$.

5.2.1.3 . Classification probabiliste

La valeur moyenne sur la ROI de la tumeur a été utilisée pour chaque caractéristique afin d'obtenir deux vecteurs par lésion, un pour toutes les images ($p = 924$), et l'autre uniquement pour les images standards à 20 minutes (TEP20, L/Smax20, et L/Smean20, $p = 231$), afin d'évaluer la valeur ajoutée de l'approche double temps.

En suivant les mêmes étapes que pour la prédiction des métastases pulmonaires en STS (Section 4.2.1.4 du Chapitre 4), nous avons construit deux modèles probabilistes, M20 et Mdt (double temps), pour identifier la radionécrose [3]. Après avoir réduit la redondance à l'aide du VIF, nous avons utilisé la sélection SFS et la régression logistique avec régularisation LASSO.

Les fonctions linéaires de décision de M20 et Mdt, D_{M20} et D_{Mdt}, ont été obtenues en suivant une stratégie de *bagging* sur l'ensemble de la base de données, avec 1000 tirages *bootstrap*.

Les RDMs $DV^{(i)}_{M20}$ et $DV^{(i)}_{Mdt}$ ont ensuite été calculées pour tout patient i en rétroprojetant les coefficients de D_{M20} et D_{Mdt} à l'échelle du voxel, après avoir normalisé la valeur de chaque voxel pour les cartes de caractéristiques sélectionnées tel que défini dans l'équation (4.13).

La Figure 5.3 illustre la chaîne de traitement et d'analyse proposée dans cette étude, de l'acquisition des images à leur analyse.

3. Dans le cas d'images TEP positives, le défi concerne la spécificité de détection des progressions. En effet, lorsque la TEP révèle un hypersignal 18F-FDOPA, la sensibilité de détection est bonne, mais de nombreux faux positifs subsistent. La radionécrose étant plus rare, donc moins fréquente dans notre base de données, nous lui avons affecté la classe positive, tel que c'est généralement le cas dans le domaine du ML où la classe positive est attribuée au groupe minoritaire.

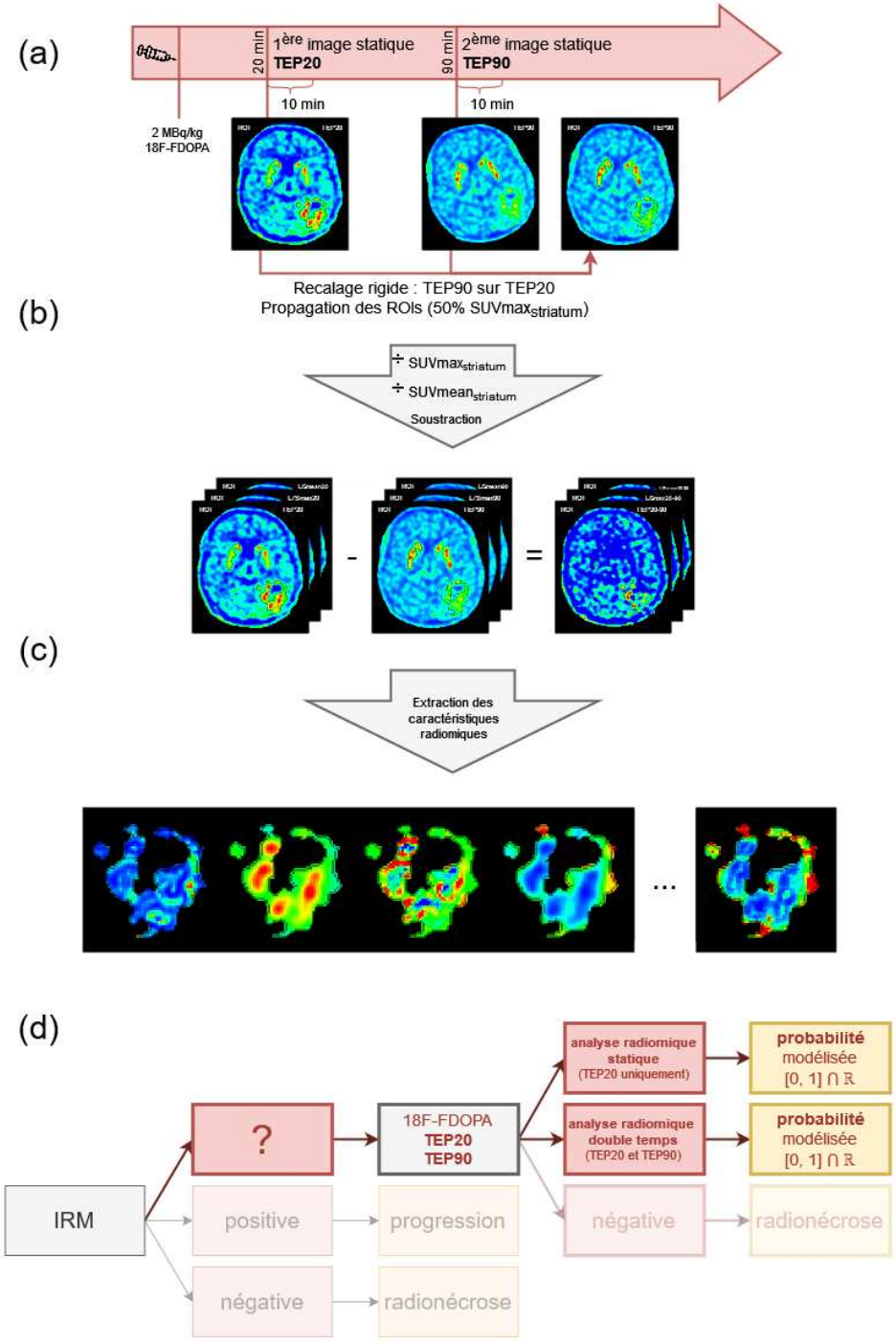


Figure 5.3 – Chaine de traitement et d'analyse proposée pour la réalisation du diagnostic différentiel entre progression tumorale et nécrose radio-induite grâce à la radiomique en imagerie TEP à la 18F-FDOPA pour des patients atteints de tumeurs cérébrales. (a) Protocole d'imagerie. (b) Prétraitement des images et obtention des images double temps. (c) Calcul des caractéristiques. (d) Classification probabiliste.

5.2.2 . Résultats

$17/231$ et $31/924$ caractéristiques ont été conservées respectivement pour la construction de M20 et Mdt après réduction de la multicollinéarité, démontrant une forte redondance dans la représentation des images par les caractéristiques radiomiques.

L'optimisation des grilles de recherche a sélectionné 3 caractéristiques avec $C = 2,2$ dans le contexte statique à 20 minutes, et 5 caractéristiques avec $C = 4,6$ pour l'approche double temps. En termes de performances, dans les deux cas, les tests de permutations pour l'ASB étaient à la limite de la significativité ($0,775 \pm 0,101$, p-*value* $= 0,050$ pour M20, et $0,793 \pm 0,115$, p-*value* $= 0,050$ pour Mdt), cependant ils étaient significatifs pour la Bacc ($0,705 \pm 0,092$, p-*value* $= 0,039$ pour M20, $0,749 \pm 0,090$, p-*value* $= 0,034$ pour Mdt). Les résultats de la construction des modèles M20 et Mdt sont résumés dans le Tableau 5.1. En comparaison, Rollet a rapporté que l'analyse visuelle réalisée par les experts médicaux du CAL aidés par l'échelle de Lizarraga a donné une $Se = 1,000$ et une $Sp = 0,300$. Ces résultats correspondent à une $Bacc = 0,650$ (statistique J de Youden $J = 0,300$), plus faible que les performances rapportées par Herrmann et al. ($Bacc = 0,788$) et Lizarraga et al. ($Bacc = 0,828$). Ceci s'explique notamment par l'exclusion des lésions négatives en TEP, augmentant le taux de faux positifs lors de l'utilisation de l'échelle de Lizarraga [384]. Sur une partie des patients étudiés lors de la création de la base de données ($n = 78$), une étude pilote a été présentée en 2019 par Orlhac et al [410]. Les auteurs ont rapporté une $Se = 0,970$ et une $Sp = 0,300$ ($Bacc = 0,635$, $J = 0,270$) lors de l'analyse visuelle. Ils ont obtenu les meilleurs résultats en utilisant des caractéristiques radiomiques classiques, provenant d'une image équivalente à l'image TEP20-90, en entrée d'un classifieur d'analyse linéaire discriminante [422]. Les performances atteintes par ce modèle correspondaient à une $Se = 0,650$ et une $Sp = 0,800$ ($Bacc = 0,725$, $J = 0,450$) en validation croisée *leave-one-out*.

Les fonctions de décision linéaires D_{M20} et D_{Mdt} sont données par les équations (5.8) et (5.9) avec l'écart-type associé à chaque caractéristique sur 1000 échantillons *bootstrap*.

Tableau 5.1 – Performances en validation croisée pour l'optimisation par grille de recherche des paramètres de régularisation LASSO et de sélection des caractéristiques SFS.

Paramètres de construction des modèles	**M20**	**Mdt**
C	$2,2$	$4,6$
Nombre de caractéristiques sélectionnées	3	5
ASB ($\pm$ 1 écart-type)	$0,775 \pm 0,101$	$\mathbf{0,793 \pm 0,115}$
Bacc ($\pm$ 1 écart-type)	$0,705 \pm 0,092$	$\mathbf{0,749 \pm 0,090}$
Se ($\pm$ 1 écart-type)	$0,656 \pm 0,189$	$\mathbf{0,763 \pm 0,182}$
Sp ($\pm$ 1 écart-type)	$\mathbf{0,754 \pm 0,120}$	$0,734 \pm 0,118$
Brier *score loss* ($\pm$ 1 écart-type)	$0,221 \pm 0,103$	$\mathbf{0,201 \pm 0,113}$
ROC AUC ($\pm$ 1 écart-type)	$0,717 \pm 0,120$	$\mathbf{0,753 \pm 0,111}$

$$\begin{aligned} D_{M20} = &- 0,326(\pm 0,375) \times L/Smax20_{1^{er}ordre_{minimum}} \\ &- 0,624(\pm 0,346) \times L/Smax20_{GLDM_{SDLGLE^*}} \\ &+ 0,844(\pm 0,385) \times L/Smean20_{GLCM_{joint\ energy}} \\ &- 0,150(\pm 0,104) \end{aligned} \tag{5.8}$$

$$\begin{aligned} D_{Mdt} = &- 0,516(\pm 0,301) \times TEP90_{GLDM_{LDLGLE^{**}}} \\ &- 1,016(\pm 0,380) \times TEP20\text{-}90_{1^{er}ordre_{médiane}} \\ &+ 0,309(\pm 0,342) \times L/Smean20\text{-}90_{1^{er}ordre_{skewness}} \\ &+ 1,319(\pm 0,472) \times L/Smean20\text{-}90_{GLDM_{LDE^{***}}} \\ &- 0,564(\pm 0,334) \times (L/Smax20_{GLCM_{IDN^{****}}} - L/Smax90_{GLCM_{IDN}}) \\ &- 0,396(\pm 0,218) \end{aligned} \tag{5.9}$$

Les caractéristiques double temps ont été préférentiellement sélectionnées par le modèle lorsqu'elles étaient disponibles, avec $4/5$ caractéristiques pour Mdt. Les caractéristiques L/S ont également été préférentiellement sélectionnées par les deux modèles, avec $3/3$ caractéristiques pour M20 et $3/5$ pour Mdt.

*Petite dépendance de faible niveau de gris (*small dependence low gray level emphasis* (SDLGLE)).

**Large dépendance de faible niveau de gris (*large dependence low gray level emphasis* (LDLGLE)).

***Large dépendance (*large dependence emphasis* (LDE)).

****Différence inverse normalisée (*inverse difference normalized* (IDN)).

Pour M20, le coefficient négatif associé à l'augmentation du minimum local dans l'image L/Smax20 suggère que la prédiction de la radionécrose était plutôt associée à une faible fixation relative par rapport au striatum. Bien que la valeur de la caractéristique SDLGLE augmente lorsque le signal diminue, elle augmente également dans les zones hétérogènes (petites dépendances). Le signe négatif associé aux faibles dépendances suggère alors un lien entre radionécrose et signal plutôt homogène, interprétation confortée par le signe positif associé à l'énergie locale de la GLCM. Pour le modèle Mdt, la médiane locale, mesurant la différence de signal SUV lissé par un filtre médian entre 20 et 90 minutes à l'échelle du voxel, était associée à un signe négatif. Cela suggère un *wash-out* faible associé à la prédiction de radionécroses. De plus, la caractéristique LDLGLE augmente à mesure que le signal SUV à 90 minutes présente des plages homogènes (larges dépendances) de faible niveau de gris. Associée à un signe négatif, cette caractéristique suggère qu'un faible signal à 90 minutes était plutôt associé à une progression. La *skewness* de premier ordre prend quant à elle des valeurs élevées dans les zones homogènes de faible signal. De façon cohérente avec les deux caractéristiques précédentes, le modèle a sélectionné cette caractéristique de l'image double temps L/Smean20-90 avec un signe positif. Quant à l'hétérogénéité du signal, les larges dépendances (LDE) de L/Smean20-90 se sont vu affecter un signe positif, bien que la différence inverse normalisée (IDN), mesurant également l'homogénéité locale du signal sur la même image, ait été associée à un signe négatif.

Ces éléments suggèrent qu'en imagerie statique, la nécrose a été prédite pour les lésions présentant une fixation limitée et homogène. Dans le cas de l'imagerie double temps, il semblerait qu'elle soit corrélée à un *wash-out* faible, également homogène. Cependant, ces interprétations demeurent très incertaines, voire spéculatives. Comme expliqué dans le chapitre précédent, la définition des caractéristiques est complexe, et il est difficile de les relier, avec les coefficients qui leur sont associés, au signal réellement capturé par les modèles. De plus, comme discuté dans le paragraphe précédent, certaines caractéristiques mettant en évidence des motifs d'homogénéité du signal pouvaient être associées à des coefficients négatifs, alors que d'autres se sont vu affecter des coefficients positifs. Comme la complexité des définitions ne permet pas de comprendre précisément comment cette homogénéité est capturée, l'association de la radionécrose à ce type de motifs reste hypothétique.

La Figure 5.4 montre un exemple de coupe pour les images TEP20 (a), L/Smax20 (b), L/Smean20 (c), TEP90 (d), L/Smax90 (e), L/Smean90 (f), TEP20-90 (g), L/Smax20-90 (h), L/Smean20-90 (i), et de RDMs pour M20 (j) et Mdt (k) pour un patient. Le diagnostic de ce patient était une progression. Les prédictions probabilistes étaient $P_{M20} = 0,41$ pour M20 ($\overline{DV_{M20}} = -0,38 \pm 2,14$), $P_{Mdt} = 0,59$ pour Mdt ($\overline{DV_{Mdt}} = 0,38 \pm 3,24$)

pour la classe radionécrose. Nous pouvons observer que pour DV_{M20}, le signal de décision était colocalisé avec les foyers de fixation de la 18F-FDOPA, avec une décision faible (bleu, progression) dans les zones de fixation, et inversement un signal fort (rouge, radionécrose) dans les régions de faible fixation. Ces observations suggèrent que dans le cas de M20, c'est essentiellement la fixation relative par rapport au striatum qui a permis de discriminer la progression de la radionécrose. Contrairement à notre lecture de la fonction de décision D_{M20}, les motifs d'homogénéité ou d'hétérogénéité semblent avoir une importance moindre. Le signal de décision de DV_{Mdt} semble mettre en évidence des zones de faible *wash-out* homogènes associées à la radionécrose (rouge). Inversement, une faible décision (bleue) plutôt associée à une progression était colocalisée avec des zones de *wash-out* substantiel, ainsi qu'avec de forts gradients. De façon cohérente avec notre lecture de l'équation de D_{Mdt}, l'aspect homogène de l'image double temps parait s'ajouter à l'évolution de la fixation entre 20 et 90 minutes pour prédire la toxicité liée au traitement.

5.3 . Représentation simplifiée à l'échelle du voxel pour l'ensemble de la cohorte

L'utilisation des RDMs suggère ici que les modèles ont capturé de l'information cohérente avec la littérature, avec une faible fixation relative de la tumeur par rapport au striatum associée à la radionécrose en imagerie statique, et un faible *wash-out* en imagerie double temps ($\approx$ cinétique). Dans le cas de l'imagerie double temps, l'homogénéité ou l'hétérogénéité du signal semble influencer la prédiction. Comment objectiver ces intuitions ? En effet, l'observation conjointe des RDMs et des images d'entrée ne permet pas une interprétation aussi claire que pour l'étude sur les STS. De plus, comme expliqué au paragraphe 3.3.2.3 de Chapitre 3, les méthodes de cartographie sont des approches d'interprétabilité et d'explicabilité locales, nécessitant d'observer les individus un par un afin de se faire une idée générale de l'information capturée globalement par les modèles.

La question qui se pose alors est la suivante : Comment pouvons-nous, à la fois, baser nos interprétations sur une représentation à l'échelle du voxel facilement compréhensible, et expliquer un modèle globalement, pour l'ensemble des patients ? Dans cette section, nous avons tenté de répondre à cette problématique en proposant un substitut simplifié du signal de décision à l'échelle du voxel, représenté dans un même repère pour l'ensemble des lésions de la cohorte.

Les signaux simples, réguliers, peuvent être décrits avec moins de paramètres que les signaux complexes. L'information présente dans le signal et les données peut généralement être décrite au moyen d'une représentation

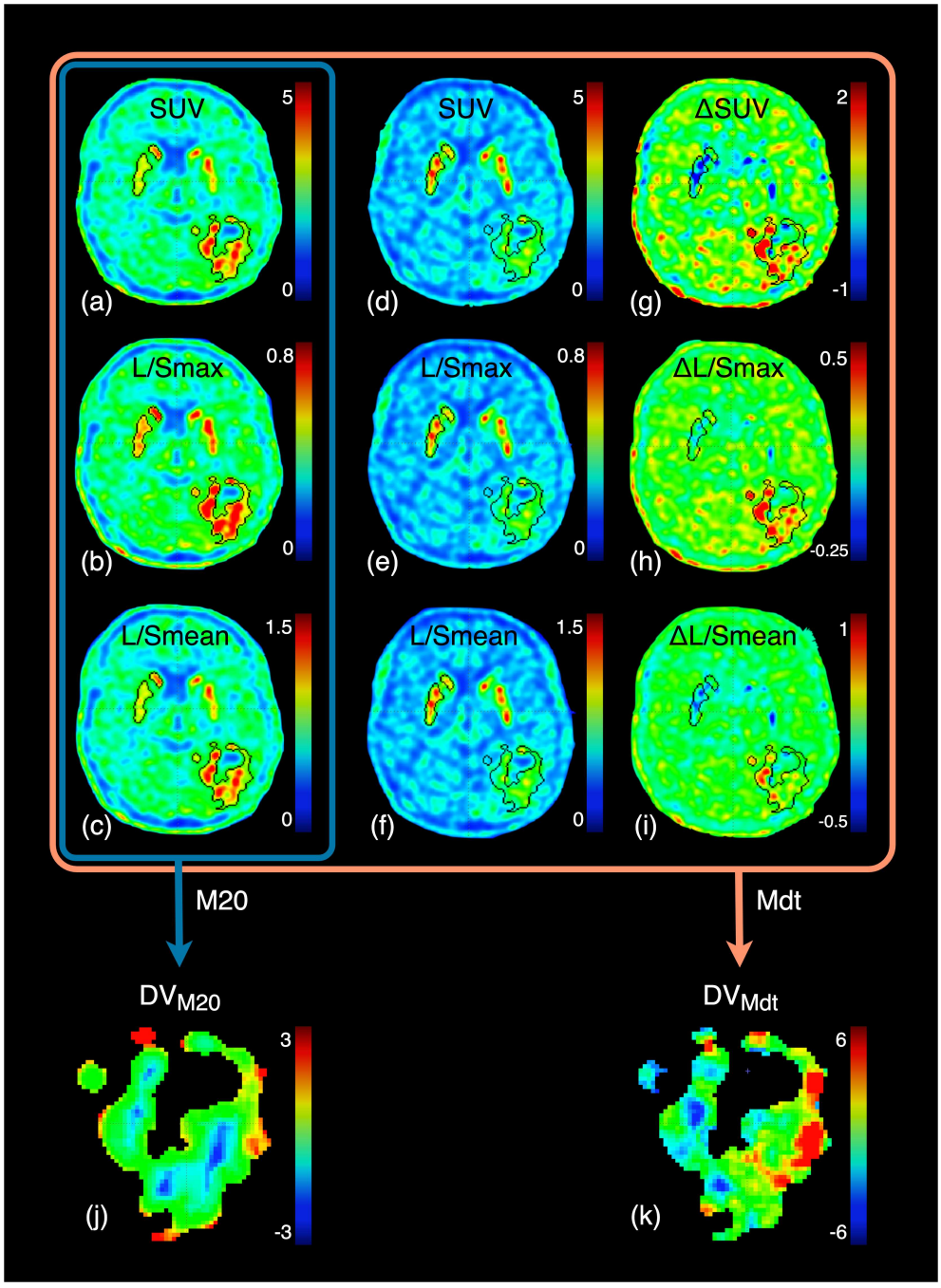


Figure 5.4 – Exemples d'images TEP20 (a), L/Smax20 (b), L/Smean20 (c), TEP90 (d), L/Smax90 (e), L/Smean90 (f), TEP20-90 (g), L/Smax20-90 (h), L/Smean20-90 (i), et de RDMs pour M20 (j) et Mdt (k) pour un patient. Le diagnostic de ce patient était une progression. Les probabilités prédites étaient $P_{M20} = 0,41$ pour M20 ($\overline{DV_{M20}} = -0,38 \pm 2,14$), et $P_{Mdt} = 0,59$ pour Mdt ($\overline{DV_{Mdt}} = 0,38 \pm 3,24$) pour la classe radionécrose.

de dimension inférieure à la représentation originale. Lorsqu'il s'agit d'interprétabilité et d'explicabilité, il apparaît également important de s'interroger sur notre capacité à comprendre les motifs présents dans un signal. Pour comprendre le signal capturé par un modèle, il est nécessaire que ce dernier capture de l'information compatible avec notre capacité d'analyse [254]. En outre, il est important de se rappeler qu'avec des ensembles de données limités tels que généralement disponibles en imagerie médicale, les performances d'un modèle s'expliquent possiblement par une part de surapprentissage dû à un ajustement au bruit et aux particularités des données. Ainsi, il est nécessaire de ne pas surinterpréter les modèles, qui peuvent avoir capturé des motifs certes prédictifs, mais d'une manière surajustée.

Dans notre application, quels types de motifs locaux interprétables peuvent être contenus dans les images ? Après inspection visuelle des images et des RDMs (cf, Figure 5.4), au-delà de la valeur des voxels en tant que telle, nous proposons d'utiliser le gradient local de l'image comme l'une des formulations a priori les plus simples et interprétables de son hétérogénéité à l'échelle du voxel.

5.3.1 . Matériels et méthodes

Calcul des gradients locaux

Pour chaque voxel de chaque image, l'intensité du gradient local a été calculée avec la bibliothèque SimpleITK en Python [423].

Représentation globale simplifiée à l'échelle du voxel

La correspondance entre les motifs capturés par les modèles et cartographiés par les RDMs, et la valeur du signal ainsi que de l'intensité locale du gradient dans les images d'entrée a été évaluée grâce à des nuages de points (graphiques de dispersion). Le nombre total de voxels contenus dans l'ensemble des ROIs pour toutes les lésions était de 933659, parmi lesquels 10000 ont été tirés aléatoirement.

Pour M20, un espace à deux dimensions a été défini, correspondant à la valeur du signal dans l'image L/Smax20 ainsi qu'à l'amplitude du gradient (gL/Smax20) pour les 10000 voxels. Pour Mdt, deux axes ont également été définis, avec d'une part la valeur du *wash-out* mesuré dans l'image TEP20-90, et d'autre part l'amplitude du gradient dans l'image L/Smean20-90 (gL/Smean20-90). Les images utilisées ici ont été choisies d'après les fonctions de décisions D_{M20} (5.8) et D_{Mdt} (5.9). Formant deux nuages de points, respectivement pour M20 et Mdt, chaque voxel v a été coloré suivant la valeur de sa décision, respectivement DV_{M20_v} et DV_{Mdt_v}.

Enfin, trois lésions typiques, avec une probabilité prédite de radionécrose élevée, faible, ou proche de $0,5$, ont été sélectionnées pour chacun des deux modèles. Une couleur leur a été affectée de sorte que les points du nuage correspondant à ces lésions soient facilement identifiables.

5.3.2 . Résultats

Calcul des gradients locaux

La Figure 5.5 montre les RDMs DV_{M20} et DV_{Mdt} présentées à la Figure 5.4, superposées sur les images d'entrée L/Smax20 et TEP20-90, ainsi que sur l'intensité des gradients locaux gL/Smax20 et gL/Smean20-90.

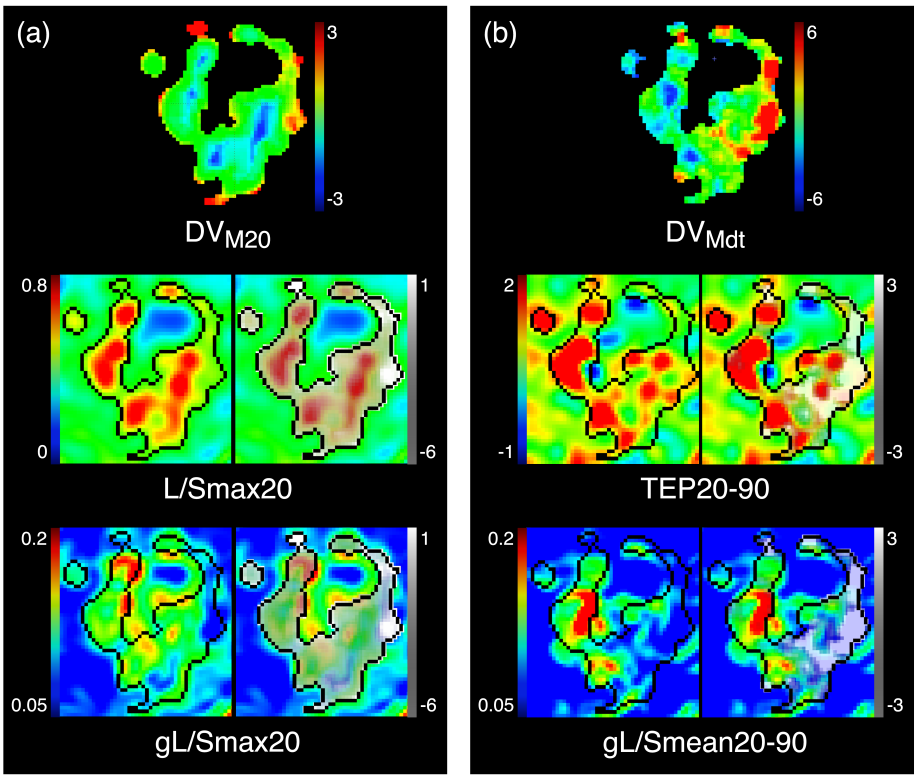


Figure 5.5 – Superposition des RDMs DV_{M20} sur les images L/Smax20 et gL/Smax20 (a), et DV_{Mdt} sur les images TEP20-90 et gL/Smean20-90 (b), pour le patient illustré en Figure 5.4. Le diagnostic de ce patient était une progression. Les probabilités prédites étaient $P_{M20} = 0,41$ pour M20 ($\overline{DV_{M20}} = -0,38 \pm 2,14$), et $P_{Mdt} = 0,59$ pour Mdt ($\overline{DV_{Mdt}} = 0,38 \pm 3,24$) pour la classe radionécrose.

Pour M20, nous observons une colocalisation bien plus marquée de la décision DV_{M20} avec le signal de l'image L/Smax20 qu'avec l'intensité de son gradient local gL/Smax20 (a). La décision DV_{Mdt} était quant à elle bien colocalisée à la fois avec le signal de l'image TEP20-90 et avec l'intensité du gradient gL/Smean20-90 (b).

