Développement et validation de techniques d'analyse d'imagerie radiologique de tumeurs solides

Benoît Magnin

UNIVERSITÉ CLERMONT AUVERGNE
ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE – AGRONOMIE - ENVIRONNEMENT

THÈSE

présentée et soutenue publiquement le 17 décembre 2024 pour l'obtention du titre de

DIPLÔME D'ÉTAT DE DOCTEUR D'UNIVERSITE

Discipline : Sciences de la vie et de la santé

Par

MAGNIN Benoît

DEVELOPPEMENT ET VALIDATION DE TECHNIQUES D'ANALYSE D'IMAGERIE RADIOLOGIQUE DE TUMEURS SOLIDES

Président	M AUBE Christophe	PU PH, Université d'Angers
Rapportrice	Mme WAGNER Mathilde	PU PH, Paris Sorbonne Université
Rapporteur	M TASU Jean-Pierre	PU PH, Université de Poitiers
Membre invité	M BARTOLI Adrien	Professeur, Université Clermont Auvergne
Membre invité	M LUCIANI Alain	PU PH, Université Paris Est Créteil
Membre invité	M PEZET Denis	PU PH, Université Clermont Auvergne
Directrice de thèse	Mme CASSAGNES Lucie	PU PH, Université Clermont Auvergne
Co-directeur de thèse	M CHABROT Pascal	PU PH, Université Clermont Auvergne

Laboratoire Institut Pascal - Axe Thérapie Guidée par l'Image
UMR 6602 CNRS - Université Clermont Auvergne - Clermont-Ferrand - France

Remerciements

A Monsieur le Professeur Christophe Aubé,
Votre compétence dans notre spécialité et votre engagement pour son développement sont exemplaires. Je vous remercie de m'avoir fait l'honneur de présider ce jury.

A Madame la Professeur Mathilde Wagner,
Ton investissement et tes compétences de jeune radiologue m'étaient déjà apparues admirables lorsque j'étais ton interne, la suite de ta carrière a confirmé ces impressions. Je te remercie de m'avoir fait l'honneur d'être rapportrice de ce travail.

A Monsieur le Professeur Jean-Pierre Tasu,
C'est sous votre présidence que j'ai eu mes premiers contacts avec la SIAD, me permettant de bénéficier de vos expériences et compétences. Je vous remercie de m'avoir fait l'honneur d'être rapporteur de ce travail.

A Monsieur le Professeur Adrien Bartoli,
C'est avec grand plaisir que je reçois la qualité scientifique de nos échanges, en essayant d'y être à la hauteur. Je te remercie de m'avoir fait l'honneur d'évaluer ce travail.

A Monsieur le Professeur Alain Luciani,
J'ai eu la chance de bénéficier de tes enseignements dès mon internat, j'essaie depuis de tendre vers ce niveau de compétence que tu représentes. Je te remercie sincèrement pour l'honneur de ta présence dans mon jury de thèse.

A Monsieur le Professeur Denis Pezet,
J'ai pu voir au quotidien avec quelles compétences et quel investissement vous preniez soin des patients de notre région. Je vous remercie pour l'honneur de votre présence dans mon jury de thèse.

A Madame la Professeur Lucie Cassagnes et à Monsieur le Professeur Pascal Chabrot,
Je vous remercie pour votre encadrement, vos conseils et votre compréhension dans ce travail. Je souhaite qu'il ne soit qu'une étape parmi d'autres dans la poursuite du développement de la radiologie hospitalo-universitaire de notre ville.

A Monsieur le Professeur Louis Boyer et Monsieur le Professeur Jean-Marc Garcier,
Vous m'avez soutenu, conseillé et aidé pour mener à bien mon projet professionnel tout en étant à l'écoute des aspects personnels depuis mon retour à Clermont-Ferrand. La réussite de cette thèse, qui est une des étapes de ce projet professionnel, vous revient.

Table des matières

Summary

Development and validation of radiologic image analysis in oncology

Our objective was to validate new image reconstruction techniques, develop and validate image analysis methods through radiomics in oncological imaging, and investigate the influence of these new image reconstruction techniques on the stability of radiomics.

Our first study focused on the impact of Deep Learning Image Reconstructions (DLIR) on the detection of liver metastases in CT scans. We reconstructed CT images of 121 patients with liver metastases using iterative reconstruction (50%-ASiR-V) and three levels of DLIR. Two double-blinded radiologists independently counted up to ten metastases for each reconstruction. One reader detected a higher number of metastases with DLIR-high: a median of 7 (range 2-10) compared to 5 (range 2-10) for DLIR-medium, DLIR-low, and ASiR-V ($p<.001$). Notably, ten patients were identified with more metastases using DLIR-high. Our results indicate that DLIR-high enhances the detection and visibility of liver metastases compared to ASiR-V and increases the number of detected liver metastases.

Our second study sought to determine whether radiomic analysis of paravertebral muscles from CT scans could aid in predicting survival in locally advanced head and neck neoplasms. We included 71 patients treated for locally advanced head and neck cancer. Radiomic features were extracted from manually segmented paravertebral muscles at the L1 level, retaining 21 features for analysis. No parameter significantly predicted survival or treatment-related toxicity in multivariate analysis. However, two radiomic features, the sum of HU densities for survival and the standard deviation of HU values for toxicity, appear to be promising.

Our third study aimed to develop a prognostic tool for patients with metastatic melanoma (MM) undergoing immunotherapy by creating radiomic models based on pretreatment CT scans. We analyzed 503 metastatic lesions in 71 patients, extracting 46 radiomic features after lesion segmentation. Predictive accuracies for overall survival (OS) and treatment response were compared across five feature selection methods and four classifiers. A fivefold cross-validation was conducted at the patient level using tumor-based classification. The highest accuracy for OS predictions was AUC=0.91, while the highest accuracy for treatment response predictions was AUC=0.88. Our findings suggest that combining pretreatment CT radiomic features from a single tumor with data selection and classifiers may accurately predict OS and treatment response in MM patients treated with immunotherapy.

Our fourth study investigated the influence of MRI DLIR on the repeatability and reproducibility of radiomic features. MRI acquisition was performed using seven common sequences across four distinct acquisition protocols on a phantom composed of 12 fruits. Each acquisition was repeated four times, resulting in 16 images. Each image was reconstructed using both a conventional non DLIR and DLIR. After segmenting the fruits, we extracted 107 radiomic features. Reproducibility was assessed for each feature using the ICC(3,1) on the 16 images, considering a feature reproducible if ICC > 0.9. A total of 399 features (53.3%) were reproducible with DLIR, compared to 326 (43.5%) with non DLIR. Thus, DLIR increased the number of reproducible radiomic features by 9.7%. This study underscores the validity of using MRI DLIR over traditional reconstruction methods in MRI radiomic investigations.

Keywords: Medical imaging; image analysis; radiomics; CT scanner; MRI; oncology

Résumé

Développement et validation de techniques d'analyse d'imagerie radiologique de tumeurs solides

Notre objectif était de valider de nouvelles techniques de reconstruction d'images, de développer et valider des techniques d'analyse d'images par radiomique en imagerie oncologique, et enfin d'étudier l'influence de ces nouvelles techniques sur la stabilité de la radiomique.

Notre première étude portait sur l'impact des reconstructions d'images par Deep Learning (DLIR) sur la détection des métastases hépatiques en scanner. 121 scanners de patients avec métastases hépatiques ont été reconstruits par reconstruction itérative (50%-ASiR-V) et trois niveaux de DLIR. Pour chaque reconstruction, deux radiologues ont recensé en double aveugle le nombre de métastases. Un plus grand nombre de métastases a été détecté par un lecteur avec DLIR-high : 7 (2-10) (médiane (Q_1-Q_3)) contre 5 (2-10) pour DLIR-medium, DLIR-low, et ASiR-V ($p<.001$). Plus de métastases ont été détectées avec DLIR-high chez 10 patients. Nos résultats montrent ainsi que DLIR améliore la détection et la visibilité des métastases hépatiques par rapport à ASiR-V.

Notre deuxième étude a cherché à déterminer si l'analyse par radiomique des muscles paravertébraux au scanner peut prédire la survie dans les néoplasies localement avancées de la tête et du cou. 71 patients traités pour un cancer avancé laryngé ont été inclus. Vingt et un paramètres de radiomique ont été extraits après segmentation manuelle des muscles paravertébraux au niveau L1. Aucun paramètre ne prédisait de façon significative en analyse multivariée la survie ou la survenue d'une toxicité du traitement. Néanmoins, deux paramètres de radiomique, la somme des densités HU pour la survie ou l'écart type des HU pour la toxicité semble prometteurs.

Notre troisième étude visait à développer un outil pronostique chez les patients atteints de mélanome métastatique traités par immunothérapie, en créant des modèles de radiomique d'analyse du scanner pré-traitement. Nous avons analysé 503 lésions métastatiques chez 71 patients ~~en~~ en extrayant 46 paramètres de radiomique après segmentation des lésions. La prédiction de la survie globale et de la réponse au traitement ont été comparées pour cinq méthodes de sélection de paramètres et quatre classificateurs, en utilisant une validation croisée à cinq plis. La meilleure prédiction de survie avait pour AUC=0.91; pour la réponse, AUC=0.88. Notre étude montre que l'utilisation de paramètres de radiomique du TDM pré traitement d'une seule tumeur via des sélections de données et des classificateurs peut prédire la survie et la réponse sous immunothérapie du mélanome métastatique.

Notre quatrième étude visait à déterminer l'influence des DLIR en IRM sur la répétabilité et la reproductibilité des paramètres de radiomique. Des acquisitions IRM ont été réalisées avec 7 séquences et 4 protocoles d'acquisition distincts sur un fantôme composé de 12 fruits. Chaque acquisition a été répétée 4 fois, aboutissant à 16 images. Chaque image a été reconstruite par la transformation inverse de Fourier (TIF) habituelle et par DLIR. Après segmentation des fruits, 107 paramètres de radiomique ont été extraits. La reproductibilité a été évaluée pour chaque paramètre en utilisant l'ICC(3,1) sur les 16 images; un paramètre était reproductible si ICC > 0.9. Au total, 399 paramètres (53,3%) étaient reproductibles avec DLIR et 326 (43,5%) avec la TIF. Ainsi, DLIR a augmenté le nombre de paramètres de radiomique reproductibles de 9,7%. Cette étude valide l'utilisation de DLIR en IRM plutôt que la TIF traditionnelle dans les études de radiomique.

Mots clés : Imagerie médicale ; analyse d'images ; radiomique ; scanner ; IRM ; oncologie

Abréviations

CNN : Convolutional Neural Networks (réseaux de neurones convolutifs)

DL : deep learning (apprentissage profonde)

DLIR : deep learning image reconstruction (reconstruction de l'image par apprentissage profond)

FBP : filtered back projection (rétroprojection filtrée)

GAN : generative adversarial networks (réseaux de neurones antagonistes génératifs)

IBSI : Image Biomarker Standardisation Initiative

Introduction

L'imagerie a une place importante et croissante dans le parcours de soins du patient, notamment en oncologie. L'imagerie prend une part de plus en plus grande dans le dépistage, depuis de nombreuses années dans le cancer du sein ou pour le carcinome hépato cellulaire chez les patients cirrhotique, et façon plus récente pour le cancer du poumon (1). L'imagerie est indispensable dans le bilan d'extension, étant un élément important dans les classifications et stadification dont découlent d'une part les possibilités thérapeutiques et d'autre part le pronostic des patients. L'imagerie est ensuite un élément clé dans l'évaluation sous traitement (2,3). On ne détaillera pas ici plus longuement l'autre versant de la radiologie, la radiologie interventionnelle, présente depuis le diagnostic mini invasif jusqu'aux possibilités de traitement curatif et de soins de support.

De plus, il existe des modifications technologiques de notre spécialité qui améliorent le service rendu au patient en oncologie. Les outils fournis au radiologue sont plus nombreux, ouvrant des perspectives d'amélioration de la prise en charge : imageries plus rapides, de meilleure qualité, logiciels signalant les anomalies, proposant un pré compte rendu…

Le radiologue reste cependant responsable de la décision qui découle de l'analyse de ces images quelque soit l'outil utilisé. Il nous incombe ainsi de comprendre et maîtriser quand cela est possible les outils technologiques qui nous sont proposés et à défaut de vérifier de façon scientifique leur validité. De nouveaux outils peuvent aussi être développés ou codéveloppés par les radiologues, utilisant à la fois leurs capacités de cliniciens attentifs au patient, et leurs compétences scientifiques pour produire un outil améliorant les soins.

Ce travail s'inscrit dans cette démarche, en présentant tout d'abord un travail de validation d'une technique récente : s'assurer qu'une reconstruction par apprentissage profond des images de scanner soit pertinente pour la détection des métastases hépatiques. Nous présenterons ensuite deux travaux de développement de modèles basés sur la radiomique, l'un sur les muscles paravertébraux dans les néoplasies ORL, l'autre dans le mélanome métastatique. Nous présenterons enfin un travail plus fondamental qui vise à valider la méthodologie de la radiomique : il s'agit d'étudier l'impact des reconstructions par apprentissage profond sur la stabilité des paramètres de radiomique en IRM. Ces travaux seront précédés de rappels des connaissances du domaine et de l'état de l'art de la recherche.

L’apprentissage profond

Présentation de l’apprentissage profond

L’apprentissage profond (deep learning) est une catégorie d’algorithmes d'apprentissage automatique caractérisée par l’utilisation de réseaux neuronaux à de nombreuses couches.

L’apprentissage profond est une forme d'apprentissage de représentation dans laquelle l'algorithme apprend une composition de caractéristiques qui reflète une hiérarchie de structures présentes dans les données. Des représentations complexes sont ainsi exprimées en fonction de représentations plus simples. Ces systèmes d’apprentissage profond adoptent une approche progressive en apprenant d'abord des caractéristiques simples (telles que l'intensité du signal, les contours et les textures), qui servent ensuite de base pour des caractéristiques plus complexes, comme les formes, les lésions ou les organes. Cela permet de tirer parti de la nature compositionnelle des images.

Les systèmes d'apprentissage profond codent les caractéristiques en utilisant une architecture de réseaux de neurones artificiels, une approche qui repose sur des nœuds connectés, tels des neurones interconnectés.

Structure des réseaux de neurones artificiels

La structure des réseaux de neurones artificiels s'inspire en effet de celle des réseaux neuronaux biologiques (Figure 1).

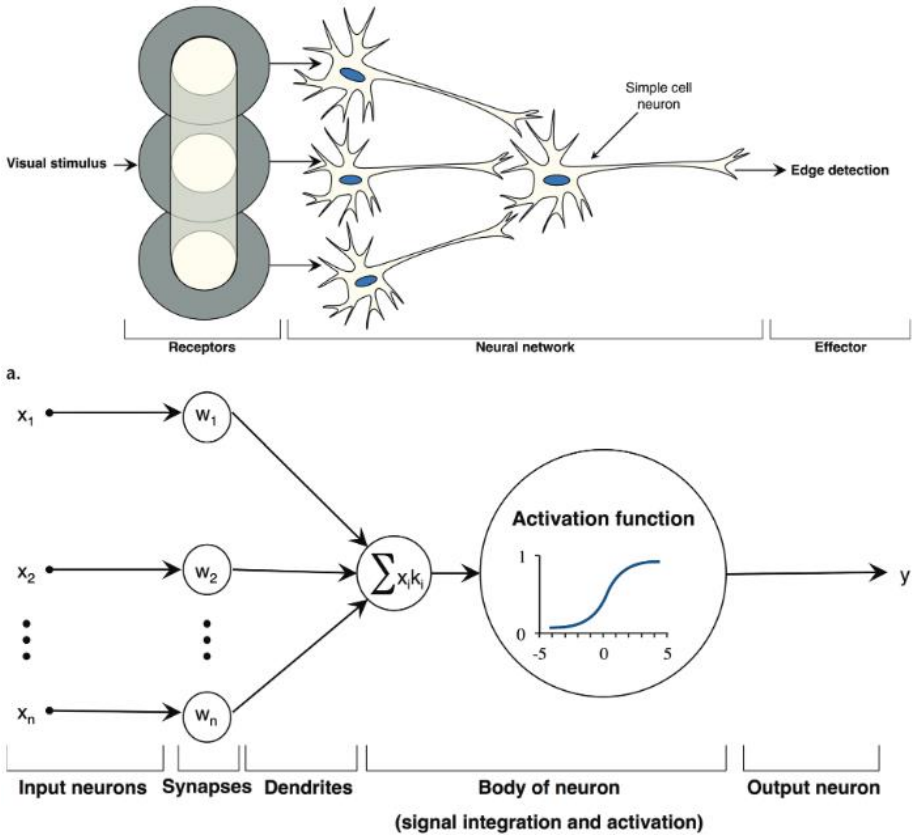


Figure 1 : Analogie entre les réseaux de neurones biologiques (a) et les réseaux de neurones artificiels (b). a : un stimulus visuel active les récepteurs avec une transmission par 3 neurones de l'information. Le neurone suivant intègre les 3 signaux et transmet un signal (par exemple la détection d'un bord). b : un réseau de neurones artificiels est composé de neurones artificiels interconnectés. Chaque neurone artificiel est un simple modèle de classification : il reçoit en entrée la somme des

signaux des neurones d'entrée, qui est évalué par la fonction d'activation donnant ainsi un signal de sortie de décision. Figure de Chartrand et coll. (4)

L'unité de base d'un réseau de neurones artificiel, le neurone artificiel ou nœud, est un modèle simplifié qui imite le mécanisme de base des neurones biologiques. Le neurone artificiel prend en entrée un ensemble de valeurs représentant des caractéristiques, chacune multipliée par un poids correspondant. Ces caractéristiques pondérées sont ensuite additionnées et transmises à une fonction d'activation non linéaire. Ainsi, le neurone artificiel peut être vu comme produisant une décision en pondérant un ensemble de preuves. Bien qu'un neurone artificiel soit relativement simple, des architectures de réseaux neuronaux appelées perceptrons multicouches, composées de milliers de neurones, peuvent représenter des fonctions non linéaires très complexes.

Ces perceptrons multicouches sont généralement construits en assemblant plusieurs neurones pour former une couche, et en empilant ces couches de manière à connecter la sortie d'une couche à l'entrée de la suivante. Le terme "profond" dans l'apprentissage profond fait référence à l'architecture multicouche des perceptrons multicouches. La première couche, appelée couche d'entrée, représente les données d'entrée telles que les intensités individuelles des pixels, tandis que la couche de sortie produit des valeurs cibles, comme un résultat de classification. Les couches intermédiaires de perceptrons multicouches sont appelées couches cachées, car elles ne produisent pas directement les sorties visibles souhaitées, mais calculent plutôt des représentations intermédiaires des caractéristiques d'entrée, utiles pour le processus d'inférence (Figure 2).

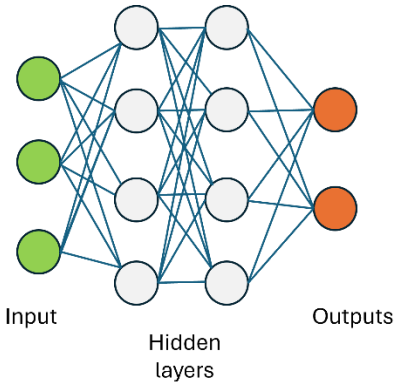


Figure 2 : Structure simplifiée d'un réseau de neurones avec 2 couches cachées

En empilant plusieurs couches, un réseau peut représenter une hiérarchie de caractéristiques qui forment une composition de plus en plus complexe des caractéristiques de bas niveau issues des données d'entrée, modélisant ainsi des niveaux d'abstraction plus élevés. Cette capacité de composition des architectures profondes permet aux réseaux neuronaux de prendre des décisions fondées sur des concepts abstraits.

Obtenir une prédiction à partir d'une observation (par exemple, une image) avec un réseau neuronal consiste à calculer, de manière séquentielle, l'activation de chaque nœud de chaque couche, en commençant par la couche d'entrée et en allant jusqu'à la couche de sortie, un processus appelé propagation avant (*forward propagation*).

Un réseau neuronal est entraîné en ajustant ses paramètres, constitués des poids et des biais de chaque nœud. Les réseaux neuronaux modernes contiennent des millions de paramètres. Partant d'une configuration initiale aléatoire, les paramètres sont ajustés par un algorithme d'optimisation appelé descente de gradient (*gradient descent*), qui cherche à trouver un ensemble de paramètres performant sur un ensemble d'apprentissage. Chaque fois qu'une prédiction est calculée à partir d'un échantillon (propagation avant), la performance du réseau est évaluée à l'aide d'une fonction de perte

(*loss function*) qui mesure quantitativement l'inexactitude de la prédiction. Chaque paramètre est ensuite ajusté par de petites augmentations dans la direction qui minimise cette perte, un processus appelé rétropropagation (*back-propagation*).

Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (Convolutional Neural Networks, CNN) sont des réseaux de neurones présentant une architecture particulière qui connaissent un succès d'utilisation croissant. Ils exploitent des caractéristiques locales pour traiter efficacement des entrées plus larges et variables, ce qui les rend plus adaptés que les perceptrons multicouches pour l'analyse d'images. Contrairement aux perceptrons multicouches, les CNN sont capables de gérer la variabilité de forme, d'orientation et de position des objets en appliquant des détecteurs de caractéristiques sur chaque partie de l'image grâce aux convolutions. Chaque détecteur de caractéristiques est limité aux zones locales, ce qui convient aux images naturelles, permettant ainsi de traiter des variations dans les objets sans compromettre la performance globale. Les CNN profonds (ou *deep CNNs*) tirent parti de la structure compositionnelle des images naturelles, ce qui rend les décalages et déformations d'objets dans les images peu pénalisants pour les performances globales du réseau. Ces réseaux sont conçus pour traiter des tâches complexes telles que la classification d'images en s'appuyant sur une architecture de modèle efficace (Figure 3) se basant sur:

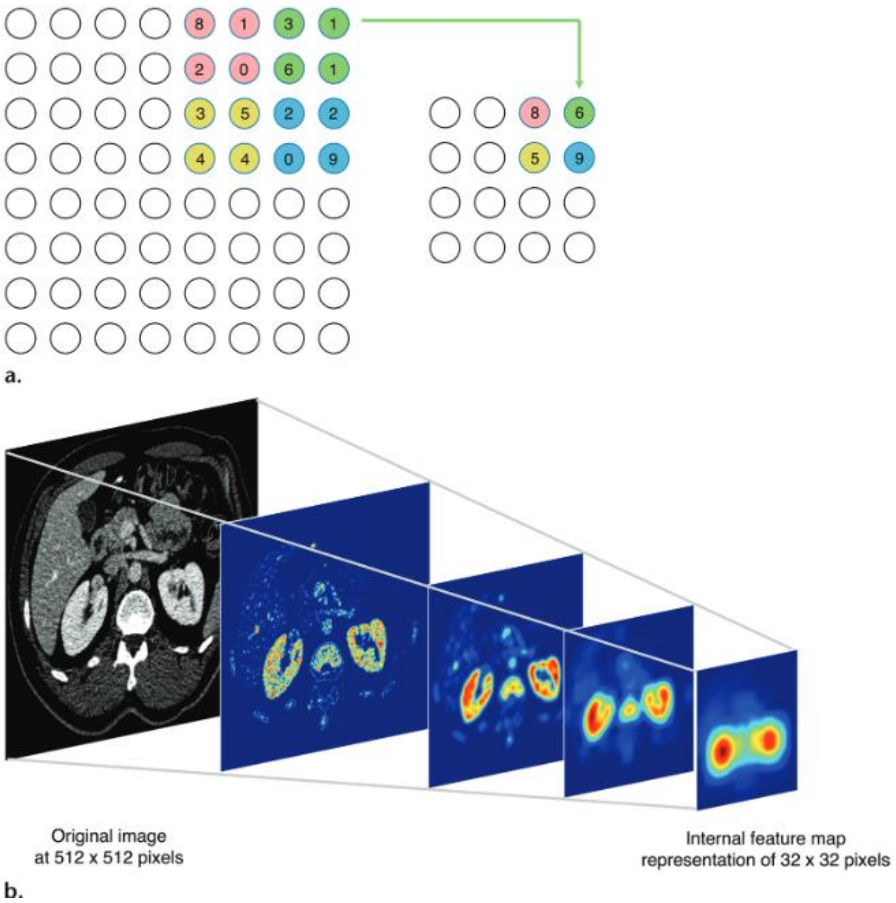


Figure 3 : Un réseau de neurones convolutifs (CNN) crée une représentation interne hiérarchique de caractéristiques visuelles en empilant des couches de convolution. Afin de capturer un champ de vision de plus en plus large, les cartes de caractéristiques sont réduites spatialement par sous-échantillonnage progressif des images. (a) La couche de « max pooling », généralement utilisée pour effectuer ce sous-échantillonnage, ne transmet que l'activation maximale à la couche

suivante. Les couches de convolution suivantes deviennent ainsi moins sensibles aux petits décalages ou déformations de l'objet dans les cartes de caractéristiques extraites.

(b) Représentations sous-échantillonnées des reins à partir d'un scanner. Cette opération réduit non seulement de manière significative les besoins en mémoire, mais permet également au réseau de rester robuste face aux variations de forme et de position des reins détectés qui sont les caractéristiques d'intérêt dans les images. Figure de Chartrand et coll. (4)

- Une couche d'entrée qui introduit les données brutes (pixels d'une image)

- Une couche de convolution qui analyse les images en entrée, détecte les motifs spécifiques dans l'image, extrait les caractéristiques de données d'entrée et forme une cartographie de la caractéristique (feature map) recherchés. La convolution est une méthode s'apparentant à un filtre de petite taille se déplaçant sur l'image réalisant des fonctions diverses (filtre dérivateur filtre moyenne, filtre gaussien), ce qui permet d'extraire des caractéristiques propres à chaque image.

- Une couche de « max pooling » qui reçoit les cartographies et réduit la taille des images en préservant les caractéristiques essentielles. Ces caractéristiques sont sous-échantillonnées, en réduisant la dimension des données extraites, ce qui diminue le nombre de paramètres à apprendre et fournit une invariance aux petites translations

- Une couche d'activation qui complexifie le modèle

- Une couche entièrement connectée (fully connected) qui stabilise les caractéristiques extraites par interconnexion neuronale au sein d'une même couche

- Une couche de sortie qui retranscrit les résultats.

Ces quelques détails sur les réseaux de neurones convolutifs ne sont qu'un aperçu de la complexité possible des architectures des réseaux de neurones. Certaines architectures de réseaux de neurones convolutifs connus et ayant démontré leur efficacité pour une tâche donnée sont décrites (VGG-16, ResNet-50, U-Net…) et pour certains disponibles librement.

Entrainement des réseaux de neurones

L'entrainement des réseaux de neurones est nécessaire pour fixer la pondération des différentes couches et les paramètres des fonctions d'activation. Il est classiquement réalisée de façon supervisée : on dispose d'une base de données souvent de très grande taille comportant des données d'entrée (des images de scanner par exemple) ainsi que une annotation (ou Ground Truth) correspondant par exemple au type de tumeur que l'on veut classifier. L'apprentissage implique souvent de séparer cette population d'apprentissage en 3 parties : apprentissage, validation et test. La population d'apprentissage est passée (propagation avant) de façon itérative dans le réseau de neurones en utilisant l'erreur prédite sur cette population (différence entre la prédiction faite par l'état actuel du réseau en comparaison à l'annotation, quantifiée par la fonction de perte) pour ajuster les paramètres de connexion du réseau (rétropropagation). La population de validation est utilisée pour surveiller les performances du modèle pendant cet apprentissage. Une fois que les paramètres du réseau de neurones sont fixés, on peut évaluer sa performance sur la population de test.

Il existe aussi un apprentissage dit non supervisé où les données de la base d'apprentissage ne sont pas annotées. L'algorithme essaie ainsi de regrouper les données en des sous-groupes selon une cohérence interne qu'il découvre.

Applications de l’apprentissage profond

Il existe une utilisation croissante des algorithmes d'apprentissage profond dans notre vie courante : dans les systèmes d'automatisation des voitures, dans l'utilisation de nos téléphones connectées ...

Il existe de façon parallèle une augmentation de la place des algorithmes avec apprentissage profond en santé, notamment en imagerie, au moins dans les activités de recherche. Un des champs d’application évident est l’analyse des images. Au sein de l’analyse d’image on peut distinguer des tâches de segmentation d’images médicales (5–7), des taches de détection (8–10), des taches de pronostic (11,12), d’aide à l’interprétation (8,13) . Il existe aussi des champs d’application sur la rédaction, la standardisation ou l’analyse des comptes rendus (14). Il existe enfin un champ d’application sur la création des images d’IRM et de scanner que nous allons détailler.

Reconstructions des images en radiologie

Reconstruction des images d'IRM

Technique classique de reconstruction des images d'IRM

Le signal capté en IRM est stocké dans une matrice dans l'espace k (ou espace de Fourier). Une reconstruction est nécessaire pour le transformer dans le domaine image, c'est-à-dire en une image anatomique visuellement interprétable. La méthode traditionnelle de reconstruction est basée sur la transformée inverse de Fourier.

Reconstruction des images d'IRM par apprentissage profond

Des constructeurs ont développé des réseaux de neurones pour la reconstruction des images d'IRM pour remplacer ou compléter la transformée inverse de Fourier. Ces réseaux de neurones peuvent être construits de façon différente (Figure 4) (15):

- Dans le domaine de l'image, c'est-à-dire qu'il prend en entrée une image (dans le domaine image c'est-à-dire une image anatomique préalablement reconstruite par transformée inverse de Fourier) et qu'il donne en sortie une image de meilleure qualité. Il s'agit ainsi d'une étape de post traitement.
- En correspondance directe (direct mapping) prenant en entrée la matrice dans l'espace k, et donnant en sortie une image dans le domaine image, remplaçant totalement la transformée inverse de Fourier.
- De façon plus anecdotique, dans l'espace k, en prenant en entrée la matrice dans l'espace k et donnant en sortie une autre matrice de l'espace k.

Les deux premières solutions peuvent être utilisées de façon non exclusive.

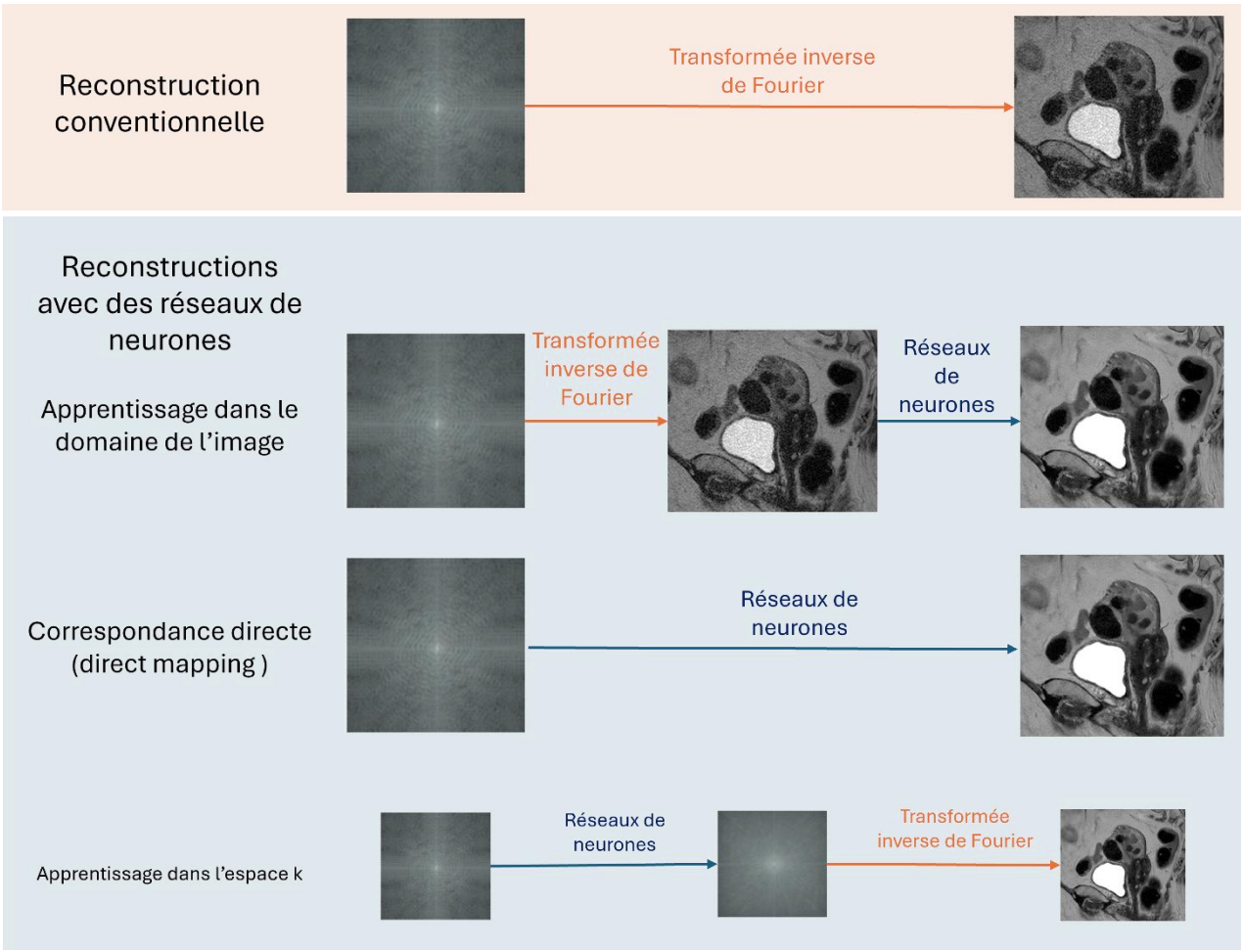


Figure 4 : Comparaison schématique des différentes reconstructions en IRM. (inspiré de (16))

Nous détaillons l'implémentation commerciale de GE Healthcare car elle a servi à la réalisation de notre 4e travail (p 99). Il s'agit d'une implémentation de type correspondance directe, c'est-à-dire qu'elle prend la matrice dans l'espace k pour fournir une image anatomique. L'architecture du réseau n'est pas dévoilée, mais quelques détails sont disponibles (17). Il s'agit d'un réseau de neurones convolutifs contenant 4,4 millions de paramètres à entrainer. La base de données d'entrainement n'est pas publique, mais on peut en appendre le principe. Il s'agit d'un apprentissage supervisé mettant correspondance des images d'IRM de très haute qualité et des données artificiellement altérées avec davantage d'artefacts et de bruit, combinées avec des techniques de multiplication d'images (augmentation, transformations géométrique) pour obtenir 4 millions de combinaisons d'images (17).

Avantage et limites des reconstructions par apprentissage profond en IRM

L'amélioration de la qualité d'image par la diminution du bruit est le premier avantage des reconstructions IRM par apprentissage profond (18,19). Une liste plus détaillée des avantages par organes est disponible (16).

Un autre avantage possible est la diminution du temps d'acquisition des IRM à qualité égale. Le principe est de ne remplir que partiellement l'espace k, ce qui nécessite moins de temps d'acquisition, en apprenant à un réseau de neurones à reconstruire l'image malgré le manque de

données dans l'espace k. Une étude rétrospective sur des IRM de glioblastome a mis en évidence la possibilité de diviser par 10 le temps d'acquisition avec cette approche (20).

Les inconvénients des reconstructions par apprentissage profond décrits dans la littérature sont en lien avec les approches de post traitement (méthodes dans le domaine de l'image) liés au démasquage d'artefacts ou l'hétérogénéité du bruit (16).

Reconstruction des images de scanner

Techniques classiques de reconstruction des images de scanner

Pour former une image au scanner, un faisceau de rayons X se dirige depuis le tube vers les détecteurs situés de l'autre côté de l'anneau du scanner. Ce faisceau traverse successivement des tissus différents tels que la peau, les muscles les organes et les os. On obtient ainsi une image de projection, qui est répétée lors de la rotation du tube du scanner. Il faut ainsi réaliser une transformation, appelée reconstruction pour passer de ces images de projection, appelées sinogrammes, à des images volumiques correspondant à l'anatomie normale.

La reconstruction historique est la rétroprojection filtrée, supplantée à partir de 2009 par les méthodes de reconstruction itérative (21), décrite à la Figure 5.

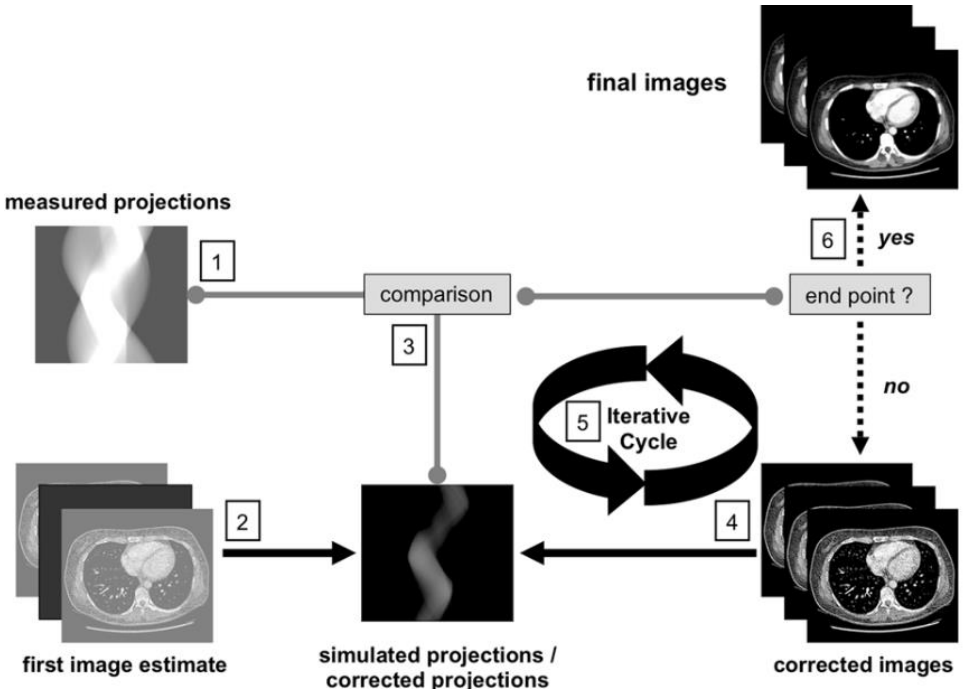


Figure 5 : Principe de la reconstruction d'images scanner par reconstruction itérative. A partir du sinogramme (1), le système crée une estimation initiale de l'image (2). Pour affiner cette estimation, le système simule le passage d'un faisceau de rayons X à travers cette image estimée. Les données de projection simulées sont ensuite comparées aux projections mesurées (3). Si des écarts existent entre ces deux ensembles de données, l'image est ajustée de manière itérative (4 et 5). Chaque mise à jour corrige ces différences, dans le but de réduire toute incohérence entre les projections mesurées et simulées. Ce processus de correction continue — avec des ajustements de l'image et des données de projection — jusqu'à ce que l'algorithme atteigne un certain niveau de précision ou remplisse une condition d'arrêt prédéfinie. À ce stade, l'image finale

est générée (6), offrant une représentation plus fidèle de l'objet scanné basée sur ces ajustements successifs. Figure de Geyer et coll. (21)

Reconstruction des images de scanner par apprentissage profond

De nouvelles méthodes de reconstruction sont apparues à la fin des années 2010, utilisant des réseaux de neurones. Les constructeurs ont développé des réseaux de neurones (dont le détail de l'architecture relève du secret industriel) qui prennent en entrée le sinogramme du patient et donne en sortie une image volumique de scanner. Ces réseaux ont été entrainés sur une quantité importante de scanner en donnant en entrée un sinogramme et en sortie comme « ground truth » un scanner reconstruit de façon optimale par rétroprojection filtrée. Une fois le réseau de neurones entrainé, il peut être appliqué pour la reconstruction en routine clinique quotidienne (Figure 6).