Représentation globale simplifiée à l'échelle du voxel

La Figure 5.6 représente les graphiques de dispersion des voxels échantillonnés aléatoirement dans l'ensemble des données, colorés en fonction de leurs valeurs de décision prédites, avec six patients mis en évidence (vert, violet, orange dans chaque graphique).

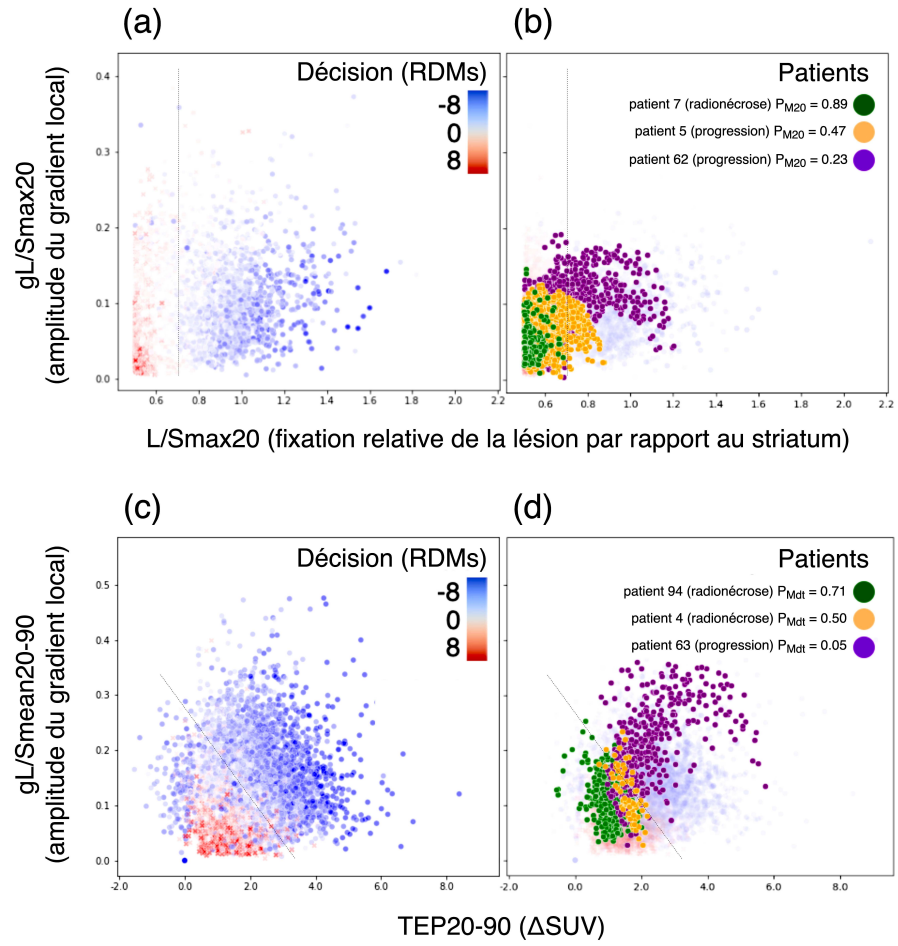


Figure 5.6 – Graphiques de dispersion des 10000 voxels échantillonnés dans l'ensemble des lésions, colorés en fonction de leurs valeurs de décision prédites, respectivement pour M20 (a, b) et Mdt (c, d), avec six patients mis en évidence (vert, violet, orange dans chaque graphique).

À 20 minutes, une probabilité élevée de radionécrose (points rouges dans la Figure 5.6 (a), points verts dans la Figure 5.6 (b)) était essentiellement associée à un faible rapport de fixation entre la lésion et le striatum (axe horizontal), avec une décision nulle autour de $L/Smax20 = 0,7$. Dans le contexte du double temps, une probabilité élevée de radionécrose (points rouges dans la Figure 5.6 (c), points verts dans la Figure 5.6 (d)) était associée à un *wash-out* lent (faible ΔSUV) et homogène (faible amplitude du gradient local), et inversement pour la prédiction de progressions.

Ces résultats correspondent aux hypothèses issues de l'observation des RDMs. Ils appuient l'importance du rapport de fixation de la lésion sur le striatum et du *wash-out* dans les prédictions des modèles. Le risque de fausse découverte lié à l'hétérogénéité du signal dans l'imagerie statique lors de la lecture de D_{M20} seule est diminué. Les nuages de points aident en outre à clarifier les explications, en s'appuyant sur les valeurs des axes définis a priori (eg, valeur de transition autour de $L/Smax20 = 0,7$ dans le contexte statique à 20 minutes).

5.4 . Modèles substituts à l'échelle du voxel et comparaisons avec des mesures simples

5.4.1 . Matériels et méthodes

Modèles substituts à l'échelle du voxel

Afin de produire deux modèles substituts simplifiés, M20' et Mdt', définis à l'échelle du voxel, nous avons utilisé la régression linéaire des moindres carrés. Pour chacun des deux modèles originaux M20 et Mdt, l'espace de décision rétroprojeté à l'échelle du voxel a été approximé par une fonction linéaire prenant en entrée les variables associées aux axes montrés en Figure 5.6 (L/Smax20 et gL/Smax20 pour M20, TEP20-90 et gL/Smean20-90 pour Mdt).

L'identification de la lésion d'appartenance de chaque voxel a permis de réaliser 1000 tirages *bootstrap* à l'échelle des lésions, tout en ajustant la régression à l'échelle du voxel. Les modèles substitut M20' et Mdt' ont ainsi été obtenus par *bagging*. Pour chaque lésion i, les prédictions globales correspondaient à la transformation logistique des valeurs moyennes des décisions prédites par M20' et Mdt' pour les voxels de la ROI tumorale telle que :

$$\hat{y}^{(i)}_{M20'} = \frac{1}{1 + e^{-\overline{DV}^{(i)}_{M20'}}} \tag{5.10}$$

et

$$\hat{y}^{(i)}_{Mdt'} = \frac{1}{1 + e^{-\overline{DV}^{(i)}_{Mdt'}}}. \tag{5.11}$$

Comparaisons avec des mesures simples

À l'échelle des ROIs, les performances OOB de M20, Mdt, M20', et Mdt' ont été comparées lors de la procédure de *bagging*, entre-elles, avec les mesures locales simples SUV, ΔSUV, L/S, et ΔL/S, ainsi qu'avec l'amplitude de leur gradient. L'AUC et la Bacc ont été choisies comme figures de mérite. Pour les modèles logistiques M20, Mdt, M20', et Mdt', le seuil de probabilité pour calculer la Bacc a été fixé à $0,5$, tandis qu'il a été optimisé sur l'échantillon *bootstrap* (entraînement) avant d'être appliqué sur l'ensemble OOB (validation) pour les autres mesures non probabilistes.

5.4.2 . Résultats

Modèles substituts à l'échelle du voxel

La Figure 5.7 montre les cartes de décision des modèles M20 et Mdt pour un patient, comparées à celles de leurs substituts simplifiés M20' et Mdt', dont les fonctions de décision sont données par les équations (5.12) et (5.13).

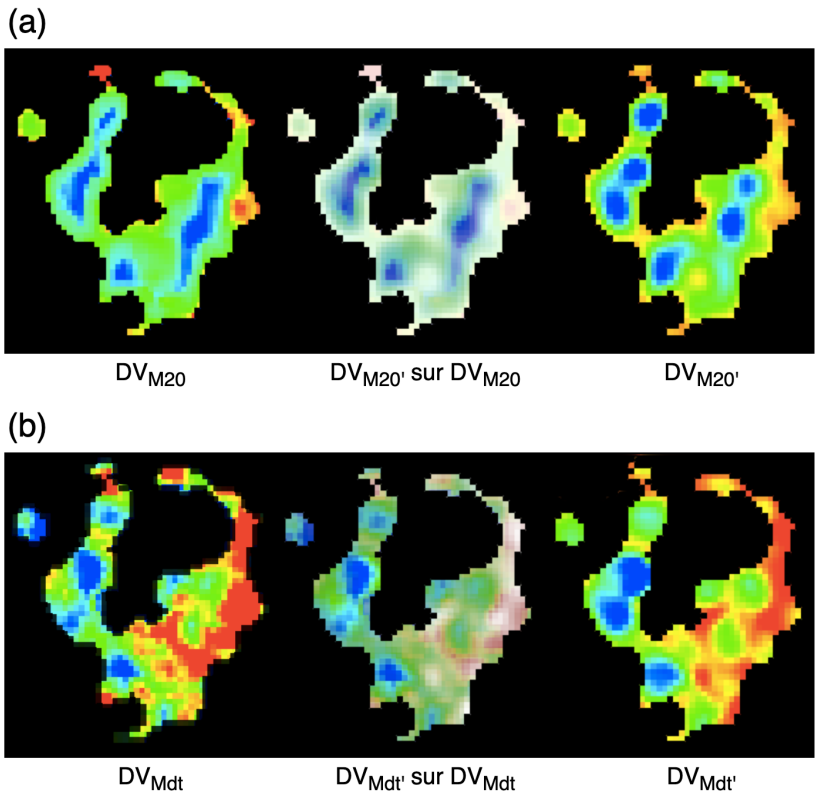


Figure 5.7 – RDMs des modèles M20 (a) et Mdt (b) pour un patient, comparées à celles de leurs substituts simplifiés M20′ et Mdt′.

$$\begin{aligned} D_{M20'} = &- 5,611(\pm 0,634) \times L/Smax20 \\ &- 3,234(\pm 0,519) \times gL/Smax20 \\ &+ 6,042(\pm 0,776) \end{aligned} \tag{5.12}$$

$$\begin{aligned} D_{Mdt'} = &- \ \ 1,450(\pm 0,160) \times TEP20\text{-}90 \\ &- 23,809(\pm 2,391) \times gL/Smean20\text{-}90 \\ &+ \ \ 4,750(\pm 0,634) \end{aligned} \tag{5.13}$$

Les coefficients de $D_{M20'}$ (5.12) et de $D_{Mdt'}$ (5.13) sont associés aux variables dans leur échelle d'origine. Leur valeur normalisée permettant d'estimer l'importance de chaque caractéristique sont reportées dans les équations (5.14) et (5.15). Elles montrent l'importance de l'hétérogénéité dans le contexte de l'imagerie double temps, avec un coefficient normalisé de $-1,479$ associé au gradient gL/Smean20-90, proche du coefficient associé à l'intensité du *wash-out* mesuré localement dans l'image TEP20-90, estimé à $-1,273$. D'autre part, le coefficient associé au gradient gL/Smax20 en imagerie standard à 20 minutes, de $-0,245$, est environ 6 fois inférieur à celui associé à la fixation relative par rapport au striatum, estimé à $-1,477$, de façon cohérente avec les observations effectuées sur la Figure 5.6, soulignant une importance minime de l'hétérogénéité pour ce modèle.

$$\begin{aligned} D^*_{M20'} = &-1,477(\pm 0,166) \times L/Smax20 \\ &-0,245(\pm 0,038) \times gL/Smax20 \\ &-0,397(\pm 0,100) \end{aligned} \tag{5.14}$$

$$\begin{aligned} D^*_{Mdt'} = &-1,273(\pm 0,160) \times TEP20\text{-}90 \\ &-1,479(\pm 0,135) \times gL/Smean20\text{-}90 \\ &-0,726(\pm 0,120) \end{aligned} \tag{5.15}$$

Comparaisons avec des mesures simples

Les performances OOB obtenues pour les modèles M20, Mdt, M20', et Mdt', ainsi que pour les variables simples SUV, ΔSUV, L/S, ΔL/S et leur gradient respectif sont reportées dans le Tableau 5.2. Globalement, les modèles M20' et Mdt' étaient fidèles aux modèles originaux en termes de performance de classification OOB. En imagerie double temps, Mdt' a donné une AUC ($-0,001$ en moyenne) et une Bacc ($-0,029$) équivalentes ou légèrement plus faibles par rapport à Mdt. Le modèle M20' a également donné une Bacc légèrement plus faible ($-0,036$), mais son AUC était quant à elle légèrement plus élevée ($+0,018$). De la même façon que pour les différences entre les modèles originaux M20 et Mdt, le nombre restreint d'individus ainsi que les faibles écarts de performances observées n'ont pas permis d'établir de différences significatives sur le plan statistique.

L'AUC OOB était supérieure à $0,7$ pour les variables L/Smean20 ($0,722$), L/Smax20 ($0,707$), gL/Smean20 ($0,719$), et gL/Smax20 ($0,710$). Cependant, à l'exeption de gL/Smean20 ($0,616$), aucune variable simple n'a atteint une Bacc supérieure à $0,6$, tel que c'est le cas pour les modèles M20 ($0,667$), Mdt ($0,695$), et leurs substituts M20' ($0,631$) et mdt' ($0,666$). Ces éléments suggèrent qu'au-delà de ranger les examens dans un ordre (*ranking*) convenable les uns par rapport aux autres (AUC), il est difficile de déterminer une valeur seuil permettant de créer deux groupes. Pour les modèles M20 et

Tableau 5.2 – Performances de classification OOB obtenues pour les modèles M20, Mdt, M20', et Mdt', ainsi que pour les variables simples SUV, ΔSUV, L/S, ΔL/S et leur gradient respectif.

OOB moyenne (± 1 écart-type) 95% CI	**AUC**	**Bacc**	**Se**	**Sp**	**Seuil (appris)**
M20	$0,692 \pm 0,093$ $[0,502\ ;0,864]$	$0,667 \pm 0,084$ $[0,486\ ;0,817]$	$0,603 \pm 0,170$ $[0,250\ ;0,901]$	$0,731 \pm 0,113$ $[0,484\ ;0,917]$	$0,5$ (non appris)
Mdt	$\mathbf{0,723 \pm 0,080}$ $[\mathbf{0,557}\ ;\mathbf{0,867}]$	$\mathbf{0,695 \pm 0,090}$ $[\mathbf{0,500}\ ;\mathbf{0,850}]$	$0,681 \pm 0,170$ $[0,310\ ;1,000]$	$0,710 \pm 0,104$ $[0,483\ ;0,885]$	$0,5$ (non appris)
SUV20	$0,501 \pm 0,102$ $[0,274\ ;0,668]$	$0,493 \pm 0,070$ $[0,332\ ;0,615]$	$0,579 \pm 0,357$ $[0,000\ ;1,000]$	$0,407 \pm 0,298$ $[0,000\ ;0,929]$	$4,143 \pm 0,658$ $[3,095\ ;5,802]$
LSmax20	$0,707 \pm 0,085$ $[0,536\ ;0,863]$	$0,595 \pm 0,071$ $[0,482,0,755]$$	$0,556 \pm 0,334$ $[0,000\ ;1,000]$	$0,633 \pm 0,328$ $[0,000\ ;1,000]$	$0,694 \pm 0,068$ $[0,553\ ;0,824]$
LSmean20	$0,722 \pm 0,077$ $[0,562\ ;0,873]$	$0,575 \pm 0,056$ $[0,490\ ;0,693]$	$0,635 \pm 0,386$ $[0,000\ ;1,000]$	$0,514 \pm 0,371$ $[0,077\ ;1,000]$	$1,041 \pm 0,118$ $[0,884\ ;1,163]$
SUV2090	$0,568 \pm 0,104$ $[0,276\ ;0,726]$	$0,536 \pm 0,075$ $[0,344\ ;0,643]$	$0,652 \pm 0,390$ $[0,000\ ;1,000]$	$0,421 \pm 0,302$ $[0,069\ ;0,920]$	$1,911 \pm 0,380$ $[1,379\ ;2,248]$
LSmax2090	$0,616 \pm 0,110$ $[0,343\ ;0,802]$	$0,563 \pm 0,079$ $[0,377\ ;0,708]$	$0,628 \pm 0,271$ $[0,000\ ;1,000]$	$0,497 \pm 0,240$ $[0,100\ ;1,000]$	$0,344 \pm 0,046$ $[0,230\ ;0,406]$
LSmean2090	$0,648 \pm 0,087$ $[0,484\ ;0,812]$	$0,576 \pm 0,070$ $[0,432\ ;0,696]$	$0,655 \pm 0,328$ $[0,083\ ;1,000]$	$0,498 \pm 0,278$ $[0,038\ ;0,966]$	$0,491 \pm 0,770$ $[0,388\ ;0,663]$
gSUV20	$0,659 \pm 0,091$ $[0,479\ ;0,823]$	$0,576 \pm 0,070$ $[0,451\ ;0,720]$	$\mathbf{0,691 \pm 0,320}$ $[\mathbf{0,000}\ ;\mathbf{1,000}]$	$0,462 \pm 0,308$ $[0,042\ ;1,000]$	$0,516 \pm 0,120$ $[0,343\ ;0,776]$
gLSmax20	$0,710 \pm 0,085$ $[0,531\ ;0,868]$	$0,594 \pm 0,072$ $[0,470\ ;0,743]$	$0,612 \pm 0,306$ $[0,000\ ;1,000]$	$0,576 \pm 0,313$ $[0,091\ ;1,000]$	$0,081 \pm 0,017$ $[0,053\ ;0,114]$
gLSmean20	$0,719 \pm 0,085$ $[0,540\ ;0,874]$	$0,616 \pm 0,085$ $[0,479\ ;0,784]$	$0,589 \pm 0,298$ $[0,000\ ;1,000]$	$0,643 \pm 0,308$ $[0,038\ ;1,000]$	$0,128 \pm 0,036$ $[0,091\ ;0,255]$
gSUV2090	$0,602 \pm 0,104$ $[0,303\ ;0,764]$	$0,555 \pm 0,080$ $[0,369\ ;0,696]$	$0,681 \pm 0,285$ $[0,000\ ;1,000]$	$0,430 \pm 0,245$ $[0,077\ ;1,000]$	$0,456 \pm 0,071$ $[0,301\ ;0,595]$
gLSmax2090	$0,692 \pm 0,081$ $[0,525\ ;0,844]$	$0,565 \pm 0,660$ $[0,476\ ;0,723]$	$0,397 \pm 0,363$ $[0,000\ ;1,000]$	$\mathbf{0,733 \pm 0,306}$ $[\mathbf{0,138}\ ;\mathbf{1,000}]$	$0,063 \pm 0,014$ $[0,046\ ;0,095]$
gLSmean2090	$0,688 \pm 0,082$ $[0,523\ ;0,849]$	$0,597 \pm 0,076$ $[0,479\ ;0,773]$	$0,565 \pm 0,325$ $[0,000\ ;1,000]$	$0,629 \pm 0,335$ $[0,038\ ;1,000]$	$0,116 \pm 0,031$ $[0,072\ ;0,208]$
M20'	$0,710 \pm 0,080$ $[0,543\ ;0,851]$	$0,631 \pm 0,077$ $[0,481\ ;0,786]$	$0,431 \pm 0,173$ $[0,100\ ;0,750]$	$0,830 \pm 0,097$ $[0,607\ ;0,966]$	$0,5$ (non appris)
Mdt'	$0,722 \pm 0,079$ $[0,556\ ;0,864]$	$0,666 \pm 0,072$ $[0,518\ ;0,803]$	$0,694 \pm 0,162$ $[0,364\ ;1,000]$	$0,638 \pm 0,114$ $[0,440\ ;0,857]$	$0,5$ (non appris)

Mdt, l'optimisation des hyperparamètres et la sélection des caractéristiques ont été faites avec l'ASB (3.26). Leur ajustement lors de l'entraînement a été fait avec l'entropie croisée équilibrée. Ces métriques pouvant être vues comme des extensions continues et probabilistes de la Bacc (3.23), il semble qu'elles aient donné à M20 et Mdt un léger avantage dans le contexte d'une classification binaire. M20' et Mdt' ayant été ajustés pour approcher la décision à l'échelle du voxel prédite respectivement par M20 et Mdt, ils ont également bénéficié de cet avantage.

De façon intéressante, le seuil moyen estimé pour la variable L/Smax20 était de $0,694 \pm 0,068$, cohérent avec la zone de transition de la décision prédite à l'échelle du voxel observée aux alentours de $0,7$ en Figure 5.6.

En outre, bien que Mdt et Mdt' apparaissent comme étant les plus performants dans l'ensemble, les variables simples associées aux performances les plus élevées sont celles issues de l'imagerie standard à 20 minutes.

5.5 . Exportabilité des résultats à des patients atteints de métastases cérébrales

Dans cette partie, nous avons testé nos modèles originaux et substituts, ainsi que les variables simples, pour réaliser le diagnostic différentiel entre radionécrose et progression sur un petit ensemble de patients atteints de métastases cérébrales. Ainsi, l'échantillon concerné n'était pas issu de la même population de patients. Nous ne pouvons donc pas parler de validation externe stricto sensu. Les patients métastatiques présentant néanmoins des phénotypes analogues aux gliomes en imagerie dans ce contexte, ce travail correspond approximativement à une évaluation de l'exportabilité hors distribution de nos résultats.

5.5.1 . Matériels et méthodes

5.5.1.1 . Patients et données

L'échantillon comportait 30 lésions provenant de 26 patients[4] atteints de métastases cérébrales suite à différents cancers, traitées par radiothérapie et chirurgie éventuelle.

Critères d'inclusion et d'exclusion

Les critères d'inclusion et d'exclusion étaient différents de ceux utilisés pour l'échantillon de patients atteints de gliomes. En effet, l'atteinte métastatique étant variable, la prise en charge des patients est moins homogène que lors d'une atteinte primaire.

Pour au moins une lésion pour chaque patient, l'imagerie de suivi était litigieuse pour la différentiation entre progression et toxicité liée à la radiothérapie. Les critères d'inclusion comprenaient ainsi une IRM cérébrale de suivi disponible et un suivi clinique minimum de trois mois.

Les patients qui suivaient un traitement par anti-VEGF au moment de l'acquisition de la TEP/TDM ont été exclus.

Enfin, comme pour les gliomes, les lésions présentant des examens IRM suspects mais sans fixation en TEP n'ont pas été prises en compte. Seules les images des lésions présentant une fixation suspecte (positive) ont été incluses.

Caractéristiques de l'échantillon

Une chirurgie a été menée pour 9 (30,0%) patients, permettant une confirmation biologique du diagnostic final de leurs lésions. Pour 2 (6,7%) patients, le suivi était limité car ils ont été orientés vers les soins palliatifs.

Diagnostic final

Basé sur un suivi de trois à six mois en l'absence de confirmation biologique, 7 (23,3%) radionécroses et 23 (76,7%) progressions ont été observées.

4. Quatre patients présentaient deux lésions litigieuses.

Dans cet échantillon, l'incertitude liée au diagnostic final inclut non seulement la possible co-existence de contingents tumoraux et inflammatoires, mais aussi les variabilités dans les manifestations et le suivi des métastases cérébrales, notamment pour les patients placés en soins palliatifs.

Protocole d'imagerie

Comme dans les gliomes, les patients devaient être à jeun de protéines depuis au moins quatre heures avant l'imagerie TEP/TDM à la 18F-FDOPA. Ceux sans contre-indication ont reçu $100mg$ de Carbidopa en prémédication une heure avant l'injection de $2MBq \times kg^{-1}$ de radiotraceur. Les images ont été acquises sur le même scanner avec le même protocole.

5.5.1.2 . Traitement et représentation des images, et classification des lésions

Les images ont été traitées exactement de la même façon que pour les gliomes, de même que pour l'extraction des caractéristiques radiomiques et leur représentation simplifiée.

Pour la classification, nous avons utilisé l'AUC et la Bacc comme figures de mérite. Les modèles M20, Mdt, M20', et Mdt' entrainés ont été utilisés pour prédire la probabilité de radionécrose pour chaque lésion. Un seuil de $0,5$ a été appliqué pour le calcul de la Bacc. Concernant les variables simples, le seuil utilisé était celui estimé sur la base de donnée de patients atteints de gliomes (Tableau 5.2). Afin d'estimer la dispersion des performances, 1000 tirages *bootstrap* ont été réalisés pour chaque modèle et chaque variable simple.

5.5.2 . Résultats

Les performances obtenues pour les modèles M20, Mdt, M20', et Mdt', ainsi que pour les variables simples SUV, ΔSUV, L/S, ΔL/S et leur gradient respectif sont reportées dans le Tableau 5.3.

Avec des performances globalement plus faibles, il est cependant difficile d'observer une cohérence forte avec l'ensemble des résultats obtenus sur les données gliomes (Tableau 5.2). Par exemple, alors que la valeur moyenne du signal dans la ROI à 20 minutes (SUV20) donnait les performances les plus basses pour les gliomes ($AUC = 0,501$ et $Bacc = 0,493$), elle est ici associée à une AUC de $0,639$ (quatrième AUC la plus élevée) et une Bacc de $0,584$. Le modèle Mdt ($AUC = 0,582$ et $Bacc = 0,644$) demeure néanmoins supérieur au modèle M20 ($AUC = 0,435$ et $Bacc = 0,304$). De façon intéressante, l'AUC la plus élevée $(0,691)$ et la seconde Bacc la plus élevée $(0,665)$ sont associées à l'intensité moyenne du *wash-out* dans la ROI (SUV20-90). En outre, la Bacc la plus élevée $(0,670)$ est associée à la fixation relative par rapport au SUVmax striatal en imagerie standard à 20 minutes (L/Smax20). Enfin, les modèles substituts M20' et Mdt' ont globalement conduit à de meilleurs scores que les modèles M20 et Mdt.

Tableau 5.3 – Performances de classification obtenues pour les patients atteints de métastases cérébrales avec les modèles M20, Mdt, M20', et Mdt' entraînés, ainsi que les variables simples SUV, ΔSUV, L/S, ΔL/S et leur gradient respectif.