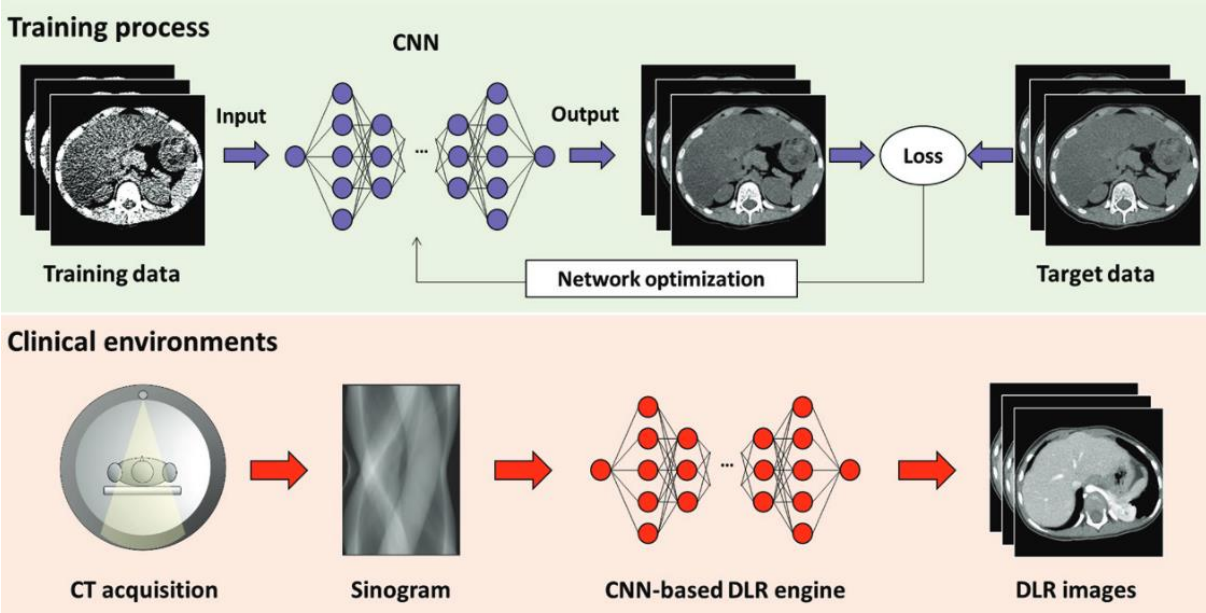


Figure 6 : Principe de la reconstruction par réseaux de neurones. En haut, phase d'entrainement du réseau pour optimisation de ses paramètres à partir de données reconstruites en rétroprojection filtrée. En bas, application clinique en utilisant le réseau de neurones pour obtenir les images de scanner à partir du sinogramme. Figure de Nagayama et coll. (22)

Ces méthodes de reconstruction dite par Deep Learning sont désormais d'utilisation quotidienne en radiologie.

Avantage et limites des reconstructions par apprentissage profond au scanner

Il est apparu assez rapidement que les reconstructions par apprentissage profond permettent une diminution importante du bruit et donc une augmentation du rapport signal sur bruit ainsi qu'une amélioration de la qualité subjective de l'image (23,24). Cette reconstruction a permis aussi de diminuer la dose délivrée tout en maintenant une qualité diagnostique comparable ou en l'améliorant (25–27). Il a été aussi démontré sur fantôme que ces reconstructions permettent une bonne détection des petites structures ou des structures avec de faibles contrastes (28,29).

Une des limites possibles de ce type de reconstruction tient à la façon dont elle a été créée. Le résultat des reconstructions découle d'une part de l'architecture du réseau construit par chaque constructeur et d'autre part de la base de données d'apprentissage sur laquelle le réseau a été entrainé. Cette base de données n'est pas publique, elle comporte d'après les constructeurs « différents types de morphologies et d'anatomies, conditions de scanner et d'indications cliniques » (30). Il est donc licite de se poser la question si la base d'entrainement est assez adaptée pour réaliser une reconstruction

de qualité dans des situations clinques qui sont possiblement peu représentée dans la base d'entrainement selon l'origine ethnique, l'âge, gabarit et spécificités de la pathologie étudiée. En d'autres termes, il n'est pas évident par exemple que la performance de la reconstruction dans le scanner cervical d'une lésion tumorale d'un nouveau-né soit aussi bonne que dans le scanner thoracique normal de l'adulte.

Ainsi, il nous a paru légitime de réaliser une étude de validation de la reconstruction par apprentissage profond quand elle a été disponible sur le scanner de notre service :
- d'une part de façon systématique comme devant toute innovation proposée par un constructeur pour s'assurer du bénéfice qu'elle peut amener à la prise en charge de nos patients, afin de rester maître des outils que nous utilisons,
- d'autre part plus spécifiquement devant le risque lié à un entrainement incomplet du réseau de neurones.

Cette étude de validation sur la détection des métastases hépatiques au scanner est l'objet de la première étude de ce travail (p 19)

Première étude : Impact des reconstructions par apprentissage profond sur la détection des métastases hépatiques au scanner

Kanan *et al. Insights into Imaging* (2024)15:167
https://doi.org/10.1186/s13244-024-01753-1

EUROPEAN SOCIETY OF RADIOLOGY
Insights into Imaging

ORIGINAL ARTICLE Open Access

Deep learning CT reconstruction improves liver metastases detection

Achraf Kanan[1], Bruno Pereira[2], Constance Hordonneau[1], Lucie Cassagnes[3,4], Eléonore Pouget[1], Léon Appolinaire Tianhoun[1,5], Benoît Chauveau[1] and Benoît Magnin[1,3,6*]

Abstract

Objectives Detection of liver metastases is crucial for guiding oncological management. Computed tomography through iterative reconstructions is widely used in this indication but has certain limitations. Deep learning image reconstructions (DLIR) use deep neural networks to achieve a significant noise reduction compared to iterative reconstructions. While reports have demonstrated improvements in image quality, their impact on liver metastases detection remains unclear. Our main objective was to determine whether DLIR affects the number of detected liver metastasis. Our secondary objective was to compare metastases conspicuity between the two reconstruction methods.

Methods CT images of 121 patients with liver metastases were reconstructed using a 50% adaptive statistical iterative reconstruction (50%-ASiR-V), and three levels of DLIR (DLIR-low, DLIR-medium, and DLIR-high). For each reconstruction, two double-blinded radiologists counted up to a maximum of ten metastases. Visibility and contour definitions were also assessed. Comparisons between methods for continuous parameters were performed using mixed models.

Results A higher number of metastases was detected by one reader with DLIR-high: 7 (2–10) (median (Q_1–Q_3); total 733) versus 5 (2–10), respectively for DLIR-medium, DLIR-low, and ASiR-V ($p < 0.001$). Ten patents were detected with more metastases with DLIR-high simultaneously by both readers and a third reader for confirmation. Metastases visibility and contour definition were better with DLIR than ASiR-V.

Conclusion DLIR-high enhanced the detection and visibility of liver metastases compared to ASiR-V, and also increased the number of liver metastases detected.

Critical relevance statement Deep learning-based reconstruction at high strength allowed an increase in liver metastases detection compared to hybrid iterative reconstruction and can be used in clinical oncology imaging to help overcome the limitations of CT.

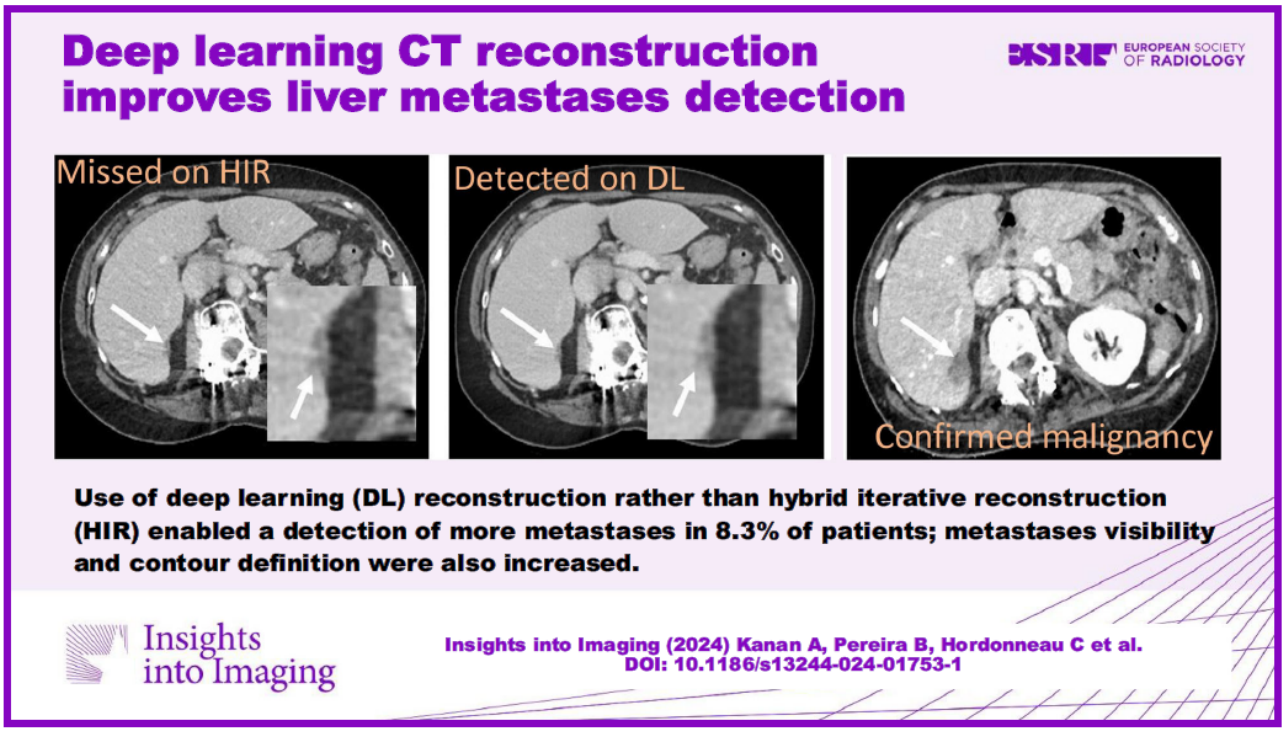
Key Points

- Detection of liver metastases is crucial but limited with standard CT reconstructions.
- More liver metastases were detected with deep-learning CT reconstruction compared to iterative reconstruction.
- Deep learning reconstructions are suitable for hepatic metastases staging and follow-up.

Keywords Liver neoplasm, Image reconstruction, Artificial intelligence, Deep learning, Computed tomography

*Correspondence:
Benoît Magnin
bmagnin@chu-clermontferrand.fr
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

Early detection of liver metastases plays a major role in management options and long-term prognosis and mostly relies on CT [1].

To obtain images from raw data, various algorithms can be used, with iterative reconstructions being the most widespread. Adaptive statistical iterative reconstruction-V (ASiR-V) (GE-HealthCare®) is a hybrid iterative reconstruction technique used in conjunction with filtered back projection in variable proportions according to user preferences. The measured value of each pixel is re-estimated and compared to an ideal predicted through an algebraic noise model. This process is repeated until there is concordance between the estimated and ideal values, thereby reducing noise while maintaining image quality [2].

Detection of liver metastases can be challenging, and CT through hybrid iterative reconstruction methods has certain limitations. It has been shown that using a high percentage of ASiR can lead to a lower image quality, giving it a plastic appearance or an unusually blurry texture, which limits the potential for noise reduction [3]. When comparing CT to MRI, 10% of liver metastases from pancreatic ductal carcinoma were missed [4], and up to 32% were noted as indeterminate [5]. Furthermore, CT scan has a low sensitivity for detecting lesions smaller than 10 mm [6].

Deep learning-based reconstructions are now available and aim to significantly reduce image noise [7, 8]. Deep learning image reconstruction (DLIR) TrueFidelity (GE-HealthCare®) is a new reconstruction method based on a convolutional neural network. The network was trained on thousands of high-quality CT datasets from patients and phantoms, acquired using filtered back projection. It enhances the raw data from a low-dose protocol by comparing it to the optimal data obtained during the training phase. Parameters such as noise, low-contrast resolution, and texture are analyzed and compared. The differences between the two datasets are minimized to achieve the best possible image. This process has been optimized through a learning phase [9]. Three selectable deep learning strength levels (DLIR-low, DLIR-medium, and DLIR-high) are configured by the manufacturer and available for use by clinicians to provide different amounts of noise reduction without impacting reconstruction speed.

These reconstructions provide high quality abdominal CT at same radiation doses compared to iterative reconstructions [10–12].

Several studies have investigated the benefits of DLIR for hepatic lesions. Jensen et al demonstrated that diagnostic confidence scores for abdominal lesions were significantly higher with DLIR compared to ASiR-V. However, their study included all solid organ lesions and did not specifically target hepatic metastases [13]. Nakamura et al found that DLIR resulted in higher scores for the conspicuity of hepatic metastases compared to adaptive iterative dose reduction 3D (AiDR 3D, Canon Medical System®) [14]. However, they did not assess lesion detection. Singh et al showed that DLIR was equivalent to AiDR for the detection of abdominal lesions in a prospective multi-institutional study [15]. Of the 31 lesions evaluated, only 13 were low-attenuating hepatic lesions, which limits the ability to draw definitive conclusions regarding metastasis detection. Therefore, the impact of DLIR on hepatic metastases detection remains unclear.

We hypothesized that the image enhancement from these new reconstruction techniques could allow an increased detection of liver metastases compared to conventional iterative reconstructions. Our main objective was to compare the number of metastases detected using three different levels of DLIR and a 50%-ASiR-V. Our secondary objective was to compare metastases conspicuity for each reconstruction.

Methods

This was a retrospective observational single institutional study conducted in our medical imaging department.

Patient selection

All CT scans of the abdomen and pelvis performed for cancer initial assessment or follow-up between November 2020 and July 2021 were selected for the inclusion process. The inclusion criterion was the presence of at least one hypoattenuating liver metastasis described in the radiology report. Exclusion criteria were the loss of at least one reconstruction (loss of raw data, at least one reconstruction not saved on picture archiving and communication system (PACS)), double energy acquisition, age less than 18 years old, hypervascular metastases, and absence of histopathological proof of cancer.

Imaging technique and CT reconstructions

CT scans were performed using the same Revolution Evo system (GE-HealthCare®) at 120 kV tube, 160 to 500 mA current range with organ dose modulation, 1.375 pitch, 40 mm detector collimation, 0.70 second rotation time, and 1.25 mm thickness. Iodine contrast material was administered with a basis of 2 mL per kilogram adapted to body weight (mean 93 ± 10 mL; range 80-130 mL) (Xenetix 350, Guerbet or Omnipaque 350) into the cubital vein at an injection rate of 2 mL per second. The acquisition was performed 90 seconds after injection. Volume computed tomography dose index ($CTDI_{VOL}$) and dose length product (DLP) were recorded.

One standard 50%-ASiR-V reconstruction and three deep-learning reconstructions were obtained using the DLIR algorithm TrueFidelity at different strength levels: DLIR-low (DLIR-L), DLIR-medium (DLIR-M), and DLIR-high (DLIR-H). All CT scans were anonymized before analysis.

CT analysis and lesion detection

Metastases number evaluation and subjective analyses and were performed independently by Reader 1, A.K., with three years of in-training experience in radiology, and Reader 2, B.C., with ten years of experience in abdominal radiology. Readers were blinded to the reconstruction method and the patient past medical history. Both readers received identical and standardized printed instructions before evaluation. All CT scans were randomly split into four equal blocks, each block containing one random reconstruction method by patients (121 scans per block). CT scans were analyzed block by block in a random and different order from July to December 2021. To avoid memory bias, an interval of one month between each block analysis was respected. The evaluation was performed on an Advantage Workstation (AW3.2, GE-HealthCare®). Readers were able to adjust the window (width and level) and use coronal or sagittal sections and minimum intensity projection as desired.

Both readers counted the number of hepatic metastases from 0 to a maximum of 10. In cases where both readers found more metastases with DLIR-H than ASiR-V or vice versa, a third independent radiologist (B.M.), with 11 years of experience in abdominal radiology, blindly evaluated the number of lesions on both reconstructions to confirm or disprove the difference. In case of discrepancies, an unblinded consensus reading was made by the three readers. They used all available data, such as MRIs, previous or subsequent CT scans, and clinical reports to verify the metastatic nature of missed lesions, mainly based on their MRI signal characteristics or size variation over time.

Both readers rated overall image quality, image noise reduction, hepatic metastases visibility and hepatic metastases contour definition using a five-point scale: 1-inacceptable; 2-low; 3-medium; 4-good; 5-excellent, based on their own subjectivity.

Attenuation measurements

Measurements were performed by reader 1, using an Advantage Workstation. Regions of interest (ROIs) were placed on nine anatomical structures (Table 1) and at the center of one randomly selected hepatic metastasis, avoiding artifacts and irregularities. The ROI was then

Table 1 Image noise and CNRs of anatomical structures and selected metastases for each reconstruction

	ASiR-V	DLIR-L	DLIR-M	DLIR-H	ROI surface (mm^2)	
Image noise	16.0 ± 2.8	17.2 ± 3.0	13.7 ± 2.7	10.5 ± 3.2		
Organs CNRs						
Paravertebral muscle	*NA*	*NA*	*NA*	*NA*	261 ± 134	
Abdominal subcutaneous fat	9.77 ± 2.2	9.13 ± 2.1	11.47 ± 2.6	5.39 ± 3.9	233 ± 146	$p < 0.001^a$
Abdominal aorta	5.41 ± 1.9	5.06 ± 1.7	6.33 ± 2.1	8.53 ± 2.9	119 ± 45.7	$p < 0.001^a$
Spleen	3.45 ± 1.2	3.21 ± 1.1	4.01 ± 1.3	5.42 ± 1.8	346 ± 146	$p < 0.001^a$
Right hepatic lobe	2.95 ± 1.2	2.73 ± 1.1	3.42 ± 1.3	4.61 ± 1.9	440 ± 185	$p < 0.001^a$
Left hepatic lobe	2.97 ± 1.2	2.75 ± 1.1	3.44 ± 1.3	4.64 ± 1.9	312 ± 136	$p < 0.001^a$
Vessels CNRs						
Main portal vein	5.79 ± 2.1	5.43 ± 1.9	6.79 ± 2.2	9.16 ± 3.2	113 ± 64.6	$p < 0.001^a$
Right portal vein	5.82 ± 2.1	5.46 ± 1.9	6.83 ± 2.3	9.21 ± 3.3	82.4 ± 42.1	$p < 0.001^a$
Left portal vein	5.72 ± 2.1	5.39 ± 2.0	6.73 ± 2.3	9.07 ± 3.4	54.4 ± 32.2	$p < 0.001^a$
Metastases CNRs	3.14 ± 1.4	2.89 ± 1.3	3.56 ± 1.7	4.66 ± 2.3	7.75 ± 1.65	$p < 0.001^a$

Values are given as mean value ± standard deviation
Image noise was defined as the standard deviation of attenuation in the paraspinal muscle
[a] Statistical difference was observed in pairwise comparison between each reconstruction ($p < 0.001$)

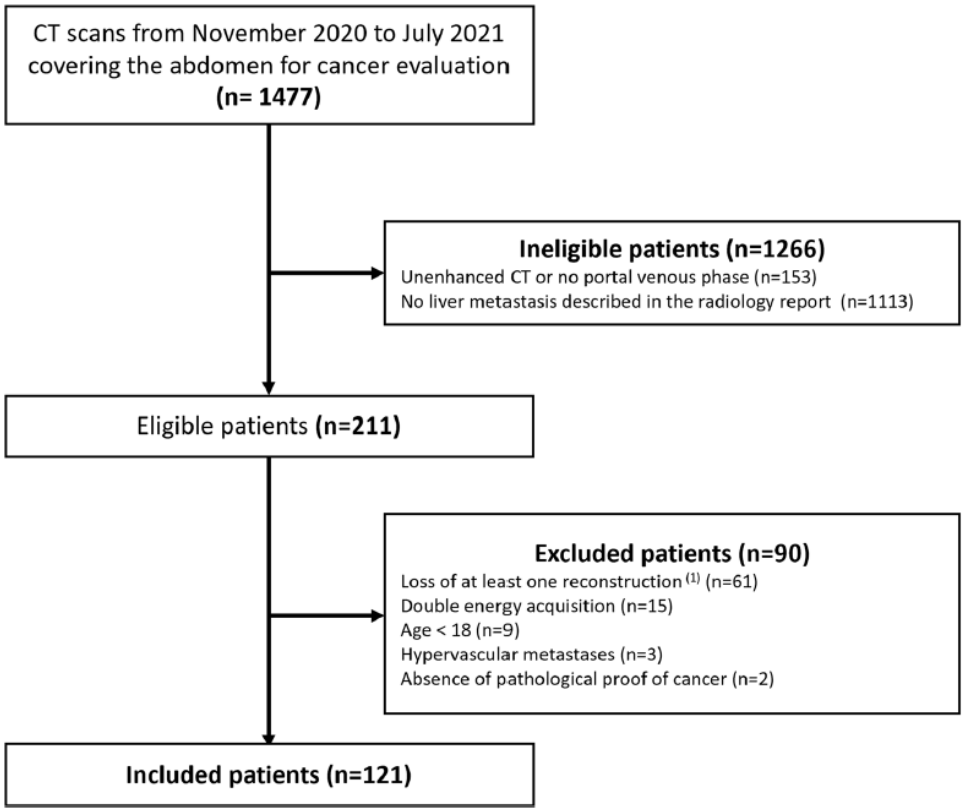


Fig. 1 Patient flow-chart. (1) Loss of raw data, at least one reconstruction not saved on PACS

cloned at the same location for each reconstruction. Image noise (N) was defined as the standard deviation of attenuation in the paraspinal muscle. The contrast-to-noise ratio (CNR) of a structure was calculated as the absolute difference between its mean attenuation and the mean attenuation of paraspinal muscle divided by

Table 2 Patients and primitive tumor characteristics

Patient characteristics	($n = 121$) (%)
Demographic	
Male	73 (60)
Female	48 (40)
Mean age, SD	65 ± 12
Cancer type	
Adenocarcinoma	88 (74)
Colic	31 (26)
Pancreatic	29 (26)
Rectal	16 (13)
Gastric	9 (7.4)
Small bowel	2 (1.7)
Ovarian	1 (0.8)
Melanoma	11 (9.1)
Neuro-endocrine	6 (5.0)
Pancreatic	3 (2.5)
Intestinal	2 (1.7)
Hepatic	1 (0.8)
Gastrointestinal stromal tumor	5 (4.1)
Intestinal	4 (3.3)
Gastric	1 (0.8)
Cholangiocarcinoma	5 (4.1)
Ampullary carcinoma	3 (2.5)
Lymphoma	1 (0.8)
Prior systemic treatment [a]	95 (78)

[a] Prior history of systemic oncologic care, such as chemotherapy and immunotherapy, before CT acquisition
SD standard deviation

image noise $CNR_a = |HU_a - HU_{muscle}|/N$, (HU: *Hounsfield unit*).

Statistical analysis

Continuous variables were expressed according to their statistical distribution with mean and standard deviation (SD). Metastasis number was, however expressed as median and interquartile range. An arbitrary limit of 10 lesions was given, and the mean number of lesions seemed less appropriate to be presented as a result. Agreement (between both readers and between reconstruction methods) was assessed using Lin's concordance correlation coefficient. The results were interpreted in relation to recommendations reported in the literature by Altman: < 0.4: no agreement, 0.4–7: poor agreement, > 0.7: moderate to strong agreement [16]. Comparisons between methods for continuous variables were completed using mixed models that allowed to consider between- and within-patient variability (i.e., subject as a random effect). The normality of residuals from these models was analyzed with the Shapiro–Wilk test and graphical presentation. When appropriate, a logarithmic transformation of the dependent variable has been applied. Statistical analyses were performed using Stata software (version 15, StataCorp, College Station) by B.P. All statistical tests were carried out based on a two-sided type I error at 5%. Sidak's type I error correction was applied for two-by-two multiple comparisons between methods.

Results

Patient and lesion characteristics

A total of 121 patients were included in the study. A flow chart of the inclusion process is shown in Fig. 1. Patients and primitive tumor characteristics are listed in Table 2.

Lesion detection

A higher number of metastases was detected by the senior reader (R2) with DLIR-high: 7 (2–10) (median (Q_1–Q_3); total 733) versus 5 (2–10) respectively for DLIR-medium, DLIR-low, and ASiR-V ($p < 0.001$) (Table 3). The junior reader (R1) found no significant difference in metastases number between reconstructions.

For 12 patients, both readers simultaneously found a higher number of metastases with DLIR-H compared to ASiR-V. This was confirmed for ten patients by the third radiologist and disproved for the other two patients (Figs. 2 and 3). Consensus reading was allowed by comparison with subsequent MRI for three patients, subsequent CT for three patients, and previous CT for four patients. In these cases, one additional metastasis was detected using DLIR-H in six patients, and two in four patients. This led to 14 missed lesions with a median size of 7 mm. (details available in Appendix 1).

For two patients, both readers detected a higher number of metastases with ASiR-V than DLIR-H. Confirmation with the third radiologist was obtained for one patient only with consensus reading using a subsequent MRI. The missed lesion was 11 mm in size (Fig. 4).

CT subjective analysis

Image quality and noise reduction were lower for ASiR-V and increased with deep learning levels (Fig. 5a). Significant differences were observed between all reconstructions for both readers ($p < 0.001$). Metastases visibility and contour definition were better for DLIR-H compared to other reconstructions for both readers (Fig. 5b).

Attenuation measurements

Image noise was significantly higher for DLIR-L, followed by ASiR-V, DLIR-M, and DLIR-H ($p < 0.001$). CNRs of anatomical structures were significantly different between all reconstructions, with the highest values for DLIR-H followed by DLIR-M, ASiR-V and DLIR-L ($p < 0.001$).

CNRs of metastases were higher using DLIR-H and DLIR-M compared to ASiR-V ($p < 0.001$) (Table 1).

Radiation dose

Mean $CTDI_{VOL}$ and DLP were 9,2 mGy ± 2,5 and 512 mGy.cm ± 158, respectively.

Table 3 Number of detected hepatic metastases by both readers for each reconstruction

	ASiR-V	DLIR-L	DLIR-M	DLIR-H	
Reader 1					
total	673	679	680	686	
median	5 (2–10)	6 (2–10)	5 (2–8)	6 (2–10)	$p = 0.78$[a]
Reader 2					
total	686	674	674	733	
median	5 (2–10)	5 (2–10)	5 (2–10)	7 (2–10)	$p < 0.001$[b]

Readers counted up to a maximum of ten lesions per patient
Data are expressed as a total number of lesions with a median per patient (and interquartile range)
[a] No significant difference was observed between each reconstruction
[b] Pairwise significant difference was observed between DLIR-H and other reconstructions only

Discussion

This study aimed to compare a recent deep learning-based reconstruction (TrueFidelity) and a standard iterative reconstruction (50%-ASiR-V) for the detection of hypoattenuating liver metastases on CT. The main objective was to determine whether DLIR would affect the number of detected lesions. High-strength DLIR led to a statistical increase in the number of detected lesions for one of the two readers. Additionally, high-strength DLIR enabled both readers to simultaneously detect more metastases in ten patients compared to ASiR-V. This statement was confirmed by a third independent reader. As a secondary objective, we compared lesion conspicuity between both reconstructions. The visibility and contour definition of hepatic metastases received better scores with DLIR-high compared to the other reconstructions for both readers.

The most common etiologies of liver metastases arise from the gastrointestinal tract, mainly colorectal and pancreatic cancers [17]. Management of patients depends on the presence of liver metastases. Their number, size, and location can guide clinicians toward curative or conservative techniques. Treatment modalities include

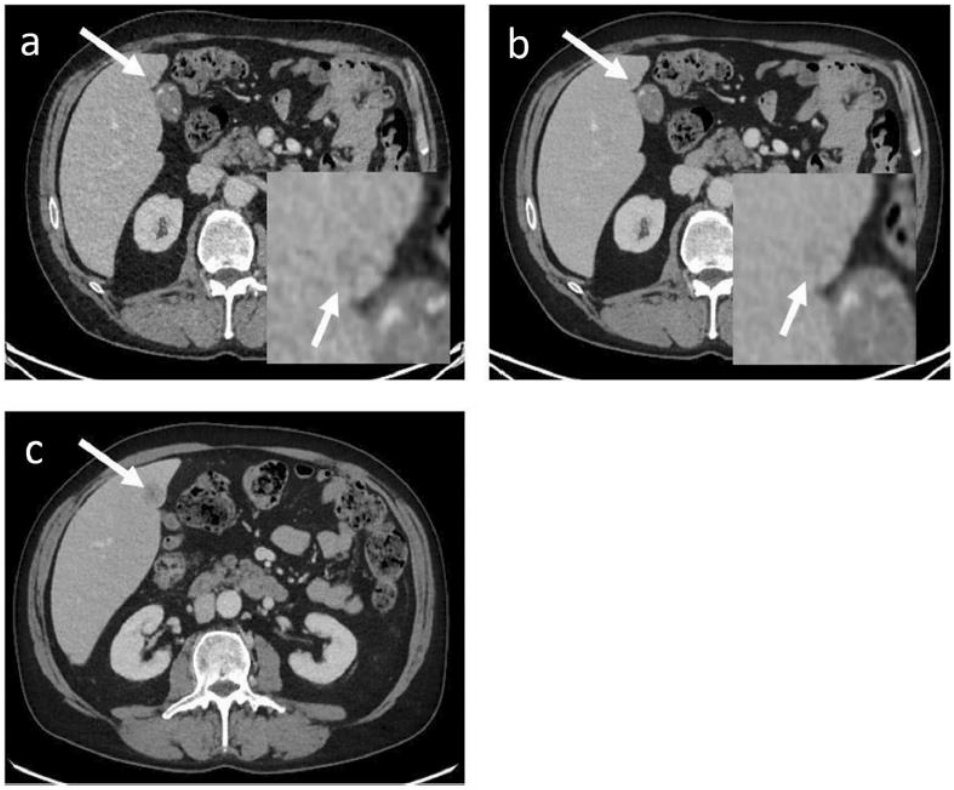


Fig. 2 A contrast-enhanced CT image obtained with ASiR-V (**a**) and DLIR-H (**b**) showing the same hypoattenuating metastasis, magnified in the right lower corner (white arrows). Both readers detected the lesion on DLIR-H and missed the diagnosis on ASiR-V, as did a third independent reader. CT image of the same patient two months later showing the growth of the lesion and confirming its malignancy (white arrow) (**c**)

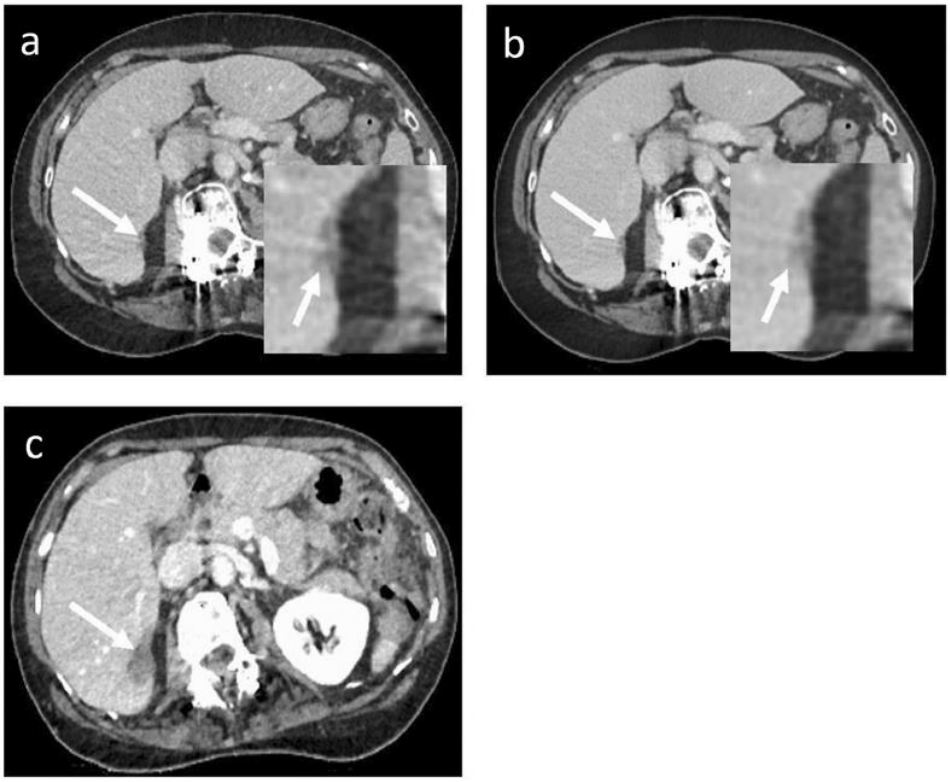


Fig. 3 A contrast-enhanced CT image obtained with ASiR-V (**a**) and DLIR-H (**b**) showing the same hypoattenuating metastasis, magnified in the right lower corner (white arrows). Both readers detected the lesion on DLIR-H and missed the diagnosis on ASiR-V, as did a third independent reader. The artifact reduction provided by DLIR can be seen in this example with osteosynthesis material artifact near the lesion significantly reduced. CT image of the same patient 18 months earlier before systemic treatment confirms lesion malignancy (**c**)

resection surgeries, thermoablation procedures, stereotactic radiotherapy, endovascular treatments, or systemic therapies [18]. Many techniques can be used for liver metastases detection as a conjunct to the initial CT staging. MRI seems to be superior to CT scan, especially for < 10 mm lesions [6, 19]. Fluorine-18-fluorodeoxyglucose positron emission tomography/CT (PET/CT) is also very sensitive but has limited performances for small lesions [20]. Other techniques have been evaluated, such as Kupffer-phase imaging in contrast-enhanced endoscopic ultrasonography, with superior results for small left liver metastases [21].

Liver metastases of digestive adenocarcinoma often appear as multiple hypoattenuating nodular lesions. As adenocarcinomas represented 74% of our study population, comparison between reconstructions were performed on a homogeneous pool of lesions. Imaging features may however change, based on histopathological characteristics and may moderate interpretation of the results. For example, desmoplastic reactions around colorectal liver metastases are closely related to peripheral enhancement [22]. Cystic components of primary tumor and severe necrosis can lead to cyst-like hepatic metastases. Other characteristics may be present, such as calcifications in mucinous adenocarcinoma, and peripheral wash-out and hypervascularity in neuro-endocrine tumors [23].

Despite the multiple modalities, CT scan remains the gold standard for gastro-intestinal cancer staging and follow up according to international recommendations. The current protocol often involves a CT scan of thorax, abdomen, and pelvis [24–26]. Obtaining high-quality images is therefore essential and implementation of artificial intelligence-based reconstruction algorithms facilitate early detection of metastases.

Deep learning methods still have certain limitations. Kaga et al showed that high levels of deep learning can reduce the conspicuity of hepatic lesions compared to ASiR-V, especially for small lesions [27]. This was also described for extrahepatic exploration such as chest CT, where small structures had lower conspicuity scores with high-strength DLIR [28]. Yang et al found no difference in liver lesion detection between DLIR and ASIR-V; however, they only involved 8 patients and 13 malignant lesions in their analysis [29].

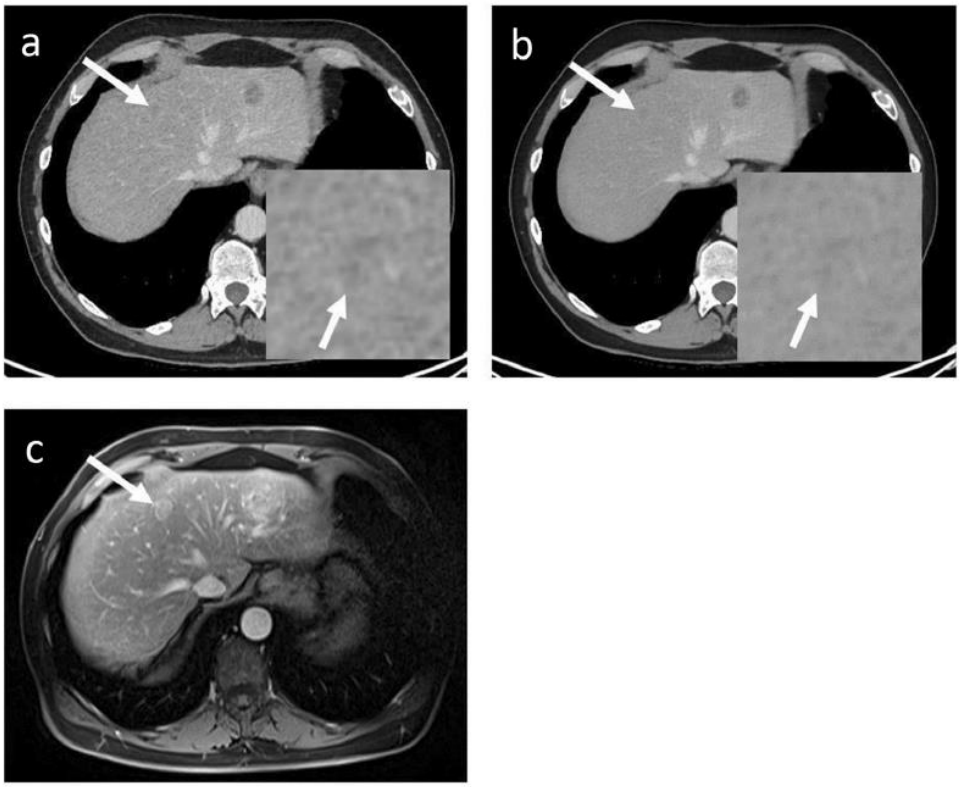


Fig. 4 A contrast-enhanced CT image obtained with ASiR-V (**a**) and DLIR-H (**b**) showing the same hypoattenuating metastasis of 11 mm, magnified in the right lower corner (white arrows). Both readers missed the lesion on DLIR-H but not on ASiR-V. MRI of the same patient six weeks later in portal venous phase T1-weigthed image showing lesion growth and confirming its malignancy (white arrow) (**c**)

There have been recent studies experimenting different low-dose protocols of DLIR in terms of quality assessment of images and lesion detection [30–36]. Wang et al found that low-dose deep learning algorithms may provide better images, signal-to-noise, and contrast-to-noise ratios of unenhanced CT scans when compared to standard-dose iterative reconstruction. They found no difference in sensitivity and diagnostic confidence for liver metastases detection [37]. When comparing a 33%-dose protocol with DLIR to a standard-dose iterative reconstruction, Lee et al found lower noise on DLIR and comparable diagnostic performance in detecting malignant liver tumors [38].

The aim of our study was to determine whether diagnostic performance was superior at the same dose level, specifically if more hepatic metastases could be detected. To our knowledge, no study has yet described an increase in the number of detected hepatic metastases using deep learning reconstructions compared to iterative reconstructions. This finding could significantly impact patients' oncologic evaluation. Our study showed that more metastases were detected in 10 out of 121 patients with DLIR. As expected, the majority of missed lesions were smaller than 10 mm, as subcentimeter lesions can often be missed on CT scans [3]. Detecting these small lesions can significantly influence the therapeutic management of patients, such as surgical or ablative planning in colorectal cancer, or a switch from curative to palliative care in pancreatic adenocarcinoma.