***bootstrap* moyenne (± 1 écart-type) 95% CI**	**AUC**	**Bacc**	**Se**	**Sp**	**Seuil**
M20	$0,435 \pm 0,114$ $[0,215 ; 0,653]$	$0,304 \pm 0,051$ $[0,204 ; 0,400]$	$0,000 \pm 0,000$ $[0,000 ; 0,000]$	$0,608 \pm 0,102$ $[0,409 ; 0,800]$	$0,5$
Mdt	$\mathbf{0,582 \pm 0,132}$ $[\mathbf{0,310} ; \mathbf{0,835}]$	$\mathbf{0,644 \pm 0,089}$ $[\mathbf{0,444} ; \mathbf{0,795}]$	$0,849 \pm 0,142$ $[0,500 ; 1,000]$	$0,440 \pm 0,108$ $[0,240 ; 0,667]$	$0,5$
SUV20	$0,639 \pm 0,110$ $[0,418 ; 0,833]$	$0,584 \pm 0,086$ $[0,405 ; 0,729]$	$0,866 \pm 0,142$ $[0,500 ; 1,000]$	$0,303 \pm 0,095$ $[0,130 ; 0,500]$	$4,143$
LSmax20	$0,538 \pm 0,102$ $[0,335 ; 0,732]$	$0,670 \pm 0,091$ $[0,479, 0,826]$$	$0,866 \pm 0,142$ $[0,500 ; 1,000]$	$0,470 \pm 0,108$ $[0,273 ; 0,692]$	$0,694$
LSmean20	$0,563 \pm 0,113$ $[0,339 ; 0,773]$	$0,653 \pm 0,049$ $[0,562 ; 0,750]$	$1,000 \pm 0,000$ $[1,000 ; 1,000]$	$0,306 \pm 0,098$ $[0,125 ; 0,500]$	$1,041$
SUV2090	$0,691 \pm 0,115$ $[0,442 ; 0,903]$	$0,665 \pm 0,089$ $[0,477 ; 0,818]$	$0,849 \pm 0,143$ $[0,500 ; 1,000]$	$0,481 \pm 0,105$ $[0,273 ; 0,682]$	$1,911$
LSmax2090	$0,385 \pm 0,123$ $[0,159 ; 0,640]$	$0,479 \pm 0,103$ $[0,259 ; 0,667]$	$0,700 \pm 0,185$ $[0,333 ; 1,000]$	$0,258 \pm 0,091$ $[0,095 ; 0,450]$	$0,344$
LSmean2090	$0,653 \pm 0,119$ $[0,394 ; 0,868]$	$0,577 \pm 0,087$ $[0,375 ; 0,717]$	$0,849 \pm 0,143$ $[0,500 ; 1,000]$	$0,305 \pm 0,278$ $[0,130 ; 0,500]$	$0,491$
gSUV20	$0,564 \pm 0,133$ $[0,296 ; 0,816]$	$0,532 \pm 0,106$ $[0,317 ; 0,731]$	$\mathbf{0,715 \pm 0,185}$ $[\mathbf{0,333} ; \mathbf{1,000}]$	$0,350 \pm 0,102$ $[0,158 ; 0,550]$	$0,516$
gLSmax20	$0,488 \pm 0,109$ $[0,279 ; 0,709]$	$0,497 \pm 0,110$ $[0,280 ; 0,708]$	$0,560 \pm 0,195$ $[0,167 ; 1,000]$	$0,434 \pm 0,195$ $[0,167 ; 1,000]$	$0,081$
gLSmean20	$0,494 \pm 0,110$ $[0,284 ; 0,715]$	$0,497 \pm 0,110$ $[0,280 ; 0,708]$	$0,560 \pm 0,195$ $[0,167 ; 1,000]$	$0,434 \pm 0,195$ $[0,167 ; 1,000]$	$0,128$
gSUV2090	$0,589 \pm 0,130$ $[0,326 ; 0,832]$	$0,510 \pm 0,107$ $[0,292 ; 0,704]$	$0,715 \pm 0,185$ $[0,333 ; 1,000]$	$0,305 \pm 0,097$ $[0,130 ; 0,520]$	$0,456$
gLSmax2090	$0,498 \pm 0,114$ $[0,278 ; 0,732]$	$0,467 \pm 0,106$ $[0,283 ; 0,683]$	$0,282 \pm 0,185$ $[0,000 ; 0,667]$	$\mathbf{0,653 \pm 0,101}$ $[\mathbf{0,454} ; \mathbf{0,850}]$	$0,063$
gLSmean2090	$0,540 \pm 0,119$ $[0,304 ; 0,770]$	$0,578 \pm 0,087$ $[0,396 ; 0,729]$	$0,849 \pm 0,143$ $[0,500 ; 1,000]$	$0,307 \pm 0,096$ $[0,136 ; 0,500]$	$0,116$
M20'	$0,550 \pm 0,112$ $[0,323 ; 0,760]$	$0,594 \pm 0,109$ $[0,373 ; 0,800]$	$0,710 \pm 0,184$ $[0,333 ; 1,000]$	$0,479 \pm 0,108$ $[0,273 ; 0,692]$	$0,5$
Mdt'	$0,655 \pm 0,127$ $[0,365 ; 0,889]$	$0,604 \pm 0,110$ $[0,380 ; 0,824]$	$0,422 \pm 0,203$ $[0,000 ; 0,833]$	$0,785 \pm 0,088$ $[0,608 ; 0,950]$	$0,5$

5.6 . Discussion

Nous avons proposé dans ce travail d'appliquer notre méthode de cartographie de décision radiomique RDM à une problématique constituant un défi majeur dans la prise en charge des patients atteints de tumeurs cérébrales. En effet, le diagnostic différentiel entre récidive tumorale et nécrose induite par le traitement peut être difficile, alors qu'il est crucial pour le choix du traitement. En plus de l'approche radiomique guidée par les données, l'introduction de l'imagerie double temps a permis une caractérisation cinétique simplifiée de la fixation de la 18F-FDOPA entre 20 minutes et 90 minutes après l'injection du radiotraceur. Afin d'étudier la plus-value apportée par cette approche, nous avons construit deux modèles : un basé uniquement sur l'imagerie standard à 20 minutes (M20), et un autre utilisant l'imagerie double temps (Mdt).

L'utilisation des RDMs pour M20 et Mdt, en plus de leurs équations linéaires de décision, a permis de faire émerger des intuitions quant au signal capturé dans les images pour prédire le diagnostic différentiel. Dans le contexte statique à 20 minutes postinjection, le modèle M20 semble principalement avoir capturé la fixation relative moyenne dans la ROI lésionnelle par rapport au striatum. En imagerie double temps (Mdt), un motif d'homogénéité ou d'hétérogénéité semble s'être ajouté à l'intensité ou la vitesse du *wash-out*.

Comparativement à l'étude sur les STS présentée au Chapitre 4, l'observation des RDMs était plus opaque, et il demeurait difficile d'interpréter les modèles. Nous nous sommes donc demandés quels motifs de signal interprétables a priori pouvaient être extraits à l'échelle du voxel. En observant les images, nous avons considéré deux caractéristiques : l'intensité locale des signaux (SUV, L/S, ΔSUV, ΔL/S), et leur homogénéité ou hétérogénéité locale, que nous avons choisi de formuler de la manière la plus simple en utilisant l'amplitude du gradient pour tous les voxels dans les ROIs. L'examen de la valeur de décision pour chaque voxel dans l'ensemble de la cohorte en fonction de ces caractéristiques à l'aide de diagrammes de dispersion a permis de renforcer nos interprétations. Au-delà d'une simplification des modèles les rendant compréhensibles plus facilement, cette substitution a permis de mitiger le risque de fausse découverte à propos de l'hétérogénéité du signal dans le contexte statique, qui aurait pu être interprété comme importante d'après la fonction de décision D_{M20}.

En termes de performance de classification, bien que la petite taille de l'échantillon de patients ne permette pas de conclure de façon significative quant à sa supériorité, le modèle Mdt a donné une Bacc, une AUC, et un ASB légèrement plus élevés que le modèle M20 en validation croisée, comme dans les résultats d'Orlhac et al. impliquant l'essentiel des patients présents dans notre étude [410]. C'est également le cas du modèle Mdt' comparativement à M20'. À l'opposé, en utilisant les variables simples extraites des images en analyse univariée, celles issues de l'imagerie standard à 20 minutes présentaient des performances légèrement meilleures. En outre, lors de leur déploiement sur les données de métastases cérébrales, les modèles substituts ont conduit à de meilleures performances que les modèles originaux desquels ils proviennent. Malgré la faible valeur statistique de ces observations, cela suggère qu'il est pertinent de simplifier les modèles radiomiques a posteriori pour les rendre plus généralisables. Plus généralement, l'utilisation de la radiomique n'a pas apporté de plus-value substantielle dans notre cas.

Ces résultats sont cohérents avec les études de Zaragori et al. ($n = 51$, Figure 5.8) [406] et Lohmann et al. ($n = 34$) [414], concluant que l'approche dynamique pouvait s'avérer discriminante pour différencier la récidive tumorale de la toxicité liée au traitement, mais qu'elle n'améliorait pas signi-

ficativement ce diagnostic différentiel par rapport à l'imagerie standard (Figure 5.8). En outre, Zaragori et al. ont discuté l'inclusion de lésions ne fixant pas significativement la 18F-FDOPA dans leur étude. En effet, elle a probablement conduit à des statistiques de performance trop optimistes, notamment en imagerie statique, là où l'analyse dynamique a pu s'avérer trompeuse en cas de trop faible signal. Bien qu'il aurait été intéressant de comparer nos résultats à ceux de travaux n'incluant que les lésions avides de radiotraceur et classant d'emblée les TEP négatives comme radionécrotiques, aucune étude correspondant à cette situation n'a encore été publiée à notre connaissance. De façon également concordante, au-delà de l'apport marginal de la radiomique globalement par rapport à des variables plus simples, Ahrari et al. ont obtenu des performances similaires entre leurs modèles statiques et ceux incluant des caractéristiques dynamiques, avec toutefois une légère supériorité pour ces derniers [415].

Les gliomes de bas grade présentent généralement un *wash-out* peu marqué [405, 424], similaire aux courbes d'activité temporelle des tissus et lésions en situation d'inflammations radiques [388, 406]. Bien que les objectifs de leur étude et leur protocole étaient légèrement différents des nôtres, nos résultats sont ainsi comparables à ceux de Lohmann et al. en TEP double temps à la 18F-FET pour l'identification de gliomes de haut grade ($n = 36$) [409]. Dans la réalisation du diagnostic différentiel, les auteurs ont en effet rapporté une performance très légèrement supérieure pour la baisse de la fixation relative par rapport au cerveau sain, que pour son intensité en imagerie standard, associée à deux faux positifs supplémentaires (Figure 5.9).

Un élément intéressant de nos résultats concerne les seuils estimés en imagerie standard pour maximiser les performances de classification. Une valeur de fixation normalisée par le SUVmax striatal proche de $0,7$ en moyenne, et proche de 1.0 lorsqu'elle est normalisée par le SUVmean striatal, appuie l'hypothèse de la séparation des patients de part et d'autre de l'isointensité entre la lésion et le striatum. Néanmoins, de façon cohérente avec la variabilité de seuils identifiés dans la littérature, nous avons observé une variabilité lors de l'estimation des seuils optimaux lors de l'analyse univariée (dernière colonne du Tableau 5.2).

Alors que les imageries dynamiques et double temps semblent donner des résultats concordants s'inscrivant dans notre progression quant à la compréhension des mécanismes associés aux cancers cérébraux traités par chirurgies et radiochimiothérapie, la question de sa valeur ajoutée en routine clinique demeure. En effet, la réalisation d'une acquisition TEP tardive, voire l'utilisation d'un protocole d'imagerie dynamique, complexifient l'examen.

Notre étude admet des limites importantes. Tout d'abord, un point commun de nos travaux avec les études abordées ici est le faible nombre d'individus.

Comme expliqué en Section 5.2.1.1, l'échantillon étudié présentait également de probables biais d'échantillonnage liés notamment à l'âge au diagnostic et au $ratio_{sexe}$. De plus, la vérité terrain utilisée n'était pas définie de façon certaine, et uniquement 27 lésions disposaient d'une preuve anatomopathologique. Même chez ces patients, l'équipe du CAL a observé que le diagnostic entre récidive tumorale et radionécrose était parfois difficile en raison de la coexistence possible de ces deux entités histologiques sur les prélèvements. Il est alors fréquent que même l'analyse anatomo-pathologique ne puisse pas classer strictement certaines lésions. C'est pourtant à l'échelle des lésions que nous avons optimisé nos modèles originaux et évalué les variables simples.

D'autres limites affectent notre étude. Toutes les images ont été acquises sur la même machine. Ceci évite plusieurs difficultés pour les études rétrospectives et de petites cohortes, mais pénalise l'aptitude des modèles à se généraliser, l'estimation des seuils et des performances. De plus, notre approche comprend plusieurs étapes, de la segmentation à l'interprétation des modèles et à leur reformulation, en passant par le recalage des images par exemple. Par conséquent, le risque de propagation d'erreurs affectant les résultats finaux est élevé. Un traitement totalement automatique permettrait d'atténuer ce risque en homogénéisant les approches, à condition d'être performant, ce qui n'est pas toujours le cas en particulier pour la segmentation des lésions.

Tel que discuté par Rollet dans sa thèse de médecine, il est possible que la 18F-FDOPA soit limitée de façon intrinsèque dans la réalisation du diagnostic différentiel étudié ici. En effet, certaines fixations non spécifiques peuvent être observées en cas d'inflammation et lors de crises d'épilepsie. Ces deux phénomènes peuvent être rencontrés dans des proportions variables en cas de toxicité, mais également dans des contextes tumoraux [384]. Cette limite souligne l'importance de la mise en perspective des résultats de l'analyse des images TEP avec ceux issus d'autres modalités et d'autres examens. L'interprétabilité joue un rôle important pour cette mise en perspective, car c'est précisément elle qui permet à l'équipe médicale d'associer toutes les informations qu'elle détient (Section 3.3.1.1 du Chapitre 3).

Enfin, de même que pour le travail présenté dans le chapitre précédent et conséquence de la taille limitée de notre échantillon, il n'est pas exclu qu'une part de surapprentissage soit associée aux modèles M20 et Mdt, et ainsi à leurs substituts M20' et Mdt'. Cette hypothèse est étayée par les performances estimées sur la base de données de métastases cérébrales. Bien qu'il soit impossible d'attribuer ces observations au surapprentissage de façon certaine, notamment du fait de la très faible taille de l'échantillon et parce qu'il ne s'agissait pas de la même pathologie, les résultats obtenus sur les gliomes n'ont pas été strictement reproduits sur les métastases cérébrales.

Néanmoins, ils demeurent cohérents, avec une progression plutôt associée à une fixation relative élevée par rapport au striatum ou un *wash-out* intense, et une plus-value marginale de l'information cinétique par rapport à l'imagerie standard.

5.7 . Conclusion

L'application de diagnostic différentiel traitée dans ce chapitre nous a permis de proposer une chaîne d'analyse radiomique interprétable, aboutissant à une compréhension des motifs utilisés par les modèles pour produire un résultat. La représentation graphique par les nuages de points a permis d'allier interprétation globale (pour tous les individus) et caractérisation à l'échelle du voxel. En utilisant les RDMs pour interpréter les modèles radiomiques construits à partir d'un nombre limité de lésions, notre approche guidée par les données a mis en évidence deux motifs discriminant la radionécrose de la progression tumorale en accord avec la littérature sur le sujet : la fixation relative de la lésion par rapport au striatum en imagerie statique et l'intensité du *wash-out* en imagerie dynamique. L'hétérogénéité locale du *wash-out* mesurée à l'aide du gradient apparaît également comme motif discriminant potentiel qu'il serait intéressant de tester à plus grande échelle.

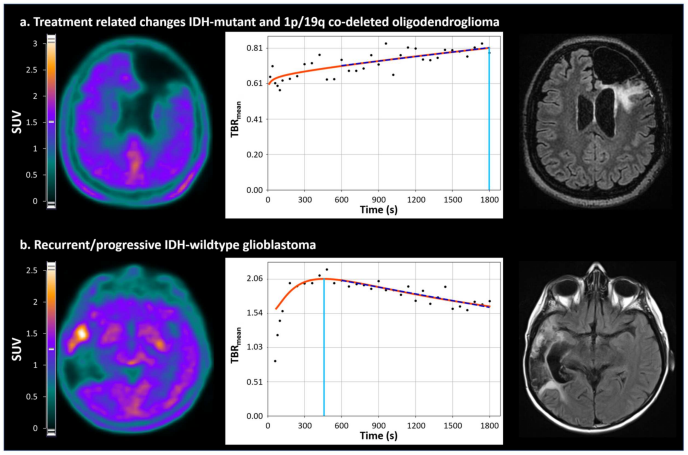
En outre, nos résultats suggèrent que l'imagerie double temps et la radiomique apportent une plus-value marginale en termes de performance de classification. Néanmoins, cette complexité n'est probablement pas justifiée compte tenu de la charge associée à l'acquisition d'une image supplémentaire tardive, et aux risques liés à la généralisation et à l'interprétation des modèles radiomiques par rapport aux variables simples extraites de l'imagerie standard. L'entraînement de modèles linéaires généralisés prenant en entrée des variables simples semble être un compromis intéressant. En plus d'être interprétables par nature, ils pourraient produire des informations probabilistes, robustes, et pertinentes à l'échelle du patient dans sa prise en charge.

Il est important de noter ici que notre méthode d'interprétation permet d'investiguer l'information capturée par le modèle, plus que celle réellement pertinente pour répondre au problème clinique ou scientifique, même si ces deux types d'information sont liés si le modèle est performant. Interpréter un modèle ne suffit pas pour conclure sur les données. Par exemple, si les données ne représentent pas la population de façon fidèle et qu'un motif apparaît discriminant de façon fortuite dans un échantillon étudié, l'interprétation de ce motif n'en fait pas une information réellement pertinente en routine.

De façon générale, les limites de cette étude sont essentiellement liées à la petite taille des cohortes étudiées. Le glioblastome est une maladie rare, dont l'incidence annuelle est estimée à 3 à 5 nouveaux cas pour 100000 habitants par an. Il est donc difficile de constituer une cohorte de grande taille

permettant une puissance statistique élevée. Dans ce contexte, il serait pertinent de constituer des cohortes plus grandes avec d'autres équipes, comme cela a été initié en France par le CAL, les centres hospitaliers universitaires et de recherche (CHU, CHRU) de Nice, Rennes, Nîmes, Montpellier, et Nancy, l'assistance publique des hôpitaux de Paris (AP-HP), et la plateforme d'imagerie moléculaire Nancyclotep [425].

(a)



(b)

Parameter	Overall	No recurrence/progression (*n* = 17)	Recurrence/progression (*n* = 34)	Adjusted *p* value
Static				
TBR_{max}	2.22 [1.36; 3.12]	1.26 [1.12; 1.37]	2.71 [2.20; 3.69]	< 0.001
TBR_{mean}	1.57 [1.02; 2.13]	0.84 [0.75; 1.04]	1.95 [1.57; 2.70]	< 0.001
TSR_{max}	1.31 [0.82; 1.90]	0.77 [0.65; 0.82]	1.64 [1.31; 2.28]	< 0.001
TSR_{mean}	0.92 [0.65; 1.43]	0.51 [0.46; 0.65]	1.19 [0.92; 1.69]	< 0.001
MTV	1.49 [0.00; 8.22]	0.00 [0.00; 0.00]	5.35 [1.49; 15.75]	< 0.001
Dynamic				
TTP	7.70 [3.35; 18.65]	14.53 [1.55; 30.00]	7.67 [4.13; 14.62]	1
Slope	− 0.14 [− 0.82; 0.13]	0.07 [− 0.07; 0.31]	− 0.59 [− 0.94; − 0.08]	< 0.001

MTV metabolic tumor volume, *TBR* tumor-to-normal brain ratio, TSR tumor-to-striatum ratio

(c)

	AUC	CI (95%) AUC	Threshold	Sensitivity	Specificity	Accuracy
TBR_{max}	0.969	(0.923–1.0)	1.61	97.1%	94.1%	96.0%
TBR_{mean}	0.983	(0.956–1.0)	1.3	94.1%	94.1%	94.1%
TSR_{max}	0.976	(0.939–1.0)	1.0	97.1%	94.1%	96.0%
TSR_{mean}	0.986	(0.964–1.0)	0.83	91.2%	100%	94.1%
MTV	0.978	(0.939–1.0)	0.045 mL	97.1%	94.1%	96.0%
Slope	0.818	(0.702–0.935)	− 0.26 h^{-1}	67.6%	94.1%	76.5%

AUC area under the curve, *CI* confidence interval, *MTV* metabolic tumor volume, *TBR* tumor-to-normal brain ratio, *TSR* tumor-to-striatum ratio

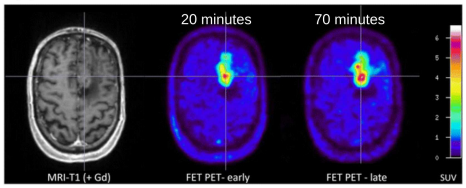
(d)

Parameter	Coefficient	*p* value
Intercept	− 10.179	0.001*
TBR_{mean}	–	0.244
TBR_{max}	–	0.605
TSR_{mean}	–	0.116
TSR_{max}	10.039	0.002*
MTV	–	0.729
Slope	–	0.380

*Indicates parameters ultimately included in the final multivariable model
MTV metabolic tumor volume, *TBR* tumor-to-normal brain ratio, *TSR* tumor-to-striatum ratio

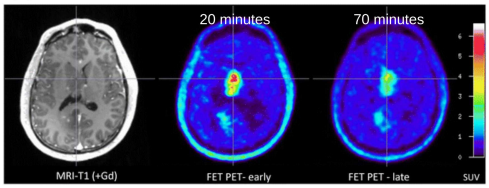
Figure 5.8 – Synthèse des résultats de l'étude de Zaragori et al. en TEP dynamique à la 18F-FDOPA pour le diagnostic différentiel entre progression et radionécrose. (a) Coupes d'images TEP à la 18F-FDOPA et IRM associées aux courbes temporelles de fixation du radiotraceur pour deux patients. (b) Médiane [intervalle interquartile] de caractéristiques TEP dans l'ensemble de l'échantillon étudié et dans les deux classes de patients. (c) Résultats des analyses de courbes ROC pour l'identification des patients présentant une progression. (d) Résultats multivariés en régression logistique pour la prédiction de la progression à 6 mois après la TEP à la 18F-FDOPA. Figure créée à partir de Zaragori et al. [406]

(a)



Oligoastrocytome de grade 2.
Variation de la fixation relative par rapport au cerveau sain : +13%

(b)



Glioblastome de haut grade.
Variation de la fixation relative par rapport au cerveau sain : -21%

(c)

Identification of high-grade glioma	TBR_{mean} > 2.0[a]	TTP < 35 min[b]	TAC patterns II or III[c]	Early vs. late PET: change of TBR_{mean} < - 8 %
True positives	20	14	21	20
False positives	5	1	3	3
True negatives	7	11	9	9
False negatives	4	10	3	4
Sensitivity	83 %	58 %	88 %	83 %
Specificity	58 %	92 %	75 %	75 %
Accuracy	75 %	69 %	83 %	81 %
AUC ± SE	0.65 ± 0.10	0.76 ± 0.08	-	0.85 ± 0.06
p value	0.15 (n.s.)	0.01	<0.001	<0.001

[a] TBR_{mean} = ratio of mean ^{18}F-FET uptake in the tumour divided by mean uptake in the brain
[b] TTP = time-to-peak of the time-activity curve of ^{18}F-FET uptake
[c] TAC = time-activity curve; for definition of patterns, see text

Figure 5.9 – Synthèse des résultats de Lohmann et al. en TEP double temps à la 18F-FET pour la détection de gliomes de haut grade. (a, b) Coupes d'images TEP à la 18F-FDOPA précoces et tardives et IRM pour deux patients. (c) Métriques de performance pour l'identification des patients atteints de gliomes de haut grade. Figure créée à partir de Lohmann et al. [409]

6 - Tumeur primaire et atteinte ganglionnaire dans le cancer du sein : Étude comparative de la radiomique classique, profonde, et de mesures conventionnelles en TEP au FDG

L'étude présentée ici porte sur des patientes atteintes de cancer du sein, prises en charge à l'Institut Curie, et dont les examens d'imagerie ont été réalisés à l'hôpital de Saint-Cloud (France). Ce chapitre compare différentes approches pour prédire l'atteinte ganglionnaire à partir des caractéristiques de la tumeur primaire lors du bilan d'extension. Les résultats ont été présentés sous la forme d'un poster au congrès 2022 de l'Association Européenne de Médecine Nucléaire (*European Association of Nuclear Medicine* (EANM)) [426].

6.1 . Introduction

Le cancer du sein est celui qui cause le plus grand nombre de décès féminins par cancer. Bien que son incidence augmente depuis les années 90, la survie nette à 5 ans normalisée en fonction de l'âge s'améliore au cours du temps. Cela s'explique en partie par l'amélioration des traitements, mais aussi par un diagnostic précoce et personnalisé, de plus en plus adapté aux patientes et à leur niveau de risque (Figure 6.1) [427].

La détermination de l'atteinte ganglionnaire axillaire est une étape clé dans la prise en charge des patientes [428]. Elle détermine si un curage doit être réalisé d'emblée. En l'absence de détection clinique ou par imagerie, la technique du ganglion sentinelle sera proposée. En outre, l'estimation de l'atteinte axillaire est une information importante pour la stratification du risque de rechute tumorale, qui permettrait une optimisation de la prise en charge des patientes [429].

Aujourd'hui, la spécificité de détection des atteintes ganglionnaires axillaires est aux alentours de 90% en TEP au FDG [430]. Cependant, sa sensibilité demeure modérée, environ 55% à 75% selon les études [431-433], ce qui peut s'expliquer en partie par le faible volume ou la faible captation de FDG des ganglions.

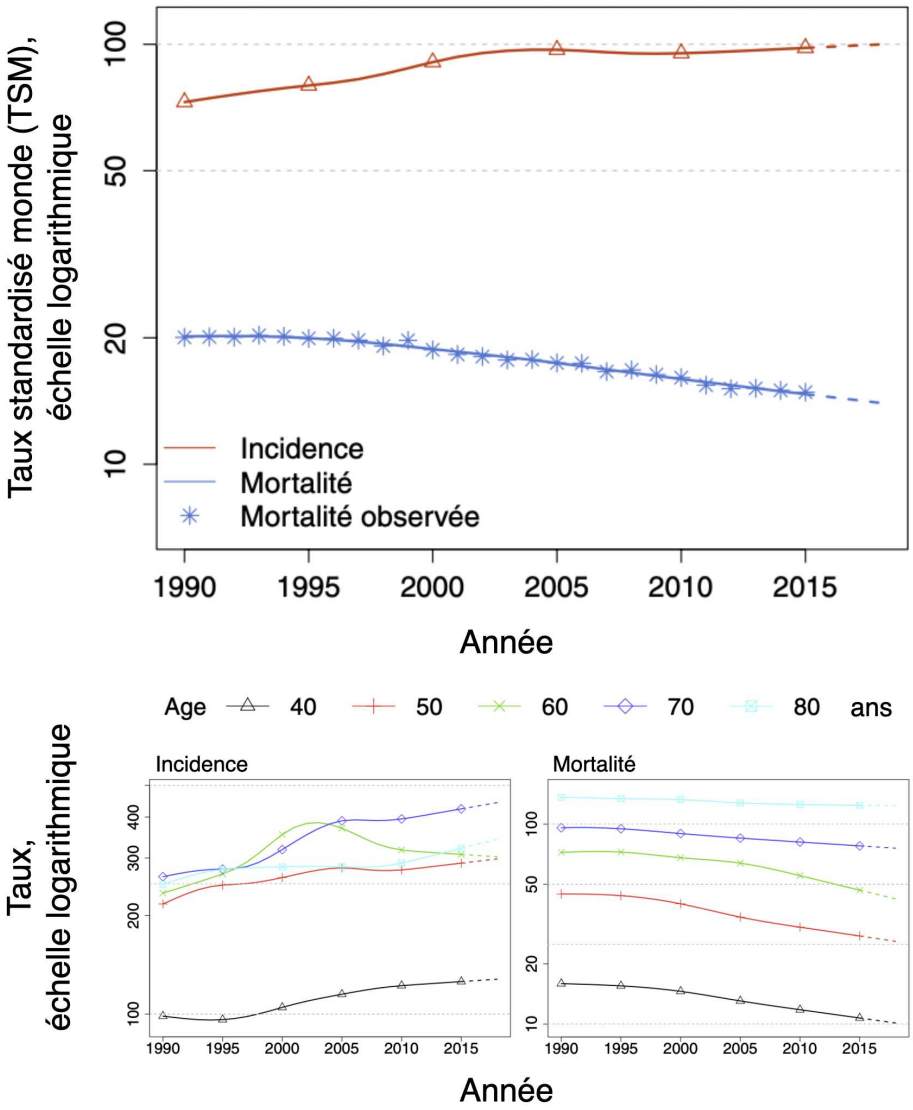


Figure 6.1 – Taux d'incidence et de mortalité en cancer du sein en France selon l'année et par âge. Le taux standardisé monde (TSM) signifie que la normalisation a été réalisée selon la pyramide des âges estimée dans la population mondiale. Figure créée à partir de la rubrique sur le cancer du sein du site web de l'Institut National du Cancer [427].

Plusieurs études en cancer du sein suggèrent que l'imagerie de la tumeur primitive contient de l'information pouvant être associée à l'atteinte ganglionnaire axillaire [434-445], et plus généralement, au risque de développement de métastases [446, 447]. À terme, une caractérisation précise impliquant l'imagerie pourrait ainsi s'intégrer, entre autres, au processus de désescalade de l'acte chirurgical axillaire.

Dans ce contexte, l'objectif de cette étude était double. Premièrement, il s'agissait de déterminer si des informations métaboliques de la tumeur primaire en TEP au FDG permettaient de prédire l'atteinte axillaire. Deuxièmement, nous voulions comparer les performances de l'approche radiomique utilisant des caractéristiques prédéfinies à l'échelle du voxel proposée précédemment, une approche employant le DL, et une modélisation logistique ne prenant que des variables d'imagerie simples ou conventionnelles en entrée.

6.2 . Matériels et méthodes

Patients et données

Les données étaient issues de patientes suivies rétrospectivement à l'Institut Curie en 2018 et 2019, ayant bénéficié d'une TEP/TDM au FDG au service de médecine nucléaire de l'hôpital de Saint-Cloud.

Critères d'inclusion et d'exclusion

Les critères d'inclusion comprenaient une confirmation histologique de cancer du sein invasif initial pour des femmes ayant bénéficié d'une TEP/TDM au bilan d'extension avant un traitement chirurgical éventuel. Toutes les patientes incluses ont fait l'objet d'une dissection du « ganglion sentinelle » lymphatique axillaire ipsilatéral afin de déterminer leur statut [448]. Le stade T (TNM) était déterminé pour toutes les patientes à partir de l'imagerie IRM ou échographique réalisée dans le mois précédent la TEP/TDM.

Les critères d'exclusion de l'étude étaient des images non interprétables dues à l'extravasation du radiotraceur, ou lorsque la tumeur primaire ne fixait pas le FDG. Les patientes dont la glycémie était anormale ont également été exclues [449]. Enfin, les patientes ayant été atteintes d'un autre cancer par le passé, ou souffrant conjointement d'un autre cancer en plus du sein, ont été exclues.

Provenant de 185 patientes, 191 lésions mammaires ont finalement été incluses [1].

Caractéristiques de l'échantillon

Les caractéristiques cliniques et pathologiques des patientes incluses sont résumées dans le Tableau 6.1 [2].

1. Six patientes étaient atteintes aux deux seins. Dans ce travail, nous avons considéré l'atteinte ganglionnaire axillaire ipsilatérale. Une patiente atteinte aux deux seins pouvait de ce fait être doublement positive ou négative, ou bien présenter une atteinte axillaire mixte.

2. Les comparaisons en fonction de l'atteinte ganglionnaire ont été réalisées l'aide du test de Mann-Whitney (avec correction des exæquos) pour les variables numériques et catégorielles ordinales [407, 450], et du test d'indépendance du $\chi 2$ pour les variables binaires et catégorielles nominales [451].

Diagnostic final

Les résultats de l'examen pathologique du ganglion sentinelle ont été utilisés pour déterminer le statut du ganglion lymphatique axillaire ipsilatéral de chaque lésion.

Protocole d'imagerie

Les patientes opérées ont bénéficié d'un examen TEP/TDM au FDG avant l'intervention chirurgicale, sans chimiothérapie néoadjuvante. Toutes les patientes incluses sont restées à jeun pendant au moins six heures avant l'acquisition des images. La glycémie mesurée avant l'injection du radiotraceur était systématiquement inférieure à $9mmol \times l^{-1}$ [449]. L'acquisition a été réalisée 60 minutes après l'injection de $3MBq \times kg^{-1}$ de FDG. Une acquisition du haut du crâne à mi-cuisse a été réalisée pendant une minute sur un même scanner pour toutes les patientes (Vereos, Philips Healthcare). La correction d'atténuation a été effectuée avec la TDM avec une énergie de $120kV$ et une intensité adaptative d'environ $60mA$. Les images TEP ont été reconstruites par la technique OSEM avec 5 sous-ensembles et 3 itérations intégrant le temps de vol, corrigées de l'atténuation, de la diffusion, des détections fortuites, et de la réponse impulsionnelle.

Traitement et représentation des images

Prétraitement des images

En utilisant LIFEx, les lésions ont été délinéées via un seuil à 40% du SUVmax tumoral. La nécrose présente dans certaines lésions a été incluse dans la région tumorale par des opérations de morphologie mathématiques. Une dilatation de trois voxels a ensuite été appliquée à la région.

Calcul des caractéristiques

Pour l'approche radiomique avec des caractéristiques prédéfinies, l'extraction suit les mêmes étapes que précédemment :

Tableau 6.1 – Caractéristiques des patientes atteintes d'un cancer du sein initial.

Caractéristique	Total	Atteinte ganglionnaire positive	Atteinte ganglionnaire négative	p-*value*
Nombre de lésions (%)	$191(100\%)$	$94(49\%)$	$97(51\%)$	-
Age au diagnostic $[années]$				
moyenne ($\pm$ 1 écart-type)	53 ± 14	52 ± 15	53 ± 13	$0,474$
médiane [intervalle]	$50[25\ ;91]$	$50[25\ ;91]$	$50[26\ ;79]$	
Poids au diagnostic $[kg]$				
moyenne ($\pm$ 1 écart-type)	$67,6 \pm 13,9$	$69,7 \pm 15,1$	$65,4 \pm 12,1$	$0,058$
médiane [intervalle]	$65,0[45,0\ ;120,0]$	$68,5[45,0\ ;120,0]$	$64,0[45,0\ ;96,0]$	
Sein atteint				
nombre (%)				
Droit	$75(39\%)$	$35(37\%)$	$40(41\%)$	$0,676$
Gauche	$116(61\%)$	$59(63\%)$	$57(59\%)$	
Atteinte bilatérale	$6(3\%)$	$3(3\%)$	$3(3\%)$	
Type				
nombre (%)				
Canalaire invasif	$161(84\%)$	$81(86\%)$	$80(82\%)$	$0,619$
Lobulaire invasif	$16(8\%)$	$6(6\%)$	$10(10\%)$	
Autre	$14(7\%)$	$7(7\%)$	$7(7\%)$	

Tableau 6.1 - Caractéristiques des patientes atteintes d'un cancer du sein initial (suite de la page précédente).

Caractéristique	Total	Atteinte ganglionnaire positive	Atteinte ganglionnaire négative	p-*value*
Grade				
nombre (%)				
I	$8(4\%)$	$5(5\%)$	$3(3\%)$	$0,033$
II	$90(47\%)$	$51(54\%)$	$39(40\%)$	
III	$91(48\%)$	$38(40\%)$	$53(55\%)$	
Inconnu	$2(1\%)$	$0(0\%)$	$2(2\%)$	
Récepteurs hormonaux				
nombre (%)				$0,014$
Positifs	$127(66\%)$	$71(76\%)$	$56(58\%)$	
Négatifs	$64(34\%)$	$23(24\%)$	$41(42\%)$	
Récepteurs aux œstrogènes				
nombre (%)				$0,034$
Positifs	$125(65\%)$	$69(73\%)$	$56(58\%)$	
Négatifs	$66(35\%)$	$25(27\%)$	$41(42\%)$	
Récepteurs à la progestérone				
nombre (%)				$0,090$
Positifs	$105(55\%)$	$58(62\%)$	$47(48\%)$	
Négatifs	$86(45\%)$	$36(38\%)$	$50(52\%)$	

Tableau 6.1 - Caractéristiques des patientes atteintes d'un cancer du sein initial (suite de la page précédente).

Caractéristique	Total	Atteinte ganglionnaire positive	Atteinte ganglionnaire négative	p-*value*
Récepteurs HER2				
nombre (%)				$0,435$
Positifs	$53(28\%)$	$29(31\%)$	$24(25\%)$	
Négatifs	$138(72\%)$	$65(69\%)$	$73(75\%)$	
Type moléculaire				
nombre (%)				$0,004$
HER2	$18(9\%)$	$11(12\%)$	$7(7\%)$	
Luminal A	$92(48\%)$	$53(56\%)$	$39(40\%)$	
Luminal B	$35(18\%)$	$18(19\%)$	$17(18\%)$	
Triple négatif	$46(24\%)$	$12(13\%)$	$34(35\%)$	
Ki-67 $[\%]$ (2 manquants)				
moyenne ($\pm$ 1 écart-type)	$37,8 \pm 24,7$	$36,3 \pm 24,9$	$39,3 \pm 24,4$	$0,148$
médiane [intervalle]	$30,0[2,0\ ;90,0]$	$30,0[2,0\ ;90,0]$	$31,0[3,0\ ;90,0]$	
Stade T (TNM)				
nombre (%)				$1,5 \times 10^{-5}$
T1	$53(28\%)$	$36(37\%)$	$17(18\%)$	
T2	$100(52\%)$	$53(55\%)$	$47(50\%)$	
T3	$24(13\%)$	$7(7\%)$	$17(18\%)$	
T4	$14(7\%)$	$1(1\%)$	$13(14\%)$	

Tableau 6.1 - Caractéristiques des patientes atteintes d'un cancer du sein initial (suite de la page précédente).

Caractéristique	Total	Atteinte ganglionnaire positive	Atteinte ganglionnaire négative	p-*value*
Glycémie $[mmol \times l^{-1}]$				
moyenne ($\pm$ 1 écart-type)	$5,3 \pm 0,9$	$5,3 \pm 0,9$	$5,2 \pm 0,9$	$0,826$
médiane [intervalle]	$5,1[2,5\ ;9,8]$	$5,1[3,6\ ;9,8]$	$5,0[2,5\ ;9,5]$	
Graisse brune				
nombre (%)				$0,882$
Positive	$22(12\%)$	$10(11\%)$	$12(12\%)$	
Négative	$169(88\%)$	$84(89\%)$	$85(88\%)$	
Chirurgie d'emblée				
nombre (%)				$0,005$
Oui	$75(39\%)$	$27(29\%)$	$48(49\%)$	
Non	$116(61\%)$	$67(71\%)$	$49(51\%)$	

- **Interpolation** : Utilisation des images d'origine sans interpolation ($2mm \times 2mm \times 2mm$)
- **Discrétisation** : taille de *bins* fixe : $0,3125SUV$
- **Caractéristiques** :
  - premier ordre ($p_{po} = 18$)
  - GLCM ($p_{GLCM} = 24$)
  - GLDM ($p_{GLDM} = 14$)
  - GLRLM ($p_{GLRLM} = 16$)
  - NGTDM ($p_{NGTDM} = 5$)
  - GLSZM [3] ($p_{GLSZM} = 16$)
  - total par image ($p_{tot} = 93$)
- **Fenêtre glissante** : $5 \times 5 \times 5$ voxels
- **Agrégation** : moyenne
- **Modèle final** : *bagging*

Pour l'approche par DL, un *zero-padding* a été utilisé de sorte que toutes les images aient les mêmes dimensions en x, y, et z. Illustrée en Figure 6.2, une architecture analogue à l'approche radiomique proposée, basée sur l'architecture U-Net, a été développée [99, 453, 454]. Permettant d'apprendre des cartes de caractéristiques dans la résolution initiale, une opération de « ROI-*average* » *pooling* (ROI-AP) a été placée entre ces dernières et une couche de classification. Agrégeant le signal des cartes de caractéristiques dans les ROIs à la manière de l'approche RDM, elle permet de former un CNN de classification à partir de l'architecture U-Net, développée initialement pour de la segmentation (classification de voxels). Notre objectif était d'évaluer la capacité du DL à apprendre la représentation (les caractéristiques) dans un contexte similaire à la méthode RDM. Par conséquent, nous avons utilisé la même segmentation des lésions, bien que ce ne soit pas strictement nécessaire en pratique. En outre, des essais préliminaires ont été réalisés en corps entier sans segmentation et en utilisant des architectures conventionnelles, mais ceux-ci n'ont pas convergé lors des entraînements.

Les variables d'imagerie simples ou conventionnelles utilisées pour la troisième modélisation étaient le SUVmax, le SUVmean, le MTV, le TLG, le volume fermé Vf (après morphologie mathématique), le plus grand diamètre de la ROI, le second plus grand diamètre perpendiculaire, le troisième diamètre perpendiculaire aux deux autres, le volume fermé non métabolique $Vf_{<40\%SUVmax}$, la proportion non métabolique $rVf_{<40\%SUVmax}$, la surface de la ROI, et le ratio de la surface de la ROI sur son volume [120, 121].

3. Les caractéristiques de la matrice GLSZM n'étaient pas disponibles à l'échelle du voxel dans Pyradiomics lors des études précédentes sur les STS et les tumeurs cérébrales (versions 2.2.0 et 3.0). La mise à jour vers la version 3.0.1 a permis de les ajouter à l'étude présentée ici [452].

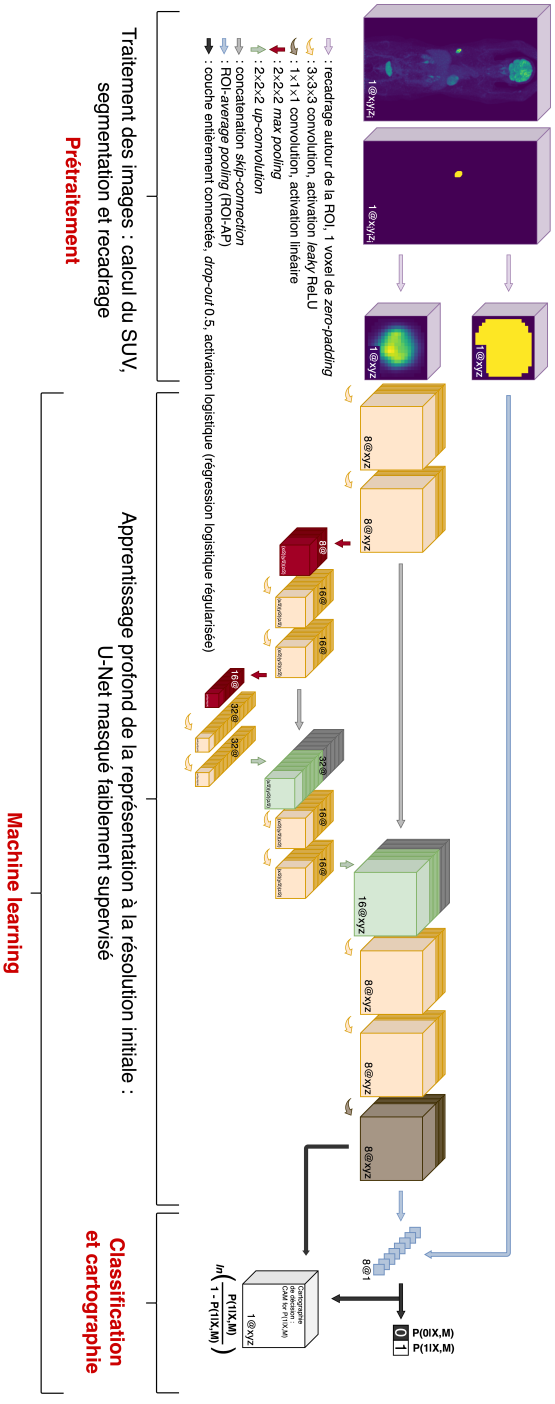


Figure 6.2 – Approche d'apprentissage profond proposée pour la classification basée sur l'architecture U-Net.

Tableau 6.2 – Hyperparamètres pour la construction du modèle M2 par DL.

Hyperparamètre	**Valeur, type**	**Choix**
Optimiseur	Adam	[171]
Taux d'apprentissage	$0,001$	Par défaut
Adam $[\beta 1, \beta 2]$	$[0,900\,;0,999]$	Par défaut
Adam ϵ	10^{-8}	Par défaut
Taille des lots (*batch size*)	8	[186]
Fonction de perte	Entropie croisée équilibrée	Par analogie à la régression logistique
Régularisation	*Drop-out* : $0,5$	[455, 456]
Nombre maximal d'*epochs*	300	Stagnation sur les ensembles de validation croisée lors d'essais préliminaires

Classification

Le modèle radiomique, M1, a été construit de la même manière que précédemment (cf, Chapitre 4). En ce qui concerne le modèle de DL, M2, les hyperparamètres ont été définis a priori ou laissés par défaut, et sont listés dans le Tableau 6.2. Le nombre d'*epochs*[4] a été déterminé par validation croisée répétée stratifiée avec 15 plis (3×5 plis). La chaine d'analyse ML pour le modèle prenant des variables simples ou conventionnelles en entrée, M3, était la même que pour M1, et incluait donc une réduction de la multicollinéarité à l'aide du VIF, une sélection des caractéristiques SFS, et un modèle logistique régularisé avec la technique LASSO.

Apport potentiel par rapport à la routine clinique

En pratique, l'estimation de l'atteinte ganglionnaire en TEP/TDM au FDG repose principalement sur l'interprétation des médecins nucléaires dans la zone axillaire. De plus, certaines informations qui se sont avérées statistiquement significatives dans notre cohorte, telles que le type moléculaire (Tableau 6.1), ne sont généralement pas disponibles lors du bilan d'extension. C'est pourquoi nous ne les avons pas incluses dans notre modèle. Notre objectif était de fournir des éléments pour interpréter les images de manière plus globale lors de la détermination de l'atteinte ganglionnaire, en se basant notamment sur la lésion primaire. Ainsi, afin d'évaluer l'apport potentiel de l'information extraite de la tumeur primitive par chacune des approches, le résultat du diagnostic réalisé en TEP directement dans la zone axillaire ipsilatérale a également été considéré pour chaque lésion. Ce dernier a été réalisé par les praticiens du service de médecine nucléaire de l'hôpital de Saint-Cloud (5 années d'expérience minimum) lors de leurs vacations. Les examens ont été notés positifs ou négatifs pour chaque côté, selon la fixation de FDG et plus généralement l'analyse visuelle de l'image TEP/TDM.

4. En DL, une époque ou « *epoch* » désigne un passage complet de l'ensemble des données d'entraînement, par lots, à l'algorithme d'optimisation.

6.3 . Résultats

Les fonctions de décision des modèles M1 et M3 sont données par les équations (6.1) et (6.2). En validation croisée via les sélections SFS et LASSO, trois caractéristiques radiomiques ont été retenues pour M1. Pour M3, les meilleures performances ont été obtenues avec une seule caractéristique : le plus grand diamètre de la ROI. D_{M3} est appris avec les variables dans leur échelle initiale. D^*_{M3} (6.3) correspond à l'estimation des coefficients standardisés, analogues à l'importance, associés aux variables exprimées en écarts-types par rapport à leur moyenne.

$$\begin{aligned} D_{M1} = &- 1,059(\pm 0,528) \times GLCM_{cluster\ prominence} \\ &+ 1,601(\pm 0,513) \times GLDM_{LDHGLE^*} \\ &+ 0,483(\pm 0,272) \times GLDM_{SDLGLE^{**}} \\ &+ 0,067(\pm 0,064) \end{aligned} \tag{6.1}$$

$$\begin{aligned} D_{M3} = &+ 0,025(\pm 0,010) \times diamètre\ maximal\ [mm] \\ &- 0,908(\pm 0,376) \end{aligned} \tag{6.2}$$

$$\begin{aligned} D^*_{M3} = &+ 0,782(\pm 0,293) \times diamètre\ maximal\ [écart\text{-}type] \\ &+ 0,058(\pm 0,051) \end{aligned} \tag{6.3}$$

Les performances de classification en validation croisée répétée stratifiée pour les modèles M1, M2, et M3, ainsi que pour l'examen visuel de la zone axillaire sont reportées dans le Tableau 6.3. Pour le modèle M2 (DL), le nombre d'*epoch* associé à un bon compromis entre les meilleures performances en validation et le plus faible surapprentissage des données d'entraînement était de 130 (Figure 6.3).

Tableau 6.3 – Performances de classification obtenues en validation croisée répétée stratifiée (3×5 plis) pour les modèles M1, M2, M3, ainsi que pour l'examen visuel de la zone axillaire.

CV moyenne ($\pm$ 1 écart-type)	**AUC**	**Bacc**	**Se**	**Sp**
M1	$0,682 \pm 0,075$	$0,646 \pm 0,057$	$0,546 \pm 0,118$	$0,746 \pm 0,098$
M2	$0,650 \pm 0,061$	$0,586 \pm 0,062$	$0,566 \pm 0,260$	$0,607 \pm 0,269$
M3	$0,701 \pm 0,078$	$0,643 \pm 0,077$	$0,714 \pm 0,143$	$0,570 \pm 0,124$
Examen visuel axillaire		$0,863 \pm 0,047$	$0,840 \pm 0,075$	$0,890 \pm 0,064$

*Large dépendance de haut niveau de gris (*large dependence high gray level emphasis* (LDLGLE)).

**Petite dépendance de faible niveau de gris (*small dependence low gray level emphasis* (LDLGLE)).

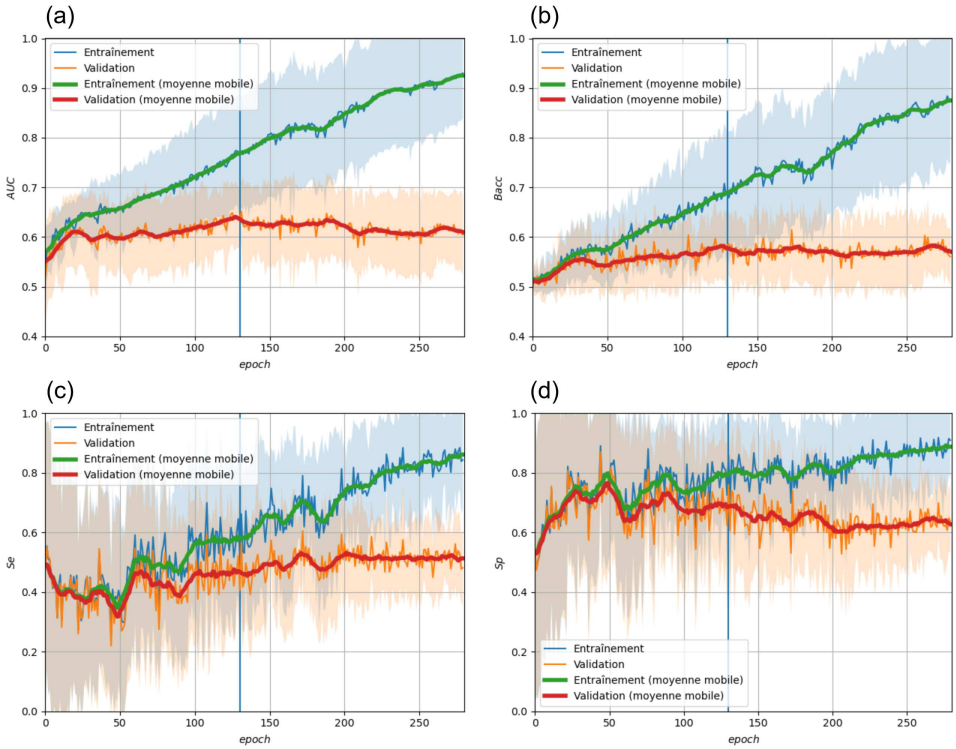


Figure 6.3 – Courbes d'apprentissage et de validation du modèle M2 en fonction du nombre d'*epochs* pour l'AUC (a), la Bacc (b), la Se (c), et la Sp (d), estimées en validation croisée répétée stratifiée (3×5 plis). La ligne horizontale représente un nombre d'*epochs* de 130 pour l'entraînement des modèles, associé à un bon compromis entre les meilleures performances en validation et le plus faible surapprentissage des données d'entraînement.

Globalement, ni le modèle radiomique M1 ni le CNN M2 n'ont surpassé l'analyse conventionnelle impliquant M3, ne retenant que la mesure de la plus grande dimension de la ROI tumorale pour prédire le statut axillaire à partir de la TEP au FDG. Comparativement à l'examen visuel de la zone axillaire par les médecins nucléaires, les trois modèles ont obtenu des Bacc, Se, et Sp plus faibles en se basant sur la tumeur primitive.

Le graphique de dispersion multiple en Figure 6.4 compare les probabilités prédites par M1, M2, et M3. Bien qu'ils ne se confondent pas, les modèles M1 et M2 sont plus corrélés entre eux (Pearson $r_{P_{M1,M2}} = 0.690$) qu'avec M3 basé uniquement sur le plus grand diamètre de la ROI ($r_{P_{M1,M3}} = 0.517$, et $r_{P_{M2,M3}} = 0.453$).

De façon cohérente avec la littérature, la spécificité de l'examen visuel de la zone axillaire était élevée dans notre étude avec $Sp = 0,890 \pm 0,064$. La sensibilité était de $Se = 0,840 \pm 0,075$, également élevée, et supérieure aux données de la littérature. Sur l'ensemble des lésions ($n = 191$), ces per-

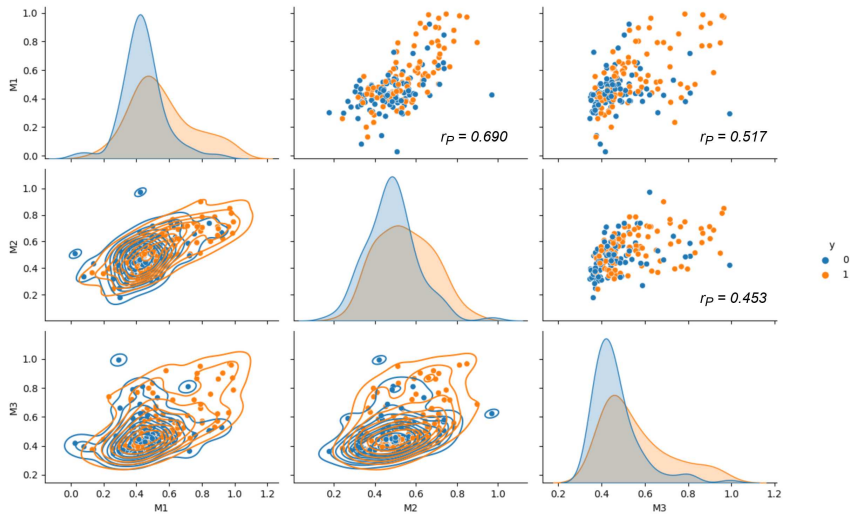


Figure 6.4 – Graphique de dispersion multiple comparant les probabilités prédites par M1, M2, et M3, qualitativement grâce aux nuages de points ainsi que quantitativement selon leurs coefficients de corrélation de Pearson r_P. Les atteintes ganglionnaires positives et négatives sont respectivement représentées par les points bleus ($y = 0$) et orange ($y = 1$).

formances correspondent à 11 faux positifs ($FP = 11$), 15 faux négatifs ($FN = 15$), 79 vrais positifs ($TP = 79$), et 86 vrais négatifs ($TN = 86$). Parmi les faux positifs et les faux négatifs de l'examen visuel axillaire réalisé par les médecins nucléaires, 8 (73%), 4 (36%), et 7 (64%) faux positifs, et 7 (47%), 6 (40%), et 6 (40%) faux négatifs ont bien été classifiés, respectivement pour M1, M2, et M3 avec plusieurs lésions concernées communes entre les modèles.

La Figure 6.5 montre des exemples des coupes de RDMs (M1), de CAMs (M2), et de TEP pour cinq lésions, associées au plus grand diamètre de leur ROI (M3). Comme attendu, le décodeur de l'architecture U-net a permis d'obtenir des CAMs (b) avec une définition relativement fine, comparables aux RDMs (a). Très similaires visuellement, les signaux des RDMs et des CAMs semblaient suivre la captation tumorale locale de FDG, et aucun autre motif n'était facilement identifiable.

Compte tenu des faibles performances de classification associées à M1 et M2 et de l'extrême simplicité de M3 par rapport à ces derniers, nous n'avons pas étudié de manière plus approfondie l'information portée par ces modèles. Néanmoins, il est intéressant de noter que malgré leurs similarités visuelles à l'échelle du voxel, M1 et M2 n'étaient pas colinéaires à l'échelle globale, bien que positivement corrélés. Cet élément appuie l'importance de l'interprétation quantitative des modèles, au-delà de leur interprétabilité spatiale impliquant une technique de cartographie (cf, Section 4.1.2, [200]).