Our study has some limitations. First, this was a retrospective study, relying on a single CT and one manufacturer's DLIR. Second, our inclusion criterion was solely based on the final CT report with no prior confirmation of our own. However, all cases had histopathological proof of the primary cancer. Most of the patients had previous MRIs, which were used for comparison by radiologists allowing reduction of potential selection bias. We only included patients with liver metastases, which suggests that our results showed DLIR to increase sensitivity rather than specificity for lesion detection. Finally, we did not perform subgroup analyses based on primary tumor type.

Although our study doesn't evaluate patient management and outcome, our results suggest a potential implication of DLIR and can be considered as an aid in selecting the most appropriate CT reconstruction and the

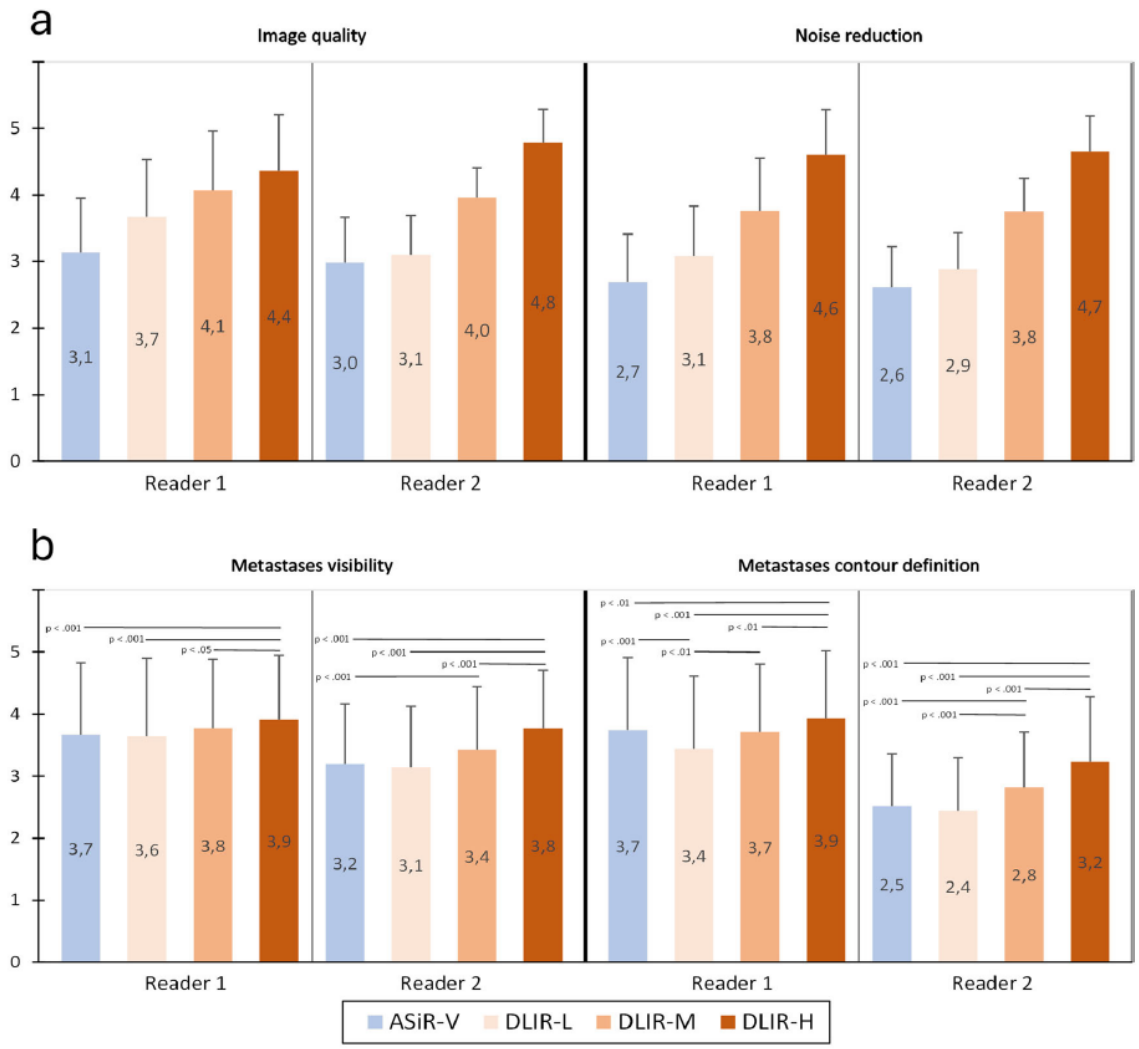


Fig. 5 **a** Subjective evaluation of CT image quality and noise. Statistically significant differences were obtained for all pairwise comparisons. **b** Subjective evaluation of hepatic metastases. All statistically significant differences of pairwise comparisons are displayed along with their *p*-values. Values are given as mean score (bars) ± standard deviation (error bars) of a five-point rating scale

level of deep learning for liver metastases detection. In conclusion, high-strength DLIR statistically increased the detection and conspicuity of liver metastases compared to ASiR-V. Additional studies should be conducted to assess the clinical impact of these findings, but our results encourage the use of deep-learning reconstructions when performing abdominal CT scans in oncology.

Abbreviations

ASiR-V	Adaptive Statistical Iterative Reconstruction-V
CNR	Contrast-to-Noise Ratio
$CTDI_{VOL}$	Computed Tomography Dose Index
DLIR	Deep Learning Image Reconstruction
DLIR-H	Deep Learning Image Reconstruction High
DLIR-L	Deep Learning Image Reconstruction Low
DLIR-M	Deep Learning Image Reconstruction Medium
DLP	Dose Length Product

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13244-024-01753-1.

ELECTRONIC SUPPLEMENTARY MATERIAL

Acknowledgements

The authors would like to thank Helen Braund for language editing and Anthony Aufaure for his help.

Author contributions

A.K. contributed to the literature search, conceptualization, collection, C.T. reading, data analysis, data interpretation, manuscript edition, and validation of the study. B.P. contributed to the literature search, conceptualization, data analysis, data interpretation, and manuscript edition. C.H. contributed to the literature search, conceptualization, collection, data analysis, data interpretation, and manuscript edition. L.C. contributed to the literature search,

conceptualization, collection, data interpretation, and manuscript edition. E.P. contributed to the literature search, conceptualization, collection, data interpretation, and manuscript edition. L.T. contributed to the literature search, conceptualization, collection, data interpretation, and manuscript edition. B.C. contributed to the literature search, conceptualization, collection, C.T. reading data analysis, data interpretation, manuscript edition, validation of the study, and study supervision. B.M. contributed to the literature search, conceptualization, collection, C.T. reading data analysis, data interpretation, manuscript edition, validation of the study, and study supervision. All the authors have read and approved the final manuscript.

Funding
The authors received no specific funding for the study.

Data availability
The CT scans analyzed during the current study are not publicly available to protect study participant privacy according to local legislation. The anonymized analyzed data are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethical Committee for Research in Medical Imaging (n° CRM-2203-250), which waived patients' prior written consent due to the retrospective design of the study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Radiology, Estaing Hospital, Clermont University Hospital, Clermont-Ferrand, France. [2]Department of Biostatistics, DRCI, Clermont University Hospital, Clermont-Ferrand, France. [3]Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, Clermont-Ferrand, France. [4]Department of Radiology, Gabriel Montpied Hospital, Clermont University Hospital, Clermont-Ferrand, France. [5]Department of Radiology, Tengandogo' Ouagadougou University Hospital Center, Ouagadougou, Burkina Faso. [6]DI2AM, DRCI, Clermont University Hospital, Clermont-Ferrand, France.

Received: 6 March 2024 Accepted: 17 June 2024
Published online: 06 July 2024

References
1. Germani MM, Borelli B, Boraschi P et al (2022) The management of colorectal liver metastases amenable of surgical resection: How to shape treatment strategies according to clinical, radiological, pathological and molecular features. Cancer Treat Rev 106:102382. https://doi.org/10.1016/j.ctrv.2022.102382
2. Tayal U, King L, Schofield R et al (2019) Image reconstruction in cardiovascular CT: Part 2 – Iterative reconstruction; potential and pitfalls. J. Cardiovasc. Comput. Tomogr 13:3–10. https://doi.org/10.1016/j.jcct.2019.04.009
3. Padole A, Ali Khawaja RD, Kalra MK, Singh S (2015) CT radiation dose and iterative reconstruction techniques. AJR Am J Roentgenol 204:W384–W392. https://doi.org/10.2214/AJR.14.13241
4. Marion-Audibert A-M, Vullierme M-P, Ronot M et al (2018) Routine MRI with DWI sequences to detect liver metastases in patients with potentially resectable pancreatic ductal carcinoma and normal liver CT: a prospective multicenter study. AJR Am J Roentgenol 211:W217–W225. https://doi.org/10.2214/AJR.18.19640
5. Kim HW, Lee J-C, Paik K-H et al (2017) Adjunctive role of preoperative liver magnetic resonance imaging for potentially resectable pancreatic cancer. Surgery 161:1579–1587. https://doi.org/10.1016/j.surg.2016.12.038
6. Tsili AC, Alexiou G, Nakal C, Argyropoulou MI (2021) Imaging of colorectal cancer liver metastases using contrast-enhanced US, multidetector CT, MRI, and FDG PET/CT: a meta-analysis. Acta Radiol 62:302–312. https://doi.org/10.1177/0284185120925481
7. McLeavy CM, Chunara MH, Gravell RJ et al (2021) The future of CT: deep learning reconstruction. Clin Radiol 76:407–415. https://doi.org/10.1016/j.crad.2021.01.010
8. Boedeker K (2019) AiCE Deep Learning Reconstruction: Bringing the power of Ultra-High Resolution CT to routine imaging. Available via https://fr.medical.canon/wp-content/uploads/sites/21/2019/11/White-paper-Kirsten-Boedeker.pdf. Accessed 3rd March 2024
9. Hsieh J, Liu E, Nett B, Tang J, Thibault J-B, Sahney S (2019) A new era of image reconstruction: TrueFidelity - Technical white paper on deep learning image reconstruction. Available via https://www.gehealthcare.com/-/jssmedia/040dd213fa89463287155151fdb01922.pdf. Accessed 3rd March 2024
10. Akagi M, Nakamura Y, Higaki T et al (2019) Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. Eur Radiol 29:6163–6171. https://doi.org/10.1007/s00330-019-06170-3
11. Park C, Choo KS, Jung Y, Jeong HS, Hwang J-Y, Yun MS (2021) CT iterative vs deep learning reconstruction: comparison of noise and sharpness. Eur Radiol 31:3156–3164. https://doi.org/10.1007/s00330-020-07358-8
12. Ichikawa Y, Kanii Y, Yamazaki A et al (2021) Deep learning image reconstruction for improvement of image quality of abdominal computed tomography: comparison with hybrid iterative reconstruction. Jpn J Radiol 39:598–604. https://doi.org/10.1007/s11604-021-01089-6
13. Jensen CT, Liu X, Tamm EP et al (2020) Image quality assessment of abdominal CT by use of new deep learning image reconstruction: initial experience. AJR Am J Roentgenol 215:50–57. https://doi.org/10.2214/AJR.19.22332
14. Nakamura Y, Higaki T, Tatsugami F et al (2021) Deep learning–based CT image reconstruction: initial evaluation targeting hypovascular hepatic metastases. Radiol Artif Intell 1:e180011. https://doi.org/10.1148/ryai.2019180011
15. Singh R, Digumarthy SR, Muse VV et al (2020) Image quality and lesion detection on deep learning reconstruction and Iterative reconstruction of submillisievert chest and abdominal CT. AJR Am J Roentgenol 214:566–573. https://doi.org/10.2214/AJR.19.21809
16. Altman DG (1990) Practical statistics for medical research. Chapman and hall, New York https://doi.org/10.1201/9780429258589
17. Horn SR, Stoltzfus KC, Lehrer EJ et al (2020) Epidemiology of liver metastases. Cancer Epidemiol 67:101760. https://doi.org/10.1016/j.canep.2020.101760
18. Cervantes A, Adam R, Roselló S et al (2023) Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 34:10–32. https://doi.org/10.1016/j.annonc.2022.10.003
19. Renzulli M, Clemente A, Ierardi AM et al (2020) Imaging of Colorectal Liver Metastases: New Developments and Pending Issues. Cancers 12:151. https://doi.org/10.3390/cancers12010151
20. Serrano PE, Gu C-S, Moulton C-A et al (2020) Effect of PET-CT on disease recurrence and management in patients with potentially resectable colorectal cancer liver metastases. Long-term results of a randomized controlled trial. J Surg Oncol 121:1001–1006. https://doi.org/10.1002/jso.25864
21. Minaga K, Kitano M, Nakai A et al (2021) Improved detection of liver metastasis using Kupffer-phase imaging in contrast-enhanced harmonic EUS in patients with pancreatic cancer (with video). Gastrointest Endosc 93:433–441. https://doi.org/10.1016/j.gie.2020.06.051
22. Eble JA, Niland S (2019) The extracellular matrix in tumor progression and metastasis. Clin Exp Metastasis 36:171–198. https://doi.org/10.1007/s10585-019-09966-1
23. Ozaki K, Higuchi S, Kimura H, Gabata T (2022) Liver Metastases: Correlation between Imaging Features and Pathomolecular Environments. Radiographics 42:1994–2013. https://doi.org/10.1148/rg.220056
24. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology Colon Cancer Version 1 (2020) Available via https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf. Accessed 20 May 2024.
25. Haria PD, Baheti AD, Palsetia D et al (2021) Follow-up of colorectal cancer and patterns of recurrence. Clin Radiol 76:908–915. https://doi.org/10.1016/j.crad.2021.07.016

26. Daamen LA, Groot VP, Intven MPW et al (2019) Postoperative surveillance of pancreatic cancer patients. Eur J Surg Oncol 45:1770–1777. https://doi.org/10.1016/j.ejso.2019.05.031
27. Kaga T, Noda Y, Fujimoto K et al (2021) Deep-learning-based image reconstruction in dynamic contrast-enhanced abdominal CT: image quality and lesion detection among reconstruction strength levels. Clin Radiol 76:710.e15–710.e24. https://doi.org/10.1016/j.crad.2021.03.010
28. Tian Q, Li X, Li J et al (2022) Image quality improvement in low-dose chest CT with deep learning image reconstruction. J Appl Clin Med Phys 23:e13796. https://doi.org/10.1002/acm2.13796
29. Yang S, Bie Y, Pang G et al (2021) Impact of novel deep learning image reconstruction algorithm on diagnosis of contrast-enhanced liver computed tomography imaging: Comparing to adaptive statistical iterative reconstruction algorithm. J Xray Sci Technol 29:1009–1018. https://doi.org/10.3233/XST-210953
30. Nakamura Y, Narita K, Higaki T, Akagi M, Honda Y, Awai K (2021) Diagnostic value of deep learning reconstruction for radiation dose reduction at abdominal ultra-high-resolution CT. Eur Radiol 31:4700–4709. https://doi.org/10.1007/s00330-020-07566-2
31. Lyu P, Liu N, Harrawood B et al (2022) Is it possible to use low-dose deep learning reconstruction for the detection of liver metastases on CT routinely? Eur Radiol 33:1629–1640. https://doi.org/10.1007/s00330-022-09206-3
32. Toia GV, Zamora DA, Singleton M et al (2023) Detectability of Small Low-Attenuation Lesions With Deep Learning CT Image Reconstruction: A 24-Reader Phantom Study. AJR Am J Roentgenol 220:283–295. https://doi.org/10.2214/AJR.22.28407
33. Lyu P, Li Z, Chen Y et al (2024) Deep learning reconstruction CT for liver metastases: low-dose dual-energy vs standard-dose single-energy. Eur Radiol 34:28–38. https://doi.org/10.1007/s00330-023-10033-3
34. Jensen CT, Gupta S, Saleh MM et al (2022) Reduced-Dose Deep Learning Reconstruction for Abdominal CT of Liver Metastases. Radiology 303:90–98. https://doi.org/10.1148/radiol.211838
35. Steuwe A, Weber M, Bethge OT et al (2020) Influence of a novel deep-learning based reconstruction software on the objective and subjective image quality in low-dose abdominal computed tomography. Br J Radiol 94:20200677. https://doi.org/10.1259/bjr.20200677
36. Cao L, Liu X, Li J et al (2021) A study of using a deep learning image reconstruction to improve the image quality of extremely low dose contrast-enhanced abdominal CT for patients with hepatic lesions. Br J Radiol 94:20201086. https://doi.org/10.1259/bjr.20201086
37. Wang X, Zheng F, Xiao R et al (2021) Comparison of image quality and lesion diagnosis in abdominopelvic unenhanced CT between reduced-dose CT using deep learning post-processing and standard-dose CT using iterative reconstruction: A prospective study. Eur J Radiol 139:109735. https://doi.org/10.1016/j.ejrad.2021.109735
38. Lee DH, Lee JM, Lee CH, Afat S, Othman A (2024) Image Quality and Diagnostic Performance of Low-Dose Liver CT with Deep Learning Reconstruction versus Standard-Dose CT. Radiol Artif Intell 6:e230192. https://doi.org/10.1148/ryai.230192

Publisher's Note

Deep learning CT reconstruction improves liver metastases detection

ELECTRONIC SUPPLEMENTARY MATERIAL

Appendix 1. Follow up data of patients with concordent differences between ASiR and DLIR-H simultaneously by both readers for detected metastases and their consensus reading

Detection step: Difference of number of detected metastases (blinded to reconstructions)

	Patient number	Reader 1 (junior) n(ASIR)-n(DLIR-H)	Reader 2 (senior) n(ASIR)-n(DLIR-H)	Reader 3 (senior) n(ASIR)-n(DLIR-H)	Differences between R1 and R2		Missed lesions size1 (mm)	size 2(mm)	Modality used
DLIR-H > ASiR	#9	2	1	1	Confirmed by R3		6	*NA*	subsequent MRI
	#25	1	3	2	Confirmed by R3		7	21	previous CT
	#27	1	1	1	Confirmed by R3		13	*NA*	previous CT
	#40	2	1	2	Confirmed by R3		7	4	previous CT
	#68	1	3	1	Confirmed by R3		7	*NA*	subsequent MRI
	#77	1	4	2	Confirmed by R3		4	5	subsequent CT
	#82	3	5	2	Confirmed by R3		8	7	subsequent CT
	#83	2	2	1	Confirmed by R3		9	*NA*	subsequent CT
	#84	3	2	0	Disproved by R3 (no consensus reading)				
	#104	1	1	1	Confirmed by R3		12	*NA*	previous CT
	#113	2	1	1	Confirmed by R3		8	*NA*	subsequent MRI
	#120	7	4	0	Disproved by R3 (no consensus reading)				
TOTAL NUMBER OF PATIENTS	12				10	2	**14 lesions missed. Median size of missed lesions on ASiR: 7 mm**		
ASiR > DLIR-H	#18	-1	-2	0	Disproved by R3 (no consensus reading)				
	#114	-1	-2	-1	Confirmed by R3		11	*NA*	subsequent MRI
TOTAL NUMBER OF PATIENTS	2				1	1	**One missed lesions on DLIR-H : 11 mm**		

Insights Imaging (2024) Kanan A, Pereira B, Hordonneau C, et al.

Radiomique

Nous allons maintenant introduire une seconde technique d'analyse d'image, la radiomique.

Présentation de la radiomique

La radiomique est une modalité d'analyse d'images qui a émergé dans le domaine de l'imagerie médicale en 2012 (31). Elle représentait alors une approche innovante aux perspectives multiples. Elle a été particulièrement étudiée ces dernières années dans le domaine oncologique, en particulier afin d'améliorer la caractérisation lésionnelle des tumeurs, ou de rechercher des facteurs pronostiques (32). En effet l'analyse des examens d'imagerie est basée essentiellement sur l'analyse visuelle par un radiologue des caractéristiques évaluant la taille, la morphologie, les contours, l'hétérogénéité, l'intensité ou densité de signal ou le rehaussement après injection d'une lésion à caractériser. La radiomique est une analyse mathématique de l'image ou d'une partie de l'image (une tumeur par exemple), qui permet de quantifier les caractéristiques de l'images, celles visibles comme celles non perceptibles à l'œil humain. Ainsi, à travers la radiomique, « les images sont plus que des photos, elles sont des données » (33), c'est-à-dire que la radiomique transforme une image de scanner en une liste de chiffres dont certains peuvent être des biomarqueurs pertinents.

Les paramètres de radiomique

Les paramètres de radiomique peuvent rangés en plusieurs catégories, selon la nature et la complexité des phénomènes qu'ils décrivent :

- les paramètres de forme : ils donnent des données quantifiées de la forme qui est étudiée, sans étudier le contenu. Ils comportent des données simplement mesurables (petit axe, grand axe, surface, volume) ainsi que des données plus élaborées (élongation, aplatissement, sphéricité)

- les paramètres de premier ordre (ou paramètre d'histogramme) : ils sont issus de l'analyse de l'histogramme des intensités des pixels, qui évalue la répartition des pixels en fonction de leur intensité de gris. Ils comprennent des paramètres simples qui peuvent être donnés par une console PACS tels que l'intensité moyenne, maximale et minimale, l'écart-type des intensités. Ils comprennent aussi des descripteurs plus complexes de l'histogramme tels que l'entropie (caractère aléatoire de la distribution des niveaux de gris), l'uniformité de la distribution (energy) et des données quantifiant les différences à la distribution normale (Figure 7). Ces données de premier ordre ne tiennent pas compte de la localisation des pixels ni de leur relation spatiale.

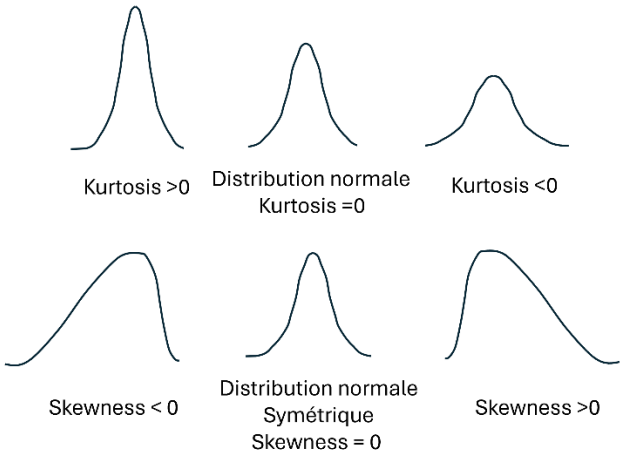


Figure 7 : Quelques paramètres de premier ordre : kurtosis, quantifiant de l'aplatissement de la courbe et skewness quantifiant l'asymétrie de la courbe. (inspiré de (34))

- les paramètres de deuxième ordre, quantifient l'organisation spatiale d'une image en analysant la relation spatiale d'une paire de pixels ou de voxels (

Figure 8). Ils déterminent la fréquence à laquelle un pixel d'intensité x se trouve dans une certaine relation spatiale avec un autre pixel d'intensité y. Ils sont basés sur des mesures de fréquence ou cooccurrence, calculées à partir de matrices spatiales des niveaux de gris.

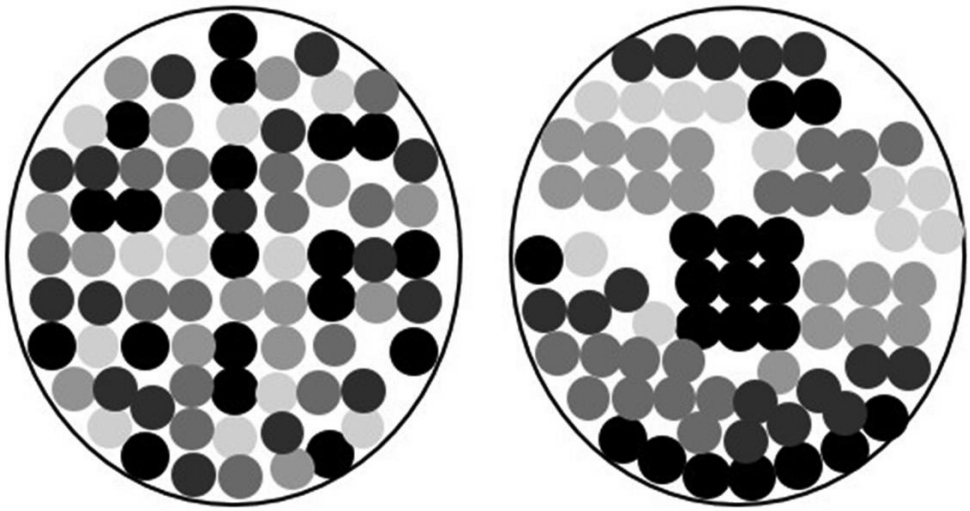


Figure 8 : Intérêt des paramètres de radiomique de 2e ordre ou supérieur. Les deux cercles contiennent chacun autant de pixels blancs, gris clairs, gris intermédiaires, gris foncés et noirs, et donc des paramètres de radiomique de premier ordre identiques. Mais il est clair visuellement que l'organisation spatiale au sein des deux images est bien différente. Les paramètres de second ordre et d'ordre supérieur permettent de quantifier cette différence d'organisation. (figure issue de (34))

Il existe différentes matrices :

-La matrice de cooccurrence (GLCM : Grey-level Cooccurrence Matrix), qui analyse l'entropie (caractère aléatoire), l'énergie (uniformité), le second angular moment (caractère répétitif), l'homogeneity

(homogénéité), la dissimilarity (mesure des différences) et la corrélation (mesure des dépendances linéaires).

- La matrice des longueurs des séries homogènes (GLRLM Grey-level Run Length Matrix) et la matrice des longueurs des zones homogènes (GLZLM Grey-level Zone Length Matrix) analysent la texture dans une direction définie (Figure 9) Un « run » est une longueur de pixels consécutifs avec la même intensité de niveaux de gris dans une direction prédéfinie. Les relations entre les longueurs de parcours donnent naissance à la texture. La texture fine présente plus de longueurs courtes avec des intensités de niveaux de gris similaires, tandis que la texture grossière présente plus de longueurs longues avec des intensités de niveaux de gris différentes. Les paramètres de la GLRLM comprennent l'emphase à court terme (SRE Short-Run Emphasis : mesure la distribution des courts tirages), l'emphase à long terme (LRE Long-Run Emphasis), la non-uniformité des niveaux de gris (GLNU Grey-level Non-Uniformity : mesure la similarité des valeurs des niveaux de gris), et la non-uniformité des longueurs de tirage (GRLNU Grey-level Length Non-Uniformity : mesure la similarité des longueurs des parcours, elle est faible si les longueurs des parcours sont similaires).

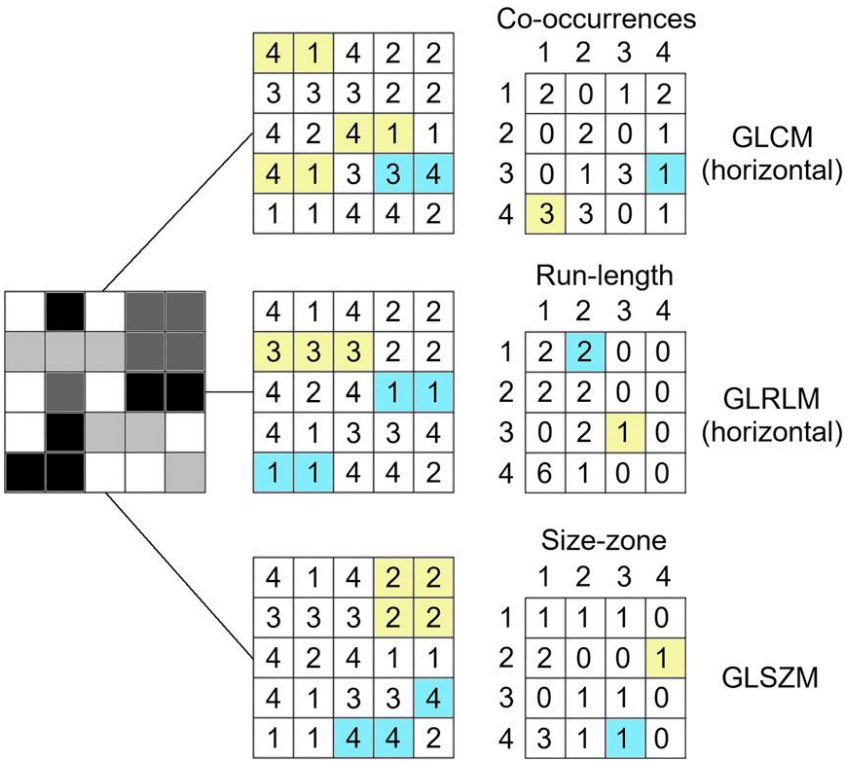


Figure 9 : Exemple de calcul de paramètres de second ordre. Le GLCM est basé sur un décompte des paires de pixels, le GLRLM relit des « runs » (ou longueurs) et le GLSZM analyse des zones. Figure issue de (35)

- Les paramètres d'ordre supérieur analysent la répartition et la relation spatiale entre 3 pixels ou plus, et sont calculées à l'aide de la matrice de différence de niveaux de gris (NGLDM Neighborhood Grey-level Difference Matrix), qui fournissent par exemple des valeurs de contraste (variation locale d'intensité dans l'image), de coarseness (taux spatial de variation d'intensité) et de busyness (fréquence de variation des niveaux de gris). Les paramètres de0 texture d'ordre supérieur permettent d'évaluer les pixels dans leur contexte local, en tenant compte de la relation avec les pixels voisins.

On note qu'il peut exister d'autres classifications des paramètres de radiomique (36).

Schémas classiques d'utilisation de la radiomique

Nous allons lister les étapes habituelles dans la mise en œuvre d'une étude de radiomique.

Détermination d'un but clinique

Il s'agit de déterminer la question clinique pour laquelle la radiomique peut apporter une solution. Il peut s'agir d'une question de caractérisation tumorale (37), de pronostic du patient (38,39), de prédiction d'une mutation tumorale (40,41)...

Acquisition des images

Une analyse de radiomique peut être fait sur tout type d'image, même non médical. A fortiori, elle peut être réalisée sur toute modalité d'imagerie médicale. Néanmoins les variations des paramètres d'acquisition (kVp, épaisseur de coupe, méthode de reconstruction, filtre de reconstruction ... pour le scanner (42,43) et TR, TE épaisseur de coupe, résolution spatiale dans le plan... pour l'IRM (44)) influent sur les valeurs de radiomique, ce qui oblige à des précautions sur les études multicentriques.

Segmentation des zones d'intérêt

Celle-ci peut être manuelle, semi automatique ou automatique.

Pré traitement des images

Plusieurs techniques de pré traitement des images sont possibles, toutes dans but de rendre comparables les paramètres de radiomique entre les patients notamment de différents centres

-discrétisation : c'est-à-dire la transformation de l'intensité des voxels de l'image pour qu'ils prennent un nombre limité de valeur. Il existe une discrétisation relative, où l'on définit un nombre de valeurs d'intensité possibles (« fixed-bin number », à privilégier pour une modalité d'imagerie à intensité calibrée comme le scanner ou la TEP TDM (45)) et une discrétisation absolue où l'on fixe l'intervalle d'intensité de signal entre deux valeurs (« fixed-bin size » à privilégier pour une modalité dont les intensités sont arbitraires comme l'IRM (45))

- rééchantillonnage : il s'agit de modifier la taille des voxels ou pixels, quitte à utiliser des méthodes d'interpolation

-normalisation d'intensité : plusieurs méthodes ont été proposées : la correspondance d'histogramme (46), la normalisation par le score Z qui est fréquemment utilisée, la normalisation par la valeur moyenne d'un organe, ou des méthodes de normalisation par apprentissage profond (47).

- correction du biais du champ de l'IRM : la méthode la plus utilisée est N4ITK (48), d'autres méthodes mathématiques (sur la distribution de gradient, expectation maximization, fuzzy C means) existent avec aussi le développement de méthodes par réseaux de neurones antagonistes génératifs (generative adversarial networks ou GAN, (49))

Extraction des paramètres de radiomique

Plusieurs logiciels (au moins 14) existent pour l'extraction de paramètres de radiomique, les plus fréquemment utilisés étant PyRadiomics (50), MaZda (51), LifeX (52) et IBEX (53).

Certains logiciels de radiomique proposent de plus d'appliquer un filtre à l'image avant d'extraire les paramètres. Cela n'a pas pour but d'harmoniser les images entre elles mais de multiplier le nombre de paramètres extraits par le nombre de filtres utilisés. On peut ainsi avec Pyradiomics obtenir une liste de paramètres sur l'image, puis une seconde liste avec une première combinaison par filtre d'ondelettes….

Le nombre de paramètres fournis par logiciel est variable, LifeX fournissant par exemple 46 paramètres, PyRadiomics pouvant fournir 107 paramètres par type de filtre, ce qui peut amener à extraire plus de 1700 paramètres (54).

Post traitement des paramètres de radiomique

Il est possible, toujours dans le but de rendre comparables les paramètres de radiomique entre les patients notamment de différents centres de réaliser des opérations sur les paramètres de radiomique.

On citera la famille la plus utilisée de méthode qui est ComBat, qui utilise les méthodes empiriques de Bayes pour corriger l'effet de « batch » initialement décrit en génomique et qui décrit la variation induite par l'obtention des données par différents techniciens dans différents laboratoires à des jours différents. La méthode ComBat (42,44) a été déclinée avec plusieurs améliorations possibles (55–57)

Il est possible aussi de normaliser les paramètres de radiomique afin que tous les paramètres soient entre 0 et 1 ou de standardiser les paramètres afin d'avoir une moyenne à 0 et un écart type à 1.

Sélection des paramètres

Avec un nombre élevé de paramètre, il y a un risque de surapprentissage dans les analyses (58) ; par conséquent, la dimensionnalité doit être réduite en priorisant les paramètres. Cela peut être fait en enlevant notamment les données redondantes, puis en ne sélectionnant que les paramètres les plus pertinents pour le but souhaité. Il existe de nombreuses méthodes de sélection des paramètres couramment utilisées, comme on le verra dans la 3^{e} étude (p 64)

Modèle de classification – combinaison avec d'autres données

Une fois le nombre de paramètre diminué, il est possible de réaliser l'exploration de données (data mining) pour mettre en avant les relations entre les paramètres de radiomique et l'objectif clinique. Il peut s'agir de méthodes statistiques classiques, de classificateur par apprentissage machine (machine learning), ou bien de méthodes de type intelligence artificielle plus avancées. Le but est d'avoir une relation entre un ou des paramètres de radiomique et le but clinique ou même souvent de pouvoir prédire cet objectif clinique.

Il est possible à cette étape d'ajouter d'autres types de données à celles de radiomique : cliniques, biologiques, génomiques….

Seconde étude : Analyse par radiomique du scanner des muscles paravertébraux comme facteur pronostique des néoplasies ORL localement avancées

Après avoir présenté la radiomique et ses méthodes d'exploitation, nous vous présentons une première étude d'application de la radiomique.

Une équipe locale de chirurgiens ORL, oncologues, radiothérapeutes et radiologues s'est intéressé aux facteurs pronostiques de néoplasies ORL localement avancées d'un point de vue nutritionnel. Une première étude a mis en évidence des paramètres cliniques et biologiques nutritionnels en lien avec la survie et la toxicité du traitement (59). Une seconde étude a montré que la mesure de masse musculaire au scanner était fortement prédictive de la toxicité du traitement (60).

L'hypothèse de départ de cette nouvelle étude sur la même population était que l'analyse par radiomique des muscles axiaux pouvait quantifier, au-delà de la masse de muscle, des qualités du muscle qui sont liées le pronostic du patient.

Cet article est prévu pour soumission à Journal of Cachexia, Sarcopenia and Muscle (IF 9.4, rang A).

CT radiomic analysis of paraspinal muscles in the prognosis of head and neck advanced cancers

Rémi Thomas-Monier [1], Alexane Lere-Chevaleyre [2], Bruno Pereira [3], Julian Biau [4], Maureen Bernadach[5], Lucie Cassagnes [1,6], Nicolas Saroul [2,7], Benoît Magnin [1,6,8]*

1 Radiology Department, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

2 Otolaryngology—Head and Neck Surgery Department, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

3 Biostatistics Unit, DRCI, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

4 Radiation Oncology Department, Centre Jean Perrin, Clermont-Ferrand, France

5 Medical Oncology Department, Centre Jean Perrin, Clermont-Ferrand, France

6 Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, Clermont-Ferrand, France

7 University of Clermont Auvergne, CHU-Clermont-Ferrand, INRAE, UNH, Clermont-Ferrand, France

8 DI2AM, DRCI, Clermont University Hospital, Clermont-Ferrand, France

* corresponding author:
Benoît Magnin
bmagnin@chu-clermontferrand.fr
Phone: 00 33 4 73 75 02 44
Fax: 00 33 4 73 75 02 39

Abstract

Background: Sarcopenia is a well-known risk factor for its poor outcome in neoplastic diseases. Body composition via CT scan is a common method for evaluating this condition. Radiomics, an automated image analysis technique that has proven effective in establishing prognostic criteria across several fields, has never been applied to the evaluation of axial muscles for predicting outcomes in head and neck cancers. Our primary objective was to determine whether radiomic analysis of the paravertebral muscles using CT imaging could aid in predicting survival in patients with locally advanced head and neck cancers.

Methods: We retrospectively included 71 patients treated for locally advanced head and neck cancer with induction chemotherapy at our center. Radiomic parameters were extracted after manual segmentation of the paravertebral muscles at the L1 level on CT scan. Only those parameters that were invariant with respect to injection timing and demonstrated high intra-observer reproducibility were retained. To explore associations between these parameters and survival, univariate and multivariate Cox regression were performed; toxicity, and treatment response were explored with Student's t-test or Mann-Whitney test and multivariate logistic regression.

Results: A total of 21 radiomic parameters were retained for analysis. None of the parameters were significantly associated with survival in multivariate analysis. However, the possibility of oral feeding at diagnosis and the histogram radiomic feature discretized HU sum were the most promising factors. Notably, the latter allowed for the separation based on its median value of patients into two groups with significantly different survival outcomes (Cox regression, $p = 0.02$). None of the parameters were significantly associated with toxicity in multivariate analysis. However, the CT subcutaneous fat index and the second-order radiomic feature GLRLM SRE showed a trend as a risk factor for toxicity.

Conclusions: No parameter, including radiomic features, was able to statistically and independently demonstrate prognostic value for locally advanced head and neck cancers. However, trends were observed with two radiomic parameters, one related to survival and the other to prognosis, which appear promising when used in conjunction with CT body composition parameters and clinico biological nutritional parameters.

Keywords

radiomics; body composition; sarcopenia; otorhinolaryngologic neoplasms

Abbreviations

CT AF index: CT abdominal fat index

CT SCF index: CT subcutaneous fat index

GLRLM SRE: grey-level run length matrix short-run emphasis

GLRLM SRLGE: grey-level run length matrix short-run low gray-level emphasis

GLZLM ZLNU: grey-level zone length matrix zone length non-uniformity

GLZLM ZP: grey-level zone length matrix zone percentage

HNC: head and neck cancer

UICC/AJCC: Union for International Cancer Control/American Joint Committee on Cancer

Introduction

Head and neck cancers (HNC) are common cancers (6th most prevalent worldwide, with approximately 650,000 new cases annually) and are responsible for around 330,000 deaths each year. The main risk factors are predominantly alcohol and tobacco consumption, as well as infections from HPV viruses. Patients with HNC frequently suffer from malnutrition at the time of their care. This malnutrition results from reduced food intake caused by the tumor (dysphagia, odynophagia, anorexia) in a context of underlying nutritional risk (depression, alcoholism), combined with an increased metabolic demand (cancer-related hypermetabolism [1]). This cancer-related malnutrition is therefore a multifactorial syndrome characterized by a loss of muscle mass, potentially accompanied by a loss of fat mass [2]. It is estimated that around 50% of patients with HNC have an altered nutritional status, and approximately 60% of patients experience muscle mass loss at the time of their care [3]. It is now well recognized that low muscle mass, and more recently a change in muscle structure (muscle infiltration by lipid droplets, known as myosteatosis), are independent prognostic factors in the management of HNC [4–6].