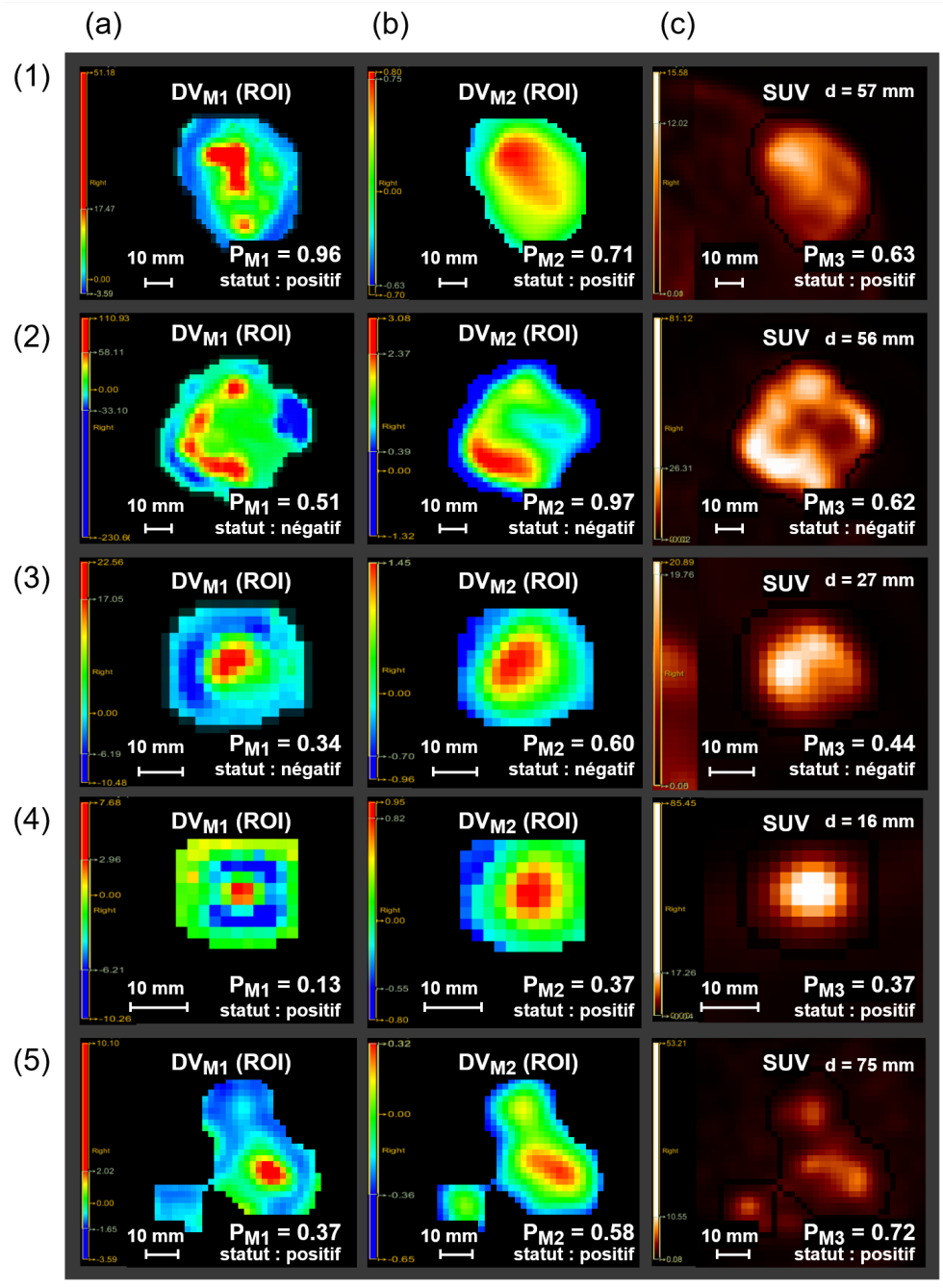


Figure 6.5 – Exemples de coupes de RDMs (DV_{M1}, $\hat{y}_{M1}$) (a), de CAM (DV_{M2}, $\hat{y}_{M2}$) (b), et d'images TEP (c) pour cinq lésions (1-5). Pour chaque lésion, le plus grand diamètre de la ROI est également reporté et associé à sa probabilité prédite par M3 (d, $\hat{y}_{M3}$) (c).

6.4 . Discussion

Dans ce travail, notre objectif était double. Premièrement nous avons étudié l'apport potentiel d'information métabolique de la tumeur primitive en imagerie TEP au FDG pour prédire l'atteinte ganglionnaire axillaire en cancer du sein. En outre, nous avons comparé trois approches de simplicité, flexibilité, et interprétabilité variables. Le but était d'examiner l'axiome communément accepté qui suggère un compromis entre l'interprétabilité du modèle et sa justesse, selon lequel les modèles plus simples et interprétables ont tendance à être moins précis, et inversement.

De façon générale, les performances de prédiction de l'atteinte ganglionnaire ipsilatérale à partir de l'image TEP de la tumeur primitive étaient limitées, avec une AUC maximale de $0,701$ pour le modèle M3 en validation croisée. Ces résultats suggèrent que le déploiement d'un tel modèle en routine clinique n'est pas envisageable à ce stade. Cependant, ils suggèrent l'existence d'un signal de la tumeur primitive associé à l'atteinte axillaire du même côté. De plus, nous avons observé que parmi les faux positifs et faux négatifs de l'examen visuel de la zone axillaire, 36% à 73% avaient été bien classifiés par les modèles. Ceci évoque la complémentarité de l'information capturée dans la tumeur par rapport à la zone axillaire directe.

De façon intéressante, le modèle le plus performant, M3, était aussi le plus simple et le plus interprétable. En effet, basé sur une seule caractéristique, il se limitait à l'activation logistique d'une transformation affine du plus grand diamètre de la ROI tumorale. À l'inverse, le modèle le moins contraint et par conséquent le plus flexible était le moins performant (M2, DL). Une explication potentielle de ce phénomène pourrait être que les modèles simples bénéficient en définitive de l'expérience et de l'intelligence humaine des spécialistes ayant conduit à la formulation et l'utilisation de variables pertinentes a priori. D'autre part, l'apprentissage plus ou moins profond de la représentation, qu'elle soit issue de l'association de caractéristiques de texture ou apprise par un réseau de neurones, est plus sensible au bruit, et semble ainsi nécessiter de grands ensembles de données [457-460]. De plus, si les approches radiomiques permettent une certaine optimisation de la représentation des images, elles possèdent tout de même une structuration intrinsèque limitant le type d'information que les modèles peuvent capturer. Par exemple, Klyuzhin et al. ont étudié la capacité de certaines architectures de CNNs à reproduire des caractéristiques prédéfinies. Leurs résultats montraient que certaines variables radiomiques pouvaient difficilement être codées par un CNN [461].

Concernant l'application clinique, nos résultats sont cohérents avec les résultats de l'étude de Sopik et al. menée en 2018 et portant sur 792123 patientes [444]. En effets, les auteurs ont montré un lien significatif entre l'atteinte axillaire et le plus grand diamètre tumoral. Bien que notre ensemble

de données soit de plusieurs ordres de grandeur plus petits, les proportions d'atteintes ganglionnaires positives et négatives que nous avons observées par paquets de valeur de plus grand diamètre suivaient précisément les leurs (Figure 6.6). Cet élément appuie encore la robustesse de résultats obtenus avec une représentation simple des données. Ainsi, on pourrait imaginer à terme mettre en place des stratégies de désescalade chirurgicale tenant compte du diamètre maximal de la lésion primaire, simple à mesurer lors du bilan d'extension sur la majorité des consoles d'interprétation actuelles.

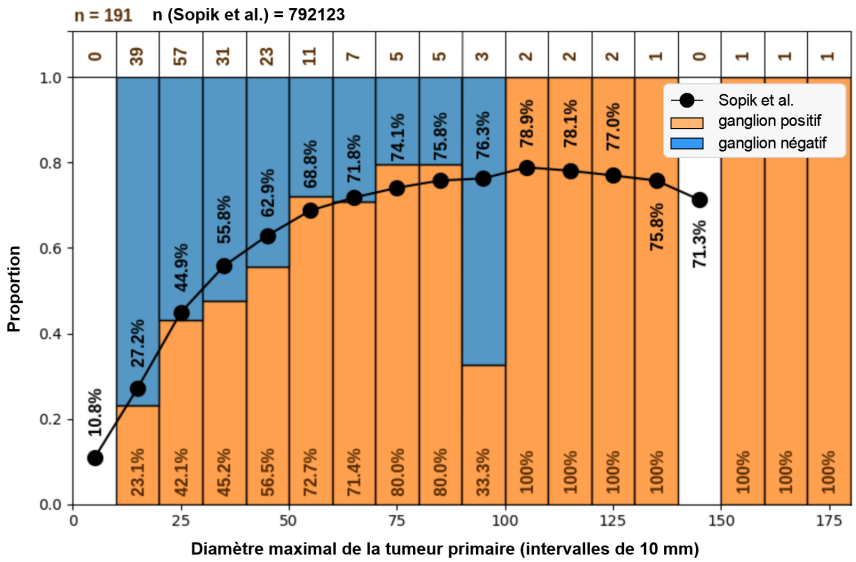


Figure 6.6 – Histogrammes cumulés des proportions de ganglions positifs et négatifs en fonction du diamètre maximal de la tumeur primaire (par pas de $10mm$), comparant nos résultats ($n = 192$) et ceux de Sopik et al. ($n = 792123$) [444].

Quant à l'analyse méthodologique, si nos résultats ne remettent pas en cause l'intérêt du DL, ils soulignent l'importance de considérer la simplicité comme une approche viable dans certains contextes.

À titre d'exemple supplémentaire, nous avons avec certains membres du LITO mené un projet annexe dans le cadre du challenge HECKTOR 2022 (*head and neck tumor segmentation and outcome prediction in PET/CT images*) [345, 462]. Lors de la 25$^{\text{ème}}$ conférence internationale sur l'imagerie médicale et l'intervention assistée par ordinateur (*Medical Image Computing and Computer Assisted Intervention* (MICCAI)) en 2022, la compétition était organisée en deux étapes : une sur la segmentation automatique des lésions et des ganglions pathologiques en cancer de la tête et du cou, et l'autre sur la prédiction de la survie sans progression (*progression free survival* (PFS)). Pour la segmentation, nous avons été classés 4$^{\text{ème}}$ en utilisant la méthode

nnUNet [100] basée sur l'architecture U-net [99], avec de minimes ajustement. Au-delà de notre utilisation simpliste, la philosophie de la méthode nnUNet est qu'en segmentation, il n'est pas nécessaire de complexifier la méthodologie d'apprentissage (eg, architecture), et que la performance d'un modèle est plus étroitement liée à la qualité et la gestion des données. Ce paradigme relativement récent dans le domaine est souvent nommé « *data-centric* », par opposition à « *model-centric* » [463]. Les compétiteurs qui se sont placés devant nous sur le podium ont tous adopté une stratégie similaire. Pour la prédiction de la PFS, nous avons décidé de proposer une approche originale [464]. L'idée centrale était que seulement la détermination du signe de la corrélation de la totalité des caractéristiques par rapport à la PFS, et leur combinaison, après sélection univariée, sous forme d'une moyenne centrée réduite, réduirait l'apprentissage à son strict minimum tout en permettant une prédiction efficace. Tout en évitant au maximum toute forme de surapprentissage, cette approche simplifiait également à l'extrême l'utilisation des données (très peu d'ajustement d'hyperparamètres). Cette méthode nous a placés en tête de la compétition. Parmi les six approches les mieux classées, une seule utilisait des caractéristiques profondes[345].

La principale limite de ce travail est qu'il gagnerait à être approfondi, que ce soit sur le plan clinique, ou méthodologique. La temporalité de mes projets doctoraux ne m'a pas permis d'étudier plus en détails certains points clés de nos résultats.

Par exemple, il serait intéressant d'augmenter la taille de la base de données. La recherche plus exhaustive de biomarqueur serait pertinente, en imagerie, mais aussi avec de l'ajout de caractéristiques cliniques, biologiques, voire génomiques, si disponibles.

Nous avons observé des performances plus faibles lors de la prédiction du statut ganglionnaire à partir de la tumeur primaire comparativement à l'analyse visuelle de la zone axillaire. Ce résultat était attendu. Néanmoins, certaines lésions mal classifiées par l'analyse visuelle standard l'ont bien été par les modèles M1, M2, ou M3. Il serait pertinent de caractériser ce phénomène, afin de progresser vers une stratégie hybride permettant d'augmenter la performance du diagnostic en combinant diverses informations.

Enfin, sur le plan méthodologique, caractériser les différences majeures entre les méthodes desquelles sont issus les modèles M1 et M2 permettrait de mieux comprendre les différences entre les deux formes d'analyse radiomique principales d'aujourd'hui.

6.5 . Conclusion

Dans ce chapitre nous avons étudié l'apport de l'imagerie TEP au FDG de la tumeur primitive dans le diagnostic de l'atteinte ganglionnaire axillaire.

Alors que cette tâche de prédiction était difficile et a abouti à des performances limitées, elle a appuyé l'hypothèse de l'existence de signaux tumoraux pertinents dans ce contexte.

De plus, nous avons obtenu des résultats en faveur des techniques simples par rapport aux méthodes actuelles, plus complexes. Il est probable que notre approche par DL puisse être améliorée. Cependant, la versatilité des ANNs et CNNs rend difficile cette optimisation, et il est difficile voire impossible d'être exhaustif dans ce travail. D'autre part, les méthodes classiques, bien que nombreuses et variées également, sont plus standards et simples, tant dans leur utilisation que dans les modèles qui en découlent. Ce chapitre représente ainsi une contribution au débat en cours sur la complexité et la précision des modèles dans le domaine de la radiomique et du ML. En remettant en question certaines hypothèses dominantes et en présentant des résultats, nous invitons à réévaluer les normes établies et encourageons à ne pas négliger les recherches sur des approches de modélisation plus simples.

7 - Conclusion et perspectives

Le but premier de mon travail de thèse était d'identifier des sous-régions tumorales associées à une prédiction ou à une classification, avec deux objectifs principaux : premièrement, faciliter l'interprétation des modèles et, deuxièmement, proposer une méthode pouvant être intégrée à terme dans une approche de traitement ciblé. Encouragés par des résultats prometteurs, nous avons toutefois observé que notre méthode pouvait s'avérer insuffisante dans le cas de certaines applications. Par conséquent, une technique de visualisation originale a été proposée afin de mieux comprendre les subtilités de l'information capturée dans les images par les modèles. Par la suite, nous avons comparé les performances de classification de notre méthode avec deux autres approches. La première était un modèle de régression logistique simple. La seconde correspondait à l'apprentissage profond, flexible, versatile, et actuellement très populaire. De façon intéressante, nos résultats penchaient plutôt en faveur des modélisations plus simples, donc plus interprétables d'emblée.

Nous résumons ici brièvement les conclusions générales tirées de cette série de résultats, et, plus généralement, de l'expérience que j'ai acquise au cours de cette thèse.

La simplicité est tout ce dont nous avons besoin

Pour chaque application, formuler un modèle prédictif simple et performant constitue un objectif clé pour son adoption. Un tel modèle répondrait à de nombreux prérequis pour son adoption dans la pratique clinique ou en recherche médicale. En effet, en plus d'être interprétables, les modèles simples présentent généralement l'avantage d'être plus robustes, stables, et donc maîtrisables, de leur développement à leur maintenance, en passant par leur déploiement et leur diffusion. Ces caractéristiques sont essentielles pour le bénéfice et la sécurité des patients, pour lesquels des décisions cruciales doivent être prises en toute confiance et de manière éclairée. Ainsi, en poursuivant cette quête de simplicité, nous espérons non seulement aboutir à des modèles interprétables et vérifiables, mais également améliorer leur fiabilité et leur pertinence pour une meilleure prise en charge en oncologie.

Cependant, nous reconnaissons que la formulation de tels modèles est une entreprise ardue. C'est pourquoi nous pensons qu'adopter une approche reposant sur la radiomique, qu'elle soit profonde ou non, constitue une voie pertinente. Nous avons alors exploré ces méthodes, flexibles, en les ajustant pour répondre à nos besoins spécifiques, afin de les interpréter par la suite pour en extraire des modèles substituts, plus simples. En empruntant de nouvelles voies dans la représentation des images, l'approche radiomique a le

potentiel de découvrir de nouveaux modèles prédictifs, mais aussi de nouvelles caractéristiques, et par conséquent de nouveaux biomarqueurs.

Dans le domaine des sciences de l'ingénieur, il est indéniable que la complexité exerce une certaine attraction. Comprendre et maîtriser le fonctionnement de systèmes nouveaux et complexes procure une satisfaction intellectuelle particulière. Cela peut parfois nous pousser à privilégier une approche en raison de sa nature plutôt que de son adéquation aux objectifs réels de nos recherches. Malheureusement, ce phénomène peut entraîner, dans les pires cas, des biais de publication en faveur de méthodes sous-optimales en pratique. Ceci peut avoir un retentissement sur l'ensemble de la communauté scientifique et médicale.

Aussi, si le modèle final est simple, son obtention s'avère parfois complexe.

L'interprétabilité n'est pas l'ennemie de la performance

Il est crucial de souligner que l'interprétabilité d'un modèle n'est pas synonyme de faibles performances. En réalité, il existe plusieurs raisons pour lesquelles un modèle interprétable peut conserver des performances élevées, voire même les améliorer.

En comprenant comment les caractéristiques des données sont utilisées pour prendre des décisions, il devient plus facile de détecter les erreurs et les biais potentiels. Cela permet d'effectuer des ajustements et d'optimiser le modèle pour des performances optimales.

De plus, l'interprétation d'un modèle facilite l'identification de ses limites. Cela permet aux chercheurs et aux praticiens de comprendre dans quelles conditions le modèle est le plus fiable et dans quelles situations il peut présenter des lacunes. Ainsi, des précautions peuvent être prises pour éviter des décisions inappropriées basées sur des prédictions peu pertinentes.

Alors que la communauté de l'apprentissage automatique a un temps pensé que l'interprétabilité et la performance étaient antagonistes, nous commençons à comprendre qu'elles ne le sont pas nécessairement !

Le théorème du « no free lunch » s'applique à l'interprétation

Ce théorème énonce que, de manière générale, il n'existe pas de méthode d'apprentissage qui puisse être la meilleure dans tous les domaines et pour tous les problèmes. Il en va de même lorsqu'il s'agit d'interpréter les décisions prises par un modèle. Il n'y a pas de méthode unique qui puisse être considérée comme satisfaisante dans toutes les situations.

Par exemple, si l'identification de sous-régions tumorales nous a permis de comprendre l'information capturée par les modèles dans le contexte des STS, elle n'a pas été suffisante dans le cas des lésions plus petites dans les cancers cérébraux. Dans ce dernier cas, il a fallu reformuler des caractéristiques pour comprendre le modèle.

Chaque méthode possède ses propres hypothèses, contraintes, limites, et par conséquent son cadre applicatif. Ainsi, le choix d'une méthode d'interprétation dépendra de la nature de la tâche, des contraintes, des données disponibles, et de la façon dont elles ont utilisées.

L'interprétation n'est pas toujours une vérité absolue, mais souvent une représentation ou une approximation de la manière dont le modèle prend ses décisions. Elle peut donc présenter des biais, des limitations, et des simplifications qui doivent être prises en compte. Il est essentiel de comprendre le fonctionnement de chaque méthode et de les considérer comme des outils pour aider à la compréhension.

Chaque caractéristique pertinente, chaque modèle, et chaque méthode d'interprétation est une vue du problème. Si nous n'avons pas accès à la vérité, multiplier ces vues et tirer parti de leur complémentarité pourraient constituer une approche pertinente pour s'en rapprocher !

Les données sont la clé

En apprentissage automatique, au-delà de la méthode d'apprentissage elle-même, ce sont les données qui revêtent une importance primordiale. Alors que les méthodes d'apprentissage aboutissent souvent à des modèles équivalents (ensembles Rashomon), les données sont le fondement sur lequel ils sont construits, éprouvés et validés. Constituant la matière première, elles jouent un rôle déterminant dans la qualité des résultats obtenus. Des données de qualité, complètes, en grande quantité, et représentatives de la population étudiée, permettent aux modèles de capturer des motifs clés, des relations, et de mettre en évidence des tendances pour répondre au problème posé.

Alors que les modèles sont entraînés sur des milliers d'images dans d'autres domaines, obtenir de grandes bases de données de qualité en imagerie médicale présente des défis majeurs. L'accès limité aux données en raison de leur confidentialité et les coûts élevés de collecte sont des obstacles. Plus précisément, l'annotation et la mise en forme des données peuvent s'avérer chronophages, sources d'erreurs, coûteuses, et difficiles en termes de logistique.

La collaboration entre institutions médicales, l'apprentissage fédéré, le partage sécurisé de données ou de modèles pré-entraînés, la constitution de bases de données publiques, le développement d'infrastructure et de solutions d'annotations automatiques, sont des approches intéressantes pour améliorer la disponibilité et la diversité des bases de données en imagerie médicale. Avec une priorité absolue pour une pratique éthique, libre, et éclairée pour le patient, il est crucial de continuer à fournir des efforts dans cette direction.

Alors que c'est l'algorithme qui brandit la médaille, ce sont essentiellement les données qui la méritent !

Opinion personnelle et conclusion générale

Je suis fermement convaincu du potentiel considérable de l'apprentissage automatique et de la fouille de données en imagerie médicale, et plus généralement en médecine. Les avancées technologiques dans ce domaine offrent des opportunités passionnantes pour améliorer le diagnostic, la détection précoce des maladies, et la prise en charge des patients. Cependant, tout en étant enthousiaste, je ressens une certaine appréhension quant à la direction que cela pourrait prendre.

Mon inquiétude se situe dans la possibilité d'une course effrénée vers l'automatisation totale, où l'optimisation deviendrait la priorité absolue. Cette orientation pourrait potentiellement déshumaniser certains aspects essentiels de la médecine. Les soins et notre relation à l'autre ne peuvent selon moi être remplacés par des machines, même si ces dernières finissent par surpasser l'humain sur certains aspects. Bien que nous n'en soyons pas encore à ce stade, il est important d'être conscient des dangers potentiels, ne serait-ce que sur nos considérations, et de veiller à ne pas perdre de vue que le soin est avant tout un accompagnement.

Comme le microscope permet de voir dans l'infiniment petit et le télescope dans l'infiniment grand, la science des données doit nous permettre de voir dans l'infiniment complexe. Il est indéniable que les outils que nous développons avancent dans cette direction, et ont le potentiel d'améliorer considérablement la prise en charge des patients en fournissant une analyse approfondie de différents aspects de différentes pathologies. Cependant, il est crucial de ne pas succomber aux avantages financiers et de confort que ces outils peuvent offrir. Notre objectif primordial doit rester l'amélioration réelle de la qualité des soins et du bien-être des patients. Il est essentiel de trouver un équilibre entre l'automatisation et l'interaction humaine. Les machines peuvent être des alliées précieuses, mais elles ne doivent pas remplacer la présence et l'empathie humaines. Préserver le lien entre le médecin et le patient est fondamental, tout autant que reconnaître l'importance des émotions, de l'intuition, du jugement clinique, et de l'expérience de l'équipe médicale.

Je suis convaincu que les nouvelles recherches continueront à repousser les limites de ce qui est possible, à s'adapter, à évoluer, et à susciter la surprise et l'émerveillement des plus passionnés. Je suis fier du modeste travail présenté dans ce manuscrit et des efforts que nous avons déployés en équipe pour y parvenir. J'espère que nos contributions, à leur mesure, seront utiles aux futurs chercheurs et médecins du domaine, mais surtout aux patients.

Production scientifique

Articles publiés dans des journaux scientifiques

1er auteur : Escobar, T., Vauclin, S., Orlhac, F., Nioche, C., Pineau, P., Champion, L., Brisse, H. & Buvat, I. Voxel-wise supervised analysis of tumors with multimodal engineered features to highlight interpretable biological patterns. *Medical Physics* **49,** 3816-3829 (2022)

3ème auteur : Khalid, F., Goya-Outi, J., Escobar, T., Dangouloff-Ros, V., Grigis, A., Philippe, C., Boddaert, N., Grill, J., Frouin, V. & Frouin, F. Multimodal MRI radiomic models to predict genomic mutations in diffuse intrinsic pontine glioma with missing imaging modalities. *Frontiers in Medicine* **10,** 1071447 (2023)

Article publié dans le cadre d'une conférence

co-1er auteur : Rebaud, L., Escobar, T., Khalid, F., Girum, K. & Buvat, I. *Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT* in *Head and Neck Tumor Segmentation and Outcome Prediction* (éd. Andrearczyk, V., Oreiller, V., Hatt, M. & Depeursinge, A.) (Springer, 2023), 121-134

Résumés de présentations dans le cadre de conférences

1er auteur : Escobar, T., Vauclin, S., Orlhac, F., Nioche, C., Pineau, P. & Buvat, I. *An original voxel-wise supervised analysis of tumors with multimodal radiomics to highlight predictive biological patterns [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **62** (SNMMI, 2021), 1404

1er auteur : Escobar, T., Orlhac, F., Rollet, A.-C., Humbert, O., Vauclin, S., Pineau, P., Darcourt, J. & Buvat, I. *Radiomic decision maps reveal patterns discriminating between glioma progression and radiation-induced necrosis in static and dual time [18F]-FDOPA PET [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **63** (SNMMI, 2022), 2520

1er auteur : Escobar, T., Provost, C., Seban, R. D., Vauclin, S., Pineau, P., Champion, L. & Buvat, I. *Predicting axillary lymph node metastasis in early-stage breast cancer using primary tumor image features on [18F]FDG PET: a comparative study of engineered radiomics, deep learning, and conventional methods [conference abstract]* in *Annual Congress of the European Association of Nuclear Medicine, October 15-19, 2022, Barcelona, Spain* (EJNMMI, 2022), S616, D25: Artificial Intelligence, EP-449

2ème auteur : Khalid, F., Escobar, T., Goya-Outi, J., Frouin, V., Boddaert, N., Grill, J. & Frouin, F. DIPG-23. Artificial intelligence for detecting ACVR1 mutations in patients with DIPG using MRI and clinical data [conference abstract]. *Neuro-Oncology* **24,** i23-i23 (2022)

4ème auteur : Girum, K., Rebaud, L., Cottereau, A.-S., Escobar, T., Clerc, J., Vercellino, L., Casasnovas, O., Morschhauser, F. & Irène Buvat. *Fully automatic segmentation of lesions in 3D using deep learning [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* (SNMMI, 2023), accepted

Distinctions

Premier prix 2022 des jeunes investigateurs du Conseil de Physique, Instrumentation, et *Data Sciences* (PIDSC *Young Investigator Award*) de la Société américaine de Médecine Nucléaire et d'Imagerie Moléculaire (*Society of Nuclear Medicine and Molecular Imaging* (SNMMI)) pour : Escobar, T., Orlhac, F., Rollet, A.-C., Humbert, O., Vauclin, S., Pineau, P., Darcourt, J. & Buvat, I. *Radiomic decision maps reveal patterns discriminating between glioma progression and radiation-induced necrosis in static and dual time [18F]-FDOPA PET [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **63** (SNMMI, 2022), 2520

Première place pour la classification et prix du meilleur papier lors de la compétition « *head and neck tumor segmentation and outcome prediction in PET/CT images, third edition* » (HECKTOR 2022) à l'occasion de la 25ème conférence internationale sur l'imagerie médicale et l'intervention assistée par ordinateur (*Medical Image Computing and Computer Assisted Intervention* (MICCAI)) pour : Rebaud, L., Escobar, T., Khalid, F., Girum, K. & Buvat, I. *Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT* in *Head and Neck Tumor Segmentation and Outcome Prediction* (éd. Andrearczyk, V., Oreiller, V., Hatt, M. & Depeursinge, A.) (Springer, 2023), 121-134

Développement logiciel

Contributeur secondaire : Rebaud, L., Escobar, T., Khalid, F., Girum, K. & Buvat, I. *Lrebaud/ICARE: Individual Coefficient Approximation for Risk Estimation (ICARE) model* 2022. https://github.com/Lrebaud/ICARE

Bibliographie

1. NOBELPRIZE.ORG. *The Nobel Prize in Physics 1901* https://www.nobelprize.org/prizes/physics/1901/speedread/ (2022).
2. HOW RADIOLOGY WORKS LLC. *History Of X-Ray Imaging* https://howradiologyworks.com/history-xray-imaging/ (2022).
3. FASS, L. Imaging and cancer: A review. *Molecular Oncology* **2,** 115-152 (2008).
4. GERLINGER, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine England Journal of Medicine* **366,** 883-892 (2012).
5. SEQUIST, L. V. *et al.* Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Science translational medicine* **3,** 75ra26 (2011).
6. GILLIES, R. J., ANDERSON, A. R., GATENBY, R. A. & MORSE, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clinical Radiology* **65,** 517-521 (2010).
7. NOBELPRIZE.ORG. *The Nobel Prize in Physiology or Medicine 1979* https://www.nobelprize.org/prizes/medicine/1979/summary/ (2022).
8. RADON, J. On the Determination of Functions from Their Integral Values along Certain Manifolds. *IEEE Transactions on Medical Imaging* **5,** 170-176 (1986).
9. ZENG, G. L. Image reconstruction - A tutorial. *Computerized Medical Imaging and Graphics* **25,** 97-103 (2001).
10. BUZUG, T. *Computed tomography: From photon statistics to modern cone-beam CT* (Springer, 2008).
11. SHAN, X. *et al.* Necrosis degree displayed in computed tomography images correlated with hypoxia and angiogenesis in breast cancer. *Journal of Computer Assisted Tomography* **37,** 22-28 (2013).
12. SANTEGOEDS, R. G., TEMEL, Y., BECKERVORDERSANDFORTH, J. C., VAN OVERBEEKE, J. J. & HOEBERIGS, C. M. State-of-the-Art Imaging in Human Chordoma of the Skull Base. *Current Radiology Reports* **6,** 16 (2018).
13. JAGANNATHAN, J. P., TIRUMANI, S. H. & RAMAIYA, N. H. Imaging in Soft Tissue Sarcomas: Current Updates. *Surgical Oncology Clinics of North America* **25,** 645-675 (2016).
14. YAN, W. *et al.* High-mobility group box 1 activates caspase-1 and promotes hepatocellular carcinoma invasiveness and metastases. *Hepatology* **55,** 1863-1875 (2012).