Sarcopenia is defined as low muscle mass. In most clinical studies, it is assessed by measuring the muscle cross-sectional area at the third lumbar vertebra level (L3) on a CT scan, adjusted for the patient's height [7]. However, CT scan standard assessment for HNC does not cover L3, and other CT scan levels have shown their relevance in HNC management [8]. Our team, for instance, demonstrated the value of using a cross-sectional level at the first lumbar vertebra (L1) to predict complications following neoadjuvant chemotherapy in patients with HNC [9].

Myosteatosis is defined as low radiodensity due to the infiltration of fat into skeletal muscle [10]. It has been reported that this structural change in skeletal muscle is itself a prognostic factor in the management of HNC [11]. Furthermore, the association of low muscle mass with structural changes in muscle appears to compound the prognosis for patients [12]. A simple assessment of muscle mass therefore seems insufficient to understand the role of skeletal muscle in the prognosis of cancer patients.

Radiomics is an image analysis method that allows for the evaluation of parameters that the human eye cannot perceive. It has demonstrated its value in tumor characterization, allowing for the determination of prognostic factors in HNC [13, 14] or many other malignancies [15, 16].

It therefore seemed relevant for us to perform a radiomic analysis of the paravertebral muscles to investigate whether radiomic parameters of muscle could serve as prognostic markers in the progression of neoplastic disease, and whether they could provide more information than conventional methods of muscle mass estimation. Radiomic analysis of body composition has been recently used in the prognosis of some malignancies [17–19] or acute pulmonary embolism [20], but never in HNC.

Our main objective was to determine whether radiomic analysis of the paravertebral muscles from CT scans aids in predicting survival in locally advanced head and neck neoplasms.

Our secondary objectives were :

- to determine whether radiomic analysis of the paravertebral muscles from CT scans aids in predicting treatment response or toxicity in locally advanced head and neck neoplasms.
- to determine whether these radiomic parameters have a better predictive power for survival than measurements of muscle cross-sectional area
- to determine if there is a correlation between radiomic parameters of the paravertebral muscles and the patient's nutritional status.

Materials and methods

Population

We conducted a single-center retrospective study at the CHU Gabriel-Montpied in Clermont-Ferrand. A local ethics committee approved this study. We included all patients treated advanced HNC with radiotherapy and induction chemotherapy between July 2009 and January 2018. The indications for this induction treatment were either an initially inoperable tumor or for laryngeal preservation. Exclusion criteria included nasopharyngeal or paranasal sinus tumor locations, metastatic disease, and histology other than squamous cell carcinoma. These patients have already been the subject of two studies [9, 21], which focused on the central role of nutritional status, particularly muscle mass, in the risk of toxicity during chemotherapy.

The specific exclusion criteria for our study included the absence of a pre-treatment CT scan or an inappropriate scan injection timing, as determined by a preliminary study.

The following baseline clinical and biological data were collected: age at diagnosis, gender, height, weight at the start of treatment, BMI at the start of treatment, weight loss at the start of treatment, albumin level, ECOG status, UICC stage of cancer and possibility of oral feeding at diagnosis. Three events were defined concerning our objectives:

- *Overall Survival (OS),* with the date of death or, if not available, the date of the last follow-up.
- *Response,* assessed via imaging at the end of treatment according to RECIST 1.1. Treatment response was categorized into two groups: a responder group combining partial and complete responses, and a non-responder group combining stable disease and progression.
- *Toxicity*, defined as the presence of renal, hematological (anemia, neutropenia, thrombocytopenia), infectious, or digestive toxicity (nausea, diarrhea, mucositis, gastrointestinal hemorrhage) of at least grade 3 during treatment.

CT body composition data

The CT body composition data were collected using the SliceOmatic software from a previous study on this cohort at the L1 level of the pre-treatment scan, divided by the square of the patient's height, thus obtaining a CT muscular index, a CT abdominal fat index (CT AF index), and a CT subcutaneous fat index (CT SCF index) [9].

Segmentation and extraction of radiomic features

On the population of interest, a key image was extracted from the pre-treatment CT scan at the level of the L1 vertebra. Manual segmentation was then performed by a radiology resident using Life-X 4.0 software [22], involving the segmentation of two regions of interest (ROIs):

- Manual segmentation of all paravertebral muscles.
- Segmentation of a circular ROI within the aortic lumen, which was as large as possible without encroaching on the aortic wall and its calcifications to reflect the injection timing.

For each of these segmentations and for each patient, 80 radiomic parameters were extracted using LifeX software with its default settings. Additionally, 44 parameters were calculated by normalizing the values of the 44 parameters extracted by LifeX to the mean value in the aorta (to compensate for the effect of injection timing). The parameters are listed in Supporting Information 1. In total, 125 parameters were extracted per patient. ComBat was used to compensate for the variability of the radiomic parameters due to differences scans performed in different centers [23].

In light of the heterogeneity of acquisition times, we conducted a preliminary study to retain among the 125 extracted parameters only the parameters invariant to injection times. This study is presented in Supporting Information 1.

Intra-observer reproducibility

To assess intra-observer reproducibility and the robustness of the extracted radiomic parameters, 10 patients (14%) were re-segmented a second time, with an interval of more than 3 months. The same radiomic parameters were extracted a second time. Lin's concordance correlation coefficient (CCC) [24] was calculated. Only parameters with a CCC > 0.9 were retained.

Statistical analysis

A Shapiro test was performed to determine the normal distribution of the data. Differences in quantitative parameters based on the events (toxicity and response) were analyzed in a exploratory manner for the radiomic parameters, CT body composition, and clinical and biological parameters using either the Student's t-test or the Mann-Whitney test, depending on the normality of the data. A Chi-square test was used for categorical variables.

Survival was studied using a Cox regression model on the radiomic parameters and sarcopenia parameters, focusing on overall survival. For parameters significantly associated with survival, patients were divided into two groups based on the median value of that parameter, and a significant difference in survival between the two groups was sought.

A multivariate regression (Cox for survival and logistic for toxicity) was conducted using covariates based on univariate results and their clinical relevance.

Statistical analyses were performed using Stata software (version 15, StataCorp, College Station, United States) and R (version 4.3.2, R Core Team, Vienna, Austria [25]). All statistical tests were carried out based on a two-sided type I error at 5%.

Results

Population selection

A total of 92 patients were screened; 13 were excluded because the baseline CT did not explore L1. The acquisition times for the remaining 79 patients were heterogeneous (3 patients with a non-contrast scan, 71 patients with a contrast-enhanced scan at arterial or portal timing, and 5 patients with a delayed phase scan). In light of the results of the preliminary study (Supporting Information 1), we decided to exclude the 8 patients whose baseline scan was performed with an injection protocol other than arterial or portal. In total, 71 patients were included in the study (Figure 1), with CT scans coming from 13 different centers.

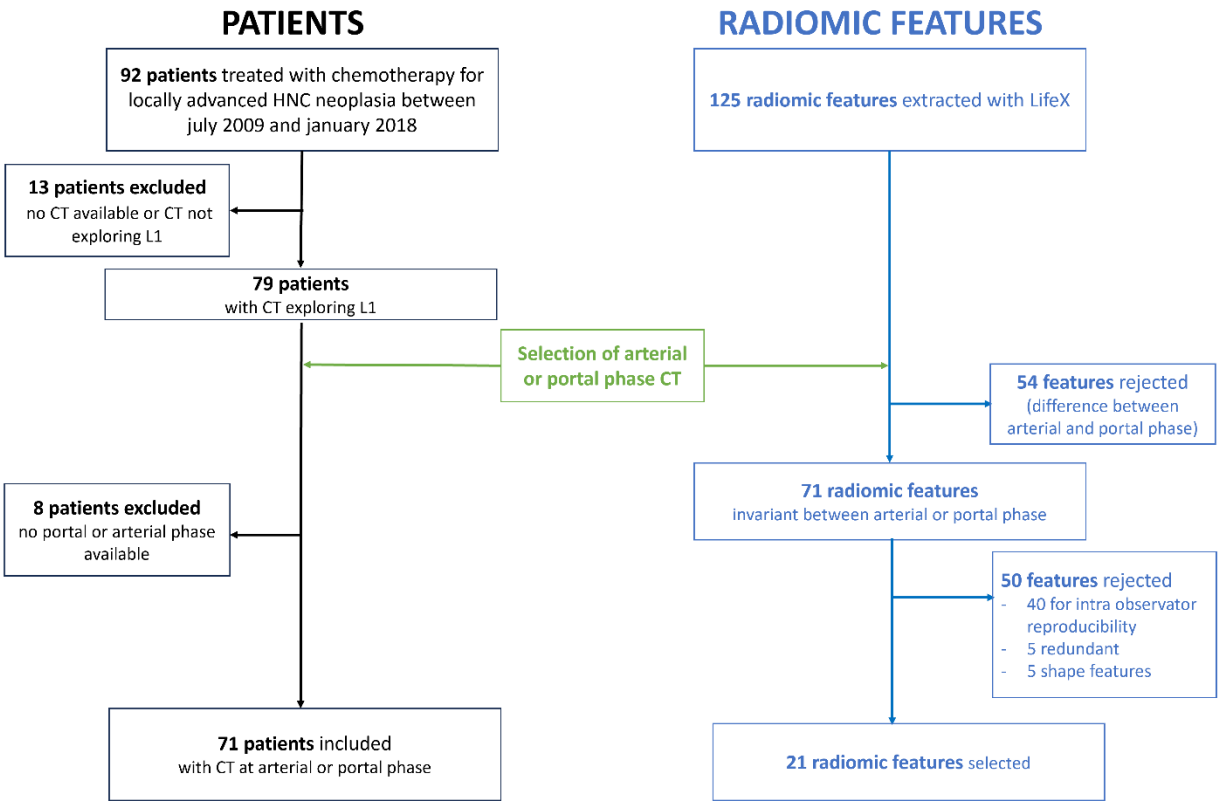


Figure 1: Flowchart of the study

Final selection of radiomic features

Among the 71 features invariant to the arterial or portal acquisition times, 31 were deemed reproducible between two segmentations (CCC > 0.9) and were retained for analysis. Among these 31 texture features, 5 were shape features that were not retained as they pertained to muscle mass measurement. Additionally, 5 features were redundant and were removed, resulting in a final total of 21 features, including 17 histogram features and 4 second-order features (Figure 1 and Supporting Information 1).

Clinical characteristics of the population

The characteristics of the population at inclusion and at follow up are summarized in Table 1. Five patients did not have an evaluable response after treatment and were considered lost to follow-up; however, all of these patients had died before the end of the follow-up.

						Total (n=71)
BASELINE						
Age (years, mean ±SD)			56 ± 7.5		**TNM stage: T**	n (%)
Male n (%)			58	T1		2 (3%)
BMI (kg/m^{2,} mean ±SD)			22.3 ± 3.8	T2		8 (11%)
Performance status n (%)				T3		37 (52%)
	0		36 (51%)	T4a		20 (30%)
	1		31 (44%)	T4b		3(4%)
	2		4 (5%)	Tx		1 (1%)
Smoker, n (%)			66 (93%)		**TNM stage: N**	n (%)
Alcohol consumption, n (%)			58 (82%)	N0		12 (17%)
UICC Stage			n (%)	N1		9 (13%)
	II		2 (3%)	N2a		2 (3%)
	III		20 (30%)	N2b		12 (17%)
	IVa		33 (65%)	N2c		21 (30%)
	IVb		16 (22%)	N3		15 (21%)
FOLLOW UP						
Toxicity ≥ grade 3						40 (56%)
Response						
	No response group					15 (21%)
		Progressive disease		7 (10%)		
		Stable disease		8 (11%)		
	Response group					51 (72%)
		Partial response		11 (15 %)		
		Complete response		40 (56%)		
	Unknown					5 (7%)
Follow-up (months)						
	Minimum			0		
	Mean			29		
	Maximum			85		
Death						37 (52%)
	1- year mortality		0.29	IC [0.15; 0.36]		
	5-year mortality		0.78	IC [0.64; 0.89]		

Table 1: Clinical data. Results are presented as n (%)

Survival analysis

Univariate analysis results for survival are presented in Table 2.

		HR	CI	p
Radiomics	HU SD	0,980	[0,9246;1,0385]	0,494
	Discretized HU Sum	1,000	[0,9999;1]	**0,016**
	Normalized HU mean	2,447	[0,1344;44,5499]	0,546
	Normalized HU Q1	3,974	[0,1609;98,178]	0,399
	Normalized HU Q2	2,048	[0,1135;36,9707]	0,627
	Normalized HU Q3	1,497	[0,1323;16,9399]	0,745
	Normalized HU SD	0,239	[0,0004;127,2753]	0,655
	Normalized HU Sum	1,000	[0,9988;1,0006]	0,483
	Normalized discretized HU min	2,845	[0;716724,9362]	0,869
	Normalized discretized HU mean	0,900	[0;63858,9125]	0,985
	Normalized discretized HU SD	0,000	[0;4,E+135]	0,930
	Normalized discretized HU max	1,296	[0;40301,5522]	0,961
	Normalized discretized HU Q1	0,762	[0;69564,3614]	0,963
	Normalized discretized HU Q2	1,123	[0;67038,6109]	0,983
	Normalized discretized HU Q3	0,843	[0;47545,5675]	0,976
	Normalized discretized Inner Rim HU min	0,665	[0;57821,2355]	0,944
	Normalized discretized Inner Rim HU mean	0,928	[0;60990,2038]	0,989
	GLRLM SRE	6,638	[0,0118;3738,8736]	0,558
	GLRLM SRLGE	6E+117	[0;Inf]	0,766
	GLZLM ZLNU	0,990	[0,971;1,0089]	0,293
	GLZLM ZP	94,928	[0;49858117950,93]	0,657
CT	CT SCF Index	0,988	[0,968;1,0085]	0,252
	CT Muscle Index	0,968	[0,9273;1,0097]	0,130
	CT AF Index	0,983	[0,9694;0,9969]	**0,016**
Clinical	Age	0,969	[0,9288;1,0113]	0,149
	Height	0,997	[0,9467;1,0489]	0,893
	Usual weight	0,984	[0,9556;1,0125]	0,264
	BMI at treatment initiation	0,894	[0,8146;0,9815]	**0,019**
	Weight at treatment initiation	0,971	[0,9444;0,9986]	**0,039**
	Weight loss	0,946	[0,9087;0,9855]	**0,008**
	Albumin	0,930	[0,8908;0,9718]	**0,001**
	ECOG	1,377	[0,7792;2,4331]	0,271
	Gender	0,838	[0,3658;1,9193]	0,676
	UICC Stage	1,830	[1,1567;2,8941]	**0,010**
	Oral feeding possible at diagnosis	0,219	[0,1107;0,433]	**1,271E-05**

Table 2: Univariate analysis of survival

After removing weight, which is redundant in relation to BMI, a multivariate regression model was conducted (Figure 2A)

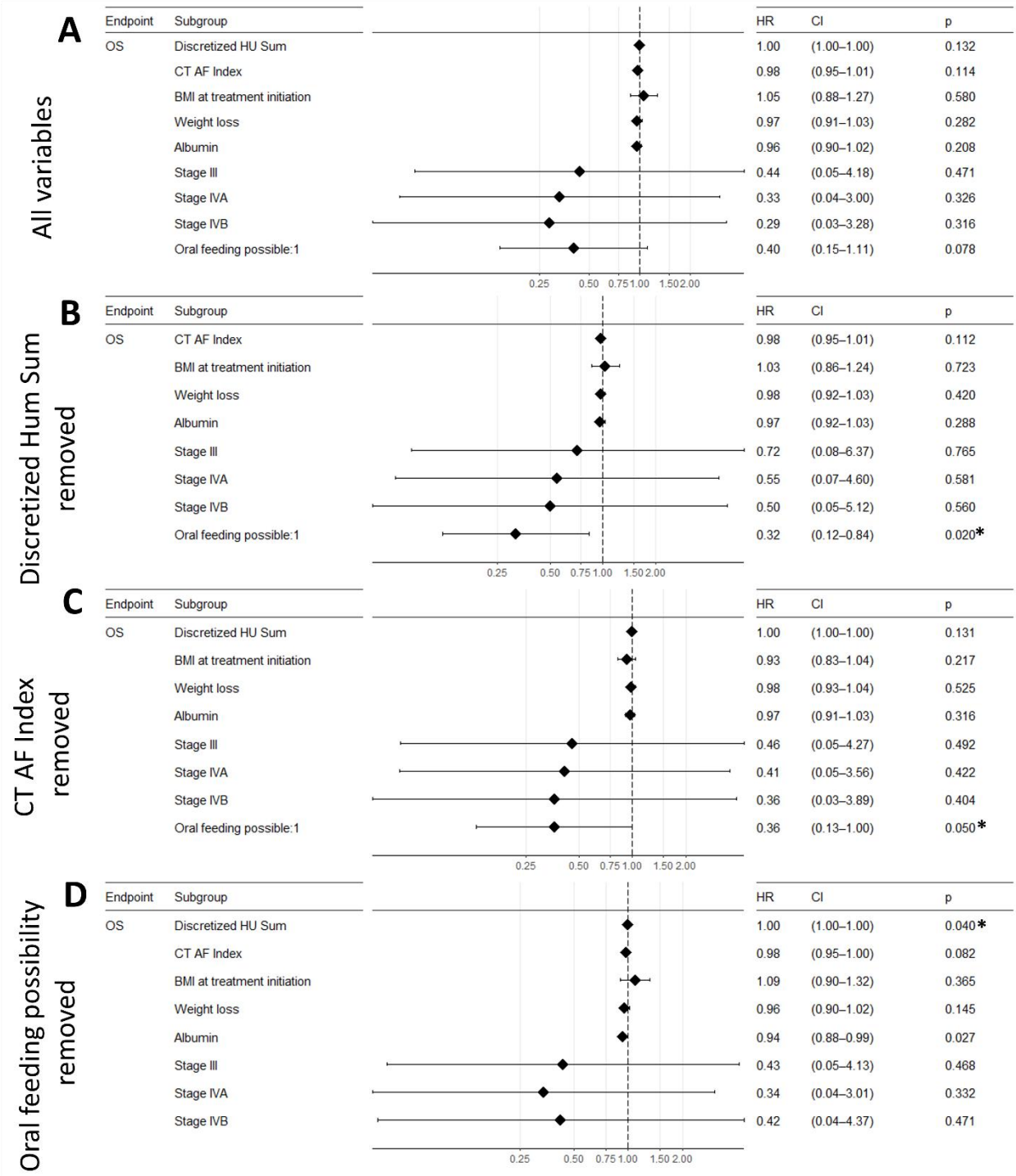


*Figure 2: Multivariate analysis of survival (A) and multivariate analysis of survival removing one variable: removing Discretized HU Sum (B), removing CT AF Index (C), removing oral feeding possibility (D). Significative variables are marked with a *.*

Given the absence of significant parameters identified in the multivariate analysis, we then performed a multivariate analysis by removing one of the three parameters closest to significance (Discretized HU Sum, CT AF Index and oral feeding possibility, Figure 2B to D)

We tested the difference in survival of the population split according to the median value of each quantitative variable assessed in the multivariate analysis (Figure 3)

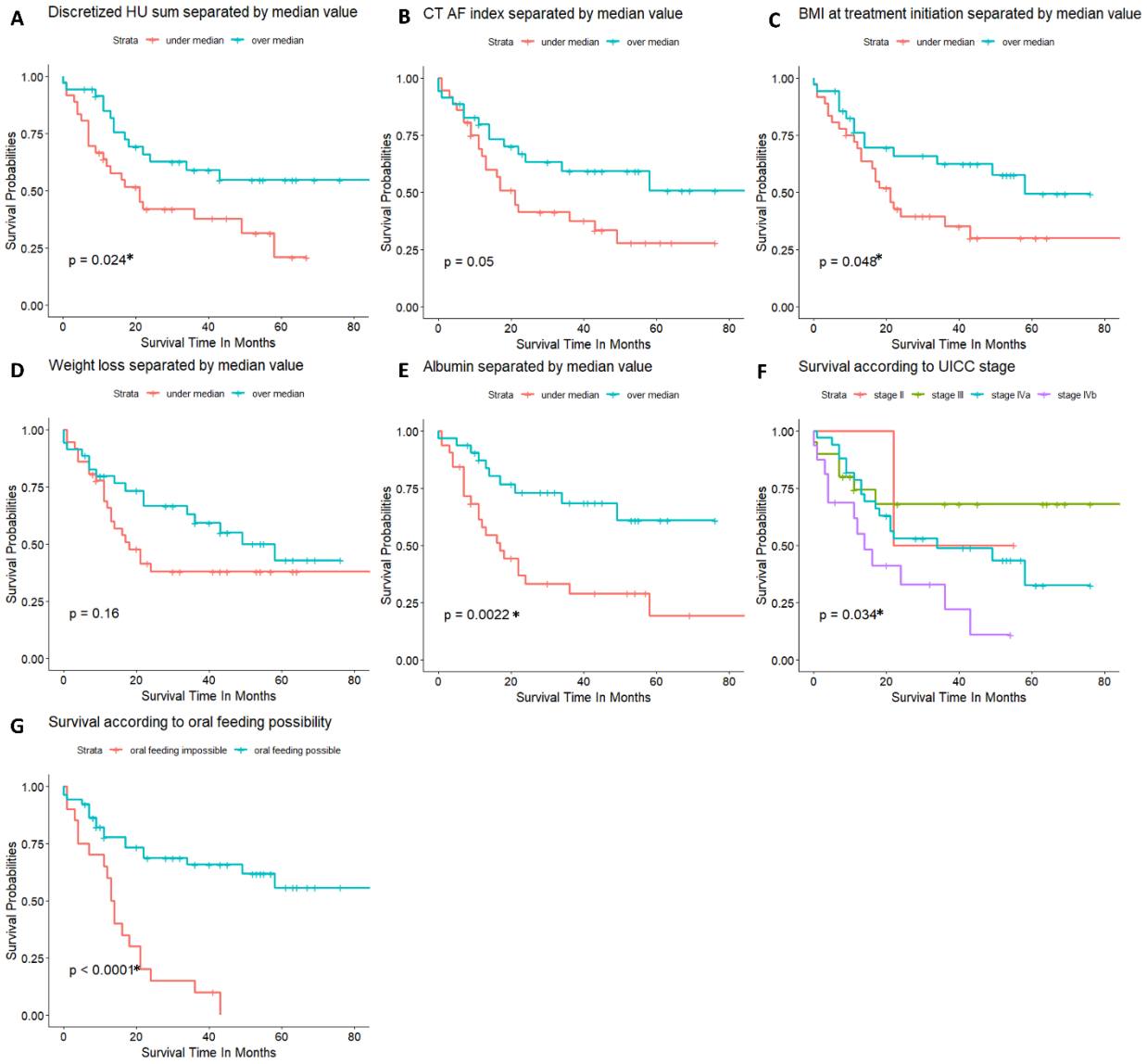


*Figure 3: Survival curve of population separated according to the median value of each quantitative variable of the multivariate analysis and by its category for categorical variables. Significative differences are marked with a *.*

Toxicity analysis

Univariate analysis results for toxicity are presented in Table 3.

			No toxicity	Toxicity	p
Radiomics	HU SD		31.47 +/- 5.01	34.31 +/-5.70	**0.032**
	Discretized HU Sum		57622.02 [48742;66838]	50983.58 [42890;58772]	0.078
	Normalized HU mean		0.19 [0.17;0.21]	0.19 [0.16;0.24]	0.745
	Normalized HU Q1		0.12 [0.09;0.15]	0.12 [0.09;0.15]	0.889
	Normalized HU Q2		0.2 [0.18;0.22]	0.21 [0.18;0.25]	0.397
	Normalized HU Q3		0.27 [0.24;0.31]	0.29 [0.26;0.36]	0.291
	Normalized HU SD		0.13 [0.11;0.15]	0.15 [0.12;0.17]	0.076
	Normalized HU Sum		579.67 [448.27;820.66]	583.03 [475.15;749.55]	0.719
	Normalized discretized HU min		0.06 [0.05;0.08]	0.06 [0.06;0.08]	0.391
	Normalized discretized HU mean		0.07 [0.06;0.09]	0.07 [0.06;0.09]	0.372
	Normalized discretized HU SD		0 [0.0020;0.0029]	0 [0.0022;0.0032]	0.123
	Normalized discretized HU max		0.08 [0.07;0.09]	0.08 [0.07;0.1]	0.417
	Normalized discretized HU Q1		0.07 [0.06;0.08]	0.07 [0.06;0.09]	0.342
	Normalized discretized HU Q2		0.07 [0.06;0.09]	0.07 [0.06;0.09]	0.378
	Normalized discretized HU Q3		0.07 [0.06;0.09]	0.08 [0.07;0.09]	0.36
	Normalized discretized Inner Rim HU min		0.07 [0.06;0.09]	0.07 [0.06;0.09]	0.366
	Normalized discretized Inner Rim HU mean		0.07 [0.06;0.09]	0.08 [0.07;0.09]	0.313
	GLRLM SRE		0.51 +/-0.0513	0.54 +/-0.0523	**0.045**
	GLRLM SRLGE		0.001 +/-0.0002	0.002 +/-0.0002	**0.031**
	GLZLM ZLNU		32.19 [25;50.84]	34.89 [29.45;46.55]	0.954
	GLZLM ZP		0.06 +/-0.0155	0.06 +/-0.0185	0.19
CT	CT SCF Index		18.38 [6.91;22.06]	25.19 [17.88;40.95]	**0.002**
	CT Muscle Index		37.91 +/-5.7807	35.17 +/-8.01	0.099
	CT AF Index		25.12 [9;43.6]	31.18 [19.93;56.58]	0.151
Clinical	Age (years)		56 [51.5;58.5]	59 [53;64.25]	**0.037**
	Height (cm)		169.94 +/-6.8796	170.5 +/-6.8985	0.733
	Usual weight (kg)		67.29 +/-10.9824	71.25 +/-11.7795	0.149
	BMI at treatment initiation		21.74 +/-3.6831	22.76 +/-3.833	0.261
	Weight at treatment initiation (kg)		63.04 +/-12.5836	66.35 +/-12.862	0.28
	ECOG				0.674
		ECOG 0	17 (54.8)	19 (47.5)	
		ECOG 1	13 (41.9)	18 (45.0)	
		ECOG 2	1 (3.2)	3 (7.5)	
	Male gender		29 (93.5)	29 (72.5)	**0.049**
	UICC Stage				0.578
		II	0 (0.0)	2 (5.0)	
		III	8 (25.8)	12 (30.0)	
		Iva	16 (51.6)	17 (42.5)	
		Ivb	7 (22.6)	9 (22.5)	

Table 3: Toxicity related to clinicopathological characteristics. Data are presented as number of patients (associated percentages), mean ± standard deviation or median [interquartile range].

Due to a strong correlation between HU SD, GLRLM SRE, and GLRLM SRLGE (Supporting Information 3), we retained only one of these three variables for the multivariate analysis. The results of the multivariate analysis for toxicity are presented in Figure 4.

Variable		N	Odds ratio		p
GLRLM SRE		71		18232.91 (0.61, 1991242545.56)	0.08
CT SCF Index		71		1.04 (1.00, 1.09)	0.07
Gender	Female	13		Reference	
	Male	58		0.32 (0.04, 1.80)	0.22
Age		71		1.05 (0.97, 1.13)	0.21

Figure 4: Multivariate analysis of toxicity

Response analysis

Univariate analysis results for toxicity are presented in Supporting Information 4. No radiomic feature showed a difference based on the response. In the absence of any radiomic feature associated with the response, we did not conduct a multivariate analysis

Correlation between radiomic features and nutritional parameters

Correlation between radiomic and CT body composition data with clinical nutritional variables is shown in Table 4.

		Weight loss	Weight	BMI	Albumin
Radiomics	HU SD	-0,12	0,069	0,069	0,196
	GLRLM SRE	-0,135	0,104	0,056	**0,017***
	GLRLM SRLGE	-0,05	0,168	0,142	**0,026***
	GLZLM ZLNU	0,017	**0,301***	0,233	0,271
	GLZLM ZP	-0,005	**0,284***	0,231	0,066
	Discretized HU Sum	0,142	**0,263***	0,217	0,267
	Normalized HU mean	-0,083	-0,001	-0,087	0,589
	Normalized HU Q1	-0,095	-0,092	-0,187	0,625
	Normalized HU Q2	-0,057	0,017	-0,054	0,609
	Normalized HU Q3	-0,035	0,074	0,006	0,745
	Normalized HU SD	0,057	0,233	0,198	0,986
	Normalized HU Sum	0,038	0,162	0,072	0,234
	Normalized discretized HU min	0,083	0,186	0,138	0,418
	Normalized discretized HU mean	0,087	0,193	0,144	0,496
	Normalized discretized HU SD	0,054	0,172	0,136	0,852
	Normalized discretized HU max	0,077	0,197	0,142	0,627
	Normalized discretized HU Q1	0,088	0,194	0,147	0,506
	Normalized discretized HU Q2	0,101	0,203	0,155	0,493
	Normalized discretized HU Q3	0,076	0,173	0,125	0,528
	Normalized discretized Inner Rim HU min	0,095	0,175	0,128	0,42
	Normalized discretized Inner Rim HU mean	0,096	0,193	0,148	0,519
CT	CT SCF Index	**0,361***	**0,567***	**0,754***	0,385
	CT Muscle Index	**0,349***	**0,535***	**0,504***	0,165
	CT AF Index	**0,487***	**0,737***	**0,808***	0,286

*Table 4: Correlation between radiomic and CT variables with clinical and biological nutritional variables. Significative correlations are marked with a *.*

Discussion

We conducted a retrospective study examining the radiomic features of the paravertebral muscles on the pre-therapeutic CT scan of a population diagnosed with locally advanced head and neck neoplasia, prior to induction chemotherapy, to identify prognostic factors for survival, toxicity, and treatment response.

Multivariate analysis for overall survival did not identify any significant parameters. Among the variables closest to significance, as in the previous study [26], is the possibility or not of oral feeding at diagnosis, confirming its crucial role in prognosis. The radiomic feature Discretized HU Sum appears promising, even though it is not significant in the multivariate analysis: it becomes significant when the possibility of oral feeding is excluded, and it splits based on its median the patients in two populations with different survival rates. This variable represents the sum of Hounsfield Units of the paravertebral muscles after a discretization step of the grayscale levels. Therefore, this parameter reflects a combination of muscle mass, as it directly depends on the size of the segmentation, as well as average muscle density, allowing for the assessment of both muscle surface area and myosteatosis. Myosteatosis is defined as abnormal fat infiltration within striated muscle or as muscle changes with intramuscular fat remodeling, which increases with age [27]. Myosteatosis is known to be negatively correlated with muscle mass, muscle strength, and mobility [28]. There is currently no standardized method for assessing myosteatosis, and the most commonly used method is the

measurement of the intramuscular fat index [29]. The only CT body composition data found to be different for survival in univariate analysis is the CT AF index, although it does not allow for the separation of patients into two distinct survival populations based on its median. Among other parameters identified in univariate analysis are those describing nutritional status (BMI, weight loss, albumin), as well as an expected parameter, cancer stage.

The multivariate analysis concerning toxicity did not identify any significant parameter. However, three radiomic features were found in the univariate analysis, including one histogram feature and two second-order features: the standard deviation of mean Hounsfield unit densities (HU SD), and the GLRLM SRE and GLRLM SRLGE features, derived from the gray-level run-length matrix. High values of these three features were predictive of a higher risk of drug toxicity. An increase in the standard deviation of mean densities can be interpreted as a feature reflecting greater muscle heterogeneity. This heterogeneity in muscle densities could indicate an increased proportion of intramuscular fat, and thus, a state of myosteatosis. The two other significant features are second-order texture indicators that depend on the distribution of pixels in the image and are based on the gray-level run-length matrix. They reflect the repetition of identical pixel sequences in multiple spatial directions, either focusing solely on low gray levels (GLRLM SRLGE) or considering all gray levels (GLRLM SRE). The GLRLM SRLGE feature is dependent on the repetition of several sequences of low-intensity pixels (likely intramuscular fat), and an increase in this feature corresponds to a higher repetition of fat pixels in adjacent regions. This may indicate muscle degeneration with a greater number of intramuscular fat islands. The GLRLM SRE feature, on the other hand, depends on the repetition of pixel sequences in various spatial directions but considers all gray levels in the image. This feature is more challenging to interpret, as higher values indicate the repetition of homogeneous areas within the image, regardless of pixel intensity. Its positivity could thus be explained solely by the repetition of low gray-level pixel sequences, similar to the GLRLM SRLGE feature. It should also be noted that there is a strong correlation between these three radiomic features. When including only one of them in the multivariate analysis, GLRLM SRE was close to significance as a risk factor for toxicity. There are differences between the study by Leyre-Chevaleyre et al. [9] and ours in the variables identified in univariate analysis, likely due to the exclusion of certain patients in our study, which caused some variables to either surpass or fall below the significance threshold. In particular, the CT muscle index, which was an independent predictive variable for toxicity in the previous study, was not found to be significant in univariate analysis, though it was close to the threshold.

Additionally, three other parameters were associated with a higher risk of developing severe treatment-related toxicity: age at diagnosis, female sex, and CT SCF index, the latter also showing a trend as a risk factor for toxicity. A low muscle area and increased fat mass are not mutually exclusive and are consistent with the concept of sarcopenic obesity, a well-known risk factor in the literature for poorer outcomes in neoplastic contexts and for an increased risk of drug toxicity during chemotherapy [30, 31]. Sarcopenic obesity could partially explain our results, as well as the higher risk of drug toxicity in patients with low muscle area and/or increased subcutaneous fat layer.

In our study, none of the radiomic features could establish a significant correlation with treatment response. However, two parameters, one clinical and one morphological, were significantly associated with treatment response (Supplementary Information 4). Firstly, higher BMI at diagnosis in responders could be explained by a higher rate of malnutrition among non-responders. More paradoxically, a larger subcutaneous fat area was associated with a better treatment response. Similar findings were observed in a study on advanced gastric cancers, where a larger subcutaneous fat area played a protective role [32]. This observation requires cautious interpretation and further studies, especially since in our study a larger subcutaneous fat area was also correlated with a higher risk of severe toxicity during chemotherapy.

The study of radiomic features and clinico-biological data reflecting the nutritional status of the population shows some statistically significant correlations, but with very modest correlation coefficients. One advanced texture feature (GLZLM ZLNU) is positively correlated with patient weight.

This relationship remains of limited interpretation due to the modest correlations but may suggest that there are connections between the fine texture of striated muscles and the nutritional parameters of patients. Some studies have shown promising results indicating correlations between muscle radiomic features and sarcopenia [33] as well as malnutrition [34]. Our results regarding the correlations between radiomic features and the clinico-biological characteristics of the population appear to be complementary, although the correlations are weak. As expected, based on the literature, there are strong correlations between CT body composition parameters, in particular intra-abdominal and subcutaneous fat indices, and clinico-biological parameters of nutritional status.

To our knowledge, no study has previously used radiomic analysis of axial muscles to get prognostic factors in HNC. Several studies have investigated correlations between axial muscle radiomic data and the presence of sarcopenia [33, 35] or the nutritional prognosis [34] of patients treated for cancer. Several studies have finally highlighted the links between radiomic analysis of axial muscles and the prognosis of neoplasms beside head and neck, such as those of the esophagus or stomach [17, 19, 36], breast [37] or HCC [18].

One of the pitfalls of radiomics studies is their low reproducibility and lack of external validation. Therefore, it is essential to apply a rigorous methodology, particularly in radiomic feature and image selection [38]. One of the limitations in the screened population for this study lies in the lack of harmonization of CT injection times, due to the heterogeneity of radiological practices in the staging of HNC. To ensure a rigorous methodology, we conducted a preliminary study on an additional population to select only those radiomic parameters that remain invariant across the selected injection times. Additionally, we performed a reproducibility study on approximately 15% of the population to retain only the data that are reproducible (CCC > 0.9) between two segmentations. This step is essential to ensure reliable and reproducible data, but it inevitably results in a significant reduction in the number of radiomic image parameters [39], ultimately retaining only robust and reproducible features. Indeed, we were forced to exclude numerous radiomic features (94 out of 125 extracted parameters, 75%) due to variability related to injection timing and intra-observer reproducibility, which led to a significant loss of information. Consequently, some features were not tested; notably, mean HU value of muscles was not tested even though it appears to be a simple metric that could potentially reflect myosteatosis.

In conclusion, no parameter, including radiomic features, was able to statistically and independently demonstrate prognostic value for locally advanced head and neck cancers. However, trends were observed with two radiomic parameters, one related to survival and the other to prognosis, which appear promising when used in conjunction with CT body composition parameters and clinico biological nutritional parameters.

Acknowledgements

None

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Baracos VE, Martin L, Korc M, et al (2018) Cancer-associated cachexia. Nat Rev Dis Primer 4:1–18. https://doi.org/10.1038/nrdp.2017.105

2. Fearon K, Strasser F, Anker SD, et al (2011) Definition and classification of cancer cachexia: an international consensus. Lancet Oncol 12:489–495. https://doi.org/10.1016/S1470-2045(10)70218-7

3. Saroul N, Pastourel R, Mulliez A, et al (2018) Which Assessment Method of Malnutrition in Head and Neck Cancer? Otolaryngol Neck Surg 158:1065–1071. https://doi.org/10.1177/0194599818755995

4. Findlay M, White K, Stapleton N, Bauer J (2021) Is sarcopenia a predictor of prognosis for patients undergoing radiotherapy for head and neck cancer? A meta-analysis. Clin Nutr 40:1711–1718. https://doi.org/10.1016/j.clnu.2020.09.017

5. Hua X, Liu S, Liao J-F, et al (2020) When the Loss Costs Too Much: A Systematic Review and Meta-Analysis of Sarcopenia in Head and Neck Cancer. Front Oncol 9:1561. https://doi.org/10.3389/fonc.2019.01561

6. Wong A, Zhu D, Kraus D, Tham T (2021) Radiologically Defined Sarcopenia Affects Survival in Head and Neck Cancer: A Meta-Analysis. The Laryngoscope 131:333–341. https://doi.org/10.1002/lary.28616

7. Shen W, Punyanitya M, Wang Z, et al (2004) Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol 97:2333–2338. https://doi.org/10.1152/japplphysiol.00744.2004

8. Swartz JE, Pothen AJ, Wegner I, et al (2016) Feasibility of using head and neck CT imaging to assess skeletal muscle mass in head and neck cancer patients. Oral Oncol 62:28–33. https://doi.org/10.1016/j.oraloncology.2016.09.006

9. Lere-Chevaleyre A, Bernadach M, Lambert C, et al (2021) Toxicity of induction chemotherapy in head and neck cancer: The central role of skeletal muscle mass. Head Neck. https://doi.org/10.1002/hed.26954

10. Aubrey J, Esfandiari N, Baracos VE, et al (2014) Measurement of skeletal muscle radiation attenuation and basis of its biological variation. Acta Physiol 210:489–497. https://doi.org/10.1111/apha.12224

11. Findlay M, Brown C, De Abreu Lourenço R, et al (2020) Sarcopenia and myosteatosis in patients undergoing curative radiotherapy for head and neck cancer: Impact on survival, treatment completion, hospital admission and cost. J Hum Nutr Diet Off J Br Diet Assoc 33:811–821. https://doi.org/10.1111/jhn.12788

12. Findlay M, White K, Brown C, Bauer JD (2021) Nutritional status and skeletal muscle status in patients with head and neck cancer: Impact on outcomes. J Cachexia Sarcopenia Muscle 12:2187–2198. https://doi.org/10.1002/jcsm.12829

13. Alabi RO, Elmusrati M, Leivo I, et al (2024) Artificial Intelligence-Driven Radiomics in Head and Neck Cancer: Current Status and Future Prospects. Int J Med Inf 188:105464. https://doi.org/10.1016/j.ijmedinf.2024.105464

14. Wang Z, Fang M, Zhang J, et al (2024) Radiomics and Deep Learning in Nasopharyngeal Carcinoma: A Review. IEEE Rev Biomed Eng 17:118–135. https://doi.org/10.1109/RBME.2023.3269776

15. Ferro A, Bottosso M, Dieci MV, et al (2024) Clinical applications of radiomics and deep learning in breast and lung cancer: A narrative literature review on current evidence and future perspectives. Crit Rev Oncol Hematol 203:104479. https://doi.org/10.1016/j.critrevonc.2024.104479

16. Badic B, Da-ano R, Poirot K, et al (2021) Prediction of recurrence after surgery in colorectal cancer patients using radiomics from diagnostic contrast-enhanced computed tomography: a two-center study. Eur Radiol. https://doi.org/10.1007/s00330-021-08104-4

17. Chen X-D, Chen W-J, Huang Z-X, et al (2022) Establish a New Diagnosis of Sarcopenia Based on Extracted Radiomic Features to Predict Prognosis of Patients With Gastric Cancer. Front Nutr 9:850929. https://doi.org/10.3389/fnut.2022.850929

18. Saalfeld S, Kreher R, Hille G, et al (2023) Prognostic role of radiomics-based body composition analysis for the 1-year survival for hepatocellular carcinoma patients. J Cachexia Sarcopenia Muscle 14:2301–2309. https://doi.org/10.1002/jcsm.13315

19. Vogele D, Mueller T, Wolf D, et al (2024) Applicability of the CT Radiomics of Skeletal Muscle and Machine Learning for the Detection of Sarcopenia and Prognostic Assessment of Disease Progression in Patients with Gastric and Esophageal Tumors. Diagnostics 14:198. https://doi.org/10.3390/diagnostics14020198

20. Shahzadi I, Zwanenburg A, Frohwein LJ, et al (2024) Short-term mortality prediction in acute pulmonary embolism: Radiomics values of skeletal muscle and intramuscular adipose tissue. J Cachexia Sarcopenia Muscle 15:1430–1440. https://doi.org/10.1002/jcsm.13488

21. Bernadach M, Lapeyre M, Dillies AF, et al (2019) [Toxicity of docetaxel, platine, 5-fluorouracil-based induction chemotherapy for locally advanced head and neck cancer: The importance of nutritional status]. Cancer Radiother J Soc Francaise Radiother Oncol 23:273–280. https://doi.org/10.1016/j.canrad.2018.08.003

22. Nioche C, Orlhac F, Boughdad S, et al (2018) LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res 78:4786–4789. https://doi.org/10.1158/0008-5472.CAN-18-0125

23. Orlhac F, Frouin F, Nioche C, et al (2019) Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 291:53–59. https://doi.org/10.1148/radiol.2019182023

24. Lin LI-K (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45:255. https://doi.org/10.2307/2532051

25. R Core Team R: A Language and Environment for Statistical Computing

26. Bernadach M, Lapeyre M, Dillies A-F, et al (2021) Predictive factors of toxicity of TPF induction chemotherapy for locally advanced head and neck cancers. BMC Cancer 21:360. https://doi.org/10.1186/s12885-021-08128-5

27. Kim H-K, Kim C-H (2021) Quality Matters as Much as Quantity of Skeletal Muscle: Clinical Implications of Myosteatosis in Cardiometabolic Health. Endocrinol Metab Seoul Korea 36:1161–1174. https://doi.org/10.3803/EnM.2021.1348

28. Correa-de-Araujo R, Addison O, Miljkovic I, et al (2020) Myosteatosis in the Context of Skeletal Muscle Function Deficit: An Interdisciplinary Workshop at the National Institute on Aging. Front Physiol 11:963. https://doi.org/10.3389/fphys.2020.00963

29. Amini B, Boyle SP, Boutin RD, Lenchik L (2019) Approaches to Assessment of Muscle Mass and Myosteatosis on Computed Tomography: A Systematic Review. J Gerontol A Biol Sci Med Sci 74:1671–1678. https://doi.org/10.1093/gerona/glz034

30. Chargi N, Bril SI, Swartz JE, et al (2020) Skeletal muscle mass is an imaging biomarker for decreased survival in patients with oropharyngeal squamous cell carcinoma. Oral Oncol 101:104519. https://doi.org/10.1016/j.oraloncology.2019.104519

31. Prado CMM, Lieffers JR, McCargar LJ, et al (2008) Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. Lancet Oncol 9:629–635. https://doi.org/10.1016/S1470-2045(08)70153-0

32. Lin Y-C, Lin G, Yeh T-S (2021) Visceral-to-subcutaneous fat ratio independently predicts the prognosis of locally advanced gastric cancer----- highlighting the role of adiponectin receptors and PPARα, β/ δ, ϒ. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol 47:3064–3073. https://doi.org/10.1016/j.ejso.2021.04.028

33. Dong X, Dan X, Yawen A, et al (2020) Identifying sarcopenia in advanced non-small cell lung cancer patients using skeletal muscle CT radiomics and machine learning. Thorac Cancer 11:2650–2659. https://doi.org/10.1111/1759-7714.13598

34. Yu W, Xu H, Chen F, et al (2023) Development and validation of a radiomics-based nomogram for the prediction of postoperative malnutrition in stage IB1-IIA2 cervical carcinoma. Front Nutr 10:1113588. https://doi.org/10.3389/fnut.2023.1113588

35. Kim YJ (2021) Machine Learning Models for Sarcopenia Identification Based on Radiomic Features of Muscles in Computed Tomography. Int J Environ Res Public Health 18:8710. https://doi.org/10.3390/ijerph18168710

36. Iwashita K, Kubota H, Nishioka R, et al (2023) Prognostic Value of Radiomics Analysis of Skeletal Muscle After Radical Irradiation of Esophageal Cancer. Anticancer Res 43:1749–1760. https://doi.org/10.21873/anticanres.16328

37. Miao S, Jia H, Cheng K, et al (2022) Deep learning radiomics under multimodality explore association between muscle/fat and metastasis and survival in breast cancer patients. Brief Bioinform 23:bbac432. https://doi.org/10.1093/bib/bbac432

38. van Timmeren JE, Cester D, Tanadini-Lang S, et al (2020) Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging 11:91. https://doi.org/10.1186/s13244-020-00887-2

39. Balagurunathan Y, Gu Y, Wang H, et al (2014) Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Transl Oncol 7:72–87. https://doi.org/10.1593/tlo.13844

Supporting Information 1

Preliminary Study for Selecting Radiomic Features Based on CT Scan Injection Times

Objective:

Since the acquisition times of the pre-treatment scans in our main study were heterogeneous, we aimed to retain only the radiomic parameters that were invariant across the acquisition times. We therefore conducted a preliminary study to assess the variations of each of the 125 parameters among the different injection times

Material and methods:

We conducted a single-center retrospective study at the CHU Gabriel-Montpied in Clermont-Ferrand. A local ethics committee approved this study.

We retrospectively included 63 consecutive patients who underwent a dedicated emergency imaging CT scan at our hospital with a multiphasic abdominal-pelvic exploration protocol: non-contrast, at arterial timing, and at portal timing. The exclusion criteria were minor patient or significant artifacts on the scan. All CT were realized on the same Revolution HD (GE Healthcare), at 100 kVp.

A key image was extracted at the L1 level for each patient at all three injection times. Segmentation of the same two regions of interest (ROIs) as in the main study (the entire paravertebral muscles and within the lumen of the aorta) was performed at this level for each injection time and for each patient.

A total of 125 radiomic features were extracted for each of the three acquisition times using the same method as in the main study using LifeX.

A statistical analysis was conducted using a mixed model to evaluate whether there was a significant difference between each feature in pairs across the non-contrast, arterial, and portal times

Results:

Out of the 125 extracted radiomic features, the number of features showing no difference between the following injection times was:

- Arterial and portal: 71 features
- Arterial and non-contrast: 36 features
- Non-contrast and portal: 49 features

Supporting Information 2

List of the 125 parameters extracted with LifeX, as referred in LifeX

CONVENTIONAL_HUmin
CONVENTIONAL_HUmean
CONVENTIONAL_HUstd
CONVENTIONAL_HUmax
CONVENTIONAL_HUQ1
CONVENTIONAL_HUQ2
CONVENTIONAL_HUQ3
CONVENTIONAL_HUSkewness
CONVENTIONAL_HUKurtosis
CONVENTIONAL_HUExcessKurtosis
CONVENTIONAL_HUpeakSphere0.5mL
CONVENTIONAL_HUpeakSphere1mL
DISCRETIZED_HUmin
DISCRETIZED_HUmean
DISCRETIZED_HUstd
DISCRETIZED_HUmax
DISCRETIZED_HUQ1
DISCRETIZED_HUQ2
DISCRETIZED_HUQ3
DISCRETIZED_HUSkewness
DISCRETIZED_HUKurtosis
DISCRETIZED_HUExcessKurtosis
DISCRETIZED_HUpeakSphere0.5mL
DISCRETIZED_HUpeakSphere1mL
DISCRETIZED_HISTO_Skewness
DISCRETIZED_HISTO_Kurtosis
DISCRETIZED_HISTO_ExcessKurtosis
DISCRETIZED_HISTO_Entropy_log10
DISCRETIZED_HISTO_Entropy_log2
DISCRETIZED_HISTO_Energy[=Uniformity]
SHAPE_Volume (=Surface)
SHAPE_Volume_indexed (=Surface_indexed)
GLCM_Homogeneity[=InverseDifference]
GLCM_Energy[=AngularSecondMoment]
GLCM_Contrast[=Variance]
GLCM_Correlation
GLCM_Entropy_log10
GLCM_Entropy_log2[=JointEntropy]
GLCM_Dissimilarity
GLRLM_SRE
GLRLM_LRE
GLRLM_LGRE
GLRLM_HGRE
GLRLM_SRLGE

GLRLM_SRHGE
GLRLM_LRLGE
GLRLM_LRHGE
GLRLM_GLNU
GLRLM_RLNU
GLRLM_RP
NGLDM_Coarseness
NGLDM_Contrast
NGLDM_Busyness
GLZLM_SZE
GLZLM_LZE
GLZLM_LGZE
GLZLM_HGZE
GLZLM_SZLGE
GLZLM_SZHGE
GLZLM_LZLGE
GLZLM_LZHGE
GLZLM_GLNU
GLZLM_ZLNU
GLZLM_ZP
NORMALIZED_CONVENTIONAL_HUmin
NORMALIZED_CONVENTIONAL_HUmean
NORMALIZED_CONVENTIONAL_HUstd
NORMALIZED_CONVENTIONAL_HUmax
NORMALIZED_CONVENTIONAL_HUQ1
NORMALIZED_CONVENTIONAL_HUQ2
NORMALIZED_CONVENTIONAL_HUQ3
NORMALIZED_DISCRETIZED_HUmin
NORMALIZED_DISCRETIZED_HUmean
NORMALIZED_DISCRETIZED_HUstd
NORMALIZED_DISCRETIZED_HUmax
NORMALIZED_DISCRETIZED_HUQ1
NORMALIZED_DISCRETIZED_HUQ2
NORMALIZED_DISCRETIZED_HUQ3
NORMALIZED_DISCRETIZED_HUpeakSphere0.5mL
NORMALIZED_DISCRETIZED_HUpeakSphere1mL
FIRST_CONVENTIONAL_RIM_HUmin
LAST_CONVENTIONAL_RIM_HUmin
FIRST_CONVENTIONAL_RIM_HUmean
LAST_CONVENTIONAL_RIM_HUmean
FIRST_CONVENTIONAL_RIM_HUstdev
LAST_CONVENTIONAL_RIM_HUstdev
FIRST_CONVENTIONAL_RIM_HUmax
LAST_CONVENTIONAL_RIM_HUmax
FIRST_CONVENTIONAL_RIM_HUVolume(mL)
LAST_CONVENTIONAL_RIM_HUVolume(mL)
FIRST_CONVENTIONAL_RIM_HUVolume(vx)
LAST_CONVENTIONAL_RIM_HUVolume(vx)
FIRST_CONVENTIONAL_RIM_HUsum

LAST_CONVENTIONAL_RIM_HUsum
FIRST_DISCRETIZED_RIM_HUmin
LAST_DISCRETIZED_RIM_HUmin
FIRST_DISCRETIZED_RIM_HUmean
LAST_DISCRETIZED_RIM_HUmean
FIRST_DISCRETIZED_RIM_HUstdev
FIRST_DISCRETIZED_RIM_HUstdev
FIRST_DISCRETIZED_RIM_HUmax
LAST_DISCRETIZED_RIM_HUmax
FIRST_DISCRETIZED_RIM_HUsum
LAST_DISCRETIZED_RIM_HUsum
NORMALIZED_FIRST_CONVENTIONAL_RIM_HUmin
NORMALIZED_LAST_CONVENTIONAL_RIM_HUmin
NORMALIZED_FIRST_CONVENTIONAL_RIM_HUmean
NORMALIZED_LAST_CONVENTIONAL_RIM_HUmean
NORMALIZED_FIRST_CONVENTIONAL_RIM_HUstdev
NORMALIZED_LAST_CONVENTIONAL_RIM_HUstdev
NORMALIZED_FIRST_CONVENTIONAL_RIM_Humax
NORMALIZED_LAST_CONVENTIONAL_RIM_Humax
NORMALIZED_FIRST_CONVENTIONAL_RIM_Husum
NORMALIZED_LAST_CONVENTIONAL_RIM_Husum
NORMALIZED_FIRST_DISCRETIZED_RIM_HUmin
NORMALIZED_LAST_DISCRETIZED_RIM_HUmin
NORMALIZED_FIRST_DISCRETIZED_RIM_HUmean
NORMALIZED_LAST_DISCRETIZED_RIM_HUmean
NORMALIZED_FIRST_DISCRETIZED_RIM_HUstdev
NORMALIZED_FIRST_DISCRETIZED_RIM_HUstdev
NORMALIZED_FIRST_DISCRETIZED_RIM_Humax
NORMALIZED_LAST_DISCRETIZED_RIM_Humax
NORMALIZED_FIRST_DISCRETIZED_RIM_Husum
NORMALIZED_LAST_DISCRETIZED_RIM_Husum

List of parameters used in our study, as referred to in LifeX and their designation in our study

CONVENTIONAL_HUstd	HU SD
FIRST_DISCRETIZED_RIM_HUsum	Discretized HU Sum
NORMALIZED_CONVENTIONAL_HUmean	Normalized HU mean
NORMALIZED_CONVENTIONAL_HUQ1	Normalized HU Q1
NORMALIZED_CONVENTIONAL_HUQ2	Normalized HU Q2
NORMALIZED_CONVENTIONAL_HUQ3	Normalized HU Q3
NORMALIZED_CONVENTIONAL_HUstd	Normalized HU SD
NORMALIZED_FIRST_CONVENTIONAL_RIM_HUsum	Normalized HU Sum
NORMALIZED_DISCRETIZED_HUmin	Normalized discretized HU min
NORMALIZED_DISCRETIZED_HUmean	Normalized discretized HU mean
NORMALIZED_DISCRETIZED_HUstd	Normalized discretized HU SD
NORMALIZED_DISCRETIZED_HUmax	Normalized discretized HU max
NORMALIZED_DISCRETIZED_HUQ1	Normalized discretized HU Q1
NORMALIZED_DISCRETIZED_HUQ2	Normalized discretized HU Q2
NORMALIZED_DISCRETIZED_HUQ3	Normalized discretized HU Q3
NORMALIZED_LAST_DISCRETIZED_RIM_HUmin	Normalized discretized Inner Rim HU min
NORMALIZED_LAST_DISCRETIZED_RIM_HUmean	Normalized discretized Inner Rim HU mean
GLRLM_SRE	GLRLM SRE
GLRLM_SRLGE	GLRLM SRLGE
GLZLM_ZLNU	GLZLM ZLNU
GLZLM_ZP	GLZLM ZP

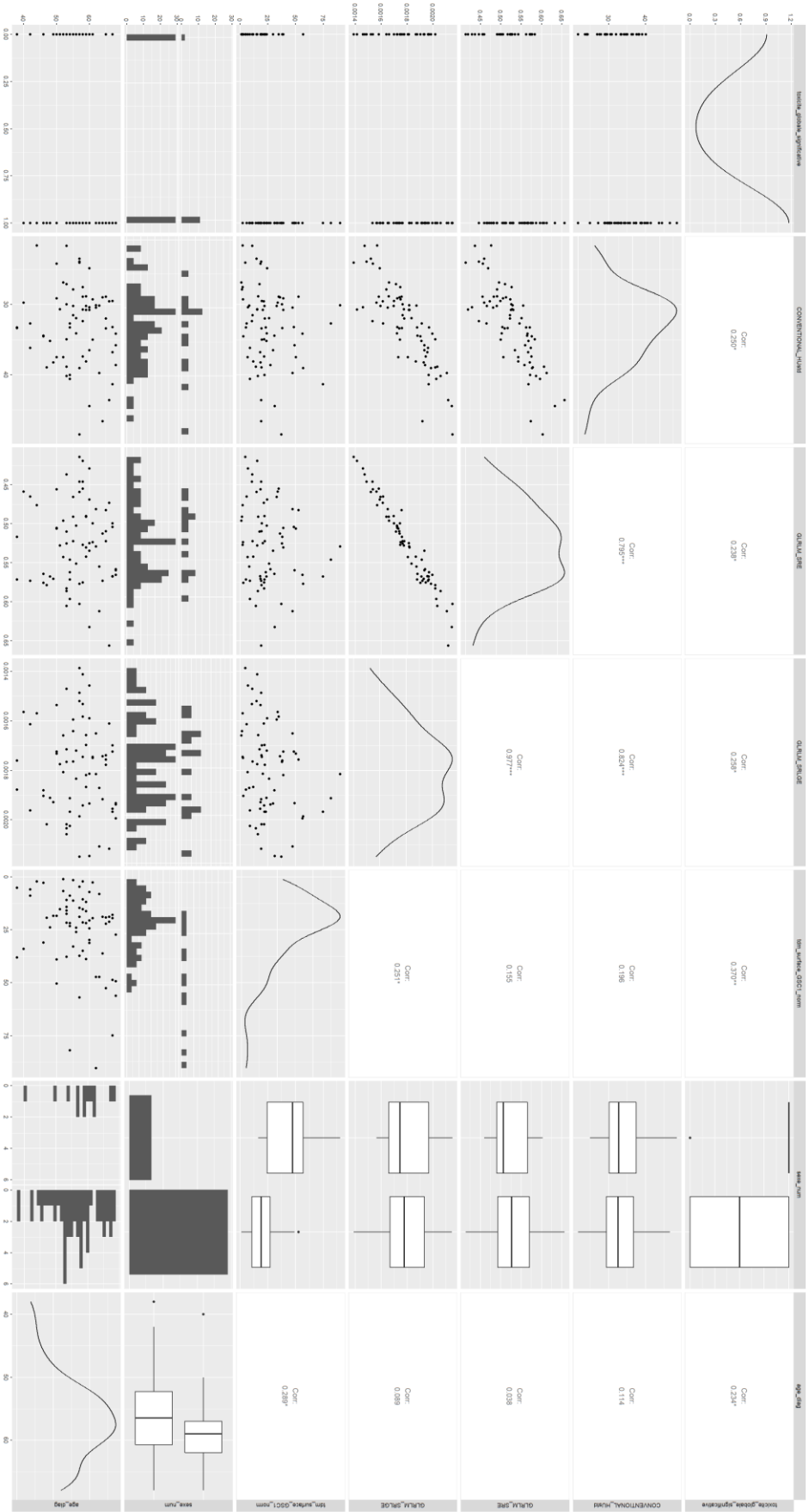
NORMALIZED_X: parameter X as extracted by LifeX divided by the mean HU value in the aorta.

FIRST_*_RIM_X: value of parameter X in the first rim of the ROI

LAST_*_RIM_X: value of parameter X in the last rim of the ROI

SHAPE_Volume _index is the parameter SHAPE_Volume divided by the square of the size of the patient.

Supporting Information 3: Correlation between variables before multivariate analyses for toxicity



Supporting Information 4

Response univariate analysis

			Response	No response	p
Radiomics	HU SD		33.723 +/- 7.8255	32.817 +/- 4.9737	0.677
	Discretized HU Sum		53826.439 +/- 12417.2736	55201.241 +/- 14139.0195	0.718
	Normalized HU mean		0.184 [0.16;0.21]	0.191 [0.16;0.25]	0.663
	Normalized HU Q1		0.121 [0.08;0.15]	0.117 [0.09;0.15]	0.608
	Normalized HU Q2		0.191 [0.18;0.22]	0.206 [0.18;0.25]	0.566
	Normalized HU Q3		0.267 [0.24;0.3]	0.283 [0.24;0.35]	0.546
	Normalized HU SD		0.128 [0.11;0.16]	0.144 [0.11;0.17]	0.535
	Normalized HU Sum		559.452 [485.04;751.81]	606.616 [415.7;805.63]	0.872
	Normalized discretized HU min		0.062 [0.06;0.07]	0.062 [0.05;0.08]	0.884
	Normalized discretized HU mean		0.073 [0.06;0.08]	0.073 [0.06;0.09]	0.813
	Normalized discretized HU SD		0.003 [0;0]	0.003 [0;0]	0.663
	Normalized discretized HU max		0.081 [0.07;0.09]	0.079 [0.07;0.1]	0.897
	Normalized discretized HU Q1		0.072 [0.06;0.08]	0.071 [0.06;0.09]	0.777
	Normalized discretized HU Q2		0.072 [0.06;0.08]	0.071 [0.06;0.1]	0.719
	Normalized discretized HU Q3		0.073 [0.06;0.08]	0.074 [0.06;0.1]	0.731
	Normalized discretized Inner Rim HU min		0.072 [0.06;0.08]	0.074 [0.06;0.09]	0.848
	Normalized discretized Inner Rim HU mean		0.074 [0.06;0.08]	0.075 [0.06;0.09]	0.872
	GLRLM SRE		0.562 [0.5;0.57]	0.517 [0.49;0.56]	0.254
	GLRLM SRLGE		0.002 +/- 0.0002	0.002 +/- 0.0002	0.531
	GLZLM ZLNU		34.044 [28.37;45.57]	34.099 [24.6;49.17]	0.994
	GLZLM ZP		0.063 [0.05;0.08]	0.053 [0.04;0.07]	0.309
CT	CT SCF Index		14.744 [3.79;20.49]	21.875 [15.33;35.53]	**0.027**
	CT Muscle Index		36.128 +/- 6.9977	36.711 +/- 7.5135	0.783
	CT AF Index		21.19 [7;33.33]	32.24 [15.84;54.84]	0.088
Clinical	Age (years)		57 [51;58.5]	57 [53;62.5]	0.597
	Height (cm)		171.867 +/- 7.1601	169.333 +/- 6.2022	0.229
	Usual weight (kg)		66.267 +/- 9.6397	69.765 +/- 11.8888	0.253
	BMI at treatment initiation		20.219 +/- 3.0019	22.884 +/- 3.8118	**0.008**
	Weight at treatment initiation (kg)		59.94 +/- 10.9595	65.806 +/- 12.5902	0.090
	Albumin		39.6 [31.8-42.9]	40.05 [37.38-42.93]	0.290
	ECOG				0.834
		ECOG 0	25 (49.0)	8 (53.3)	
		ECOG 1	24 (47.1)	6 (40.0)	
		ECOG 2	2 (3.9)	1 (6.7)	
	Male gender		42 (82.3)	13 (86.7)	1.000
	UICC Stage				0.831
		II	2 (3.9)	0 (0.0)	
		III	14 (27.5)	4 (26.7)	
		Iva	25 (49.0)	7 (46.7)	
		Ivb	10 (19.6)	4 (26.7)	
	Oral feeding possible		39 (76.5)	10 (66.7)	0.669

Response related to clinicopathological characteristics. Data are presented as number of patients (associated percentages), mean ± standard deviation or median [interquartile range].

Troisième étude : Comparaison de méthodes de sélection de données et de classification des paramètres de radiomique dans le pronostic du mélanome métastatique traité par immunothérapie

Nous vous présentons ensuite une seconde étude d'application de la radiomique.

Cette étude a pour but de trouver des facteurs pronostics à partir du scanner pré thérapeutique des patients devant bénéficier d'immunothérapie pour un mélanome métastatique. Nous avons essayé de faire une étude méthodologiquement plus exhaustive en combinant 5 méthodes de sélection de paramètres, 4 méthodes de classification et une méthode d'augmentation des données.

International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-022-02662-8

ORIGINAL ARTICLE

Metastatic melanoma treated by immunotherapy: discovering prognostic markers from radiomics analysis of pretreatment CT with feature selection and classification

Gulnur Ungan[1] · Anne-Flore Lavandier[2] · Jacques Rouanet[3] · Constance Hordonneau[2] · Benoit Chauveau[2] · Bruno Pereira[4] · Louis Boyer[2] · Jean-Marc Garcier[2,5] · Sandrine Mansard[3] · Adrien Bartoli[1] · Benoit Magnin[1,2,5]

Received: 15 December 2021 / Accepted: 26 April 2022

Abstract

Purpose Immunotherapy has dramatically improved the prognosis of patients with metastatic melanoma (MM). Yet, there is a lack of biomarkers to predict whether a patient will benefit from immunotherapy. Our aim was to create radiomics models on pretreatment computed tomography (CT) to predict overall survival (OS) and treatment response in patients with MM treated with anti-PD-1 immunotherapy.

Methods We performed a monocentric retrospective analysis of 503 metastatic lesions in 71 patients with 46 radiomics features extracted following lesion segmentation. Predictive accuracies for OS < 1 year versus > 1 year and treatment response versus no response was compared for five feature selection methods (sequential forward selection, recursive, Boruta, relief, random forest) and four classifiers (support vector machine (SVM), random forest, K-nearest neighbor, logistic regression (LR)) used with or without SMOTE data augmentation. A fivefold cross-validation was performed at the patient level, with a tumour-based classification.

Results The highest accuracy level for OS predictions was obtained with 3D lesions (0.91) without clinical data integration when combining Boruta feature selection and the LR classifier, The highest accuracy for treatment response prediction was obtained with 3D lesions (0.88) without clinical data integration when combining Boruta feature selection, the LR classifier and SMOTE data augmentation. The accuracy was significantly higher concerning OS prediction with 3D segmentation (0.91 vs 0.86) while clinical data integration led to improved accuracy notably in 2D lesions (0.76 vs 0.87) regarding treatment response prediction. Skewness was the only feature found to be an independent predictor of OS (HR (CI 95%) 1.34, *p*-value 0.001).

Conclusion This is the first study to investigate CT texture parameter selection and classification methods for predicting MM prognosis with treatment by immunotherapy. Combining pretreatment CT radiomics features from a single tumor with data selection and classifiers may accurately predict OS and treatment response in MM treated with anti-PD-1.

Keywords Metastatic melanoma · Immunotherapy · Texture analysis · Survival · Biomarker

Abbreviations

CAD	Computer-aided diagnosis
CI	Confidence interval
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4
HR	Hazard ratio
KNN	K-nearest neighbor
LDH	Serum lactate dehydrogenase

✉ Benoit Magnin
bmagnin@chu-clermontferrand.fr

1 EnCoV, Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, 28 place Henri Dunant, 63000 Clermont-Ferrand, France

2 Department of Medical Imaging, CHU Clermont Ferrand, 1 place Lucie Aubrac, 63100 Clermont-Ferrand, France

3 Dermatology Department, CHU Clermont Ferrand, 1 place Lucie Aubrac, 63100 Clermont-Ferrand, France

4 Biostatistics Unit, DRCI, CHU Clermont Ferrand, 58 rue Montalembert, 63000 Clermont-Ferrand, France

5 Anatomy Department, Université Clermont Auvergne, 28 place Henri Dunant, 63000 Clermont-Ferrand, France

Published online: 02 June 2022

MM	Metastatic melanoma
OS	Overall survival
PD-1	Program cell death 1
PFS	Progression-free survival
RECIST	Response Evaluation Criteria In Solid Tumours
RF	Random forest
ROI	Region of interest
SFS	Sequential forward selection
SMOTE	Synthetic Minority Oversampling TEchnique
SVM	Support vector machine

Introduction

Melanoma is a primary skin cancer causing approximately 55,500 deaths annually [1]. Metastatic melanoma (MM) has the highest mortality rate with an estimated three-year survival rate of between 5 and 32% [2]. An early breakthrough in the treatment of MM involved targeted therapy used to block BRAF and MEK, a treatment available to only 40% of patients, for whom the tumor presents a BRAF V600 mutation [3]. The introduction of immunotherapy, (CTLA-4 checkpoint inhibitor: ipilimumab, and later PD-1 checkpoint inhibitor: pembrolizumab and nivolumab) associated with long overall survival (OS) in MM [4, 5] constituted a second major breakthrough and has become a common first-line therapy in MM.

Improvement in survival rates, following treatment by checkpoint inhibitor therapy, however, remains heterogeneous. Predictors of immunotherapy response are essential as they enable clinicians to evaluate benefits of immunotherapy, to spare patients unnecessary risk of toxicity [6], to select the most suitable targeted therapy and to enrol patients in clinical trials [7].

Predictors in current use include the LDH value, visceral tumour burden (notably the presence of liver lesions), the relative eosinophil count, the relative lymphocyte count and age [8, 9]. In addition, visual analysis of CT images performed by a radiologist may provide prognostic predictors such as the number, size and shape of lesions and number of metastatic sites. The heterogeneity of lesions, unquantifiable by the human eye, is reported to be a prognostic factor in some tumors [10]. Texture analysis, a technique used to quantify tumor heterogeneity [11, 12], provides an analysis of the relationship between and distribution of pixel gray levels in the tumor, thus revealing the spatial variation of gray levels in image patches.

Radiomics involves the extraction of high-dimensional sets of imaging features that characterize intra-tumoral heterogeneity. Those features can be used to build models providing clinicians with key information for clinical diagnosis and assessment of prognosis and therapeutic effects. The extracted image features can be combined with other clinical, biological and genomic data, thus increasing the power of decision support systems.

A large body of studies have reported on the benefit of texture analysis in many types of cancer, namely colorectal cancer [13, 14], hepatocellular carcinoma [15], Hodgkin lymphoma [11], non-small-cell lung cancer [16–18], soft tissue sarcoma [19], oesophageal cancer [20, 21] and head and neck cancer [22], and provided information on survival rates, evaluation of treatment response [21, 23–27] and histological characterization of lesions [28–30].

A few studies have reported on the use of Computer Aided Diagnosis (CAD) including radiomics features, in MM.

Some studies identified prognostic factors: Smith et al. [31] used radiomics features from 23 CT obtained before and following initiation of treatment by bevacizumab (which is no longer administered), to identify prognostic factors of survival; Durot et al. [32] analysed average radiomics features on multiple lesions of 31 patients before administering pembrolizumab and found predictors of OS.

Some studies built prognostic models: Schraag et al. [33] extracted seven first-order texture parameters from the largest lesion in 103 patients prior to immunotherapy and built a multivariable Cox regression model with clinical and texture parameters to enable prediction of OS; Wang et al. [34] selected radiomics features from the largest lesion in 50 patients before immunotherapy and built a model with a support vector machine (SVM, [35]) able to predict first cycle response based on validation of data from 16 patients.

Feature selection is a key step in building a radiomics model. As some radiomics software can extract up to a thousand features [36], dimensionality reduction is crucial to prevent overfitting. The most common methods are logistic regression and the Least Absolute Shrinkage and Selection Operator (LASSO). From the selected features, a regression or a classification model can be built, the most common ones being SVM and Random Forest (RF).

Some studies used several feature selection methods and classifiers in other pathologies [37–39]. Our study is the first to study feature selection and classification methods for evaluation of MM prognosis.

The purpose of this study was to compare feature selection and classification methods, from radiomics features taken from contrast-enhanced CT images before initiation of treatment, so as to predict OS and treatment response in patients with MM treated by pembrolizumab or nivolumab.

Patients and methods

Study population

This monocentric retrospective cohort study was approved by our institutional review board (authorization number CRM-1905-008).

All patients treated with anti-PD-1 immunotherapy (pembrolizumab or nivolumab, identified from the hospital pharmacy database) between July 2014 and September 2018 for MM were included in the study. Exclusion criteria were: no metastasis visible on CT scans, no delineable lesion, no baseline CT scan performed within 2.5 months before beginning of treatment and a clinical follow-up period lower than one year (unless death occurred during the first year).

Recorded data included age, gender, BRAF mutation status, date of metastatic status, start date of immunotherapy treatment, date of pretreatment CT scan, number of metastatic sites, presence of hepatic, cerebral and lung metastases, number of segmented lesions, 3-month follow-up iRECIST conclusion, death, decision for supportive care or change of treatment.

Follow-up and endpoints

All patients underwent clinical, biological, and radiological follow-up every 3 months, in accordance with the hospital protocol. The radiological follow-up comprised contrast-enhanced CT scans of the brain, thorax, abdomen and pelvis. In cases of suspected progression, an additional CT was performed 6 weeks later to rule out pseudo progression.

Two endpoints were used for classification. The first endpoint was OS, for which patients were allocated to two groups based on the survival period; lower than one year after treatment initiation and longer than one year. The second endpoint was treatment response, obtained from the first CT scan taken 3 months after treatment initiation using iRECIST. Patients were allocated to two groups: favorable prognosis for stable disease or partial response and unfavorable prognosis for progression.

CT examination

The majority of CT scans (65/71, 92%) were performed on a 64-section contrast-enhanced CT scanner (Discovery CT 750 HD, GE Healthcare). A volume of 1.5 mL/kg body weight of non-ionic contrast material was injected into the peripheral upper limb vein at a flow rate of 3 $mL.s^{-1}$. Chest, abdominal and pelvic images were obtained at a portal venous phase (80 s), and cerebral images were obtained at a late phase (5 min after injection).

The acquisition parameters were as follows: 100 kVp tube voltage; helical pitch of 1.375; image reconstruction thickness of 2.5 mm. The images were reconstructed using 30% adaptive statistical reconstruction (ASiR, GE Healthcare).

Six CTs were performed on other scanners located outside our hospital. Visual assessment allowed us to ensure the quality of acquisition. The time of injection was checked before inclusion.

Data segmentation

Lesion segmentation was performed manually with the pretreatment CT scanner, using LIFEx (version 4.00, www.lifexsoft.org) [40] an ISBI-compliant feature extraction platform. Segmentation was performed following consensus by two radiologists: a senior radiologist with 8-year experience in radiology (6 years in oncological imaging) and a radiology resident with 4-year experience.

All segmentable lesions (sufficient size and definition) were segmented on the CT. 3D ROI corresponded to segmentation of the whole lesion, while the axial slice with the largest area of the segmented lesion was saved as the 2D ROI.

Feature extraction

For each segmented ROI (3D and 2D ROI for each lesion), 46 parameters were extracted by LIFEx, (Supplementary Information 1) and divided into three categories according to shape (volume, sphericity, compacity), histogram of grey level (skewness, kurtosis, entropy, energy) and texture parameters.

Quantification of noise of the radiomics features

To quantify the distribution of noise in the data, we evaluated for each radiomic feature separately on 2D and 3D ROIs the coefficient of variation, defined as the ratio between standard deviation and absolute value of the mean value of the feature $\left(\frac{\sigma}{|\bar{x}|}\right)$.

Data augmentation, feature selection and classification methods

To adjust for imbalanced datasets, data augmentation was used using the Synthetic Minority Oversampling TEchnique (SMOTE) [41].

A total of five different feature selection algorithms were used to select features extracted from CT data including sequential forward selection (SFS), Boruta, relief, recursive and RF feature selection algorithms.

Additional clinical data included age; sex; previous treatment by other immunotherapy; BRAF status; and presence

of lymph node, liver, brain, lung, adrenal, spleen, bone or gastrointestinal tract metastasis.

The four classification methods used to classify the selected features from CT imagery and clinical data were SVM (using a linear kernel) [35], RF, K-nearest neighbor (KNN) and logistic regression.

A total of 40 combinations were tested (with or without data augmentation × 5 feature selection × 4 classification methods).

Classification was firstly performed with radiomics features only, and secondly along with clinical data; the features from 3D and 2D ROIs were processed separately.

A fivefold cross-validation algorithm was used for each classification task (OS and treatment response). For each model, data were split into a training set (80% of the patients) and a test data set (the remaining 20% of the patients). The split between different folds was done on the patient level. Tumours in the test data set were then classified, resulting in a tumour-based classification. The split train–test procedure was repeated five times and allowed calculation of mean values for accuracy, sensitivity and specificity.

In addition to accuracy, sensitivity and specificity values were used in performance analysis.

Influence of feature selection and classification methods on performance

To quantify the influence of feature selection methods, we evaluated the mean performance on all combinations (classifiers, with or without data augmentation) of each feature selection method.

The same evaluation was made to quantify the influence of classifiers.

The Shapiro test rejected the normality of the distribution of accuracy. Hence, non-parametric tests were used. The Kruskal–Wallis test was used to search for significant differences of accuracy between the five feature selection methods first and the four classifiers then. In case of significant differences, a pairwise comparison was made with the Wilcoxon test, applying Bonferroni's correction for multiple comparisons.

Quantification of fit

In order to evaluate the models' fit, we calculated the train and test accuracies for each model and evaluated the train/test accuracy ratio.

Statistical tests

We used the Cox proportional hazards regression model to assess the association between survival parameters and covariates of interest. The proportional hazard hypothesis was tested using the Schoenfeld's test, and the results were expressed as hazard ratios and 95% confidence intervals (95%CI). All statistical tests were two-sided, with p-values under 0.05 considered statistically significant.

Data extraction, feature selection and classification were performed with Python using a custom script.

Inter-observer reproducibility

The assessment of intra-observer reproducibility was based on an initial analysis of 10% of segmentations following random selection and on a repeat analysis, blinded to the first, performed at a six-month interval. Assessment included comparison of features from both segmentations, extracted using LIFEx and an estimation of the Lin concordance correlation. The results were analised according to conventional rules defined in the literature [42, 43]: 0–0.2 (negligible agreement), 0.2–0.4 (low/weak agreement), 0.4–0.6 (moderate agreement), 0.6–0.8 (substantial/good agreement) and > 0.8 (strong agreement).

Results

Patient characteristics

Of 79 eligible patients, eight were excluded (five with difficult-to-define lesions, two without metastasis, one without pretreatment CT evaluation). A total of 71 patients (41 men, 30 women) of median age 66 years (interquartile range 34–90) were included.

The main baseline patient clinical and radiological characteristics are given in Table 1.

A total of 906 lesions (503 3D lesions and 403 2D lesions; minimum 1 per patient, mean 7 and maximum 31) were segmented. Thirty-five percent (25/71) of patients presented with oligo metastatic lesions (less than three metastatic lesions), 38% (27/71) had less than ten lesions and 27% (19/71) more than 10, 35% (25/71) presented with hepatic lesions, 31% (22/71) with cerebral lesions and 55% (39/71) with pulmonary lesions.

The main follow-up data are given in Table 2, with a mean follow-up of 882 days.

Reproducibility

Lin's concordance correlation coefficient was > 0.8 for 92% of the features, while for the remaining features, the coefficient was ≤ 0.20, 0.21–0.40, 0.41 − 0.60, and 0.61–0.80 for 2, 0, 4, and 2% of the features, respectively.