15. Abu, N., Rus Bakarurraini, N. A. A. & Nasir, S. N. Extracellular Vesicles and DAMPs in Cancer: A Mini-Review. *Frontiers in Immunology* **12,** 740548 (2021).

16. Brant, W. E., Webb, W. R. & Major, N. M. *Fundamentals of Body CT: 4th Edition* (Elsevier, 2014).

17. Rinck, P. A. in *Magnetic Resonance in Medicine* (BoD, 2018).

18. Liu, J. *et al.* A survey of MRI-based brain tumor segmentation methods. *Tsinghua Science and Technology* **7,** 179 (2014).

19. García-Figueiras, R. *et al.* How clinical imaging can assess cancer biology. *Insights into Imaging* **10,** 28 (2019).

20. Enochs, W. S., Petherick, P., Bogdanova, A., Mohr, U. & Weissleder, R. Paramagnetic metal scavenging by melanin: MR imaging. *Radiology* **204,** 417-423 (1997).

21. Weinmann, H. J., Brasch, R. C., Press, W. R. & Wesbey, G. E. Characteristics of gadolinium-DTPA complex: A potential NMR contrast agent. *American Journal of Roentgenology* **142,** 619-624 (1984).

22. Lauffer, R. B. Paramagnetic Metal Complexes as Water Proton Relaxation Agents for NMR Imaging: Theory and Design. *Chemical Reviews* **87,** 901-927 (1987).

23. Horowitz, A. L. *MRI Physics for Physicians* (Springer, 1989).

24. Elster, A. D. *2D vs 3D MRA - Questions and Answers in MRI* 2021. https://www.mriquestions.com/2d-vs-3d-mra.html (2022).

25. Tofts, P. S. *Quantitative MRI of the Brain* (Wiley, 2018).

26. Glover, G. H. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America* **22,** 133-139 (2011).

27. Nyú, L. G. & Udupa, J. K. On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine* **42,** 1072-1081 (1999).

28. Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9-19 (2014).

29. De Nunzio, G., Cataldo, R. & Carlà, A. Robust Intensity Standardization in Brain Magnetic Resonance Images. *Journal of Digital Imaging* **28,** 727-737 (2015).

30. Fortin, J. P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M. & Shinohara, R. T. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* **132,** 198-212 (2016).

31. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10,** 12340 (2020).

32. Lacroix, M. *et al.* Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer. *Frontiers in Oncology* **10,** 43 (2020).

33. Reinhold, J. C., Dewey, B. E., Carass, A. & Prince, J. L. *Evaluating the impact of intensity normalization on MR image synthesis* in *Proceedings of the International Society for Optical Engineering, vol 10949* (NIH Public Access, 2019), 126.

34. Anderson, C. D. The positive electron. *Physical Review* **43,** 491 (1933).

35. Fernandez, B. *De l'atome au noyau : Une approche historique de la physique atomique et de la physique nucléaire* (Ellipses, 2018).

36. Jones, T. & Townsend, D. History and future technical innovation in positron emission tomography. *Journal of Medical Imaging* **4,** 011013 (2017).

37. Slart, R. H. *et al.* Long axial field of view PET scanners: a road map to implementation and new possibilities. *European Journal of Nuclear Medicine and Molecular Imaging* **48,** 4236-4245 (2021).

38. Warburg, O., Wind, F. & Negelein, E. The metabolism of tumors in the body. *Journal of General Physiology* **8,** 519-530 (1927).

39. Koppenol, W. H., Bounds, P. L. & Dang, C. V. Otto Warburg's contributions to current concepts of cancer metabolism. *Nature Reviews Cancer* **11,** 325-337 (2011).

40. Pimlott, S. L. & Sutherland, A. Molecular tracers for the PET and SPECT imaging of disease. *Chemical Society Reviews* **40,** 149-162 (2011).

41. Moreau, A., Febvey, O., Mognetti, T., Frappaz, D. & Kryza, D. Contribution of Different Positron Emission Tomography Tracers in Glioma Management: Focus on Glioblastoma. *Frontiers in Oncology* **9,** 1134 (2019).

42. Bodanis, D. *E=mc2: A Biography of the World's Most Famous Equation* (Berkley Trade, 2000).

43. Cherry, S., Sorenson, J. & Phelps, M. *Physics in Nuclear Medicine* (Saunders, 2012).

44. Zanzonico, P. Positron Emission Tomography: A Review of Basic Principles, Scanner Design and Performance, and Current Systems. *Seminars in Nuclear Medicine* **34,** 87-111 (2004).

45. Wernick, M. N. & Aarsvold, J. N. *Emission Tomography: The Fundamentals of PET and SPECT* (Elsevier, 2004).

46. Maus, J. *Event-Driven Motion Compensation in Positron Emission Tomography: Development of a Clinically Applicable Method* PhD thesis (University of Technology Dresden, 2008).

47. Defrise, M. & Kinahan, P. in *The Theory and Practice of 3D PET* (Springer, 1998).

48. Surti, S. & Karp, J. S. Advances in time-of-flight PET. *Physica Medica* **31,** 12-22 (2016).

49. Townsend, D. W. Combined Positron Emission Tomography-Computed Tomography: The Historical Perspective. *Seminars in Ultrasound, CT and MRI* **29,** 232-235 (2008).

50. Vandenberghe, S. & Marsden, P. K. PET-MRI: a review of challenges and solutions in the development of integrated multimodality imaging. *Physics in medicine and biology* **60,** R115-R154 (2015).

51. Vaquero, J. J. & Kinahan, P. Positron Emission Tomography: Current Challenges and Opportunities for Technological Advances in Clinical and Preclinical Imaging Systems. *Annual Review of Biomedical Engineering* **17,** 385-414 (2015).

52. Unterrainer, M. *et al.* Recent advances of PET imaging in clinical radiation oncology. *Radiation Oncology* **15,** 88 (2020).

53. Rakheja, R. *et al.* Correlating metabolic activity on 18F-FDG PET/CT with histopathologic characteristics of osseous and soft-tissue sarcomas: A retrospective review of 136 patients. *American Journal of Roentgenology* **198,** 1409-1416 (2012).

54. Rakheja, R. *et al.* Necrosis on FDG PET/CT correlates with prognosis and mortality in sarcomas. *American Journal of Roentgenology* **201,** 170-177 (2013).

55. Lundemann, M. *et al.* Feasibility of multi-parametric PET and MRI for prediction of tumour recurrence in patients with glioblastoma. *European Journal of Nuclear Medicine and Molecular Imaging* **46,** 603-613 (2019).

56. Humbert, O. *et al.* 18F-DOPA PET/CT in brain tumors: impact on multidisciplinary brain tumor board decisions. *European Journal of Nuclear Medicine and Molecular Imaging* **46,** 558-568 (2019).

57. Herrmann, K. *et al.* Comparison of visual and semiquantitative analysis of 18F-FDOPA- PET/CT for recurrence detection in glioblastoma patients. *Neuro-Oncology* **16,** 603-609 (2014).

58. Lizarraga, K. J. *et al.* 18F-FDOPA PET for differentiating recurrent or progressive brain metastatic tumors from late or delayed radiation injury after radiation treatment. *Journal of Nuclear Medicine* **55,** 30-36 (2014).

59. Dimitrakopoulou-Strauss, A., Pan, L. & Sachpekidis, C. Kinetic modeling and parametric imaging with dynamic PET for oncological applications: general considerations, current clinical applications, and future perspectives. *European Journal of Nuclear Medicine and Molecular Imaging* **48,** 21-39 (2021).

60. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine and Biology* **60,** 5471-5496 (2015).

61. Escobar, T. *et al.* Voxel-wise supervised analysis of tumors with multimodal engineered features to highlight interpretable biological patterns. *Medical Physics* **49,** 3816-3829 (2022).

62. Mabray, M. C., Barajas, R. F. & Cha, S. Modern Brain Tumor Imaging. *Brain Tumor Research and Treatment* **3,** 8-23 (2015).

63. Verma, N., Cowperthwaite, M. C., Burnett, M. G. & Markey, M. K. Differentiating tumor recurrence from treatment necrosis: A review of neuro-oncologic imaging strategies. *Neuro-Oncology* **15,** 515-534 (2013).

64. Atkinson, A. J. *et al.* Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* **69,** 89-95 (2001).

65. Therasse, P. *et al.* New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute* **92,** 205-216 (2000).

66. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **50 Suppl,** 122S-150S (2009).

67. Organization, W. H. WHO handbook for reporting results of cancer treatment. *World Health Organization Offset Publication* **48** (1979).

68. Suzuki, C. *et al.* Radiologic measurements of tumor response to treatment: Practical approaches and limitations. *Radiographics* **28,** 329-344 (2008).

69. Andrearczyk, V. & Oreiller, V. *HECKTOR 2022 - Grand Challenge* 2022. https://hecktor.grand-challenge.org/ (2022).

70. Wahl, R. L., Jacene, H., Kasamon, Y. & Lodge, M. A. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *Journal of Nuclear Medicine* **50 Suppl,** 122S-150S (2009).

71. Vanderhoek, M., Perlman, S. B. & Jeraj, R. Impact of the definition of peak standardized uptake value on quantification of treatment response. *Journal of Nuclear Medicine* **53,** 4-11 (2012).

72. O'Connor, J. P. *et al.* Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology* **14,** 169-186 (2017).

73. Goldmacher, G. V. & Conklin, J. The use of tumour volumetrics to assess response to therapy in anticancer clinical trials. *British Journal of Clinical Pharmacology* **73,** 846-54 (2012).

74. Egner, J. R. AJCC Cancer Staging Manual. *JAMA* **304,** 1726-1727 (2010).

75. Im, H. J., Bradshaw, T., Solaiyappan, M. & Cho, S. Y. Current Methods to Define Metabolic Tumor Volume in Positron Emission Tomography: Which One is Better? *Nuclear Medicine and Molecular Imaging* **52,** 5-15 (2018).

76. Larson, S. M. *et al.* Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. *Clinical Positron Imaging* **2,** 159-171 (1999).

77. Meignan, M., Cottereau, A. S., Specht, L. & Mikhaeel, N. G. Total tumor burden in lymphoma – an evolving strong prognostic parameter. *British Journal of Radiology* **94,** 20210448 (2021).

78. Cottereau, A. S. *et al.* Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. *Annals of Oncology* **32,** 404-411 (2021).

79. Carbone, P. P., Kaplan, H. S., Musshoff, K., Smithers, D. W. & Tubiana, M. Report of the Committee on Hodgkin's Disease Staging Classification. *Cancer research* **31,** 1860-1861 (1971).

80. Ceccon, G. *et al.* Dynamic O-(2-18F-fluoroethyl)-L-tyrosine positron emission tomography differentiates brain metastasis recurrence from radiation injury after radiotherapy. *Neuro-Oncology* **19,** 281-288 (2017).

81. Van Helden, P. Data-driven hypotheses. *EMBO Reports* **14,** 104 (2013).

82. Langs, G. *et al.* Machine learning: from radiomics to discovery and routine. *Radiologe* **58,** 1-6 (2018).

83. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **278,** 563-577 (2016).

84. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48,** 441-446 (2012).

85. Vande Perre, S. *et al.* Radiomics: instructions for use. Methodology and examples of applications in women's imaging. *Imagerie de la Femme* **29,** 25-33 (2019).

86. Mayerhoefer, M. E. *et al.* Introduction to radiomics. *Journal of Nuclear Medicine* **61,** 488-495 (2020).

87. Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights into Imaging* **11,** 91 (2020).

88. Kocher, M., Ruge, M. I., Galldiks, N. & Lohmann, P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlentherapie und Onkologie* **196,** 856-867 (2020).

89. Orlhac, F., Nioche, C., Klyuzhin, I., Rahmim, A. & Buvat, I. Radiomics in PET Imaging: A Practical Guide for Newcomers. *PET Clinics* **16,** 597-612 (2021).

90. Google Developers. *Representation - Machine Learning* 2022. https://developers.google.com/machine-learning/crash-course/representation/ (2022).

91. Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A. & Benali, H. From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Processing Magazine* **36,** 132-160 (2019).

92. Avanzo, M. *et al.* Machine and deep learning methods for radiomics. *Medical Physics* **47,** e185-e202 (2020).

93. Van Griethuysen, J. J. *Welcome to pyradiomics documentation!* 2016. https://pyradiomics.readthedocs.io/ (2022).

94. Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **77,** e104-e107 (2017).

95. Nioche, C. *LIFEx* 2022. https://www.lifexsoft.org/.

96. Nioche, C. *et al.* Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Research* **78,** 4786-4789 (2018).

97. Seban, R. D. *et al.* Total metabolic tumor volume and spleen metabolism on baseline [18F]-FDG PET/CT as independent prognostic biomarkers of recurrence in resected breast cancer. *European Journal of Nuclear Medicine and Molecular Imaging* **48,** 3560-3570 (2021).

98. Shiyam Sundar, L. K. *et al.* Fully-automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. *Journal of Nuclear Medicine* **122,** 264063 (2022).

99. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *Lecture Notes in Computer Science, MICCAI 2015, vol 9351* (Springer, 2015), 234-241.

100. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18,** 203-211 (2021).

101. Billot, B. *et al.* SynthSeg: Domain Randomisation for Segmentation of Brain Scans of any Contrast and Resolution. *arXiv.* https://arxiv.org/abs/2107.09559v2 (2021).

102. Van Griethuysen, J. J. *What about gray value discretization? Fixed bin width? Fixed bin count? - pyradiomics documentation* 2016. https://pyradiomics.readthedocs.io/ (2022).

103. IBSI. *Intensity discretisation - Image processing - IBSI documentation* 2022. https://ibsi.readthedocs.io/ (2022).

104. Yip, S. S. & Aerts, H. J. Applications and limitations of radiomics. *Physics in Medicine and Biology* **61,** 150-166 (2016).

105. Tixier, F. *et al.* Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine* **52,** 369-378 (2011).

106. Tixier, F. *et al.* Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *Journal of Nuclear Medicine* **53,** 693-700 (2012).

107. Orlhac, F. *et al.* Tumor texture analysis in 18F-FDG PET: Relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *Journal of Nuclear Medicine* **55,** 414-422 (2014).

108. Leijenaar, R. T. *et al.* The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis. *Scientific Reports* **5,** 11075 (2015).

109. Schwier, M. *et al.* Repeatability of Multiparametric Prostate MRI Radiomics Features. *Scientific Reports* **9,** 9441 (2019).

110. Zwanenburg, A. *et al.* The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295,** 328-338 (2020).

111. IBSI. *IBSI – Image Biomarker Standardisation Initiative* 2022. https://theibsi.github.io/ (2022).

112. Reuzé, S. *et al.* Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges. *International Journal of Radiation Oncology Biology Physics* **102,** 1117-1142 (2018).

113. IBSI. *Image features - IBSI documentation* 2022. https://ibsi.readthedocs.io/ (2022).

114. Haralick, R. M., Dinstein, I. & Shanmugam, K. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics* **3,** 610-621 (1973).

115. Xu, D. H., Kurani, A. S., Furst, J. D. & Raicu, D. S. *Run-length encoding for volumetric texture* in *Proceedings of the Fourth IASTED International Conference on Visualization, Imaging, and Image Processing* (WASET, 2004), 452-458.

116. Thibault, G. *et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification* in *Pattern Recognition and Information Processing* (Springer, 2009), 140-145.

117. Amadasun, M. & King, R. Texural Features Corresponding to Texural Properties. *IEEE Transactions on Systems, Man and Cybernetics* **19,** 1264-1274 (1989).

118. Sun, C. & Wee, W. G. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics and Image Processing* **23,** 341-352 (1983).

119. Lorensen, W. E. & Cline, H. E. *Marching cubes: A high resolution 3D surface construction algorithm* in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH, vol 21* (ACM, 1987), 163-169.

120. Decazes, P. *et al.* Correction to: Tumor fragmentation estimated by volume surface ratio of tumors measured on (18)F-FDG PET/CT is an independent prognostic factor of diffuse large B-cell lymphoma. *European Association of Nuclear Medicine* **45,** 1838-1839 (2018).

121. Rebaud, L. *et al. Evaluation of the prognostic value of tumor fragmentation on [18F]-FDG PET/CT on an independent cohort of diffuse large B-cell lymphoma patients [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **63** (SNMMI, 2022), 3172.

122. Van Velden, F. H. *et al.* Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Molecular Imaging and Biology* **18,** 788-795 (2016).

123. Guezennec, C. *et al.* Inter-observer and segmentation method variability of textural analysis in pretherapeutic FDG PET/CT in head and neck cancer. *PLoS ONE* **14,** e0214299 (2019).

124. Yang, F. *et al.* Impact of contouring variability on oncological PET radiomics features in the lung. *Scientific Reports* **10,** 369 (2020).

125. Whybra, P., Parkinson, C., Foley, K., Staffurth, J. & Spezi, E. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. *Scientific Reports* **9,** 9649 (2019).

126. Papp, L., Rausch, I., Grahovac, M., Hacker, M. & Beyer, T. Optimized feature extraction for radiomics analysis of 18F-FDG PET imaging. *Journal of Nuclear Medicine* **60,** 864-872 (2019).

127. Crandall, J. P. *et al.* Repeatability of 18F-FDG PET Radiomic Features in Cervical Cancer. *Journal of Nuclear Medicine* **62,** 707-715 (2021).

128. Orlhac, F., Soussan, M., Chouahnia, K., Martinod, E. & Buvat, I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS ONE* **10,** e0145063 (2015).

129. Buvat, I., Orlhac, F. & Soussan, M. Tumor texture analysis in PET: Where do we stand? *Journal of Nuclear Medicine* **56,** 1642-1644 (2015).

130. Saint Martin, M. J. *et al.* A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. *Magnetic Resonance Materials in Physics, Biology and Medicine* **34,** 355-366 (2021).

131. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436-444 (2015).

132. Chen, X. *et al.* Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis* **79,** 102444 (2022).

133. Bai, K. *A Comprehensive Introduction to Different Types of Convolutions in Deep Learning - Towards Data Science* https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215 (2022).

134. Hao, W., Yizhou, W., Yaqin, L. & Zhili, S. *The Role of Activation Function in CNN* in *Proceedings of the 2nd International Conference on Information Technology and Computer Application, ITCA 2020* (IEEE, 2020), 429-432.

135. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. *Striving for Simplicity: The All Convolutional Net* in *Workshop Track Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015* (ICLR, 2015).

136. Graham, B. Fractional Max-Pooling. *arXiv.* https://arxiv.org/abs/1412.6071v4 (2014).

137. Ioffe, S. & Szegedy, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* in *32nd International Conference on Machine Learning, ICML 2015* **1** (IMLS, 2015), 448-456.

138. Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42,** 2011-2023 (2017).

139. Lin, M., Chen, Q. & Yan, S. *Network In Network* in *Conference Track Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014* (ICLR, 2014).

140. Biscione, V. & Bowers, J. S. Convolutional neural networks are not invariant to translation, but they can learn to be. *The Journal of Machine Learning Research* **22,** 1-28 (2021).

141. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65,** 386-408 (1958).

142. Vandeput, N. *A Brief History Of Neural Networks - Medium, Analytics Vidhya* 2021. https://medium.com/analytics-vidhya/a-brief-history-of-neural-networks-c234639a43f1 (2022).

143. Tang, Y. Deep Learning using Linear Support Vector Machines. *arXiv.* https://arxiv.org/abs/1306.0239v4 (2013).

144. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet classification with deep convolutional neural networks* in *Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS 2017* **60** (ACM, 2017), 84-90.

145. Wallis, D. *A study of machine learning and deep learning methods and their application to medical imaging* PhD thesis (Université Paris-Saclay, 2021).

146. Younes, B. *Machine Learning applications : 10 cas d'usage pratiques* 2017. https://mrmint.fr/machine-learning-applications (2022).

147. Deo, R. C. Machine Learning in Medicine. *Circulation* **132,** 1920 (2015).

148. Gupta, R. *et al.* Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity* **25,** 1315 (2021).

149. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology* **23,** 40-55 (2022).

150. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115-118 (2017).

151. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical image analysis* **42,** 60-88 (2017).

152. Borstelmann, S. M. Machine Learning Principles for Radiology Investigators. *Academic radiology* **27,** 13-25 (2020).

153. Giger, M. L. Machine Learning in Medical Imaging. *Journal of the American College of Radiology* **15,** 512-520 (2018).

154. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. Artificial intelligence in radiology. *Nature Reviews Cancer* **18,** 500-510 (2018).

155. Bradshaw, T. J. *et al.* Nuclear Medicine and Artificial Intelligence: Best Practices for Algorithm Development. *Journal of Nuclear Medicine* **63,** 500-510 (2021).

156. Yang, J., Sohn, J. H., Behr, S. C., Gullberg, G. T. & Seo, Y. Ct-less direct correction of attenuation and scatter in the image space using deep learning for whole-body fdg pet: Potential benefits and pitfalls. *Radiology: Artificial Intelligence* **3,** e200137 (2021).

157. Reader, A. J. *et al.* Deep Learning for PET Image Reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences* **5,** 1-25 (2020).

158. Katsari, K. *et al.* Artificial intelligence for reduced dose 18F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment. *EJNMMI Physics* **8,** 25 (2021).

159. Wallis, D. *et al.* An [18F]FDG-PET/CT deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients. *European Journal of Nuclear Medicine and Molecular Imaging* **49,** 881-888 (2022).

160. Weisman, A. J. *et al.* Automated quantification of baseline imaging PET metrics on FDG PET/CT images of pediatric Hodgkin lymphoma patients. *EJNMMI physics* **7,** 76 (2020).

161. Capobianco, N. *et al.* Deep-Learning 18F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma. *Journal of Nuclear Medicine* **62,** 30-36 (2021).

162. Girum, K. B. *et al.* 18F-FDG PET Maximum-Intensity Projections and Artificial Intelligence: A Win-Win Combination to Easily Measure Prognostic Biomarkers in DLBCL Patients. *Journal of Nuclear Medicine* **63,** 1925-1932 (2022).

163. Escobar, T. *et al. Radiomic decision maps reveal patterns discriminating between glioma progression and radiation-induced necrosis in static and dual time [18F]-FDOPA PET [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **63** (SNMMI, 2022), 2520.

164. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* **16,** 199-231 (2001).

165. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **17,** 37 (1996).

166. Schmauch, B. *et al.* A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications* **11,** 3877 (2020).

167. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1,** 67-82 (1997).

168. Hastie, T., Tibshirani, R. & Friedman, J. *Elements of Statistical Learning, 2nd ed.* (Springer, 2009).

169. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349,** 255-260 (2015).

170. Wright, S. J. Coordinate descent algorithms. *Mathematical Programming* **151,** 3-34 (2015).

171. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv.* https://arxiv.org/abs/1609.04747 (2017).

172. Harrel, F. *Classification vs. Prediction - Statistical Thinking* 2017. https://www.fharrell.com/post/classification/ (2022).

173. Google Developers. *Classification: Thresholding - Machine Learning* 2022. https://developers.google.com/machine-learning/crash-course/classification/thresholding (2022).

174. Escobar, T. *et al. An original voxel-wise supervised analysis of tumors with multimodal radiomics to highlight predictive biological patterns [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **62** (SNMMI, 2021), 1404.

175. Schmidt, M. *A Note on Structural Extensions of SVMs* 2009. https://www.cs.ubc.ca/{~}schmidtm/Documents/2009{_}Notes{_}StructuredSVMs.pdf.

176. Bennett, K. & Bredensteiner, E. *Duality and geometry in SVM classifiers* in *Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000* (IMLS, 2000), 57-64.

177. Van Erp, N. & Van Gelder, P. *Bayesian logistic regression analysis* in *Conference Proceedings of the American Institute of Physics, AIP 2013* **1553** (AIP, 2013), 147.

178. Schölkopf, B. & Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (The MIT Press, 2018).

179. Jordan, M. I. & Thibaux, R. *The Kernel Trick* 2004. https://people.eecs.berkeley.edu/{~}jordan/courses/281B-spring04/lectures/lec3.pdf.

180. Olah, C. *Neural Networks, Manifolds, and Topology* 2014. https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/ (2022).

181. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* **5,** 115-133 (1943).

182. Yadav, N., Yadav, A. & Kumar, M. in *Briefs in Applied Sciences and Technology* 13-15 (Springer, 2015).

183. Google Developers. A neural Network Playground. https://playground.tensorflow.org/{\#}activation=sigmoid{\&}batchSize=30{\&}dataset=circle{\&}regDataset=reg-plane{\&}learningRate=0.03{\&}regularizationRate=0{\&}noise=0{\&}networkShape=4{\&}seed=0.21011{\&}showTestData=true{\&}discretize=false{\&}percTrainData=90{\&}x=true{\&}y=true{\&}xTimesY=fal (2022).

184. Konar, J., Khandelwal, P. & Tripathi, R. *Comparison of Various Learning Rate Scheduling Techniques on Convolutional Neural Network* in *Proceedings of the International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2020* (IEEE, 2020), 1-5.

185. Wang, Q., Ma, Y., Zhao, K. & Tian, Y. A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science* **9,** 187-212 (2020).

186. Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D. & Smelyanskiy, M. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima* in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017* (ICLR, 2017).

187. Gupta, T. K. & Raza, K. in *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging* 159-182 (Academic Press, 2019).

188. Szandała, T. Review and comparison of commonly used activation functions for deep neural networks. *Studies in Computational Intelligence* **903,** 203-224 (2021).

189. Feng, J. & Lu, S. Performance Analysis of Various Activation Functions in Artificial Neural Networks. *Journal of Physics: Conference Series* **1237,** 022030 (2019).

190. Sukanya Bag. *Activation Functions - All You Need To Know! - Medium, Analytics Vidhya* 2021. https://medium.com/analytics-vidhya/activation-functions-all-you-need-to-know-355a850d025e (2022).

191. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 1989 2:4* **2,** 303-314 (1989).

192. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4,** 251-257 (1991).

193. Hastie, T., Tibshirani, R., James, G. & Witten, D. *An Introduction to Statistical Learning, 2nd ed.* (Springer, 2021).