Table 1 Patients' baseline characteristics ($n = 71$)

Characteristics	Results
Nivolumab (*n* (%))	32 (45)
Pembrolizumab (*n* (%))	39 (55)
Men (*n* (%))	41 (58)
Women (*n* (%))	30 (42)
Mean age (mean, (min–max) yr)	66 (34–*90*)
BRAF mutation	
Yes (*n* (%))	23 (32)
No (*n* (%))	48 (68)
Time between CT and treatment (mean, (min–max) days)	12.96 (0–*73*)
Spread of lesions	
Lymph node (*n* (%))	35 (50)
Brain (*n* (%))	22 (31)
Lung (*n* (%))	39 (55)
Liver (*n* (%))	25 (35)
Adrenal gland (*n* (%))	16 (23)
Spleen (*n* (%))	6 (8)
Bone (*n* (%))	11 (15)
Number of segmented lesions	906
3D (*n* (%))	503 (56)
2D (*n* (%))	403 (44)
2D segmented lesions volumes (mL)	
2D minimal volume	0.002
2D median volume	0.284
2D maximal volume	115.817
3D segmented lesions volumes (mL)	
3D minimal volume	0.007
3D median volume	1.608
3D maximal volume	530.73

Overall survival prediction

The best three combinations for 2D and 3D lesions with and without clinical data are shown in Table 3. The performance of all combinations are presented in Supplementary Information 2.

For 2D radiomics features, the best results were obtained using recursive feature selection combined with logistic regression classification and SMOTE data augmentation (Acc 0.86, Sen 0.7, Spe 0.69).

For 3D radiomic features, the best results were obtained using Boruta feature selection and logistic regression classification (Acc 0.91, Sen 0.79, Spe 0.39).

Table 2 Follow-up characteristics of the patient ($n = 71$) (PFS: Progression-Free Survival)

Characteristics	Results
Clinical follow-up (mean, (min–max) days)	882 (15–42,299)
Death at the end of follow-up	
Yes (*n* (%))	46 (65)
No (*n* (%))	25 (35)
Median survival	
OS (median (min–max), days)	502 (15–1356)
< 12 months (*n* (%))	27 (38)
< 6 months (*n* (%))	14 (20)
6–12 months (*n* (%))	13 (18)
> 12 months (*n* (%))	44 (62)
12–18 months (*n* (%))	17 (25)
18–24 months (*n* (%))	12 (16)
> 24 months (*n* (%))	15 (22)
PFS (median (min–max), days)	166 (43–1330)
< 12 months (*n* (%))	50 (70)
< 6 months (*n* (%))	34 (48)
6–12 months (*n* (%))	16 (22)
> 12 months (*n* (%))	21 (30)
12–18 months (*n* (%))	13 (18)
18–24 months (*n* (%))	3 (4)
> 24 months (*n* (%))	5 (8)
iRECIST evaluation at 3 months	
Partial response (*n* (%))	17 (24)
Stable disease (*n* (%))	16 (22)
Progression (*n* (%))	38 (54)

Treatment response prediction

The best three combinations for 2D and 3D lesions without and with clinical data are shown in Table 4.

The best results were obtained for 2D radiomics features and clinical data integration with random forest selection and SVM classification (Acc 0.87, Sen 0.44, Spe 0.8).

For 3D radiomic features, the best combination was obtained using recursive feature selection combined with logistic regression classification and SMOTE, with clinical data integration (Acc 0.83, Sen 0.65, Spe 0.6).

Overall survival analysis

Cox proportional hazard models were calculated in three ways: as a whole radiomics covariate (shape, histogram and texture features), histogram features only and texture features only.

Table 3 Best OS predictions (Acc, Sen and Spe are mean values of the fivefold CV, resulting in some cases in accuracy outside the values of Spe and Sen)

ROI type	Combination (Classifier + Feature Selection ± SMOTE Data Augmentation)	Acc	Sen	Spe
3D	Logistic Regression + Boruta	0.91	0.79	0.39
	Logistic Regression + Boruta + SMOTE	0.88	0.76	0.4
	Logistic Regression + Random Forest	0.88	0.77	0.36
3D + clinical data	SVM + SFS	0.84	0.95	0.6
	SVM + RF + SMOTE	0.78	0.4	0.85
	Logistic Regression + Recursive + SMOTE	0.7	0.86	0.6
2D	Logistic Regression + RF + SMOTE	0.81	0.6	0.75
	RF + Boruta	0.75	0.74	0.48
	Logistic Regression + Boruta + SMOTE	0.74	0.68	0.54
2D + clinical data	Logistic Regression + Recursive + SMOTE	0.86	0.7	0.69
	KNN + Relief + SMOTE	0.87	0.5	0.55
	SVM + Boruta	0.84	0.56	0.62

Table 4 Best treatment response predictions (Acc, Sen and Spe are mean values of the fivefold CV, resulting in some cases in accuracy outside the values of Spe and Sen)

ROI type	Combination (Classifier + Feature Selection ± SMOTE Data Augmentation	Acc	Sen	Spe
3D	Logistic Regression + Boruta + SMOTE	0.8	0.66	0.88
	SVM + SFS + SMOTE	0.83	0.46	0.81
	Logistic Regression + SFS + SMOTE	0.81	0.46	0.66
3D + clinical data	KNN + Relief + SMOTE	0.75	0.72	0.6
	Logistic Regression + Recursive + SMOTE	0.83	0.65	0.6
	RF + RF + SMOTE	0.74	0.53	0.66
2D	Logistic Regression + Relief + SMOTE	0.81	0.66	0.85
	Logistic Regression + RF + SMOTE	0.81	0.6	0.75
	RF + Recursive + SMOTE	0.76	0.68	0.57
2D + clinical data	SVM + RF	0.87	0.44	0.8
	SVM + Relief + SMOTE	0.78	0.46	0.88
	RF + Recursive + SMOTE	0.76	0.68	0.57

For whole radiomics covariate, significant features were shape sphericity ($p = 0.012$), GLZLM-LZE ($p = 0.041$), GLZLM-LZHGE ($p = 0.037$) and GLZLM-ZLNU ($p = 0.003$), but CI was non-significant.

Concerning histogram features (Table 5), skewness was significantly correlated with OS ($p = 0.012$, CI 95% = 1.07–1.7).

Concerning texture features (Table 5), the p-values were significant for GLRLM-SRE ($p = 0.022$), NGLDM-contrast ($p = 0.032$), GLZM-LZE ($p = 0.037$), GLZM-LZHGE ($p = 0.041$) and GLZLM-ZLNM ($p = 0.011$, but only NGLDM-contrast ($p = 0.032$, CI 95% = 1.9 10^{-3}–0.7 10^{-1}) had a significant CI.

Quantification of noise of the radiomics features

The median coefficient of variation was 1.076 for 2D features and 0.634 for 3D features. The complete set of values is given in Supplementary Information 3.

Influence of feature selection and classification methods on performance

The statistics of performance on all classifications with one feature selection method or with one classifier are given in Supplementary Information 4 (Tables and Figs. 4.1 and 4.2).

Table 5 Univariate Cox proportional hazard models of histogram and texture parameters for survival analysis

	HR	CI 95%	*p* value
Histogram parameters			
Skewness	1.34	1.07–1.7	0.012*
Kurtosis	0.99	0.92–1.1	0.684
Entropy_log10	1.48	0.34–6.3	0.6
Energy	4.16	0.22–78.4	0.342
Texture parameters			
GLRLM_SRE	1.5	1.1–2.1	0.022*
GLRLM_LRE	14	$1.3\ 10^{-3}$–$1.6\ 10^{-5}$	0.578
GLRLM_HGRE	1.3	0.05–34	0.87
GLRLM_SRHGE	1	1–1	0.308
GLRLM_LRHGE	1	1–1	0.839
GLRLM_GLNU	1	1–1	0.992
GLRLM_RLNU	1	1–1	0.757
GLRLM_RP	1	1–1	0.418
NGLDM_Coarseness	0.99	0.54–1.8	0.974
NGLDM_Contrast	0.038	0.0019–0.76	0.032*
GLZLM_SZE	0.00075	$6.5\ 10^{-8}$–8.7	0.132
GLZLM_LZE	1	1–1	0.037*
GLZLM_HGZE	1	1–1	0.381
GLZLM_SZHGE	1	1–1	0.247
GLZLM_LZHGE	1	1–1	0.041*
GLZLM_GLNU	1	1–1	0.807
GLZLM_ZLNU	1	1–1	0.011*
GLZLM_ZP	1	1–1	0.181

*Significant difference ($p < 0.05$)

There was no difference between the mean accuracies of all feature selection methods (Kruskal–Wallis test, $p = 0.635$).

There was a statistical difference between the mean accuracies of classifiers (Kruskal–Wallis test, $p = 1.3 * 10^{-10}$). Pairwise comparison showed that LR and SVM each had a significantly higher accuracy than RF and KNN (Supplementary Information 4, Table 4.3).

Quantification of fit

The mean training error was 0.18 for the 12 best OS prediction models and for the 12 best response prediction models (Supplementary Information 5, Tables 5.1 and 5.2) The mean ratio of train/test accuracy was 1.01 for the 12 best OS prediction models and 1.06 for the 12 best response prediction models (Supplementary Information 5, Tables 5.1 and 5.2). The mean ratios of train/test accuracy and the values of train and test accuracy on all classifications are given in Supplementary Information 5, Table 5.3 and 5.4.

Discussion

We investigated the performance of different radiomics models as a prognostic tool to predict OS and treatment response, in patients with metastatic melanoma treated by anti-PD-1, on pretreatment CT images. We combined five feature selection methods with four classification methods with or without SMOTE data augmentation on any segmentable lesion, resulting in a tumour-based classification. The accuracy of the 10 best classification methods for predicting OS up to and beyond one year and treatment response was found to be good (> 0.80).

To date, only four studies have reported on the prognosis of patients with MM, based on radiomic parameters.

Smith et al. [31] used radiomics features from a CT obtained before treatment and modifications in the features from a CT taken after the initiation of treatment, to identify prognostic factors of survival. This study was, however, based on a small number of patients (23), the use of a treatment which is no longer administered (bevacizumab) and data recording at 3 months after initiation of treatment.

Durot et al. [32] investigated the association of pretreatment CT scan texture parameters with OS and progression-free survival, in patients treated with pembrolizumab. The model was built using LASSO-penalized Cox regression from five histogram features. They found that skewness values above − 0.55 at coarse texture scale were significantly associated with both lower OS and lower PFS. The study however has several limitations. Firstly, the low number of patients (31 compared to 71 in our study) and a limited number of lesions per patient (5 maximum). They reported on 74 lesions in total compared to 906 in our study. Secondly, lesion contours concerned single axial sections only rather than the whole tumor in 3D which impedes assessment of tumor heterogeneity and contour replication. Thirdly, few texture parameters were extracted (compared to 46 in our study) with reporting of only average parameter values and including values from other organs. Fourthly, Durot et al. used RECIST 1.1 to establish treatment response, without taking into account the pseudo progression phenomenon. Hodi et al. [44] noted that RECIST 1.1 may lead to underestimating responses in 15% of patients and result in early discontinuation of treatment. Finally, the absence of a validation process (validation cohort or cross-validation) weakens the strength of the main result, as the threshold of skewness coarse texture scale was determined and evaluated on the same population. Yet, the only radiomics feature we found to be significantly correlated with OS was skewness, as in the study by Durot et al. [32].

Schraag et al. [33] extracted seven first-order texture parameters from the largest lesion in 103 patients prior to immunotherapy. Their model, built on clinical and texture parameters with a multivariable Cox regression, enabled the

prediction of OS (C-index 0.716) but texture parameters did not allow the prediction of treatment response.

In a recent publication, Wang et al. [34] extracted 497 radiomics features from the largest lesion in 50 patients prior to immunotherapy. On the basis of a selection of features by T-test and redundancy, their model using SVM was shown to predict first cycle response in a validation cohort of 16 patients (accuracy 75%), but without survival predictions.

The majority of the above-mentioned studies recorded data from one lesion only (usually the largest) to predict OS or treatment response, with the exception of Durot et al. who reported an average value of various lesions. By performing a tumour-based classification, our study enabled us to evaluate the ability to predict OS or treatment response from any single lesion of a patient, regardless of its size.

Moreover, all those studies used only one model to assess the prognosis in MM. Other studies have compared radiomics model performance regarding prediction of clinical event occurrence in several other cancers by using various feature selection and classification methods: lung [39, 45, 46], preoperative differentiation of sacral chordoma and sacral giant cell tumor [38], head and neck cancer [47], and rectal cancer [37]. Palmar et al. noted that the choice of classification method is the major factor driving the performance variation [47].

The present study compares a total of 160 combinations (five feature selection methods, four classification methods, SMOTE data augmentation, for 2D and 3D lesions, integration or not of clinical data, response therapy and OS), using methods shown in previous studies to provide the best performance. This is the first study to investigate texture parameter selection and classification methods for predicting MM prognosis with treatment by immunotherapy.

The highest performance for OS prediction (accuracy 0.95) was found when combining recursive feature selection with a logistic regression classifier, while for treatment response (accuracy 0.90) this was found when combining SFS selected features, a RF classifier and integration of clinical data.

For all methods, 3D segmentation provided better results than 2D segmentation (12% accuracy increase). This supports evidence reported by Ortiz-Ramon et al. [29] and Ng et al. [13]. Clinical data integration led to a greater increase in accuracy for 2D features, than for 3D features, notably concerning prediction of treatment response.

No combination of feature selection and classification method emerged clearly as the best for the different data (2D vs. 3D, with or without clinical data integration, treatment response or survival). The reasons for the variability of results depending on the "pipeline" (i.e., combination of feature selection, data augmentation and classifier) are difficult to investigate and rarely fully addressed in the literature. Yet, a few articles try to address the point, among which are the articles by Parmar et al. [45, 47] which attempt to quantify the impact of the methods on the results. They, however, do not give a detailed explanation about the characteristics that may explain the differences between different methods. The other articles attempting to address this point simply recall general properties of the methods [38].

Generally speaking, the model's performance can first be explained by a possible unusually good or bad model fitting. The repetition of experiments and the fivefold cross-validation make sure the model fitting is not a special lucky or unlucky case. Second, the model performance can be explained by the model's lack of fit. Our data concerning the fit on the training data show that the model did not underfit, as the mean train error was 0.18 for the 12 best OS prediction model and 0.18 for the 12 best response prediction model (Supplementary Information 5, Supplementary Tables 5.1 and 5.2). Lastly, the performance can be explained by the model's expressivity. To assess the model's expressivity, we computed the train/test ratio, which was 1.01 for the 12 best OS prediction model and 1.06 for the 12 best response prediction model (Supplementary Information 5, Supplementary Tables 5.1 and 5.2), showing the model's ability to generalise and its sufficient expressivity.

However, the influence of feature selection methods appears to be moderate in our models, as the performances of all feature selection methods are similar, without any significant difference (Supplementary Information 4, Table and Figure 4.1). It has been indeed noted that the influence of feature selection methods can vary depending on the type of cancer. Namely, feature selection has high impact in lung cancer and twice less in head and neck cancer [47], hence leaving the possibility that feature selection can have little impact on some cancer type, including melanoma, as per our findings.

SMOTE data augmentation had a mixed effect on the results, positive for some combinations and negative for others. It is, however, worth of note that 11 of the 12 best results for prediction of treatment response involved SMOTE data augmentation.

Concerning the classifier, LR and SVM emerged as the two best classifiers: out of the 24 best results, 12 were performed with LR and 6 with SVM. Moreover, mean performances on all classifications of those two classifiers were significantly better than those of the two other classifiers (Supplementary Information 4, Table and Figure 4.2).

Therefore, in our study, the variability depends mainly on the choice of the classifier, the two best being LR and SVM. LR is a supervised machine learning classification algorithm that does not make any assumptions regarding the distribution of independent variables. It is a commonly used classifier, providing average performances in classification tasks [48].

SVM is a supervised machine learning algorithm. It aims to find the best hyperplane to split a dataset into two classes.

It is often reported to be more robust than LR, with a lower risk of overfitting. It is one of the most used classifiers and appears to be one of the best classifiers, notably in disease prediction studies [48]. However, its behaviour depends on the type of kernel used; a linear kernel was used in our study. It can be said that linear SVM and LR have similar behaviours to find a margin between different classes. Moreover, Musa et al. demonstrated that SVM and LR can work similarly for different scenarios such as balanced and unbalanced datasets [49].

It has been shown that LR consistently performs with a higher overall accuracy as compared to RF when increasing the variance of the noise data [50]. The median coefficient of variation was high (1.076 for 2D features and 0.634 for 3D features), showing noise in our data, hence explaining the improved performances of LR compared to RF. The degree of noise can also explain why LR outperforms SVM in our data [51]. Concerning the performances of KNN, one of the reasons explaining the lack of performances is the fact that the data were not normalised or rescaled.

The limitations of this study include, firstly that six pretreatment CT were performed on a different CT scan, requiring prior to inclusion, visual assessment to ensure scan quality and time of injection. Secondly, this study was a retrospective monocentric study that despite the large number of analysed lesions (503 in 2D and 403 in 3D), relied on a relatively small number of patients. We were therefore unable to split the population into training and validation cohorts, with cross validation ensured by a fivefold cross validation algorithm. Our patient number remains, however, larger than that of most previous studies and reflects the relative rareness of the disease. Finally, our model could be improved by including more biological or genetic features such as LDH level, which was not possible due to insufficient patient data.

Our study involved the use of five commonly used feature selection and four classifier methods with encouraging results. Future research is required to evaluate the performance of more complex classification methods such as those built on deep learning.

Conclusion

Our study showed that the combination of CT texture analysis, data selection and classification algorithms may accurately predict treatment response and overall survival for patients starting anti-PD-1 immunotherapy for metastatic melanoma.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-022-02662-8.

Acknowledgements The authors would like to thank Helen Braund for language editing.

Funding The authors confirm that no funding was received for this research.

References

1. Schadendorf D, van Akkooi ACJ, Berking C, Griewank KG, Gutzmer R, Hauschild A, Stang A, Roesch A, Ugurel S (2018) Melanoma. The Lancet 392:971–984. https://doi.org/10.1016/S0140-6736(18)31559-9
2. Song X, Zhao Z, Barber B, Farr AM, Ivanov B, Novich M (2015) Overall survival in patients with metastatic melanoma. Curr Med Res Opin 31:987–991. https://doi.org/10.1185/03007995.2015.1021904
3. Larkin J, Ascierto PA, Dréno B, Atkinson V, Liszkay G, Maio M, Mandalà M, Demidov L, Stroyakovskiy D, Thomas L, de la Cruz-Merino L, Dutriaux C, Garbe C, Sovak MA, Chang I, Choong N, Hack SP, McArthur GA, Ribas A (2014) Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N Engl J Med 371:1867–1876. https://doi.org/10.1056/NEJMoa1408868
4. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, Daud A, Carlino MS, McNeil C, Lotem M, Larkin J, Lorigan P, Neyns B, Blank CU, Hamid O, Mateus C, Shapira-Frommer R, Kosh M, Zhou H, Ibrahim N, Ebbinghaus S, Ribas A (2015) KEYNOTE-006 investigators, Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med 372:2521–2532. https://doi.org/10.1056/NEJMoa1503093
5. Robert C, Long GV, Brady B, Dutriaux C, Maio M, Mortier L, Hassel JC, Rutkowski P, McNeil C, Kalinka-Warzocha E, Savage KJ, Hernberg MM, Lebbé C, Charles J, Mihalcioiu C, Chiarion-Sileni V, Mauch C, Cognetti F, Arance A, Schmidt H, Schadendorf D, Gogas H, Lundgren-Eriksson L, Horak C, Sharkey B, Waxman IM, Atkinson V, Ascierto PA (2015) Nivolumab in previously untreated melanoma without BRAF mutation. N Engl J Med 372:320–330. https://doi.org/10.1056/NEJMoa1412082
6. Martin-Liberal J, Kordbacheh T, Larkin J (2015) Safety of pembrolizumab for the treatment of melanoma. Expert Opin Drug Saf 14:957–964. https://doi.org/10.1517/14740338.2015.1021774
7. Hiniker SM, Maecker HT, Knox SJ (2015) Predictors of clinical response to immunotherapy with or without radiotherapy. J Radiat Oncol 4:339–345. https://doi.org/10.1007/s13566-015-0219-2
8. Weide B, Martens A, Hassel JC, Berking C, Postow MA, Bisschop K, Simeone E, Mangana J, Schilling B, Di Giacomo AM, Brenner N, Kähler K, Heinzerling L, Gutzmer R, Bender A, Gebhardt C, Romano E, Meier F, Martus P, Maio M, Blank C, Schadendorf D, Dummer R, Ascierto PA, Hospers G, Garbe C, Wolchok JD (2016) Baseline biomarkers for outcome of melanoma patients treated with pembrolizumab. Clin Cancer Res 22:5487–5496. https://doi.org/10.1158/1078-0432.CCR-16-0127
9. Nosrati A, Tsai KK, Goldinger SM, Tumeh P, Grimes B, Loo K, Algazi AP, Nguyen-Kim TDL, Levesque M, Dummer R, Hamid O, Daud A (2017) Evaluation of clinicopathological factors in PD-1 response: derivation and validation of a prediction scale for response to PD-1 monotherapy. Br J Cancer 116:1141–1147. https://doi.org/10.1038/bjc.2017.70
10. Rao S-X, Lambregts DM, Schnerr RS, Beckers RC, Maas M, Albarello F, Riedl RG, Dejong CH, Martens MH, Heijnen LA, Backes WH, Beets GL, Zeng M-S, Beets-Tan RG (2016) CT texture analysis in colorectal liver metastases: a better way than size and volume measurements to assess response to chemotherapy?

United Eur Gastroenterol J 4:257–263. https://doi.org/10.1177/2050640615601603
11. Ganeshan B, Miles KA, Babikir S, Shortman R, Afaq A, Ardeshna KM, Groves AM, Kayani I (2017) CT-based texture analysis potentially provides prognostic information complementary to interim fdg-pet for patients with hodgkin's and aggressive non-hodgkin's lymphomas. Eur Radiol 27:1012–1020. https://doi.org/10.1007/s00330-016-4470-8
12. Verma V, Simone CB, Krishnan S, Lin SH, Yang J, Hahn SM (2017) The rise of radiomics and implications for oncologic management. JNCI J Natl Cancer Inst. https://doi.org/10.1093/jnci/djx055
13. Ng F, Ganeshan B, Kozarski R, Miles KA, Goh V (2013) Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. Radiology 266:177–184. https://doi.org/10.1148/radiol.12120254
14. Miles KA, Ganeshan B, Griffiths MR, Young RCD, Chatwin CR (2009) Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival. Radiology 250:444–452. https://doi.org/10.1148/radiol.2502071879
15. Mulé S, Thiefin G, Costentin C, Durot C, Rahmouni A, Luciani A, Hoeffel C (2018) Advanced hepatocellular carcinoma: pretreatment contrast-enhanced CT texture parameters as predictive biomarkers of survival in patients treated with sorafenib. Radiology 288:445–455. https://doi.org/10.1148/radiol.2018171320
16. Miles KA (2016) How to use CT texture analysis for prognostication of non-small cell lung cancer, Cancer Imaging Off. Publ Int Cancer Imaging Soc 16:10. https://doi.org/10.1186/s40644-016-0065-5
17. Ahn SY, Park CM, Park SJ, Kim HJ, Song C, Lee SM, McAdams HP, Goo JM (2015) Prognostic value of computed tomography texture features in non-small cell lung cancers treated with definitive concomitant chemoradiotherapy. Invest Radiol 50:719–725. https://doi.org/10.1097/RLI.0000000000000174
18. Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K (2012) Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol 22:796–802. https://doi.org/10.1007/s00330-011-2319-8
19. Hayano K, Tian F, Kambadakone AR, Yoon SS, Duda DG, Ganeshan B, Sahani DV (2015) Texture analysis of non-contrast-enhanced computed tomography for assessing angiogenesis and survival of soft tissue sarcoma. J Comput Assist Tomogr 39:607–612. https://doi.org/10.1097/RCT.0000000000000239
20. Ganeshan B, Skogen K, Pressney I, Coutroubis D, Miles K (2012) Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival. Clin Radiol 67:157–164. https://doi.org/10.1016/j.crad.2011.08.012
21. Yip C, Landau D, Kozarski R, Ganeshan B, Thomas R, Michaelidou A, Goh V (2014) Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy. Radiology 270:141–148. https://doi.org/10.1148/radiol.13122869
22. Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, Pitman K, Caudell JJ, Hamilton RD, Ganeshan B, Smith AD (2013) Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. Radiology 269:801–809. https://doi.org/10.1148/radiol.13130110
23. Ramella S, Fiore M, Greco C, Cordelli E, Sicilia R, Merone M, Molfese E, Miele M, Cornacchione P, Ippolito E, Iannello G, D'Angelillo RM, Soda P (2018) A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients. PLoS ONE 13:e0207455. https://doi.org/10.1371/journal.pone.0207455
24. Ahn SJ, Kim JH, Park SJ, Han JK (2016) Prediction of the therapeutic response after FOLFOX and FOLFIRI treatment for patients with liver metastasis from colorectal cancer using computerized CT texture analysis. Eur J Radiol 85:1867–1874. https://doi.org/10.1016/j.ejrad.2016.08.014
25. Tian F, Hayano K, Kambadakone AR, Sahani DV (2015) Response assessment to neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion, Abdom. Imaging 40:1705–1712. https://doi.org/10.1007/s00261-014-0318-3
26. Ravanelli M, Farina D, Morassi M, Roca E, Cavalleri G, Tassi G, Maroldi R (2013) Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: prediction of the response to the first-line chemotherapy. Eur Radiol 23:3450–3455. https://doi.org/10.1007/s00330-013-2965-0
27. Knogler T, Thomas K, El-Rabadi K, Karem E-R, Weber M, Michael W, Karanikas G, Georgios K, Mayerhoefer ME, Marius Erik M (2014) Three-dimensional texture analysis of contrast enhanced CT images for treatment response assessment in Hodgkin lymphoma: comparison with F-18-FDG PET. Med Phys 41:121904. https://doi.org/10.1118/1.4900821
28. Kniep HC, Madesta F, Schneider T, Hanning U, Schönfeld MH, Schön G, Fiehler J, Gauer T, Werner R, Gellissen S (2019) Radiomics of Brain MRI: utility in prediction of metastatic tumor type. Radiology 290:479–487. https://doi.org/10.1148/radiol.2018180946
29. Ortiz-Ramón R, Larroza A, Ruiz-España S, Arana E, Moratal D (2018) Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study. Eur Radiol 28:4514–4523. https://doi.org/10.1007/s00330-018-5463-6
30. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA (2013) Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. Radiology 266:326–336. https://doi.org/10.1148/radiol.12112428
31. Smith AD, Gray MR, del Campo SM, Shlapak D, Ganeshan B, Zhang X, Carson WE (2015) Predicting overall survival in patients with metastatic melanoma on antiangiogenic therapy and recist stable disease on initial posttherapy images using CT texture analysis. Am J Roentgenol 205:W283–W293. https://doi.org/10.2214/AJR.15.14315
32. Durot C, Mulé S, Soyer P, Marchal A, Grange F, Hoeffel C (2019) Metastatic melanoma: pretreatment contrast-enhanced CT texture parameters as predictive biomarkers of survival in patients treated with pembrolizumab. Eur Radiol 29:3183–3191. https://doi.org/10.1007/s00330-018-5933-x
33. Schraag A, Klumpp B, Afat S, Gatidis S, Nikolaou K, Eigentler TK, Othman AE (2019) Baseline clinical and imaging predictors of treatment response and overall survival of patients with metastatic melanoma undergoing immunotherapy. Eur J Radiol 121:108688. https://doi.org/10.1016/j.ejrad.2019.108688
34. Wang Z, Mao L, Zhou Z, Si L, Zhu H, Chen X, Zhou M, Sun Y, Guo J (2020) Pilot study of CT-based radiomics model for early evaluation of response to immunotherapy in patients with metastatic melanoma. Front Oncol 10:1524. https://doi.org/10.3389/fonc.2020.01524
35. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1023/A:1022627411411
36. Chu H, Liu Z, Liang W, Zhou Q, Zhang Y, Lei K, Tang M, Cao Y, Chen S, Peng S, Kuang M (2021) Radiomics using CT images for preoperative prediction of futile resection in intrahepatic cholangiocarcinoma. Eur Radiol 31:2368–2376. https://doi.org/10.1007/s00330-020-07250-5
37. Badic B, Da-ano R, Poirot K, Jaouen V, Magnin B, Gagnière J, Pezet D, Hatt M, Visvikis D (2021) Prediction of recurrence after surgery in colorectal cancer patients using radiomics from diagnostic contrast-enhanced computed tomography: a two-center study. Eur Radiol. https://doi.org/10.1007/s00330-021-08104-4

38. Yin P, Mao N, Zhao C, Wu J, Sun C, Chen L, Hong N (2019) Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. Eur Radiol 29:1841–1847. https://doi.org/10.1007/s00330-018-5730-6
39. Sun W, Jiang M, Dang J, Chang P, Yin F-F (2018) Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiat Oncol 13:197. https://doi.org/10.1186/s13014-018-1140-9
40. Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, Pellot-Barakat C, Soussan M, Frouin F, Buvat I (2018) LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 78:4786–4789. https://doi.org/10.1158/0008-5472.CAN-18-0125
41. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
42. Altman DG (1990) Practical statistics for medical research. CRC Press, Florida
43. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012
44. Hodi FS, Hwu W-J, Kefford R, Weber JS, Daud A, Hamid O, Patnaik A, Ribas A, Robert C, Gangadhar TC, Joshua AM, Hersey P, Dronca R, Joseph R, Hille D, Xue D, Li XN, Kang SP, Ebbinghaus S, Perrone A, Wolchok JD (2016) Evaluation of immune-related response criteria and RECIST v1.1 in patients with advanced melanoma treated with pembrolizumab. J Clin Oncol 34:1510–1517. https://doi.org/10.1200/JCO.2015.64.0391
45. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL (2015) Machine learning methods for quantitative radiomic biomarkers. Sci Rep 5:13087. https://doi.org/10.1038/srep13087
46. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby RA, Balagurunathan Y, Goldgof D, Schabath MB, Hall L, Gillies RJ (2016) Predicting malignant nodules from screening CTs. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer 11:2120–2128. https://doi.org/10.1016/j.jtho.2016.07.002
47. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL (2015) Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol 5:272. https://doi.org/10.3389/fonc.2015.00272
48. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181
49. Musa AB (2013) Comparative study on classification performance between support vector machine and logistic regression. Int J Mach Learn Cybern 4:13–24. https://doi.org/10.1007/s13042-012-0068-x
50. Kirasich K, Smith T, Sadler B (2018) Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review 1:25
51. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19:281. https://doi.org/10.1186/s12911-019-1004-8

Supplementary Information 1

Supp Table 1.1: 46 radiomic features extracted from LIFEx

Radiomic features	Brief explanation
First order features: Shape	
Volume (mL)	Volume of ROI in mL
Volume (#vx)	Volume of ROI in number of voxels
Sphericity	Sphericity of the volume. 1 for a perfect sphere
Compacity	Compactness of ROI
First order features: Histogram	
Conventional_TLG (mL)	Total lesion glycolysis inside the ROI
Skewness	Asymmetry of the grey-level distribution in the histogram
Kurtosis	Shape of the grey-level distribution (peaked or flat) relative to a normal distribution
Entropy_log10	Randomness of the distribution
Entropy_log2	Randomness of the distribution
Energy	Uniformity of the distribution
minValue	Minimum pixel value of the ROI
meanValue	Average of pixel values
stdValue	Standard deviation of pixel values
maxValue	Maximum pixel value
Second order features	
GLCM	Arrangements of voxel pairs with same grey-level intensity
GLCM_Homogeneity	Homogeneity of grey-level voxel pairs
GLCM_Energy (Uniformity)	Uniformity of grey-level voxel pairs
GLCM_Contrast (Variance)	Local variations in GLCM
GLCM_Correlation	Linear dependency of grey-level voxel pairs
GLCM_Entropy_log10 and Entropy_log2	Randomness of grey-level voxel pairs
GLCM_Dissimilarity	Variation of grey-level voxel pairs
GLZLM	Size of homogeneous zones for each grey-level intensity
GLZLM_SZE	Short-zone emphasis: Distribution of the short homogeneous zones
GLZLM_LZE	Long-zone emphasis: Distribution of the long homogeneous zones
GLZLM_HGZE	High grey-level zone emphasis: Distribution of the high grey-level zones
GLZLM_LGZE	Low grey-level zone emphasis: Distribution of the low grey-level zones
GLZLM_SZHGE; GLZLM_SZLGE	Distribution of the short homogeneous zones with low or high grey-levels
GLZLM_LZHGE; GLZLM_LZLGE	Distribution of the long homogeneous zones with low or high grey-levels
GLZLM_GLNU	Grey-level non-uniformity: Nonuniformity of the grey-levels of the homogeneous zones
GLZLM_ZLNU	Zone length non-uniformity: Nonuniformity of the length of the homogeneous zones
GLZLM_ZP	Zone percentage: Homogeneity of the homogeneous zones
GLRLM	Size of homogeneous runs for each grey-level intensity
GLRLM_SRE	Short-run emphasis: Distribution of the short homogeneous runs

GLRLM_LRE	Long-run emphasis: Distribution of the long homogeneous runs
GLRLM_HGRE	High grey-level run emphasis: Distribution of the high grey-level runs
GLRLM_LGRE	Low grey-level run emphasis: Distribution of the low grel-level runs
GLRLM_SRHGE; GLRLM_SRLGE	Distribution of the short homogeneous runs with low or high grey-levels
GLRLM_LRHGE; GLRLM_LRLGE	Distribution of the long homogeneous runs with low or high grey-levels
GLRLM_GLNU	Grey-level non-uniformity: Nonuniformity of the grey-levels of the homogeneous runs
GLRLM_RP	Run percentage: Homogeneity of the homogeneous runs
GLRLM_RLNU	Run length non-uniformity: Nonuniformity of the length of the homogeneous runs
NGLDM	Difference of grey-level between one voxel and its 26 neighbours in 3 dimensions
NGLDM_Coarseness	Level of spatial rate of change in intensity
NGLDM_Contrast	Intensity difference between neighbouring regions
NGLDM_Busyness	Spatial frequency of changes in intensity

GLCM: Grey-level co-occurrence matrix; NGLDM: Neighborhood grey-level different matrix; GLZLM: Grey-level zone length matrix; GLRLM: Grey-level run length matrix

Supplementary Information 2

Supp Table 2.1: OS prediction with 2D features without clinical data (Acc, Sen and Spe are mean values of the 5-fold CV, resulting in some cases in Accuracy outside the values of Spe and Sen)

SFS= Sequential Forward Selection
KNN= K nearest Neighbour
RF= Random Forest
SVM= Support Vector Machine

ROI	AIM	Classifier	Selection	SMOTE?	Acc	Sen	Spe
2D	OS	KNN	Boruta	SMOTE	0.58	0.5	0.46
2D	OS	KNN	Boruta		0.58	0.63	0.29
2D	OS	KNN	Random Forest	SMOTE	0.69	0.54	0.61
2D	OS	KNN	Random Forest		0.68	0.79	0.29
2D	OS	KNN	Recursive	SMOTE	0.53	0.28	0.63
2D	OS	KNN	Recursive		0.63	0.62	0.4
2D	OS	KNN	Relief	SMOTE	0.63	0.52	0.51
2D	OS	KNN	Relief		0.64	0.74	0.27
2D	OS	KNN	SFS	SMOTE	0.7	0.62	0.54
2D	OS	KNN	SFS		0.64	0.75	0.26
2D	OS	Logistic Regression	Boruta	SMOTE	0.74	0.68	0.54
2D	OS	Logistic Regression	Boruta		0.64	0.95	0.018
2D	OS	Logistic Regression	Random Forest	SMOTE	0.81	0.6	0.75
2D	OS	Logistic Regression	Random Forest		0.7	0.98	0.09
2D	OS	Logistic Regression	Recursive	SMOTE	0.63	0.57	0.49
2D	OS	Logistic Regression	Recursive		0.62	0.84	0.11
2D	OS	Logistic Regression	Relief	SMOTE	0.63	0.21	0.9
2D	OS	Logistic Regression	Relief		0.65	0.98	0
2D	OS	Logistic Regression	SFS	SMOTE	0.54	0	1
2D	OS	Logistic Regression	SFS		0.63	0.21	0.9
2D	OS	RF	Boruta	SMOTE	0.71	0.65	0.52
2D	OS	RF	Boruta		0.75	0.74	0.48
2D	OS	RF	Random Forest	SMOTE	0.6	0.62	0.35
2D	OS	RF	Random Forest		0.68	0.81	0.26
2D	OS	RF	Recursive	SMOTE	0.76	0.68	0.57
2D	OS	RF	Recursive		0.71	0.75	0.39
2D	OS	RF	Relief	SMOTE	0.65	0.51	0.57
2D	OS	RF	Relief		0.75	0.84	0.35
2D	OS	RF	SFS	SMOTE	0.69	0.6	0.54
2D	OS	RF	SFS		0.69	0.8	0.26
2D	OS	SVM	Boruta	SMOTE	0.63	0.78	0.2
2D	OS	SVM	Boruta		0.62	0.92	0.02
2D	OS	SVM	Random Forest	SMOTE	0.77	0.59	0.44
2D	OS	SVM	Random Forest		0.66	1	0

2D	OS	SVM	Recursive	SMOTE	0.64	0.81	0.19
2D	OS	SVM	Recursive		0.61	0.909	0.01
2D	OS	SVM	Relief	SMOTE	0.63	0.21	0.9
2D	OS	SVM	Relief		0.66	1	0
2D	OS	SVM	SFS	SMOTE	0.66	1	0
2D	OS	SVM	SFS		0.61	0.9	0.03

Supp Table 2.2: OS prediction with 2D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	OS	KNN	Boruta	SMOTE	0.59	0.62	0.65
2D	OS	KNN	Boruta		0.75	0.7	0.5
2D	OS	KNN	Random Forest	SMOTE	0.45	0.4	0.44
2D	OS	KNN	Random Forest		0.49	0.41	0.24
2D	OS	KNN	Recursive	SMOTE	0.59	0.6	0.45
2D	OS	KNN	Recursive		0.72	0.6	0.7
2D	OS	KNN	Relief	SMOTE	0.87	0.5	0.55
2D	OS	KNN	Relief		0.75	0.86	0.3
2D	OS	KNN	SFS	SMOTE	0.61	0.4	0.6
2D	OS	KNN	SFS		0.42	0.85	0.12
2D	OS	Logistic Regression	Boruta	SMOTE	0.76	0.83	0.5
2D	OS	Logistic Regression	Boruta		0.8	0.89	0.1
2D	OS	Logistic Regression	Random Forest	SMOTE	0.68	0.76	0.72
2D	OS	Logistic Regression	Random Forest		0.7	0.5	0.5
2D	OS	Logistic Regression	Recursive	SMOTE	0.75	0.1	0.97
2D	OS	Logistic Regression	Recursive		0.61	1	0
2D	OS	Logistic Regression	Relief	SMOTE	0.59	0.59	0.59
2D	OS	Logistic Regression	Relief		0.76	0.75	0.2
2D	OS	Logistic Regression	SFS	SMOTE	0.8	0.7	0.6
2D	OS	Logistic Regression	SFS		0.7	0.6	0.4
2D	OS	RF	Boruta	SMOTE	0.56	0.79	0.57
2D	OS	RF	Boruta		0.38	0.45	0.13
2D	OS	RF	Random Forest	SMOTE	0.44	0.53	0.28
2D	OS	RF	Random Forest		0.3	0.8	0.03
2D	OS	RF	Recursive	SMOTE	0.47	0.43	0.33
2D	OS	RF	Recursive		0.57	0.7	0.4
2D	OS	RF	Relief	SMOTE	0.4	0.18	0.83
2D	OS	RF	Relief		0.42	1	0.22
2D	OS	RF	SFS	SMOTE	0.46	0.8	0.33
2D	OS	RF	SFS		0.42	0.66	0.2
2D	OS	SVM	Boruta	SMOTE	0.54	0.79	0.12
2D	OS	SVM	Boruta		0.84	0.56	0.62
2D	OS	SVM	Random Forest	SMOTE	0.71	0.23	0.92
2D	OS	SVM	Random Forest		0.9	1	0
2D	OS	SVM	Recursive	SMOTE	0.77	0.87	0.2
2D	OS	SVM	Recursive		0.66	0.75	0.25
2D	OS	SVM	Relief	SMOTE	0.63	0.5	0.5
2D	OS	SVM	Relief		0.75	0.57	0.4
2D	OS	SVM	SFS	SMOTE	0.48	0.25	0.77
2D	OS	SVM	SFS		0.77	0.83	0.7

Supp Table 2.3: OS prediction with 3D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	OS	KNN	Boruta	SMOTE	0.62	0.43	0.6
3D	OS	KNN	Boruta		0.65	0.46	0.6
3D	OS	KNN	Random Forest	SMOTE	0.71	0.52	0.6
3D	OS	KNN	Random Forest		0.65	0.44	0.66
3D	OS	KNN	Recursive	SMOTE	0.68	0.47	0.66
3D	OS	KNN	Recursive		0.63	0.46	0.54
3D	OS	KNN	Relief	SMOTE	0.7	0.56	0.45
3D	OS	KNN	Relief		0.64	0.54	0.33
3D	OS	KNN	Relief		0.64	0.54	0.33
3D	OS	KNN	SFS	SMOTE	0.4	0.5	0.5
3D	OS	KNN	SFS		0.56	0.34	0.7
3D	OS	Logistic Regression	Boruta	SMOTE	0.88	0.76	0.4
3D	OS	Logistic Regression	Boruta		0.91	0.79	0.39
3D	OS	Logistic Regression	Random Forest	SMOTE	0.72	0.47	0.75
3D	OS	Logistic Regression	Random Forest		0.88	0.77	0.36
3D	OS	Logistic Regression	Recursive	SMOTE	0.51	0.81	0.5
3D	OS	Logistic Regression	Recursive		0.95	0.87	0.3
3D	OS	Logistic Regression	Relief	SMOTE	0.75	0.8	0.16
3D	OS	Logistic Regression	Relief		0.86	0.98	0
3D	OS	Logistic Regression	SFS	SMOTE	0.43	0.38	0.69
3D	OS	Logistic Regression	SFS		0.94	1	0
3D	OS	RF	Boruta	SMOTE	0.64	0.44	0.63
3D	OS	RF	Boruta		0.73	0.86	0.32
3D	OS	RF	Random Forest	SMOTE	0.74	0.53	0.66
3D	OS	RF	Random Forest		0.7	0.75	0.26
3D	OS	RF	Recursive	SMOTE	0.56	0.28	0.55
3D	OS	RF	Recursive		0.44	0.75	0.36
3D	OS	RF	Relief	SMOTE	0.54	0.59	0.44
3D	OS	RF	Relief		0.62	0.75	0.46
3D	OS	RF	SFS	SMOTE	0.66	0.42	0.57
3D	OS	RF	SFS		0.6	0.62	0.44
3D	OS	SVM	Boruta	SMOTE	0.35	0.03	0.96
3D	OS	SVM	Boruta		0.73	0.98	0
3D	OS	SVM	Random Forest	SMOTE	0.5	0.2	0.96
3D	OS	SVM	Random Forest		0.71	0.98	0.02
3D	OS	SVM	Recursive	SMOTE	0.35	0.03	0.96
3D	OS	SVM	Recursive		0.97	0.9	0.24
3D	OS	SVM	Relief	SMOTE	0.89	0.11	0.9
3D	OS	SVM	Relief		0.98	1	0
3D	OS	SVM	SFS	SMOTE	0.98	0.95	0.12
3D	OS	SVM	SFS		0.98	1	0

Supp Table 2.4: OS prediction with 3D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	OS	KNN	Boruta	SMOTE	0.68	0.55	0.55
3D	OS	KNN	Boruta		0.7	0.79	0.41
3D	OS	KNN	Random Forest	SMOTE	0.82	0.64	0.4
3D	OS	KNN	Random Forest		0.54	0.84	0.25
3D	OS	KNN	Recursive	SMOTE	0.61	0.6	0.6
3D	OS	KNN	Recursive		0.8	0.79	0.3
3D	OS	KNN	Relief	SMOTE	0.72	0.62	0.57
3D	OS	KNN	Relief		0.76	1	0.1
3D	OS	KNN	Relief		0.61	0.58	0.48
3D	OS	KNN	SFS	SMOTE	0.57	0.74	0.46
3D	OS	KNN	SFS		0.64	0.8	0.4
3D	OS	Logistic Regression	Boruta	SMOTE	0.73	0.16	0.88
3D	OS	Logistic Regression	Boruta		0.71	1	0.11
3D	OS	Logistic Regression	Random Forest	SMOTE	0.77	1	0.03
3D	OS	Logistic Regression	Random Forest		0.83	0.56	0.62
3D	OS	Logistic Regression	Recursive	SMOTE	0.7	0.86	0.6
3D	OS	Logistic Regression	Recursive		0.71	1	0.11
3D	OS	Logistic Regression	Relief	SMOTE	0.63	0.89	0.05
3D	OS	Logistic Regression	Relief		0.88	1	0
3D	OS	Logistic Regression	SFS	SMOTE	0.69	0.91	0.66
3D	OS	Logistic Regression	SFS		0.7	1	0
3D	OS	RF	Boruta	SMOTE	0.56	0.79	0.57
3D	OS	RF	Boruta		0.38	0.45	0.13
3D	OS	RF	Random Forest	SMOTE	0.44	0.53	0.28
3D	OS	RF	Random Forest		0.3	0.8	0.03
3D	OS	RF	Recursive	SMOTE	0.47	0.43	0.33
3D	OS	RF	Recursive		0.57	0.7	0.4
3D	OS	RF	Relief	SMOTE	0.4	0.18	0.83
3D	OS	RF	Relief		0.42	1	0.22
3D	OS	RF	SFS	SMOTE	0.46	0.8	0.33
3D	OS	RF	SFS		0.42	0.66	0.2
3D	OS	SVM	Boruta	SMOTE	0.65	0.51	0.57
3D	OS	SVM	Boruta		0.9	0.6	0.5
3D	OS	SVM	Random Forest	SMOTE	0.78	0.4	0.85
3D	OS	SVM	Random Forest		0.78	0.97	0.15
3D	OS	SVM	Recursive	SMOTE	0.79	0.95	0.01
3D	OS	SVM	Recursive		0.7	0.2	1
3D	OS	SVM	Relief	SMOTE	0.7	0.2	0.8
3D	OS	SVM	Relief		0.76	1	0
3D	OS	SVM	SFS	SMOTE	0.37	0.03	0.97
3D	OS	SVM	SFS		0.84	0.95	0.6

Supp Table 2.5: Therapy response prediction with 2D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	Response	KNN	Boruta	SMOTE	0.58	0.42	0.63
2D	Response	KNN	Boruta		0.85	0.56	0.62
2D	Response	KNN	Random Forest	SMOTE	0.61	0.51	0.45
2D	Response	KNN	Random Forest		0.51	0.5	0.47
2D	Response	KNN	Recursive	SMOTE	0.62	0.47	0.53
2D	Response	KNN	Recursive		0.71	0.2	0.9
2D	Response	KNN	Relief	SMOTE	0.68	0.48	0.77
2D	Response	KNN	Relief		0.65	0.73	0.42
2D	Response	KNN	SFS	SMOTE	0.62	0.47	0.53
2D	Response	Logistic Regression	Boruta	SMOTE	0.8	0.2	0.9
2D	Response	Logistic Regression	Boruta		0.74	0.17	0.75
2D	Response	Logistic Regression	Random Forest	SMOTE	0.8	0.23	0.88
2D	Response	Logistic Regression	Random Forest		0.73	0.09	0.89
2D	Response	Logistic Regression	Recursive	SMOTE	0.86	0.8	0.42
2D	Response	Logistic Regression	Recursive		0.64	0.45	0.77
2D	Response	Logistic Regression	Relief	SMOTE	0.81	0.66	0.85
2D	Response	Logistic Regression	Relief		0.7	0.3	0.83
2D	Response	Logistic Regression	SFS	SMOTE	0.72	0.2	0.7
2D	Response	Logistic Regression	SFS		0.7	0.57	0.73
2D	Response	RF	Boruta	SMOTE	0.65	0.43	0.59
2D	Response	RF	Boruta		0.44	0.142	0.83
2D	Response	RF	Random Forest	SMOTE	0.59	0.25	0.8
2D	Response	RF	Random Forest		0.34	0.48	0.54
2D	Response	RF	Recursive	SMOTE	0.91	0.89	0.24
2D	Response	RF	Recursive		0.42	0.29	0.5
2D	Response	RF	Relief	SMOTE	0.52	0.37	0.61
2D	Response	RF	Relief		0.4	0.32	0.74
2D	Response	RF	SFS	SMOTE	0.62	0.48	0.57
2D	Response	RF	SFS		0.38	0.125	0.9
2D	Response	SVM	Boruta	SMOTE	0.74	0.54	0.63
2D	Response	SVM	Boruta		0.64	0.1	1
2D	Response	SVM	Random Forest	SMOTE	0.6	0.33	0.79
2D	Response	SVM	Random Forest		0.65	0.1	0.9
2D	Response	SVM	Recursive	SMOTE	0.66	0.25	0.82
2D	Response	SVM	Recursive		0.82	0	1
2D	Response	SVM	Relief	SMOTE	0.77	0.4	0.88
2D	Response	SVM	Relief		0.69	0	1
2D	Response	SVM	SFS	SMOTE	0.65	0.44	0.72
2D	Response	SVM	SFS		0.42	0	1

Supp Table 2.6: Therapy response prediction with 2D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	Response	KNN	Boruta	SMOTE	0.54	0.47	0.48
2D	Response	KNN	Boruta		0.59	0.64	0.5
2D	Response	KNN	Random Forest	SMOTE	0.57	0.407	0.57
2D	Response	KNN	Random Forest		0.65	1	0
2D	Response	KNN	Recursive	SMOTE	0.83	0.45	0.62
2D	Response	KNN	Recursive		0.59	0.01	0.85
2D	Response	KNN	Relief	SMOTE	0.5	0.61	0.46
2D	Response	KNN	Relief		0.5	0.5	0.55
2D	Response	KNN	SFS	SMOTE	0.55	0.42	0.48
2D	Response	Logistic Regression	Boruta	SMOTE	0.59	0.25	0.76
2D	Response	Logistic Regression	Boruta		0.71	0.3	0.8
2D	Response	Logistic Regression	Random Forest	SMOTE	0.7	1	0
2D	Response	Logistic Regression	Random Forest		0.58	0.25	0.81
2D	Response	Logistic Regression	Recursive	SMOTE	0.8	0.1	0.85
2D	Response	Logistic Regression	Recursive		0.75	0.25	0.82
2D	Response	Logistic Regression	Relief	SMOTE	0.71	0.26	0.81
2D	Response	Logistic Regression	Relief		0.72	0.3	0.82
2D	Response	Logistic Regression	SFS	SMOTE	0.77	0.3	0.9
2D	Response	Logistic Regression	SFS		0.8	0.37	0.84
2D	Response	RF	Boruta	SMOTE	0.71	0.65	0.52
2D	Response	RF	Boruta		0.75	0.74	0.48
2D	Response	RF	Random Forest	SMOTE	0.7	0.8	0.5
2D	Response	RF	Random Forest		0.68	0.81	0.26
2D	Response	RF	Recursive	SMOTE	0.76	0.68	0.57
2D	Response	RF	Recursive		0.71	0.75	0.39
2D	Response	RF	Relief	SMOTE	0.65	0.51	0.57
2D	Response	RF	Relief		0.75	0.84	0.35
2D	Response	RF	SFS	SMOTE	0.69	0.6	0.54
2D	Response	RF	SFS		0.69	0.8	0.26
2D	Response	SVM	Boruta	SMOTE	0.64	0.37	0.85
2D	Response	SVM	Boruta		0.55	0.01	0.88
2D	Response	SVM	Random Forest	SMOTE	0.88	0.27	0.92
2D	Response	SVM	Random Forest		0.87	0.44	0.8
2D	Response	SVM	Recursive	SMOTE	0.51	0.3	0.78
2D	Response	SVM	Recursive		0.54	0.01	0.85
2D	Response	SVM	Relief	SMOTE	0.78	0.46	0.88
2D	Response	SVM	Relief		0.47	1	0
2D	Response	SVM	SFS	SMOTE	0.75	0	1
2D	Response	SVM	SFS		0.57	0.2	0.9

Supp Table 2.7: Therapy response prediction with 3D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	Response	KNN	Boruta	SMOTE	0.8	0.45	0.72
3D	Response	KNN	Boruta		0.6	0.375	0.57
3D	Response	KNN	Random Forest	SMOTE	0.57	0.52	0.57
3D	Response	KNN	Random Forest		0.75	0.2	0.83
3D	Response	KNN	Recursive	SMOTE	0.62	0.57	0.75
3D	Response	KNN	Recursive		0.7	0.52	0.75
3D	Response	KNN	Relief	SMOTE	0.68	0.45	0.72
3D	Response	KNN	Relief		0.73	0.38	0.7
3D	Response	KNN	SFS	SMOTE	0.78	0.9	0.5
3D	Response	KNN	SFS		0.57	0.48	0.54
3D	Response	Logistic Regression	Boruta	SMOTE	0.8	0.66	0.88
3D	Response	Logistic Regression	Boruta		0.76	0.4	0.83
3D	Response	Logistic Regression	Random Forest	SMOTE	0.69	0.68	0.65
3D	Response	Logistic Regression	Random Forest		0.88	0.2	0.8
3D	Response	Logistic Regression	Recursive	SMOTE	0.55	0.52	0.66
3D	Response	Logistic Regression	Recursive		0.61	0	1
3D	Response	Logistic Regression	Relief	SMOTE	0.76	0.67	0.32
3D	Response	Logistic Regression	Relief		0.74	0.5	0.57
3D	Response	Logistic Regression	SFS	SMOTE	0.81	0.46	0.66
3D	Response	Logistic Regression	SFS		0.69	0.24	0.92
3D	Response	RF	Boruta	SMOTE	0.71	0.28	0.77
3D	Response	RF	Boruta		0.69	0.31	0.72
3D	Response	RF	Random Forest	SMOTE	0.47	0.275	0.705
3D	Response	RF	Random Forest		0.6	0.3	0.58
3D	Response	RF	Recursive	SMOTE	0.31	0.51	0.54
3D	Response	RF	Recursive		0.42	0.35	0.48
3D	Response	RF	Relief	SMOTE	0.51	0.6	0.66
3D	Response	RF	Relief		0.94	0.14	0.92
3D	Response	RF	SFS	SMOTE	0.71	0.38	0.62
3D	Response	RF	SFS		0.79	0.66	0.52
3D	Response	SVM	Boruta	SMOTE	0.72	0.19	0.92
3D	Response	SVM	Boruta		0.62	0	1
3D	Response	SVM	Random Forest	SMOTE	0.66	0.78	0.2
3D	Response	SVM	Random Forest		0.75	0	1
3D	Response	SVM	Recursive	SMOTE	0.74	0	0.92
3D	Response	SVM	Recursive		0.52	0	1
3D	Response	SVM	Relief	SMOTE	0.8	0.1	0.82
3D	Response	SVM	Relief		0.85	0	1
3D	Response	SVM	SFS	SMOTE	0.83	0.46	0.81
3D	Response	SVM	SFS		0.7	0.12	0.95

Supp Table 2.8: Therapy response prediction with 3D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	Response	KNN	Boruta	SMOTE	0.52	0.33	0.51
3D	Response	KNN	Boruta		0.51	0.48	0.4
3D	Response	KNN	Random Forest	SMOTE	0.66	0.32	0.7
3D	Response	KNN	Random Forest		0.78	0.2	0.8
3D	Response	KNN	Recursive	SMOTE	0.62	0.47	0.48
3D	Response	KNN	Recursive		0.63	0.2	0.69
3D	Response	KNN	Relief	SMOTE	0.75	0.72	0.6
3D	Response	KNN	Relief		0.78	0.12	0.94
3D	Response	KNN	SFS	SMOTE	0.76	0.41	0.7
3D	Response	KNN	SFS		0.6	0.2	0.7
3D	Response	Logistic Regression	Boruta	SMOTE	0.75	0.48	0.65
3D	Response	Logistic Regression	Boruta		0.9	0.3	0.9
3D	Response	Logistic Regression	Random Forest	SMOTE	0.58	0.44	0.52
3D	Response	Logistic Regression	Random Forest		0.53	0.1	0.95
3D	Response	Logistic Regression	Recursive	SMOTE	0.83	0.65	0.6
3D	Response	Logistic Regression	Recursive		0.78	0.3	0.8
3D	Response	Logistic Regression	Relief	SMOTE	0.7	0.77	0.3
3D	Response	Logistic Regression	Relief		0.79	0.12	0.86
3D	Response	Logistic Regression	SFS	SMOTE	0.87	0.33	0.74
3D	Response	Logistic Regression	SFS		0.68	0.01	0.94
3D	Response	RF	Boruta	SMOTE	0.64	0.44	0.63
3D	Response	RF	Boruta		0.73	0.86	0.32
3D	Response	RF	Random Forest	SMOTE	0.74	0.53	0.66
3D	Response	RF	Random Forest		0.7	0.75	0.26
3D	Response	RF	Recursive	SMOTE	0.74	0.62	0.43
3D	Response	RF	Recursive		0.44	0.75	0.36
3D	Response	RF	Relief	SMOTE	0.54	0.59	0.44
3D	Response	RF	Relief		0.62	0.75	0.46
3D	Response	RF	SFS	SMOTE	0.66	0.42	0.57
3D	Response	RF	SFS		0.6	0.62	0.44
3D	Response	SVM	Boruta	SMOTE	0.57	0.25	0.67
3D	Response	SVM	Boruta		0.8	0	1
3D	Response	SVM	Random Forest	SMOTE	0.74	0.2	0.9
3D	Response	SVM	Random Forest		0.7	0	1
3D	Response	SVM	Recursive	SMOTE	0.53	0.98	0.1
3D	Response	SVM	Recursive		0.69	0.01	0.81
3D	Response	SVM	Relief	SMOTE	0.74	0.59	0.35
3D	Response	SVM	Relief		0.71	0	1
3D	Response	SVM	SFS	SMOTE	0.89	0.01	0.9
3D	Response	SVM	SFS		0.88	0.3	0.9

Supplementary Information 3

Table 3.1: Coefficient of variation ($\frac{\sigma}{|x|}$). of radiomics features

Radiomics Feature	2D coefficient of variation	3D coefficient of variation
minValue	21.90	3.75
meanValue	0.94	1.08
stdValue	0.88	0.78
maxValue	0.51	0.54
HISTO_Skewness	60.41	4.43
HISTO_Kurtosis	0.28	0.81
HISTO_Entropy_log10	0.24	0.23
HISTO_Entropy_log2	0.24	0.23
HISTO_Energy	0.62	0.61
GLCM_Homogeneity	0.55	0.38
GLCM_Energy	1.64	1.63
GLCM_Contrast	3.58	2.80
GLCM_Correlation	0.70	0.36
GLCM_Entropy_log10	0.55	0.61
GLCM_Entropy_log2	0.55	0.36
GLCM_Dissimilarity	0.90	0.85
GLRLM_SRE	0.52	0.31
GLRLM_LRE	1.08	0.63
GLRLM_LGRE	2.81	0.40
GLRLM_HGRE	3.22	0.30
GLRLM_SRLGE	0.50	0.43
GLRLM_SRHGE	0.52	0.31
GLRLM_LRLGE	1.70	0.63
GLRLM_LRHGE	1.08	0.63
GLRLM_GLNU	3.73	4.45
GLRLM_RLNU	4.04	4.26
GLRLM_RP	0.53	0.31
NGLDM_Coarseness	1.53	1.33
NGLDM_Contrast	1.15	1.78
NGLDM_Busyness	2.59	1.60
GLZLM_SZE	0.55	0.36
GLZLM_LZE	4.90	4.43
GLZLM_LGZE	2.87	0.40
GLZLM_HGZE	0.50	0.30
GLZLM_SZLGE	4.34	0.53
GLZLM_SZHGE	0.56	0.36
GLZLM_LZLGE	4.86	4.48
GLZLM_LZHGE	4.94	4.40
GLZLM_GLNU	1.97	4.47
GLZLM_ZLNU	2.42	4.42
GLZLM_ZP	0.79	0.87
Median	**1.076**	**0.634**

Table 3.2: Mean and standard value (SD) of radiomic features

Radiomics Feature	2D		3D	
	Mean	SD	Mean	SD
minValue	3.9309	86.1002	-29.024	108.8098
meanValue	62.50549	58.5625	57.4985	61.9848
stdValue	22.05224	19.4946	26.7891	20.8367
maxValue	120.1881	61.2225	131.7911	70.8837
HISTO_Skewness	-0.00776	0.4688	-0.1296	0.5747
HISTO_Kurtosis	3.04419	0.8395	3.4344	2.7804
HISTO_Entropy_log10	0.85633	0.2095	0.925	0.2088
HISTO_Entropy_log2	2.84465	0.696	3.0727	0.6937
HISTO_Energy	0.18295	0.1131	0.158	0.0956
GLCM_Homogeneity	0.42394	0.234	0.4585	0.1733
GLCM_Energy	0.03809	0.0625	0.04	0.065
GLCM_Contrast	5.60013	20.0311	10.6954	29.8961
GLCM_Entropy_log10	0.37249	0.2591	1.5538	0.5623
GLCM_Correlation	1.29573	0.7082	0.35	0.2148
GLCM_Entropy_log2	4.30432	2.3527	5.1618	1.868
GLCM_Dissimilarity	1.32653	1.1922	1.8596	1.5728
GLRLM_SRE	0.6668	0.3454	0.8023	0.245
GLRLM_LRE	1.93866	2.0868	1.7452	1.0997
GLRLM_LGRE	7.94E-05	2.23E-04	8.19E-05	3.24E-05
GLRLM_HGRE	6.74E-05	2.17E-04	10594.4566	3182.0921
GLRLM_SRLGE	9262.53537	4647.2292	7.15E-05	3.11E-05
GLRLM_SRHGE	7697.86466	4025.1031	9209.5776	2898.8417
GLRLM_LRLGE	1.79E-04	3.04E-04	1.54E-04	9.76E-05
GLRLM_LRHGE	22384.51334	24141.3585	20081.1129	12740.7599
GLRLM_GLNU	104.41274	389.0792	566.0545	2521.7686
GLRLM_RLNU	562.60987	2270.5234	3169.6877	13504.7678
GLRLM_RP	0.63276	0.3328	0.7705	0.2405
NGLDM_Coarseness	0.06395	0.0976	0.016	0.0213
NGLDM_Contrast	0.02827	0.0325	0.0866	0.1543
NGLDM_Busyness	1.15815	2.9946	2.6829	4.2944
GLZLM_SZE	0.45316	0.2508	0.5249	0.1898
GLZLM_LZE	1355.38957	6637.224	8078.6047	35828.3955

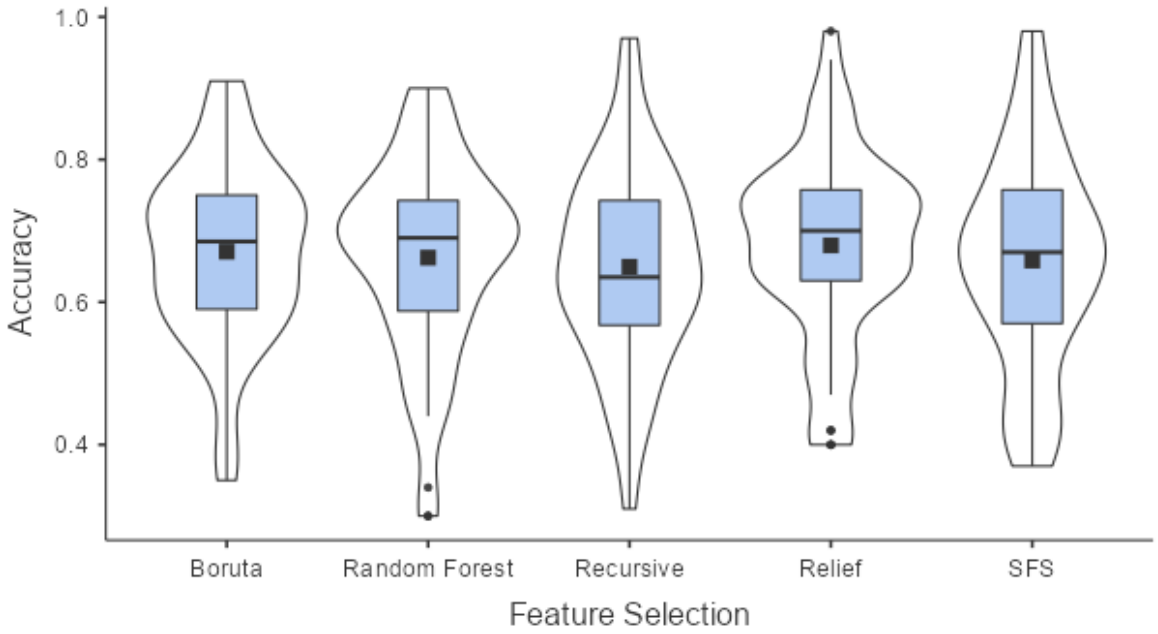
GLZLM_LGZE	7.97E-05	2.29E-04	8.23E-05	3.30E-05
GLZLM_HGZE	9270.86164	4654.5501	10575.6073	3193.705
GLZLM_SZLGE	4.86E-05	2.11E-04	4.75E-05	2.54E-05
GLZLM_SZHGE	5232.65207	2919.8812	5979.3646	2156.8144
GLZLM_LZLGE	0.11794	0.5731	0.6992	3.1326
GLZLM_LZHGE	1.56E+07	7.70E+07	9.35E+07	4.11E+08
GLZLM_GLNU	17.3782	34.1841	46.0707	206.1363
GLZLM_ZLNU	61.54293	149.164	199.1607	879.8972
GLZLM_ZP	0.25801	0.2046	0.1627	0.1412

Supplementary Information 4

Supp Table 4.1:
Descriptive statistics of performances by feature selection method.

Feature selection method	Acc		Sen		Spe	
	Mean	SD	Mean	SD	Mean	SD
Boruta	0.671	0.123	0.514	0.262	0.563	0.260
RF	0.662	0.136	0.530	0.280	0.524	0.305
Recursive	0.650	0.138	0.515	0.302	0.544	0.273
Relief	0.680	0.130	0.554	0.293	0.521	0.302
SFS	0.658	0.149	0.517	0.295	0.570	0.285

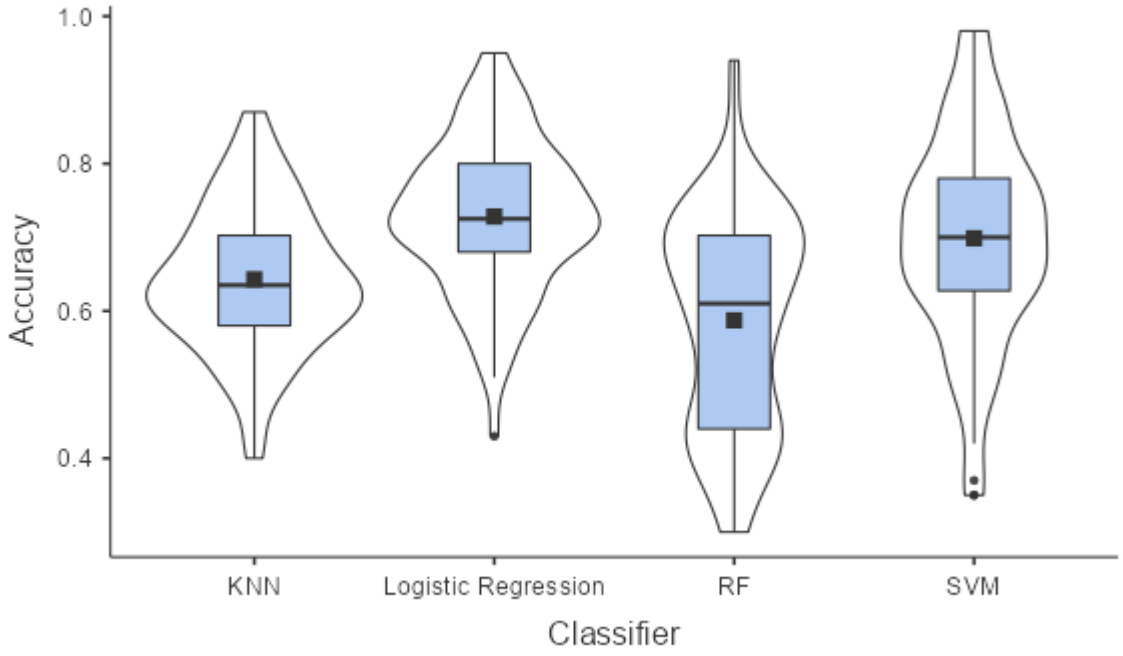
Supp Figure 4.1: Violin plot of performances by feature selection method



Supp Table 4.2: Descriptive statistics of performances by classifier

Classifier	Acc		Sen		Spe	
	Mean	SD	Mean	SD	Mean	SD
KNN	0.643	0.098	0.529	0.191	0.529	0.178
LR	0.728	0.103	0.545	0.309	0.571	0.317
RF	0.587	0.146	0.585	0.214	0.468	0.195
SVM	0.699	0.142	0.446	0.375	0.609	0.381

Supp Figure 4.2: Violin plot of performances by classifier



Supp Table 4.3 : Pair wise comparison of accuracy by classifier (Wilcoxon test, Bonferroni's correction for multiple comparison, *=significant difference)

	KNN		
LR	2.86E-06*	LR	
RF	0.218	6.54E-09*	RF
SVM	0.01*	1	5.35E-05*

Supplementary Information 5

Supp Table 5.1: Test and train accuracies of the 12 best models in our study for OS prediction (mean over the 5-fold cross validation)

ROI type	Combination (Classifier + Feature Selection +/- Smote Data Augmentation)	Train Acc	Train error	Test Acc	Train/test Acc ratio
3D	Logistic Regression + Boruta	0.8	0.2	0.91	0.88
	Logistic Regression + Boruta + SMOTE	0.9	0.1	0.88	1.02
	Logistic Regression + Random Forest	0.75	0.25	0.88	0.85
3D + clinical data	SVM + SFS	0.9	0.1	0.84	1.07
	SVM + RF + SMOTE	0.8	0.2	0.78	1.03
	Logistic Regression + Recursive + SMOTE	0.75	0.25	0.7	1.07
2D	Logistic Regression + RF + SMOTE	0.85	0.15	0.81	1.05
	RF + Boruta	0.8	0.2	0.75	1.07
	Logistic Regression + Boruta + SMOTE	0.8	0.2	0.74	1.08
2D + clinical data	Logistic Regression + Recursive + SMOTE	0.9	0.1	0.86	1.05
	KNN + Relief + SMOTE	0.77	0.23	0.87	0.89
	SVM + Boruta	0.85	0.15	0.84	1.01
			Mean Train Error 0.18		**Mean Ratio 1.01**

Supp Table 5.2: Test and train accuracies of the 12 best models in our study for response prediction (mean over the 5-fold cross validation)

ROI type	Combination (Classifier + Feature Selection +/- SMOTE Data Augmentation	Train Acc	Train Error	Test Acc	Train/test Acc ratio
3D	Logistic Regression + Boruta + SMOTE	0.88	0.12	0.8	1.1
	SVM + SFS + SMOTE	0.8	0.2	0.83	0.96
	Logistic Regression + SFS + SMOTE	0.85	0.15	0.81	1.05
3D + clinical data	KNN + Relief + SMOTE	0.83	0.17	0.75	1.11
	Logistic Regression + Recursive + SMOTE	0.9	0.1	0.83	1.08
	RF + RF + SMOTE	0.8	0.2	0.74	1.08
2D	Logistic Regression + Relief + SMOTE	0.9	0.1	0.81	1.11
	Logistic Regression + RF + SMOTE	0.8	0.2	0.6	1.33
	RF + Recursive + SMOTE	0.75	0.25	0.76	0.99
2D + clinical data	SVM + RF	0.9	0.1	0.87	1.03
	SVM + Relief + SMOTE	0.7	0.3	0.78	0.9
	RF + Recursive + SMOTE	0.7	0.3	0.76	0.92
			Mean Train Error 0.18		**Mean ratio 1.06**

Supp Table 5.3: Mean train/test accuracy ratio on all classifications by each classifier for OS prediction

Classifier	Mean train/test Acc ratio
KNN	1.04
LR	1.01
RF	1.07
SVM	0.96

Supp Table 5.4: Mean train/test accuracy ratio on all classification by each classifier for response prediction

Classifier	Mean train/test Acc ratio
KNN	1.02
LR	1.01
RF	1.05
SVM	0.95

Limites de la radiomique et solutions possibles

Après avoir vu deux études d'application de la radiomique, nous allons lister les limites de cette approche et de présenter les solutions adaptées.

La radiomique a connu un essor important quant au nombre de publications (Figure 10), mais contrastant avec les applications cliniques en pratique clinique qui sont limitées. Il a même été posé la question provocante de savoir si « les images étaient vraiment des données ou simplement des motifs dans le bruit » (61).

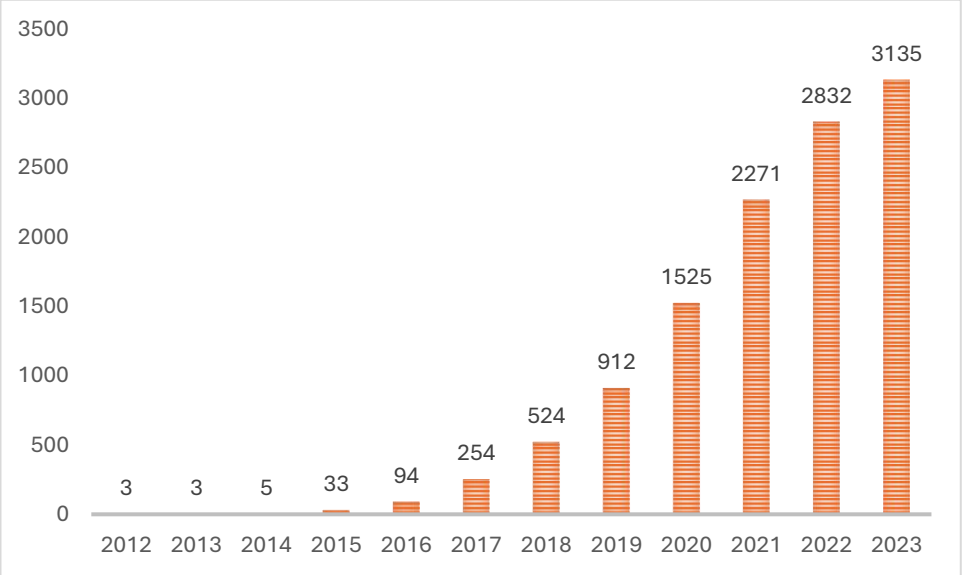


Figure 10 : Nombre de publications sur PubMed par année (d'après la recherche « radiomics » sur PubMed)

Les causes de cette difficulté d'application ont été étudiées (61–63). Parmi les causes figurent des manques de précautions méthodologiques pour s'assurer de la reproductibilité des mesures qui sont faites. En effet, différentes étapes dans la création d'un modèle de radiomique peuvent causer un manque de reproductibilité et donc de généralisation.

Nous allons citer certaines causes de façon non exhaustive :

- La variabilité test – retest. Elle est multifactorielle et doit être idéalement testée puis analysée par paramètre afin de ne garder que des paramètres invariants dans les conditions de l'examen
- La variabilité liée à la segmentation. Il est donc conseillé de vérifier la reproductibilité (inter- et intra- observateurs) des paramètres, et de sélectionner des paramètres robustes à la segmentation.
- La variabilité engendrée par des images provenant de machines différentes. Elle est liée à l'ensemble des paramètres d'acquisition modifiables et de méthode de reconstruction mais aussi à des spécificités intrinsèques des machines. Une approche peut être d'essayer de standardiser les paramètres d'acquisition. Une approche complémentaire va être un pré traitement des images (normalisation, discrétisation, rééchantillonnage) ou un post traitement des paramètres de radiomique (ComBat notamment) pour minimiser les différences liées aux spécificités intrinsèques de machines.
- La comparabilité des logiciels de radiomique. D'une part, le nombre et le nom des paramètres de radiomique n'est pas uniforme sur les logiciels, et les valeurs peuvent même varier selon les logiciels. Des recommandations ont été faites par l l'IBSI (Image Biomarker Standardization

Initiative) sur la définition des paramètres, la nomenclature et le calcul des valeurs. Il est donc recommandé d'utiliser un logiciel suivant ces recommandations (c'est le cas de PyRadiomics et LifeX, qui font partie des 4 logiciels les plus utilisés). Il est recommandé aussi de préciser la version du logiciel utilisé ainsi que les paramètres (rééchantillonnage, discrétisation…) utilisés.

- Le risque de sur apprentissage. Pour éviter ce risque, il faut pouvoir distinguer une population d'apprentissage sur lequel est développé le modèle et une population de validation, idéalement d'un autre centre, qui évalue le modèle. Des modèles de validation croisée sont possibles (64).

On peut obtenir plus de détails ainsi d'autres causes pour les difficultés d'application, ainsi que les solutions proposées en lisant les conseils et recommandations de l'IBSI (36,65) (https://theibsi.github.io/) ou très récemment de l' European Society of Medical Imaging Informatics (66).

Radiomique et reconstruction des images par apprentissage profond

Nous avons vu que les paramètres d'acquisition ou de reconstruction de l'image, en scanner comme en IRM, influent sur les paramètres de radiomique. Il se pose la question de savoir si les reconstructions d'image par apprentissage profond modifient aussi les paramètres de radiomique. Il y a en fait deux sous questions pour l'application pratique :

- Les paramètres extraits avec reconstruction par apprentissage profond sont ils comparables à ceux avec les reconstruction classiquement utilisées ? Cela permettrait de pouvoir comparer des paramètres extraits de machines aux méthodes de reconstruction différentes. Les travaux faits au scanner sur la précédente génération de reconstruction ont montré que la rétroprojection filtré et la reconstruction itérative produisent des paramètres de radiomique différents (42). Il est donc très probable que les paramètres issus des reconstructions par apprentissage profond soient différents de ceux issus d'autres méthodes de reconstruction. Néanmoins si les paramètres sont bien différents, il existe théoriquement des moyens de combler ces différences en utilisant les méthodes déjà existantes (ComBat..., cf. p 36).

- Les paramètres extraits avec une reconstruction par apprentissage profond sont ils plus stables ? Cela veut dire que l'on cherche à savoir si la variabilité des paramètres de radiomique sur des acquisitions répétées (dans les mêmes conditions ou des conditions proches) est plus faible avec la reconstruction par apprentissage profond. L'impact pratique est plus important puisqu'il n'existe pas de moyen de compenser l'instabilité liée à une étape de la formation de l'image. En cas de stabilité moindre avec les nouvelles reconstructions, les études de radiomique seraient déconseillées avec ces reconstructions.