194. Geman, S., Bienenstock, E. & Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **4,** 1-58 (1992).

195. Neal, B. *et al.* A Modern Take on the Bias-Variance Tradeoff in Neural Networks. *arXiv.* https://arxiv.org/abs/1810.08591 (2018).

196. Tong, Y. & Hong, Z. Hyper-Parameter Optimization: A Review of Algorithms and Applications. *arXiv.* https://arxiv.org/abs/2003.05689 (2020).

197. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* **13,** 281-305 (2012).

198. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. *Algorithms for Hyper-Parameter Optimization* in *Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS 2011* (ACM, 2011), 2546-2554.

199. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On Calibration of Modern Neural Networks. *arXiv.* https://arxiv.org/abs/1706.04599 (2017).

200. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1,** 206-215 (2019).

201. VanderPlas, J. in *Python Data Science Handbook* chap. 5 (O'Reilly Media, Inc., 2016). https://jakevdp.github.io/PythonDataScienceHandbook/05.08-random-forests.html.

202. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees* (CRC Press, 1984).

203. Cunningham, P. & Delany, S. J. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys (CSUR)* **54,** 1-25 (2021).

204. Dudani, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man and Cybernetics* **6,** 325-327 (1976).

205. Wu, X. *et al.* Top 10 algorithms in data mining. *Knowledge and Information Systems* **14,** 1-37 (2007).

206. Fernández, F. & Isasi, P. Local feature weighting in nearest prototype classification. *IEEE Transactions on Neural Networks* **19,** 40-53 (2008).

207. Pölsterl, S. scikit-survival. *The Journal of Machine Learning Research* **21,** 1-6 (2020).

208. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53,** 457-481 (1958).

209. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187-202 (1972).

210. Efron, B. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72,** 557-565 (1977).

211. Abbott, R. D. Logistic regression in survival analysis. *American journal of epidemiology* **121,** 465-471 (1985).

212. Annesi, I., Moreau, T. & Lellouch, J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Statistics in medicine* **8,** 1515-1521 (1989).

213. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics* **2,** 841-860 (2008).

214. Lowsky, D. J. *et al.* A K-nearest neighbors survival probability prediction method. eng. *Statistics in medicine* **32,** 2062-2069 (2013).

215. Scikit-Learn. *3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.2.0 documentation* 2022. https://scikit-learn.org/stable/modules/cross{_}validation.html{\#}cross-validation-iterators (2022).

216. Pedregosa, F *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825-2830 (2011).

217. Buitinck, L. *et al. API design for machine learning software: experiences from the scikit-learn project* in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013), 108-122.

218. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* 70 (Springer, 2013).

219. Wallace, B. C. & Dahabreh, I. J. *Class probability estimates are unreliable for imbalanced data (and how to fix them)* in *Proceedings of the IEEE International Conference on Data Mining, ICDM 2012* (IEEE, 2012), 695-704.

220. King, G. *et al.* Logistic Regression in Rare Events Data. *Political Analysis* **9,** 137-163 (2001).

221. Bergstra, J., Yamins, D. & Cox, D. D. *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures* in *International Conference on Machine Learning, ICML 2013* (IMLS, 2013), 115-123.

222. Wainer, J. & Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications* **182,** 115222 (2021).

223. Gavin, C. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research* **11,** 2079-2107 (2010).

224. Brownlee, J. *Data Leakage in Machine Learning - MachineLearningMastery.com* https://machinelearningmastery.com/data-leakage-machine-learning/ (2022).

225. Ojala, M. & Garriga, G. C. *Permutation tests for studying classifier performance* in *Proceedings of the IEEE International Conference on Data Mining, ICDM 2009* (IEEE, 2009), 908-913.

226. Butvinik, D. *Feature Selection - Exhaustive Overview - Medium, Analytics Vidhya* 2021. https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c (2023).

227. Haury, A. C., Gestraud, P. & Vert, J. P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* **6,** e28210 (2011).

228. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **40,** 16-28 (2014).

229. Jović, A., Brkić, K. & Bogunović, N. *A review of feature selection methods with applications* in *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics* (MIPRO, 2015), 1200-1205.

230. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific reports* **5,** 13087 (2015).

231. Demircioğlu, A. Benchmarking Feature Selection Methods in Radiomics. *Investigative radiology* **57,** 433-443 (2022).

232. Brownlee, J. *How to Reduce Variance in a Final Machine Learning Model - MachineLearningMastery.com* https://machinelearningmastery.com/how-to-reduce-model-variance/ (2023).

233. Zwanenburg, A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging* **46,** 2638-2655 (2019).

234. Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental* **2,** 36 (2018).

235. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **350,** g7594 (2015).

236. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14,** 749-762 (2017).

237. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2,** 305-311 (2020).

238. Pinto dos Santos, D., Dietzel, M. & Baessler, B. A decade of radiomics research: are images really data or just patterns in the noise? *European radiology* **31,** 1-4 (2021).

239. Buvat, I. & Orlhac, F. The T.R.U.E. Checklist for Identifying Impactful Artificial Intelligence-Based Findings in Nuclear Medicine: Is It True? Is It Reproducible? Is It Useful? Is It Explainable? *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **62,** 752-754 (2021).

240. Zamanipoor Najafabadi, A. H. *et al.* TRIPOD statement: A preliminary pre-post analysis of reporting and methods of prediction models. *BMJ* **10,** e041537 (2020).

241. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine* **26,** 1320-1324 (2020).

242. Jha, A. K. *et al.* Nuclear Medicine and Artificial Intelligence: Best Practices for Evaluation (the RELAINCE Guidelines). *Journal of Nuclear Medicine* **63,** 1288-1299 (2022).

243. Saboury, B. *et al.* Artificial Intelligence in Nuclear Medicine: Opportunities, Challenges, and Responsibilities Toward a Trustworthy Ecosystem. *Journal of Nuclear Medicine,* 121.263703 (2022).

244. Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5,** 1-9 (2014).

245. Sanduleanu, S. *et al.* Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **127,** 349-360 (2018).

246. Welch, M. L. *et al.* Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology* **2,** 2-9 (2019).

247. Nickerson, R. S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* **2,** 175-220 (1998).

248. Easterbrook, P. J., Gopalan, R, Berlin, J. A. & Matthews, D. R. Publication bias in clinical research. *The Lancet* **337,** 867-872 (1991).

249. Buvat, I. & Orlhac, F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. *Journal of Nuclear Medicine* **60,** 1543-1544 (2019).

250. Song, F *et al.* Dissemination and publication of research findings : an updated review of related biases. *Health Technology Assessment* **14,** 1-220 (2010).

251. Kim, B., Khanna, R. & Koyejo, O. *Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability* in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016* (ACM, 2016), 2288-2296.

252. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267,** 1-38 (2019).

253. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. *Definitions, methods, and applications in interpretable machine learning* in *Proceedings of the National Academy of Sciences of the United States of America* **116** (NAS, 2019), 22071-22080.

254. Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* (2019).

255. Vinuesa, R. *et al.* The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* **11,** 230 (2020).

256. Tobia, K., Nielsen, A. & Stremitzer, A. When Does Physician Use of AI Increase Liability? *Journal of Nuclear Medicine* **62,** 17-21 (2021).

257. Price, W. N., Gerke, S. & Cohen, I. G. How Much Can Potential Jurors Tell Us About Liability for Medical Artificial Intelligence? *Journal of Nuclear Medicine* **62,** 15-16 (2021).

258. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* **24,** 423-431 (2017).

259. Gretto, C. *The dangers of AI in health care: risk homeostasis and automation bias - Towards Data Science* 2017. https://towardsdatascience.com/the-dangers-of-ai-in-health-care-risk-homeostasis-and-automation-bias-148477a9080f (2023).

260. Challen, R. *et al.* Artificial intelligence, bias and clinical safety. *BMJ quality & safety* **28,** 231-237 (2019).

261. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* **38,** 50-57 (2016).

262. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **1,** 389-399 (2019).

263. Rudin, C. *et al.* Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys* **16,** 1-85 (2021).

264. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine* **15,** e1002683 (2018).

265. Wallis, D. & Buvat, I. Clever Hans effect found in a widely used brain tumour MRI dataset. *Medical image analysis* **77,** 102368 (2022).

266. Yan, J. *et al.* Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *Journal of Nuclear Medicine* **56,** 1667-1673 (2015).

267. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118-127 (2007).

268. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* **6,** e17238 (2011).

269. Fortin, J. P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161,** 149-170 (2017).

270. Fortin, J. P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167,** 104-120 (2018).

271. Orlhac, F. *et al.* A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *Journal of Nuclear Medicine* **59,** 1321-1328 (2018).

272. Orlhac, F., Frouin, F., Nioche, C., Ayache, N. & Buvat, I. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* **291,** 53-59 (2019).

273. Orlhac, F. *et al.* How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European radiology* **31,** 2272-2280 (2021).

274. Orlhac, F. *et al.* A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine* **63,** 172-179 (2022).

275. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv.* https://arxiv.org/abs/1702.08608v2 (2017).

276. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv.* https://arxiv.org/abs/1606.03490v3 (2016).

277. Glorot, X. & Bengio, Y. *Understanding the difficulty of training deep feedforward neural networks* in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* **9** (PMLR, 2010), 249-256.

278. Zech, J. R., Forde, J. Z. & Littman, M. L. Individual predictions matter: Assessing the effect of data ordering in training fine-tuned CNNs for medical imaging. *arXiv.* https://arxiv.org/abs/1912.03606v1 (2019).

279. Chang, C.-H., Rampasek, L. & Goldenberg, A. Dropout Feature Ranking for Deep Learning Models. *arXiv.* https://arxiv.org/abs/1712.08645v1 (2017).

280. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26,** 1340-1347 (2010).

281. MathWorks. *What Is Interpretability - MATLAB & Simulink* https://fr.mathworks.com/discovery/interpretability.html (2023).

282. Ribeiro, M. T., Singh, S. & Guestrin, C. *"Why should i trust you?" Explaining the predictions of any classifier* in *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 1135-1144.

283. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. *arXiv.* https://arxiv.org/abs/1705.07874v2 (2017).

284. Shapley, L. S. in *Notes on the N-Person Game* (RAND Corporation, 1951). https://www.rand.org/pubs/research{_}memoranda/RM0670.html.

285. Zeiler, M. D. & Fergus, R. in *Computer Vision – ECCV 2014. Lecture Notes in Computer Science* 818-833 (Springer, 2014).

286. Simonyan, K., Vedaldi, A. & Zisserman, A. *Deep inside convolutional networks: Visualising image classification models and saliency maps* in *Workshop Track Proceedings of the 2rd International Conference on Learning Representations, ICLR 2014* (ICLR, 2014).

287. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. *Learning Deep Features for Discriminative Localization* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 2921-2929.

288. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **128,** 336-359 (2020).

289. Shrikumar, A., Greenside, P. & Kundaje, A. *Learning important features through propagating activation differences* in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017* (IMLS, 2017), 3145-3153.

290. Kumar, D., Wong, A. & Taylor, G. W. *Explaining the Unexplained: A CLass-Enhanced Attentive Response (CLEAR) Approach to Understanding Deep Neural Networks* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR-W 2017* (IEEE, 2017), 1686-1694.

291. Hosny, A., Aerts, H. J. & Mak, R. H. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *The Lancet Digital health* **1,** e106-e107 (2019).

292. Reyes, M. *et al.* On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence* **2,** e190043 (2020).

293. Barrett, H. H. Is there a role for image science in the brave new world of artificial intelligence? *Journal of Medical Imaging* **7,** 12702 (2019).

294. Oren, O., Gersh, B. J. & Bhatt, D. L. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health* **2,** e486-e488 (2020).

295. Lin, K., Cidan, W., Qi, Y. & Wang, X. Glioma grading prediction using multiparametric magnetic resonance imaging-based radiomics combined with proton magnetic resonance spectroscopy and diffusion tensor imaging. *Medical physics* **49,** 4419-4429 (2022).

296. Wang, Z.-H., Xiao, X.-L., Zhang, Z.-T., He, K. & Hu, F. A Radiomics Model for Predicting Early Recurrence in Grade II Gliomas Based on Preoperative Multiparametric Magnetic Resonance Imaging. *Frontiers in oncology* **11,** 684996 (2021).

297. Shi, Y. *et al.* Ultrasound-based radiomics XGBoost model to assess the risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual application of SHAP. *Frontiers in Oncology* **12,** 897596 (2022).

298. Park, C. J. *et al.* An interpretable radiomics model to select patients for radiotherapy after surgery for WHO grade 2 meningiomas. *Radiation oncology (London, England)* **17,** 147 (2022).

299. Giraud, P. *et al.* Interpretable Machine Learning Model for Locoregional Relapse Prediction in Oropharyngeal Cancers. *Cancers* **13,** 57 (2020).

300. Rundo, L. *et al.* Clinically Interpretable Radiomics-Based Prediction of Histopathologic Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma. *Frontiers in oncology* **12,** 868265 (2022).

301. Zhou, L.-Q. *et al.* Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. *Radiology* **294,** 19-28 (2020).

302. Kumar, D., Sankar, V., Clausi, D., Taylor, G. W. & Wong, A. SISC: End-to-End Interpretable Discovery Radiomics-Driven Lung Cancer Prediction via Stacked Interpretable Sequencing Cells. *IEEE Access* **7,** 145444-145454 (2019).

303. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Medicine* **15,** e1002711 (2018).

304. Pereira, S., Meier, R., Alves, V., Reyes, M. & Silva, C. A. in *Lecture Notes in Computer Science* 106-114 (Springer, 2018).

305. Wei, L. *et al.* A deep survival interpretable radiomics model of hepatocellular carcinoma patients. *Physica Medica* **82,** 295-305 (2021).

306. Lou, B. *et al.* An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health* **1,** e136-e147 (2019).

307. Ho Cho, H. *et al.* Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from CT scans. *Communications Biology* **4,** 1-12 (2021).

308. Tomaszewski, M. R. & Gillies, R. J. The Biological Meaning of Radiomic Features. *Radiology* **298,** 505 (2021).

309. Kindermans, P. J. *et al.* in *Lecture Notes in Computer Science* 267-280 (Springer, 2019).

310. Ghorbani, A., Abid, A. & Zou, J. *Interpretation of neural networks is fragile* in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019* (2019), 3681-3688.

311. Zhang, X. *et al.* Interpretable Deep Learning under Fire. *arXiv.* https://arxiv.org/abs/1812.00891 (2018).

312. Heo, J., Joo, S. & Moon, T. Fooling neural network interpretations via adversarial model manipulation. *arXiv.* https://arxiv.org/abs/1902.02041 (2019).

313. Mittelstadt, B., Russell, C. & Wachter, S. *Explaining Explanations in AI* in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, FAT* 2019* (ACM, 2019), 279-288.

314. Gosiewska, A. & Biecek, P. Do Not Trust Additive Explanations. *arXiv.* https://arxiv.org/abs/1903.11420v3 (2019).

315. Adebayo, J. *et al.* in *arXiv* (2018). https://arxiv.org/abs/1810.03292v3.

316. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P. & Preece, A. *Sanity Checks for Saliency Metrics* in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020* (AAAI, 2020), 6021-6029.

317. Wu, J. *et al.* Tumor Subregion Evolution-Based Imaging Features to Assess Early Response and Predict Prognosis in Oropharyngeal Cancer. *Journal of Nuclear Medicine* **61,** 327-336 (2020).

318. Wu, J., Gong, G., Cui, Y. & Li, R. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging* **44,** 1107-1115 (2016).

319. Wu, J. *et al.* Robust Intratumor Partitioning to Identify High-Risk Subregions in Lung Cancer: A Pilot Study. *International Journal of Radiation Oncology Biology Physics* **95,** 1504-1512 (2016).

320. Xu, H. *et al.* Subregional Radiomics Analysis of PET/CT Imaging with Intratumor Partitioning: Application to Prognosis for Nasopharyngeal Carcinoma. *Molecular Imaging and Biology* **22,** 1414-1426 (2020).

321. Even, A. J. *et al.* Clustering of multi-parametric functional imaging to identify high-risk subvolumes in non-small cell lung cancer. *Radiotherapy and Oncology* **125,** 379-384 (2017).

322. Beaumont, J *et al.* Voxel-based identification of local recurrence subregions from pre-treatment PET/CT for locally advanced head and neck cancers. *EJNMMI Research* **9,** 90 (2019).

323. Vuong, D. *et al.* Radiomics Feature Activation Maps as a New Tool for Signature Interpretability. *Frontiers in Oncology* **10,** 578895 (2020).

324. Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging* **29,** 1310-1320 (2010).

325. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1,** 80 (1945).

326. Orlhac, F., Soussan, M., Chouahnia, K., Martinod, E. & Buvat, I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS ONE* **10,** e0145063 (2015).

327. Brooks, F. J. & Grigsby, P. W. Current measures of metabolic heterogeneity within cervical cancer do not predict disease outcome. *Radiation Oncology* **6,** 69 (2011).

328. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *Journal of Nuclear Medicine* **56,** 38-44 (2015).

329. Outi, J. G. *Développements en radiomique pour une meilleure caractérisation du gliome infiltrant du tronc cérébral à partir d'imagerie par résonance magnétique* PhD thesis (Université Paris-Saclay, 2019).

330. Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R. & Lin, C. J. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* **9,** 1871-1874 (2008).

331. Kurosawa, A. *Rashōmon* 1950.

332. Semenova, L., Rudin, C. & Parr, R. On the Existence of Simpler Machine Learning Models. *arXiv.* http://arxiv.org/abs/1908.01755 (2019).

333. Semenova, L., Rudin, C. & Parr, R. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv.* https://arxiv.org/abs/1908.01755v1 (2019).

334. D'Amour, A. Revisiting Rashomon: A Comment on "The Two Cultures". *arXiv* **7,** 59-63. https://arxiv.org/abs/2104.02150v1 (2021).

335. Box, G. E. P. & Draper, N. *Empirical Model-Building and Response Surfaces* 424 (Wiley, 1987).

336. Bernatowicz, K. *et al.* Robust imaging habitat computation using voxel-wise radiomics features. *Scientific Reports* **11,** 20133 (2021).

337. Amores, J. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* **201,** 81-105 (2013).

338. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115,** 211-252 (2014).

339. Wang, H. *et al.* Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI research* **7** (2017).

340. Sun, Q. *et al.* Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. *Frontiers in Oncology* **10,** 53 (2020).

341. Kumar, M. D., Babaie, M., Zhu, S., Kalra, S. & Tizhoosh, H. R. A Comparative Study of CNN, BoVW and LBP for Classification of Histopathological Images. *arXiv.* https://arxiv.org/abs/1710.01249v1 (2017).

342. Bizzego, A. *et al.* Integrating deep and radiomics features in cancer bioimaging. *bioRxiv* (2019).

343. Gore, S., Chougule, T., Jagtap, J., Saini, J. & Ingalhalikar, M. A Review of Radiomics and Deep Predictive Modeling in Glioma Characterization. *Academic radiology* **28,** 1599-1621 (2021).

344. Xiao, T., Hua, W., Li, C. & Wang, S. *Glioma grading prediction by exploring radiomics and deep learning features* in *Proceedings of the 3th International Symposium on Image Computing and Digital Medicine, ISICDM 2019* (ACM, 2019), 208-213.

345. Andrearczyk, V. *et al.* in *Head and Neck Tumor Segmentation and Outcome Prediction* (éd. Andrearczyk, V., Oreiller, V., Hatt, M. & Depeursinge, A.) 1-30 (Springer Nature, 2023).

346. Demircioğlu, A. Are deep models in radiomics performing better than generic models? A systematic review. *European Radiology Experimental* **7,** 11 (2023).

347. Jensen, J. L. W. V. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* **30,** 175-193 (1906).

348. Efron, B. & Tibshirani, R. Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association* **92,** 548 (1997).

349. Sahiner, B., Chan, H. P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Medical physics* **35,** 1559-1570 (2008).

350. Eary, J. F. *et al.* Sarcoma tumor FDG uptake measured by PET and patient outcome: A retrospective analysis. *European Journal of Nuclear Medicine and Molecular Imaging* **29,** 1149-1154 (2002).

351. Crombé, A. *et al.* Soft-tissue sarcomas: Assessment of MRI features correlating with histologic grade and patient outcome. *Radiology* **291,** 181659 (2019).

352. Crombé, A. *et al.* T2 -based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging* **50,** 497-510 (2019).

353. Crombé, A. *et al.* High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models? *Journal of Magnetic Resonance Imaging* **52,** 282-297 (2020).

354. Benz, M. R. *et al.* Combined assessment of metabolic and volumetric changes for assessment of tumor response in patients with soft-tissue sarcomas. *Journal of Nuclear Medicine* **49,** 1579-1584 (2008).

355. Guillou, L. *et al.* Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. *Journal of Clinical Oncology* **15,** 350-362 (1997).

356. SNMMI. *PIDSC Young Investigator Award* https://www.snmmi.org/Membership/Content.aspx?ItemNumber=10673 (2023).

357. Fédération pour la Recherche sur le Cerveau. *Les tumeurs cérébrales* https://www.frcneurodon.org/comprendre-le-cerveau/le-cerveau-malade-et-ses-maladies-neurologiques/les-tumeurs-cerebrales/ (2023).

358. Cavaliere, R. & Schiff, D. Cerebral metastases—a therapeutic update. *Nature Clinical Practice Neurology* **2,** 426-436 (2006).

359. Weller, M. *et al.* EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *The Lancet. Oncology* **15,** e395-403 (2014).

360. Chinot, O. L. *et al.* Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *The New England journal of medicine* **370,** 709-722 (2014).

361. Gerstner, E. R., Sorensen, A. G., Jain, R. K. & Batchelor, T. T. Advances in neuroimaging techniques for the evaluation of tumor growth, vascular permeability, and angiogenesis in gliomas. *Current Opinion in Neurology* **21,** 728-735 (2008).

362. Gerstner, E. R. & Batchelor, T. T. Imaging and response criteria in gliomas. *Current opinion in oncology* **22,** 598-603 (2010).

363. Wen, P. Y. *et al.* Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *Journal of Clinical Oncology* **28,** 1963-1972 (2010).

364. Vellayappan, B. *et al.* Diagnosis and Management of Radiation Necrosis in Patients With Brain Metastases. *Frontiers in oncology* **8,** 395 (2018).

365. Suh, J. H. *et al.* Current approaches to the management of brain metastases. *Nature Reviews Clinical Oncology* **17,** 279-299 (2020).

366. Vogelbaum, M. A. *et al.* Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline. *Journal of Clinical Oncology* **40,** 492-516 (2022).

367. Leibel, S. A. & Sheline, G. E. in *Radiation injury to the nervous system* (1991).

368. Verma, N., Cowperthwaite, M. C., Burnett, M. G. & Markey, M. K. Differentiating tumor recurrence from treatment necrosis: a review of neuro-oncologic imaging strategies. *Neuro-oncology* **15,** 515-534 (2013).

369. Lee, D., Riestenberg, R. A., Haskell-Mendoza, A. & Bloch, O. Brain Metastasis Recurrence Versus Radiation Necrosis: Evaluation and Treatment. *Neurosurgery clinics of North America* **31,** 575-587 (2020).

370. Lyubimova, N. & Hopewell, J. W. Experimental evidence to support the hypothesis that damage to vascular endothelium plays the primary role in the development of late radiation-induced CNS injury. *The British journal of radiology* **77,** 488-492 (2004).

371. Shah, R. *et al.* Radiation necrosis in the brain: imaging features and differentiation from tumor recurrence. *Radiographics* **32,** 1343-1359 (2012).

372. Albert, N. L. *et al.* Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. *Neuro-oncology* **18,** 1199-1208 (2016).

373. Verger, A. & Langen, K.-J. in *Glioblastoma* 155-174 (Codon Publications, 2017).

374. Treglia, G. *et al.* Diagnostic Performance and Prognostic Value of PET/CT with Different Tracers for Brain Tumors: A Systematic Review of Published Meta-Analyses. *International Journal of Molecular Sciences* **20,** 4669 (2019).

375. Furuse, M. *et al.* Radiological diagnosis of brain radiation necrosis after cranial irradiation for brain tumor: a systematic review. *Radiation Oncology* **14,** 28 (2019).

376. Langleben, D. D. & Segall, G. M. PET in differentiation of recurrent brain tumor from radiation injury. eng. *Journal of Nuclear Medicine* **41,** 1861-1867 (2000).

377. Chao, S. T., Suh, J. H., Raja, S., Lee, S. Y. & Barnett, G. The sensitivity and specificity of FDG PET in distinguishing recurrent brain tumor from radionecrosis in patients treated with stereotactic radiosurgery. *International journal of cancer* **96,** 191-197 (2001).

378. Ohtani, T. *et al.* Brain tumour imaging with carbon-11 choline: comparison with FDG PET and gadolinium-enhanced MR imaging. *European Journal of Nuclear Medicine and Molecular Imaging* **28,** 1664-1670 (2001).

379. Hara, T., Kondo, T., Hara, T. & Kosaka, N. Use of 18F-choline and 11C-choline as contrast agents in positron emission tomography imaging-guided stereotactic biopsy sampling of gliomas. *Journal of neurosurgery* **99,** 474-479 (2003).

380. Gao, L., Xu, W., Li, T., Zheng, J. & Chen, G. Accuracy of 11C-choline positron emission tomography in differentiating glioma recurrence from radiation necrosis: A systematic review and meta-analysis. *Medicine* **97,** e11556 (2018).

381. Shields, A. F. PET Imaging with 18F-FLT and Thymidine Analogs: Promise and Pitfalls. *Journal of Nuclear Medicine* **44,** 1432-1434 (2003).

382. Li, Z., Yu, Y., Zhang, H., Xu, G. & Chen, L. A meta-analysis comparing 18F-FLT PET with 18F-FDG PET for assessment of brain tumor recurrence. *Nuclear medicine communications* **36,** 695-701 (2015).

383. Guglielmo, P. *et al.* [18F] Fluorothymidine Positron Emission Tomography Imaging in Primary Brain Tumours: A Systematic Review. *Current medical imaging* **18,** 363-371 (2022).

384. Rollet, A.-C. *Analyse radiomique de TEP-TDM à la 18F-DOPA pour différencier la progression tumorale vraie de lésions radio-induites chez des patients atteints de glioblastomes traités par radio-chimiothérapie.* MD (Université de Nice Sophia Antipolis, 2020).

385. Law, I. *et al.* Joint EANM/EANO/RANO practice guidelines/SNMMI procedure standards for imaging of gliomas using PET with radiolabelled amino acids and [18F]FDG: version 1.0. *European journal of nuclear medicine and molecular imaging* **46,** 540-557 (2019).

386. Tomura, N., Kokubun, M., Saginoya, T., Mizuno, Y. & Kikuchi, Y. Differentiation between Treatment-Induced Necrosis and Recurrent Tumors in Patients with Metastatic Brain Tumors: Comparison among 11C-Methionine-PET, FDG-PET, MR Permeability Imaging, and MRI-ADC-Preliminary Results. *American Journal of Neuroradiology* **38,** 1520-1527 (2017).

387. Yu, J. *et al.* Accuracy of 18F-FDOPA Positron Emission Tomography and 18F-FET Positron Emission Tomography for Differentiating Radiation Necrosis from Brain Tumor Recurrence. *World neurosurgery* **114,** e1211-e1224 (2018).

388. Chen, W. *et al.* 18F-FDOPA PET imaging of brain tumors: comparison study with 18F-FDG PET and evaluation of diagnostic accuracy. eng. *Journal of Nuclear Medicine* **47,** 904-911 (2006).

389. Chen, W. Clinical applications of PET in brain tumors. *Journal of Nuclear Medicine* **48,** 1468-1481 (2007).

390. Youland, R. S. *et al.* The role of LAT1 in (18)F-DOPA uptake in malignant gliomas. *Journal of Neuro-Oncology* **111,** 11-18 (2013).

391. Papin-Michault, C. *et al.* Study of LAT1 Expression in Brain Metastases: Towards a Better Understanding of the Results of Positron Emission Tomography Using Amino Acid Tracers. *PloS one* **11,** e0157139 (2016).

392. Papin-Michault, C. *et al.* Study of LAT1 Expression in Brain Metastases: Towards a Better Understanding of the Results of Positron Emission Tomography Using Amino Acid Tracers. *PLOS ONE* **11,** e0157139 (2016).

393. Dadone-Montaudié, B. *et al.* [18F] FDOPA standardized uptake values of brain tumors are not exclusively dependent on LAT1 expression. *PloS one* **12,** e0184625 (2017).

394. Ważyńska, M. A., Haveman, L. Y., Windhorst, A. D., Elsinga, P. H. & Vugts, D. J. State of the art of radiochemistry for 11C and 18F PET tracers. *Nuclear Medicine and Molecular Imaging* **1,** 107-120 (2022).

395. Laverman, P., Boerman, O. C., Corstens, F. H. & Oyen, W. J. Fluorinated amino acids for tumour imaging with positron emission tomography. *European Journal of Nuclear Medicine and Molecular Imaging* **29,** 681-690 (2002).

396. Haute Autorité de Santé. *IASOdopa* https://www.has-sante.fr/jcms/c{_}538447/fr/iasodopa-fdopa-18f.

397. Haute Autorité de Santé. *DOPACIS* https://www.has-sante.fr/jcms/c{_}962479/fr/dopacis-6-fluoro-18f-l-dopa-ou-fdopa (2023).

398. Haute Autorité de Santé. *DOPAVIEW* https://www.has-sante.fr/jcms/c{_}2624087/fr/dopaview-fluorodopa-18f.

399. Galldiks, N. *et al.* Comment on "Hypometabolic gliomas on FET-PET—is there an inverted U-curve for survival?". *Neuro-Oncology* **21,** 1612-1613 (2019).

400. Sala, Q. *et al.* 18F-DOPA, a clinically available PET tracer to study brain inflammation? *Clinical nuclear medicine* **39,** e283-e285 (2014).

401. Galldiks, N. *et al.* The use of dynamic O-(2-18F-fluoroethyl)-l-tyrosine PET in the diagnosis of patients with progressive and recurrent glioma. eng. *Neuro-oncology* **17,** 1293-1300 (2015).

402. Maurer, G. D. *et al.* (18)F-FET PET Imaging in Differentiating Glioma Progression from Treatment-Related Changes: A Single-Center Experience. eng. *Journal of Nuclear Medicine* **61,** 505-511 (2020).

403. Lapa, C. *et al.* Comparison of the amino acid tracers 18F-FET and 18F-DOPA in high-grade glioma patients. eng. *Journal of Nuclear Medicine* **55,** 1611-1616 (2014).

404. Schiepers, C., Chen, W., Cloughesy, T., Dahlbom, M. & Huang, S.-C. 18F-FDOPA kinetics in brain tumors. eng. *Journal of Nuclear Medicine* **48,** 1651-1661 (2007).

405. Ginet, M. *et al.* Integration of dynamic parameters in the analysis of (18)F-FDopa PET imaging improves the prediction of molecular features of gliomas. eng. *European Journal of Nuclear Medicine and Molecular Imaging* **47,** 1381-1390 (2020).

406. Zaragori, T. *et al.* Use of static and dynamic [18F]-F-DOPA PET parameters for detecting patients with glioma recurrence or progression. *European Journal of Nuclear Medicine and Molecular Imaging Research* **10,** 56 (2020).

407. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50-60 (1947).

408. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289-300 (1995).

409. Lohmann, P. *et al.* Dual-time-point O-(2-[(18)F]fluoroethyl)-L-tyrosine PET for grading of cerebral gliomas. eng. *European Radiology* **25,** 3017-3024 (2015).

410. Orlhac, F. *et al. Identification of a radiomic signature to distinguish recurrence from radiation-induced necrosis in treated glioblastomas using machine learning methods on dual-point 18F-FDOPA PET images [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* **60** (2019), 57.

411. Orlhac, F. *et al. Identifying a reliable radiomic signature from scarce data: illustration for 18F-FDOPA PET images in glioblastoma patients [conference abstract]* in *European Association of Nuclear Medicine annual congress abstracts* (2020).

412. Lohmann, P. *et al.* Radiation injury vs. recurrent brain metastasis: combining textural feature radiomics analysis and standard parameters may increase (18)F-FET PET accuracy without dynamic scans. eng. *European radiology* **27,** 2916-2927 (2017).

413. Lohmann, P. *et al.* Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. eng. *NeuroImage. Clinical* **20,** 537-542 (2018).

414. Lohmann, P. *et al.* Fet pet radiomics for differentiating pseudoprogression from early tumor progression in glioma patients post-chemoradiation. *Cancers* **12,** 3835 (2020).

415. Ahrari, S. *et al.* Relevance of Dynamic (18)F-DOPA PET Radiomics for Differentiation of High-Grade Glioma Progression from Treatment-Related Changes. eng. *Biomedicines* **9,** 1924 (2021).

416. Carrano, A., Juarez, J. J., Incontri, D., Ibarra, A. & Guerrero Cazares, H. Sex-Specific Differences in Glioblastoma. eng. *Cells* **10,** 1783 (2021).

417. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. eng. *Neuro-oncology* **22,** iv1-iv96 (2020).

418. Baldi, I, Huchet, A, Bauchet, L & Loiseau, H. Épidémiologie des glioblastomes. *Neurochirurgie* **56,** 433-440 (2010).

419. Bros, M. *et al.* Effects of Carbidopa Premedication on (18)F-FDOPA PET Imaging of Glioma: A Multiparametric Analysis. eng. *Cancers* **13,** 5340 (2021).

420. Johnson, H. J., Harris, G. & Kent, W. BRAINSFit: Mutual Information Registrations of Whole-Brain 3D Images, Using the Insight Toolkit. *The Insight Journal* (2007).

421. Fedorov, A. *et al.* 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **30,** 1323-1341 (2012).

422. Bouveyron, C., Girard, S. & Schmid, C. High-Dimensional Discriminant Analysis. *Communications in Statistics - Theory and Methods* **36,** 2607-2623 (2007).

423. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of simpleITK. *Frontiers in Neuroinformatics* (2013).

424. Pöpperl, G. *et al.* FET PET for the evaluation of untreated gliomas: correlation of FET uptake and uptake kinetics with tumour grading. eng. *European journal of nuclear medicine and molecular imaging* **34,** 1933-1942 (2007).

425. Darcourt, J. *et al.* Added value of [18F]FDOPA PET to the management of high-grade glioma patients after their initial treatment: a prospective multicentre study. *European Journal of Nuclear Medicine and Molecular Imaging* (2023).

426. Escobar, T. *et al. Predicting axillary lymph node metastasis in early-stage breast cancer using primary tumor image features on [18F]FDG PET: a comparative study of engineered radiomics, deep learning, and conventional methods [conference abstract]* in *Annual Congress of the European Association of Nuclear Medicine, October 15-19, 2022, Barcelona, Spain* (EJNMMI, 2022), S616, D25: Artificial Intelligence, EP-449.

427. Institut National Du Cancer. *Institut National Du Cancer, "Le cancer du sein - Les cancers les plus fréquents."* 2022. https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein.

428. Marino, M. A., Avendano, D., Zapata, P., Riedl, C. C. & Pinker, K. Lymph Node Imaging in Patients with Primary Breast Cancer: Concurrent Diagnostic Tools. *The Oncologist* **25,** e231-e242 (2020).

429. Cha, J. *et al.* A hierarchical prognostic model for risk stratification in patients with early breast cancer according to 18 F-fludeoxyglucose uptake and clinicopathological parameters. *Cancer Medicine* **7,** 1127-1134 (2018).

430. Cooper, K. L. *et al.* Positron emission tomography (PET) and magnetic resonance imaging (MRI) for the assessment of axillary lymph node metastases in early breast cancer: systematic review and economic evaluation. eng. *Health technology assessment* **15,** III-IV, 1-134 (2011).

431. Maxwell, F *et al.* Diagnostic strategy for the assessment of axillary lymph node status in breast cancer. eng. *Diagnostic and interventional imaging* **96,** 1089-1101 (2015).

432. Riegger, C. *et al.* Comparison of the diagnostic value of FDG-PET/CT and axillary ultrasound for the detection of lymph node metastases in breast cancer patients. *Acta Radiologica* **53,** 1092-1098 (2012).

433. Rousseau, C. *et al.* FDG PET evaluation of early axillary lymph node response to neoadjuvant chemotherapy in stage II and III breast cancer patients. *European Journal of Nuclear Medicine and Molecular Imaging* **38,** 1029-1036 (2011).

434. Zhou, L.-Q. *et al.* Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. *Radiology* **294,** 19-28 (2020).

435. Tan, Y. Y. *et al.* Primary tumor characteristics predict sentinel lymph node macrometastasis in breast cancer. *Breast Journal* **11,** 338-343 (2005).

436. Song, B.-I. A machine learning-based radiomics model for the prediction of axillary lymph-node metastasis in breast cancer. *Breast Cancer* **28,** 664-671 (2021).

437. Dong, Y. *et al.* Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. *European Radiology* **28,** 582-591 (2018).

438. Han, L. *et al.* Radiomic nomogram for prediction of axillary lymph node metastasis in breast cancer. *European Radiology* **29,** 3820-3829 (2019).

439. Gurleyik, G., Aker, F., Aktekin, A. & Saglam, A. Tumor characteristics influencing non-sentinel lymph node involvement in clinically node negative patients with breast cancer. *Journal of Breast Cancer* **14,** 124-128 (2011).

440. Wu, J. L. *et al.* Prediction of axillary lymph node metastases in breast cancer patients based on pathologic information of the primary tumor. *Medical Science Monitor* **20,** 577-581 (2014).

441. Tseng, H. S. *et al.* Tumor characteristics of breast cancer in predicting axillary lymph node metastasis. *Medical Science Monitor* **20,** 1155-1161 (2014).

442. Yang, J. *et al.* Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. *Scientific Reports* **9,** 1-11 (2019).

443. Rasponi, A. *et al.* Breast cancer: primary tumor characteristics related to lymph node involvement. *Tumori Journal* **67,** 19-26 (1981).

444. Sopik, V. & Narod, S. A. The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. *Breast Cancer Research and Treatment* **170,** 647 (2018).

445. Altundag, K. Larger Tumor Size Detected by Sonography Might Not Always Reflect Increased Risk of Axillary Lymph Node Metastasis in Patients With Breast Cancer. *Journal of ultrasound in medicine* **38,** 2521 (2019).

446. Bartella, L., Smith, C. S., Dershaw, D. D. & Liberman, L. Imaging Breast Cancer. *Radiologic Clinics of North America* **45,** 45-67 (2007).

447. Singh, H. in *Preventive Oncology for the Gynecologist* 341-351 (Springer, 2019).

448. Institut National Du Cancer. *Institut National Du Cancer, "Exérèse du ganglion sentinelle."* https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Chirurgie-tumorectomie-et-mastectomie/Exerese-du-ganglion-sentinelle.

449. Evangelista, L., Gori, S., Rubini, G. & Gallo, M. Management of hyperglycemia in oncological patients scheduled for an FDG-PET/CT examination. *Clinical and Translational Imaging* **7,** 447-450 (2019).

450. Lehmann, E. L. & D'Abrera, H. J. M. *Nonparametrics: Statistical methods based on ranks* (Holden-Day, 1975).

451. Stigler, S. M. Karl Pearson's Theoretical Errors and the Advances They Inspired. *Statistical Science* **23,** 261-271 (2008).

452. Van Griethuysen, J. J. *Pyradiomics - Release Notes* https://pyradiomics.readthedocs.io/en/latest/changes.html.

453. Wang, S., Chen, W., Xie, S. M., Azzari, G. & Lobell, D. B. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing* **12,** 207 (2020).

454. Dubost, F. *et al. Gp-Unet: Lesion detection from weak labels with a 3D regression network* in *Lecture Notes in Computer Science* **10435 LNCS** (2017), 214-221.

455. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv.* https://arxiv.org/abs/1207.0580 (2012).

456. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15,** 1929-1958 (2014).

457. Van Der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* **14,** 1-13 (2014).

458. Xie, X. *et al.* A Survey on Incorporating Domain Knowledge into Deep Learning for Medical Image Analysis. *Medical Image Analysis* **69,** 101985 (2020).

459. Papanikolaou, N., Matos, C. & Koh, D. M. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging* **20,** 1-10 (2020).

460. Huynh, B. N. *et al. Comparing Deep Learning and Conventional Machine Learning for Outcome Prediction of Head and Neck Cancer in PET/CT* in *Lecture Notes in Computer Science* **13209 LNCS** (Springer, 2022), 318-326.

461. Klyuzhin, I. S. *et al.* Testing the Ability of Convolutional Neural Networks to Learn Radiomic Features. *Computer Methods and Programs in Biomedicine* **219,** 106750 (2022).

462. Rebaud, L., Escobar, T., Khalid, F., Girum, K. & Buvat, I. *Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT* in *Head and Neck Tumor Segmentation and Outcome Prediction* (éd. Andrearczyk, V., Oreiller, V., Hatt, M. & Depeursinge, A.) (Springer, 2023), 121-134.

463. MIT. *Data-Centric AI vs. Model-Centric AI: Introduction to Data-Centric AI* https://dcai.csail.mit.edu/lectures/data-centric-model-centric/ (2023).

464. Rebaud, L., Escobar, T., Khalid, F., Girum, K. & Buvat, I. *Lrebaud/ICARE: Individual Coefficient Approximation for Risk Estimation (ICARE) model* 2022. https://github.com/Lrebaud/ICARE.

465. Khalid, F. *et al.* Multimodal MRI radiomic models to predict genomic mutations in diffuse intrinsic pontine glioma with missing imaging modalities. *Frontiers in Medicine* **10,** 1071447 (2023).

466. Khalid, F. *et al.* DIPG-23. Artificial intelligence for detecting ACVR1 mutations in patients with DIPG using MRI and clinical data [conference abstract]. *Neuro-Oncology* **24,** i23-i23 (2022).

467. Girum, K. *et al. Fully automatic segmentation of lesions in 3D using deep learning [conference abstract]* in *Journal of Nuclear Medicine, the Society of Nuclear Medicine and Molecular Imaging annual meeting abstracts* (SNMMI, 2023), accepted.

Annexes

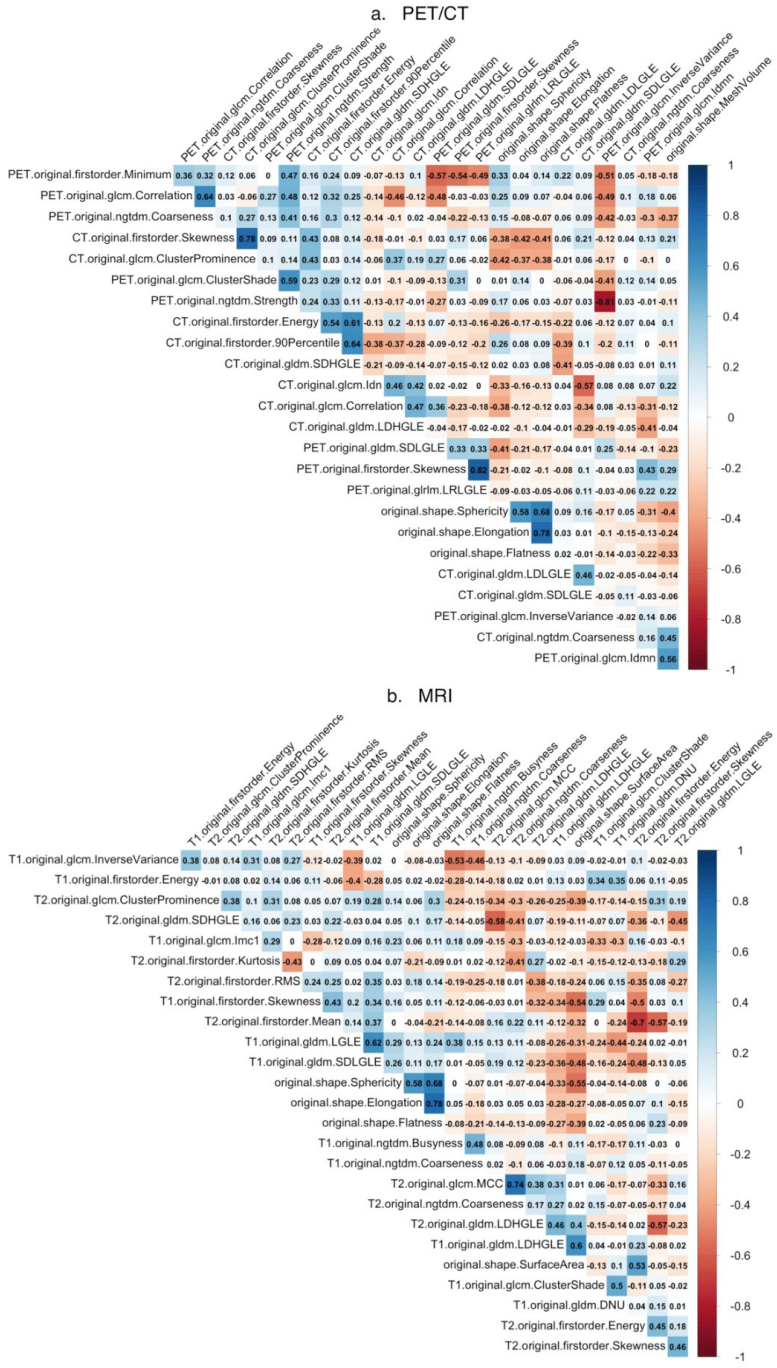
Annexe I : Tableau S1

VIF des caractéristiques sélectionnées en fonction de la multicollinéarité pour la TEP/TDM et l'IRM (le fond plus foncé correspond aux caractéristiques qui ont finalement été sélectionnées dans les modèles). Tableau provenant d'Escobar et al. [61].

Modality	Feature	VIF
PET/CT	PET-FIRST-ORDER-SKEWNESS	9.45
	CT-GLCM-CLUSTER-PROMINENCE	8.01
	PET-GLCM-INVERSE-VARIANCE	7.09
	CT-FIRST-ORDER-SKEWNESS	6.87
	PET-GLRLM-LRLGLE	6.85
	CT-GLCM-CORRELATION	6.60
	PET-GLCM-CORRELATION	6.60
	SHAPE-SHPERICITY	6.55
	CT-FIRST-ORDER-90-PERCENTILE	6.35
	PET-NGTDM-STRENGTH	6.33
	PET-NGTDM-COARSENESS	6.03
	CT-FIRST-ORDER-ENERGY	5.80
	PET-GLDM-SDLGLE	5.39
	PET-FIRST-ORDER-MINIMUM	5.16
	SHAPE-MESH-VOLUME	4.93
	SHAPE-MESH-FLATNESS	4.84
	SHAPE-MESH-ELONGATION	3.60
	PET-GLCM-CLUSTER-SHADE	3.48
	PET-GLCM-IDMN	3.38
	CT-GLCM-IDN	3.26
	CT-GLDM-LDLGLE	3.05
	CT-GLDM-SDHGLE	2.94
	CT-GLDM-SDLGLE	2.92
	CT-GLDM-LDHGLE	2.59
	CT-NGTDM-COARSENESS	2.15
MRI	fat-supressed-T2-GLCM-MCC	8.87
	SHAPE-FLATNESS	7.56
	fat-supressed-T2-FIRST-ORDER-SKEWNESS	7.29
	SHAPE-SURFACE-AREA	6.75
	fat-supressed-T2-FIRST-ORDER-ENERGY	6.29
	fat-supressed-T2-GLDM-SDHGLE	6.07
	fat-supressed-T2-GLDM-LDHGLE	5.78
	SHAPE-ELONGATION	5.78
	fat-supressed-T2-NGTDM-COARSENESS	5.48
	fat-supressed-T2-FIRST-ORDER-MEAN	5.44
	fat-supressed-T2-FIRST-ORDER-KURTOSIS	5.04
	T1-GLCM-INVERSE-VARIANCE	4.98
	T1-GLDM-LGLE	4.41
	T1-GLDM-SDLGLE	3.96
	fat-supressed-T2-GLCM-CLUSTER-PROMINENCE	3.73
	fat-supressed-T2-FIRST-ORDER-RMS	3.66
	fat-supressed-T2-GLDM-LGLE	3.65
	SHAPE-SPHERICITY	3.47
	T1-FIRST-ORDER-ENERGY	3.30
	T1-GLDM-DNU	3.27
	T1-GLDM-LDHGLE	3.25
	T1-FIRST-ORDER-SKEWNESS	3.17
	T1-GLCM-IMC1	2.78
	T1-NGTDM-BUSYNESS	2.76
	T1-NGTDM-COARSENESS	2.16
	T1-GLCM-CLUSTER-SHADE	2.13

Annexe II : Figure S1

Matrices de corrélation de Pearson pour les caractéristiques sélectionnées en fonction du VIF pour la TEP/TDM (a) et l'IRM (b). Figure provenant d'Escobar et al. [61].



Annexe III : Tableau S2

Caractéristiques pour la création du modèle de substitution. Tableau provenant d'Escobar et al. [61].

Feature	Definition	Modality
Anatomical tumor volume (ATV)	$Nv \times x \times y \times z$ *	CT
SUVmax	$\max(PET_{ROI})$	PET
Metabolic tumor volume (MTV)	$Nv_{>40\,\%\,SUVmax} \times x \times y \times z$	PET
Total lesion glycolysis (TLG)	$MTV \times \frac{1}{Nv_{>40\,\%\,SUVmax}} \times \sum_{Nv_{>40\,\%\,SUVmax}} PET_{ROI_{>40\,\%\,SUVmax}}$	PET
Non-metabolic volume (INACTIVE-FDG V)	$Nv_{<40\,\%\,SUVmax} \times x \times y \times z$	PET
Non-metabolic relative volume (INACTIVE-FDG rV)	$\frac{INACTIVE_{FDG}\,V}{ATV}$	PET/CT
Hypodense volume < 20 HU (HYPODENSE-20 HU V)	$Nv_{<20\,HU} \times x \times y \times z$	CT
Hypodense relative volume < 20 HU (HYPODENSE-20 HU rV)	$\frac{HYPODENSE_{20\,HU}\,V}{ATV}$	CT
Hypodense volume < 30 HU (HYPODENSE-30 HU V)	$Nv_{<30\,HU} \times x \times y \times z$	CT
Hypodense relative volume < 30 HU (HYPODENSE-30 HU rV)	$\frac{HYPODENSE_{30\,HU}\,V}{ATV}$	CT
Non-metabolic or hypodense < 20 HU volume (INACTIVE-FDG ∪ HYPODENSE-20 HU V)	$Nv_{<40\,\%\,SUVmax\,or\,<20\,HU} \times x \times y \times z$	PET/CT
Non-metabolic or hypodense < 20 HU relative volume (INACTIVE-FDG ∪ HYPODENSE-20 HU rV)	$\frac{INACTIVE_{FDG} \cup HYPODENSE_{20\,HU}\,V}{ATV}$	PET/CT
Non-metabolic or hypodense < 30 HU volume (INACTIVE-FDG ∪ HYPODENSE-30 HU V)	$Nv_{<40\,\%\,SUVmax\,or\,<30\,HU} \times x \times y \times z$	PET/CT
Non-metabolic or hypodense < 30 HU relative volume (INACTIVE-FDG ∪ HYPODENSE-30 HU rV)	$\frac{INACTIVE_{FDG} \cup HYPODENSE_{30\,HU}\,V}{ATV}$	PET/CT
Non-metabolic and hypodense < 20 HU volume (INACTIVE-FDG ∩ HYPODENSE-20 HU V)	$Nv_{<40\,\%\,SUVmax\,and\,<20\,HU} \times x \times y \times z$	PET/CT
Non-metabolic and hypodense < 20 HU relative volume (INACTIVE-FDG ∩ HYPODENSE-20 HU rV)	$\frac{INACTIVE_{FDG} \cap HYPODENSE_{20\,HU}\,V}{ATV}$	PET/CT
Non-metabolic and hypodense < 30 HU volume (INACTIVE-FDG ∩ HYPODENSE-30 HU V)	$Nv_{<40\,\%\,SUVmax\,and\,<30\,HU} \times x \times y \times z$	PET/CT
Non-metabolic and hypodense < 30 HU relative volume (INACTIVE-FDG ∩ HYPODENSE-30 HU rV)	$\frac{INACTIVE_{FDG} \cap HYPODENSE_{30\,HU}\,V}{ATV}$	PET/CT

*To allow for the computation of PET/CT-based features, we resampled all PET images on a common grid with their corresponding CT using nearest neighbor interpolation. Thus, x, y, and z are common to both modalities and represent the dimensions of the CT image. $Nv_{<condition>}$ is the number of voxels in the ROI that meet the condition noted as index.

Annexe IV : Logiciels et matériels utilisés

Nous présentons une vue d'ensemble des principaux logiciels et matériels utilisés au cours de cette thèse. Il convient de noter que les versions des logiciels ont été mises à jour à différents moments de la thèse. Bien que nous ne l'ayons pas spécifiquement documenté, les versions mentionnées ici sont les plus pertinentes et permettent de reproduire les résultats.

LOGICIELS

Python 3.9 : Langage de programmation majoritairement utilisé tout au long de la thèse.

R 4.0 : Langage de programmation statistique utilisé de façon sporadique lorsque certaines bibliothèques étaient pertinentes.

Visual Studio Code 1.60 : Environnement de développement pour l'utilisation du langage Python.

RStudio 1.1.456 : Environnement de développement pour l'utilisation du langage R.

LIFEx 6.49 : Logiciel de visualisation et de manipulation d'images médicales.

3D Slicer 5.2.2 : Logiciel de visualisation et de manipulation d'images médicales.

Simple ITK 2.1.1 : Bibliothèque Python encapsulant de nombreuses fonctions pour la manipulation, l'analyse, et le traitement d'images.

Pydicom 2.2.2 : Bibliothèque Python pour la manipulation de fichiers au format DICOM.

Nibabel 3.2.2 : Bibliothèque Python pour la manipulation d'images médicales au format NIFTI.

Nilearn 0.10 : Bibliothèque liée à Nibabel, encapsulant des fonctions supplémentaires de plus haut niveau.

Pyradiomics 2.2 : Bibliothèque Python pour l'extraction de caractéristiques radiomiques.

Numpy 1.23 : Bibliotèque Python de manipulation de matrices et tenseurs.

Scipy 1.7.3 : Bibliothèque Python pour le calcul scientifique.

Scikit-Learn 1.0 : Bibliothèque Python pour le ML classique.

Pytorch 1.11 : Bibliothèque Python pour le DL.

nnUNet 1.7.1 : Bibliothèque Python basée sur Pytorch et Simple ITK pour la segmentation sémantique supervisée d'images médicales.

Scickit-Survival 0.17.2 : Bibliothèque Python pour l'analyse de données de survie.

Matplotlib 3.3.1 : Bibliothèque Python pour la réalisation de figures scientifiques et mathématiques.

Seaborn 0.11 : Bibliothèque liée à Matplotlib encapsulant des fonctions de plus haut niveau.

Ray 2.5.1 : Bibliothèque Python pour la parallélisation.

Pandas 1.3.4 : Bibliothèque Python pour la gestion des données structurées sous forme de tableaux.

Orange Data Mining 3.32 : Logiciel de programation visuelle pour l'analyse exploratoire et la fouille de données.

Anaconda 1.10.0 : Gestionnaire d'environnements virtuels.

Nvidia CUDA 11.3 : Moteur de calcul sur carte graphique pour le DL.

MATERIELS

Station personnelle : Dell Precision Tower 7920, Linux Ubuntu 20.04.2, $128Go$ de mémoire RAM, 2×12 coeurs Intel Xeon Silver 4214 $64bits$, Nvidia Quadro RTX 5000 $16Go$.

Serveur de calcul partagé : Dell Precision Rack, Linux Ubuntu 20.04.2, 1To de mémoire RAM, 8×12 coeurs Intel Xeon Silver 4214 $64bits$, $3\times$ Nvidia A6000 $48Go$.