Nous allons détailler l'état de l'art des travaux sur ces deux questions.

Au scanner

- Les paramètres extraits avec deux méthodes différentes de reconstruction sont ils comparables ?

Xue et coll. (67) ont comparé les paramètres de radiomique de 60 patients dans des tumeurs hépatiques, en péri tumoral et dans le foie, après reconstruction par rétroprojection filtrée (FBP), deux niveaux de reconstruction itérative (ASIR-V 30 % et 70 %) et 3 niveaux de reconstruction par apprentissage profond (DLIR-L, DLIR-M, DLIR-H). Le pourcentage de paramètres statiquement non différents de ceux en FBP était de 71% pour l'ASIR-V 30% et 24 % pour l'ASIR-V 70%, ce qui va avec les résultats déjà connus (42). En comparant toujours FBP avec les reconstructions par apprentissage profond, ce pourcentage était de 32% pour DLIR-L, 28% pour DLIR-M et 24% pour DLIR-H. Les données comparant les reconstructions ASIR-V et DLIR sont précisées uniquement par catégorie de paramètres, comprises entre 26% (paramètres de 2^{e} ordre entre ASIR-V 30% et DLIR-H) et 100 % (paramètres de 1er ordre entre ASIR-V 70% et DLIR-M)

Zhong et coll. (68) ont réalisé une expérience sur fantôme, avec une acquisition en double énergie reconstruite en ASIR-V puis par apprentissage profond (DLIR), montrant aussi une différence entre les paramètres de radiomique extraits.

Ces études montrent donc, comme attendu, que les paramètres de radiomique ne sont pas identiques entre une reconstruction par apprentissage profond et une reconstruction itérative (ou une rétroprojection filtrée)

- Les paramètres extraits avec une reconstruction par apprentissage profond sont ils plus stables ?

Zhong et coll. (68) ont étudié partiellement le problème de répétabilité, en répétant l'acquisition. Leurs données de répétabilité (dans leur Supplementary Material 6) montre des indices de répétabilité plus élevés en DLIR qu'en ASIR-V.

Michallek et coll. (69) ont réalisé une étude sur fantôme simulant un foie siège de lésions, répétant 20 acquisitions en faisant varier les doses, avec une reconstruction par rétroprojection filtrée (FBP), deux reconstructions itératives (FIRST et AIDR 3D) ainsi qu'une par apprentissage profond (AiCE). La reproductibilité des paramètres était la plus élevée pour FBP pour des doses basses (CTDIvol<2.7 mGy environ), et pour AiCE au-delà ; les reconstructions itératives étant moins reproductibles quelque soit la dose. En ajoutant des critères de cohérence sur un même type de tissu et de discrimination entre tissus, les auteurs trouvent une supériorité des paramètres en AiCE.

Les études tendent donc à montrer une meilleure stabilité (répétabilité ou reproductibilité) des paramètres de radiomique après reconstruction du scanner par apprentissage profond que les autres méthodes de reconstruction.

En IRM

Les études sur les paramètres de radiomique en IRM avec reconstruction par apprentissage profond sont rares.

Seule une étude par Li et coll. (70) aide à répondre à la première des deux questions sur l'influence des reconstructions par apprentissage profond sur les paramètres de radiomique en IRM. Ils ont étudié les paramètres de radiomique de l'IRM de 17 patients reconstruits de façon classique (C SENSE) et par 2 méthodes par apprentissage profond (SmartSpeed et SmartSpeed-SuperRes). Les paramètres n'étaient pas identiques, avec 50 sur 86 paramètres corrélés entre C SENSE et Smart Speed et 15 sur 86 paramètres corrélés entre C SENSE et Smart Speed-SuperRes.

Cette étude tend à montrer, comme attendu, que les paramètres de radiomique ne sont pas identiques avec une reconstruction par apprentissage profond et avec une reconstruction classique en IRM.

La deuxième question, la stabilité des paramètres de radiomique après reconstruction par apprentissage profond, n'a pas été étudiée. Elle est importante car elle analyse des facteurs rendant reproductibles les études de radiomique. Elle est d'autant plus importante que l'utilisation de telles reconstructions se généralise en pratique clinique courante. Il faut donc savoir si l'utilisation de cette nouvelle technique va rendre les résultats de radiomique plus stables ou non. Cette question sera l'objet de la quatrième étude.

Quatrième étude : Impact des reconstructions par apprentissage profond en IRM sur la stabilité des paramètres de radiomique

Cet article est prévu pour une première soumission au journal Radiology (IF 12.1, rang A).

MRI Deep Learning reconstruction enhances radiomic feature stability

Gabriel Ballout [1], Navid Rabbani PhD [2,3], Bruno Pereira PhD [4], Lucie Cassagnes MD PhD[1,2], Pascal Chabrot MD PhD [1,2], Adrien Bartoli PhD [2,3], Benoit Magnin MD MSc MEng[1,2,3]*

1 Radiology Department, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

2 Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, Clermont-Ferrand, France

3 DI2AM, DRCI, Clermont University Hospital, Clermont-Ferrand, France

4 Biostatistics Unit, DRCI, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

* corresponding author

Benoît Magnin
Radiology Department CHU Estaing
1 place Lucie Aubrac 63100 Clermont Ferrand
00 33 4 73 75 02 44
bmagnin@chu-clermontferrand.fr

Funding Information
No specific funding

Manuscript type
Original research

Word count for text
2972 words

Abstract

Background
Radiomic analysis is impacted by the parameters used in image acquisition and reconstruction. Deep Learning (DL) reconstruction methods are increasingly used in MRI to effectively reduce noise, leading to enhanced image quality. However, their impact on radiomic features, particularly regarding stability, has not yet been explored.

Purpose
To compare the reproducibility and repeatability of MRI radiomic features between conventional and DL reconstructions.

Materials and Methods

An MRI protocol comprising seven sequences was conducted on a 1.5T GE Healthcare™ MRI scanner using 12 fruits as phantom models. Each sequence was acquired with four distinct protocols, repeated four times. Images were reconstructed using both conventional (non-DL) and DL techniques (AIR Recon DL, GE Healthcare™). After semi-automated segmentation, 107 radiomic features were extracted. Features were defined as reproducible if the intraclass correlation coefficient (ICC(3,1)) exceeded 0.9 and repeatable if the overall concordance correlation coefficient (OCCC) was above 0.9. Additional statistical criteria and a 0.75 threshold were also tested.

Results
DL reconstruction yielded 53.27% reproducible features, compared to 43.52% with conventional reconstruction. Among the feature types, DL improved reproducibility by 7.94% for first-order features and by 12.00% for higher-order features. The difference in repeatability between reconstructions was 0.4%, favoring DL reconstruction.

The use of other statistical criteria or threshold showed the same trend.

Conclusion
DL reconstruction improves the reproducibility of radiomic features relative to conventional methods, supporting the use of DL-based reconstruction in MRI radiomic studies.

Summary

Deep Learning based MRI reconstruction improves the number of reproducible radiomic features compared to conventional reconstruction by 7.94 %.

Key Results

- In this phantom study, we assessed reproducibility and repeatability of radiomic features after Deep Learning and conventional MRI reconstructions.
- DL reconstructions improved the percentage of reproducible features by 7.94% and of repeatable features by 0.4%.

Abbreviations

DLIR: Deep learning-based image reconstructions

DR: Dynamic Range

GLCM: gray-level co-occurrence matrix

GLDM: gray-level dependence matrix

GLRLM: gray-level run-length matrix

GLSZM: gray-level size zone matrix

OCCC: Overall concordance correlation coefficient

IBSI: Image Biomarker Standardization Initiative

ICC: Intraclass correlation coefficient

NGTDM: neighborhood gray-tone difference matrix

Introduction

Radiomics aims at extracting quantitative information, known as radiomic features, from medical images (1). Radiomic features can be used either alone or in combination with other data (clinical, biological, genetical) for clinical purpose, such as diagnosis, treatment response or prognostic prediction. Numerous studies aimed to produce a predictive model for a given pathology, based on radiomic parameters extracted from a training population (2,3).

However, despite the growing number of publications on radiomics, its practical application remains limited (4), due to the difficulty in reproducing results. One factor contributing to this lack of reproducibility is the complexity of the radiomics study methodology. To address this, efforts have been made to standardize methodologies, particularly through initiatives such as the Image Biomarker Standardization Initiative (IBSI) (4),

Thus, methodological studies have highlighted the factors influencing the stability of radiomic parameters in CT scans (5), in PET imaging, and to a lesser extent in MRI. Hence, it has been shown that the stability of MRI-derived radiomic features is influenced by the scanner and its manufacturer (7,8), the acquisition parameters (8,9), image post processing (10,11) and by the observer (10).

Deep learning-based image reconstructions (DLIR) are now available in CT and MRI. They have demonstrated advantages over traditional reconstruction methods in terms of noise reduction (12,13) and diagnostic performance (14) and are therefore now used in daily practice. The influence of DLIR on radiomic features has been studied in CT imaging. As expected, CT radiomic features after DLIR differ from those obtained with standard iterative reconstruction (15,16), but also improves radiomics reproducibility (17,18). Only one study focused on the DLIR impact on MRI radiomic feature, showing that 41% of MRI DLIR radiomic features are have moderate or low concordance to non DLIR radiomic features (19). However, the reproducibility of these features has never been tested.

The objective of our study is to evaluate the influence of DLIR on the stability of radiomic parameters in MRI. More specifically, we want to assess the repeatability (comparison under constant condition) and reproducibility (comparison under varying conditions) of radiomic features after DLIR compared to a conventional non DLIR.

Methods

The design of the study is shown on Figure 1.

Phantom

We chose to use a phantom composed of different fruits, as organic components are known to have features close to human tissues and lesions: we selected 3 oranges, 3 kiwis, 3 onions, and 3 apples, as has been done by several teams (11,20).

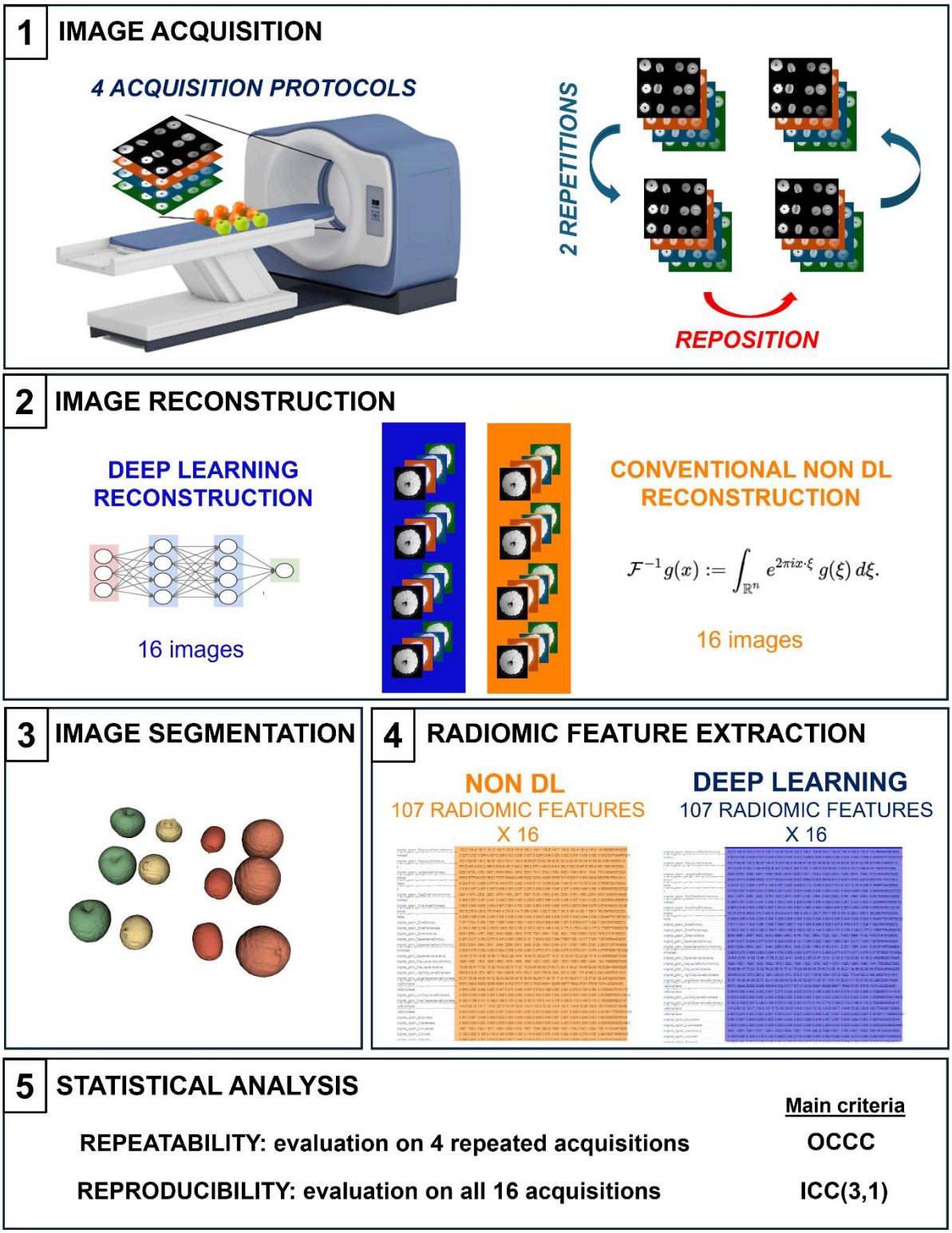


Figure 1: Study design

Image acquisition

All images were performed on a 1,5T Signa Artist MRI (GE Healthcare). We performed an MRI protocol consisting of 7 sequences commonly used during routine scans: 2D T1, 2D T2, 3D T1, 3D T2, 2D FLAIR, Diffusion, T2 In Phase.

In order to introduce variability and simulate inter scanner differences, each sequence was performed with a reference protocol and then repeated with 3 other protocols, each with a variation of one acquisition parameter: resolution, slice thickness, or TE. Acquisition parameters are detailed in Table 1.

		In plane resolution (mm x mm)	Matrix	Slice thickness (mm)	TR (ms)	TE (ms)	Flip angle (°)	NEX	b
Reference	**FLAIR**	0.7 x 0.9	352 x 256	3	12000	120	160	1	
	2D T1	0.6 x 0.8	384 x 320	3	337	Min Full	160	2	
	2D T2	0.8 x 1.2	288 x 192	3	Min	Min Full		1	
	3D T1	0.8 x 0.8	300 x 300	1.2	Min	Min Full	15	1	
	3D T2	0.8 x 0.8	320 x 320	0.8	1800	140		2	
	Diffusion	1.9 x 1.2	128 x 192	3	600	80			1000
	T2 IP	0.8 x 0.8	288 x 288	3	5759	108	140	2	
Modified in plane resolution	**FLAIR**	0.6 x 0.8	400 x 300	3	12000	120	160	1	
	2D T1	0.6 x 0.6	416 x 416	3	337	Min Full	160	2	
	2D T2	0.7 x 1	352 x 240	3	Min	Min Full		1	
	3D T1	0.6 x 0.6	400 x 400	1.2	Min	Min Full	15	1	
	3D T2	0.6 x 0.6	380 x 380	0.8	1800	140		2	
	Diffusion	1.5 x 1.0	160 x 240	3	600	80			1000
	T2 IP	0.8 x 0.8	320 x 320	3	6536	108	140	2	
Modified TE	**FLAIR**	0.7 x 0.9	352 x 256	3	12000	145	160	1	
	2D T1	0.6 x 0.8	384 x 320	3	337	18	160	2	
	2D T2	0.8 x 1.2	288 x 192	3	Min	170		1	
	3D T1	0.8 x 0.8	300 x 300	1.2	Min	Out of Phase	15	1	
	3D T2	0.8 x 0.8	320 x 320	0.8	1800	170		2	
	Diffusion	1.9 x 1.2	128 x 192	3	8500	Min			2000
	T2 IP	0.8 x 0.8	288 x 288	3	5759	130	140	2	
Modified slice thickness	**FLAIR**	0.7 x 0.9	352 x 256	5	12000	120	160	1	
	2D T1	0.6 x 0.8	384 x 320	5	337	Min Full	160	2	
	2D T2	0.8 x 1.2	288 x 192	5	Min	Min Full		1	
	3D T1	0.8 x 0.8	300 x 300	2	Min	Min Full	15	1	
	3D T2	0.8 x 0.8	320 x 320	1.6	1800	140		2	
	Diffusion	1.9 x 1.2	128 x 192	5	600	80			1000
	T2 IP	0.8 x 0.8	288 x 288	5	5759	108	140	2	

Table 1: Acquisitions parameters

The images were acquired twice consecutively; the phantom was then removed and repositioned, and another series of 2 two acquisitions were performed (scan-rescan-reposition-scan-rescan sequence). This resulted in 4 repeated acquisitions, each composed of 7 sequences, each sequence acquired with 4 parameters variations. In total we obtained 16 different acquisitions by sequence. All images were reconstructed using both standard reconstruction and a vendor-provided DL algorithm (AIR Recon DL, GE Healthcare).

Segmentation

A semi-automatic segmentation was performed by a radiology resident (G.B.) of each fruit on every sequence of the different protocols using the open-source software 3D Slicer (https://www.slicer.org) (21).

Radiomic feature extraction

PyRadiomic (22) (version 3.0.1) was used to extract radiomic features with a 1x1x1 mm resampling; all other settings were left at their default values. A total of 107 radiomic parameters were extracted including 14 shape based features, 18 first-order histogram features and 75 second-order or higher-order radiomic features (24 features from the gray-level co-occurrence matrix (GLCM), 16 features from the gray-level run-length matrix (GLRLM), 16 features from the gray-level size zone matrix (GLSZM), 14 features from the gray-level dependence matrix (GLDM) and 5 features from the neighborhood gray-tone difference matrix (NGTDM), all will be referred as “higher-order” features in contrast to first-order).

Statistical analysis

Reproducibility and repeatability of radiomic features have been evaluated with various statistical tests in the literature. We chose to evaluate repeatability and reproducibility each with a main statistical test and with two or three other secondary statistical tests, to try to evaluate the differences between the statistical methods.

To evaluate reproducibility, i.e. comparison under varying conditions, the features were evaluated on all 16 acquisitions (4 repetitions times 4 different acquisition parameters) with ICC(3,1) as defined in (27) as the main test as previously used (25,28). The secondary tests to evaluate reproducibility were ICC(2,1) and OCCC.

To evaluate repeatability, i.e. comparison under constant condition, the features were evaluated on the 4 repetitions (scan-rescan-reposition-scan-rescan) with the same acquisition parameters using overall concordance correlation coefficient (OCCC) as previously used (11,18,23–25). OCCC is a generation of twofold concordance correlation coefficient (26). The secondary tests to evaluate repeatability were ICC(1,1) and ICC(2,1) (27).

A feature was considered stable (reproducible or repeatable) when ICC or OCCC>0.9 (29). A secondary threshold of 0.75 was also tested.

Finally, similar to the combination of CCC and dynamic range (DR) previously used (9,11), we used a generalized combination of OCCC>0.9 and DR>0.9 to assess a reproducible feature. Statistical

analysis was performed in R (version 4.3.2, R Core Team, Vienna, Austria (30)) using the psych package for ICC calculation and the epiR package for OCCC.

Results

Reproducibility

Overall results

The number of reproducible features (using ICC(3,1)>0.9 as a reproducibility criteria, i.e. excellent reproducibility) by feature category and sequences are shown in Table 2. The detailed results by feature class among the higher order features are shown in Table 3.

The overall percentage of features exhibiting excellent reproducibility was in favor of DLIR compared to non-DLIR with a 9.75% difference (i.e. 73 features).

By feature categories

For shape features, 76.53% were reproducible regardless of the reconstruction method. First-order and higher-order features showed greater reproducibility in DLIR compared to non-DLIR, with a difference of 7.94% for first-order features and 12% for higher-order features. An example of the distribution of a second order feature is presented in Figure 2.

By sequences

There remains some heterogeneity in the results depending on the sequences; however, all sequences yielded more reproducible features with DLIR (Figure 3). The most significant difference was observed in the diffusion sequence, with a delta of 31.78% favoring DLIR.

For the majority of the sequences, specifically 4 out of 7 (FLAIR, 2D T1, 2D T2, and T2 In Phase), the results predominantly favored DLIR in higher-order features, particularly in the FLAIR sequence, where a difference of 10.66% was measured. In the 3D T1 and diffusion sequences, the differences were predominantly observed in first-order features, with a difference of 16.66% in 3D T1 and 44.44% in the diffusion sequence. In 3D T2, there was no overall difference between the reconstruction methods. The reproducibility of first-order features in this sequence was better with non-DLIR, representing a difference of 5.55%. Conversely, for higher-order features, the results favored DLIR, with a delta of 1.33%.

Figure 4 provides a visual overview comparing the reproducibility of each feature between the two reconstruction methods.

3D vs 2D acquisitions in DLIR

When comparing 3D and 2D acquisitions, the overall reproducibility rate in DLIR for T2 was slightly higher in 3D T2 compared to 2D T2, with a difference of 3.74%. For T1, there was a notable advantage for 2D, with 41.12% of features being reproducible in 2D compared to 28.04% in 3D, indicating a difference of 13.08%. For shape features, reproducibility remained high in 3D, with 92.86% of features in both 3D T1 and 3D T2, compared to only 50% in 2D T1 and 2D T2.

		DL	Non-DL
All sequences	**All Features**	**53.27**(399)	**43.52 (**326)
	Shape Features	76.53 (75)	76.53 (75)
	First Order Features	61.11 (77)	53.17 (67)
	Higher Order Features	47.04 (247)	35.04 (184)
FLAIR	All Features	**62.62** (67)	**55.14** (59)
	Shape Features	78.57 (11)	78.57 (11)
	First Order Features	72.22 (13)	72.22 (13)
	Higher Order Features	57.33 (43)	46.66 (35)
2DT1	All Features	**41.12** (44)	**34.57** (37)
	Shape Features	50 (7)	50 (7)
	First Order Features	55.55 (10)	55.55 (10)
	Higher Order Features	36.00 (27)	26.66 (20)
2DT2	All Features	**50.47** (54)	**44.86** (48)
	Shape Features	50 (7)	57.14 (8)
	First Order Features	66.66 (12)	66.66 (12)
	Higher Order Features	45.33 (34)	37.33 (28)

		DL	Non-DL
3DT1	All Features	**28.04** (30)	**15.89** (17)
	Shape Features	92.86 (13)	92.86 (13)
	First Order Features	16.66 (3)	0
	Higher Order Features	18.66 (14)	5.33 (4)
3DT2	All Features	**54.21** (58)	**54.21** (58)
	Shape Features	92.86 (13)	92.86 (13)
	First Order Features	50.00 (9)	55.55 (10)
	Higher Order Features	48.00 (36)	46.66 (35)
Diffusion	All Features	**49.53** (53)	**17.76** (19)
	Shape Features	71.43 (10)	71.43 (10)
	First Order Features	72.22 (13)	27.78 (5)
	Higher Order Features	40.00 (30)	5.33 (4)
T2IP	All Features	**86.92** (93)	**82.25** (88)
	Space Features	92.86 (13)	92.86 (13)
	First Order Features	94.44 (17)	94.44 (17)
	Higher Order Features	84.00 (63)	77.33 (58)

Table 2: Number of reproductible features defined as an ICC(3,1)>0.9 by feature category, sequence and reconstruction method. Data are presented as %(n)

	DL	Non-DL
All features	**53.27** (399)	**43.52** (326)
Shape	**76.53** (75)	**76.53** (75)
First order	**61.11** (77)	**53.17** (67)
Higher order	**47.05** (247)	**35.05** (184)
GLCM	59.58 (171)	48.78 (140)
GLRLM	41.96 (47)	26.79 (30)
GLSZM	38.39 (43)	28.57 (32)
GLDM	43.88 (43)	31.63 (31)
NGTDM	42.86 (15)	37.14 (13)

Table 3: Number of reproductible features defined as an ICC(3,1)>0.9 by feature category, feature class and reconstruction method. Data are presented as % (n)

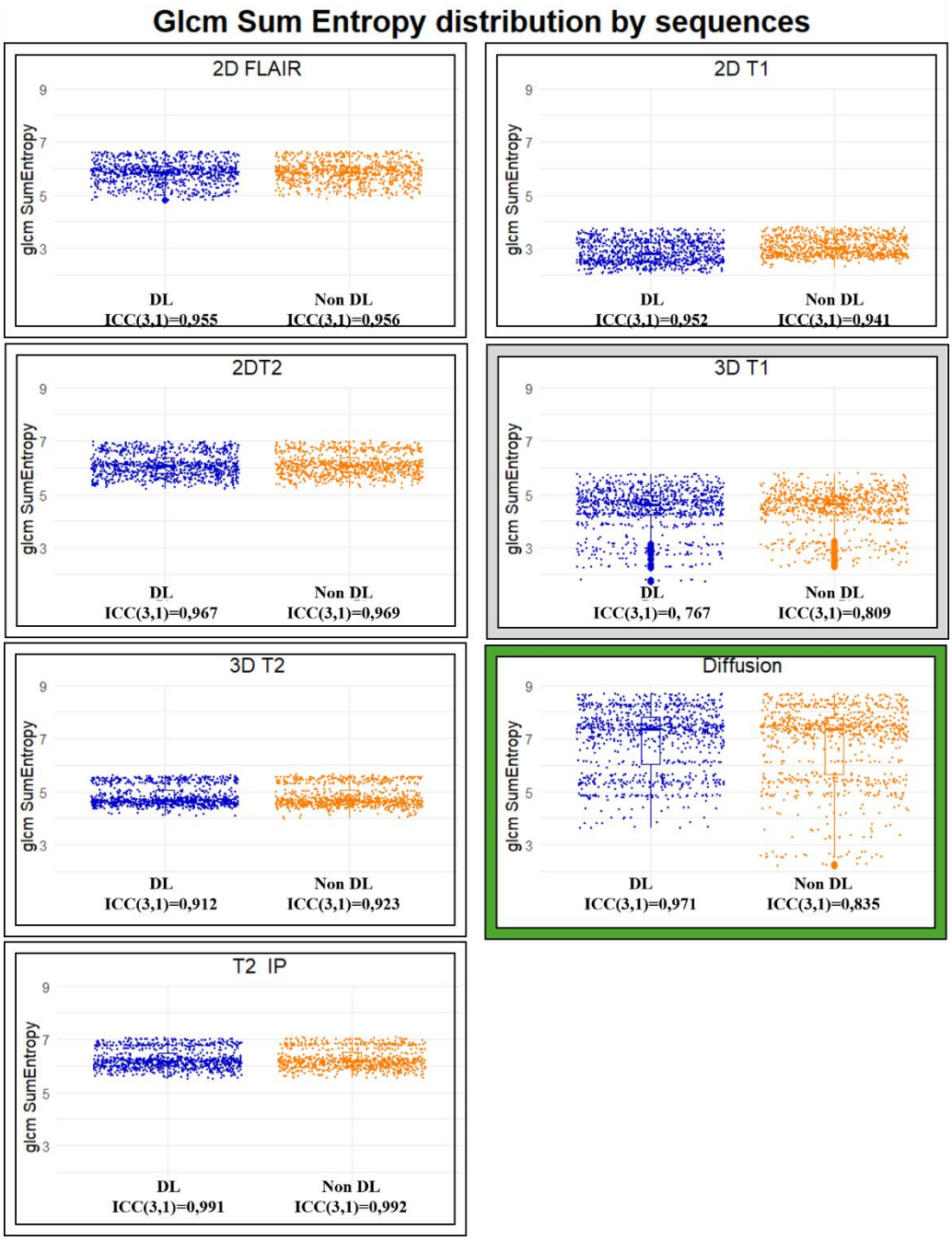


Figure 2: Distribution of one second order feature, Glcm Sum Entropy, by reconstruction and by sequence. Scatter plot and boxplot are represented, with corresponding ICC(3,1). The color frame indicates the reproducibility in both reconstructions (white: DL and non-DL reproducible, grey: DL and non DL non reproducible, green: DL reproducible, non DL not reproducible)

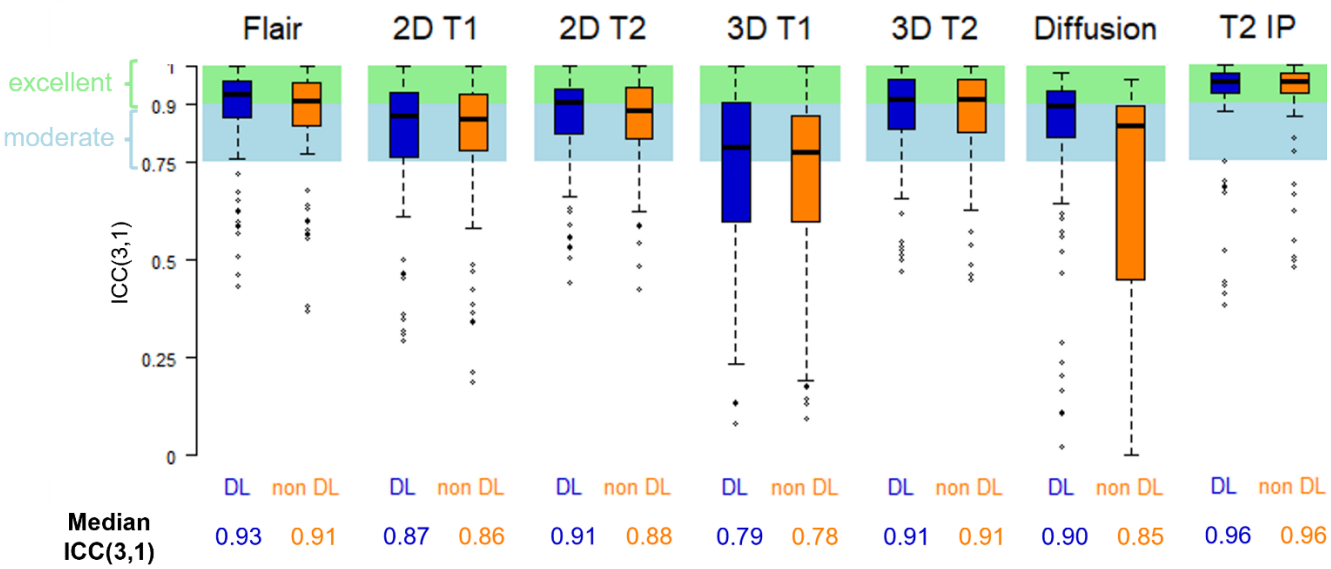


Figure 3: Boxplot of feature reproducibility evaluation by ICC(3,1) by reconstruction.

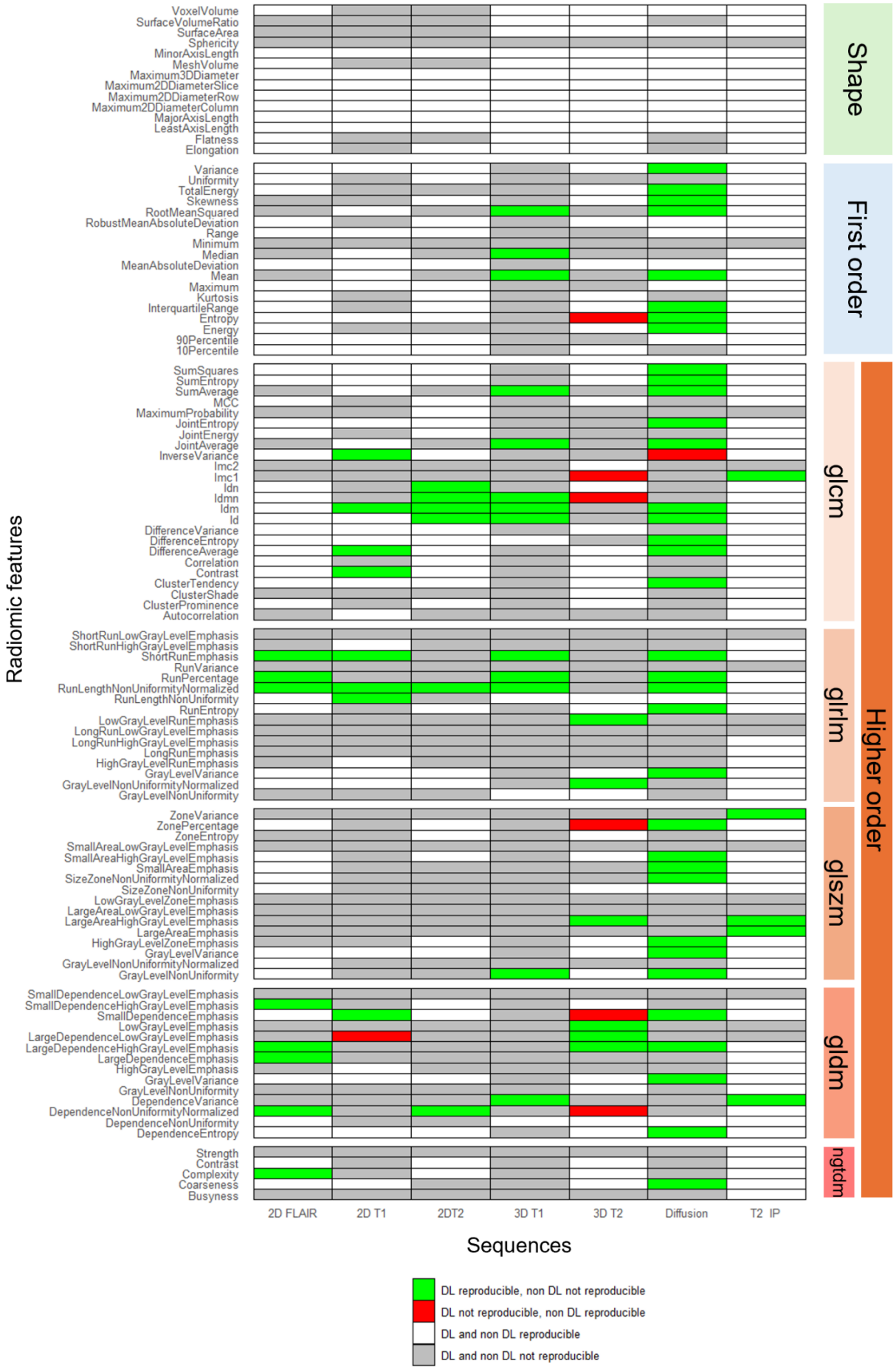


Figure 4: Heatmap comparing the reproducibility (defined as an ICC(3,1)>0.9) of each feature between the two reconstructions

Secondary criteria for reproducibility

Repeatability was also assessed using 3 other methods (ICC (2,1), OCCC and the combination of DR and OCCC), which revealed consistent trends (Figure 5 and Supplemental Material 1). Combining DR and OCCC did not modify the results of OCCC alone (Supplemental Material 1, hence not represented in Figure 5).

Additionally, applying a threshold of 0.75 identified a greater number of features meeting the criteria, but the overall trend remained unchanged. Using ICCC (3,1) with a threshold of 0.75 resulted however in a smaller difference between DLIR and non DLIR (1.73 %) than with a threshold of 0.9 (9.75%).

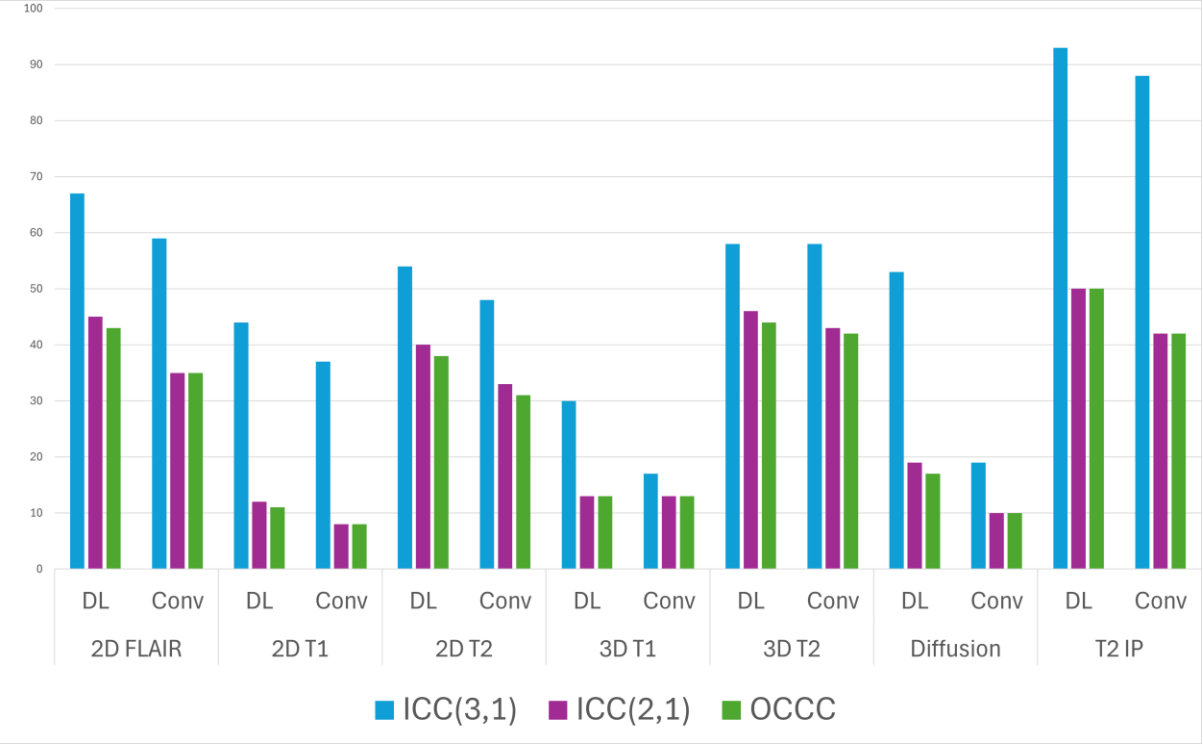


Figure 5: Comparison of the number of reproducible features by sequence, depending on the statistical test used (threshold =0.9)

Repeatability

Overall results

The number of repeatable features by feature category (using OCCC>0.9 as a repeatability criteria, i.e. excellent repeatability) is shown in Table 4.

The overall percentage of features exhibiting excellent repeatability was in favor of DLIR compared to non-DLIR with a 0.4% difference (i.e. 3 features).).

By feature categories

For shape features, repeatability was slightly higher with non-DLIR than with DLIR with a 1% difference (equivalent to one feature). First-order features appeared more repeatable with DLIR (53.97%), with a difference of 3.17% compared to non-DLIR. In contrast, higher-order features demonstrate only 43.43% with a satisfactory OCCC, regardless of whether DLIR or non-DLIR is used.

By sequences

The effect of reconstruction was not uniform across the sequences. Specifically, most sequences generally favor DLIR except for 2D T1 and 2D T2 with a 9.35% difference in 2D T1 and a 2.80% difference in 2D in favor of non-DLIR.

3D vs 2D acquisitions in DLIR

When comparing 3D and 2D acquisitions, the overall repeatability in DLIR rate was slightly higher in 3D T2 compared to 2D T2, with a difference of 2.8%. For T1, the 2D sequence was more repeatable, 39.25% of repeatable features in 2D compared to only 14.95% in 3D, showing a disparity of 24.3%. Regarding shape features, the repeatability rate was 92.86% for both 3D T1 and 3D T2, compared to 50% for 2D T1 and 2D T2.

Secondary criteria for repeatability

Repeatability was also assessed using two other methods (ICC(1,1) and ICC(2,1)), which revealed consistent trends (Supplemental Material 2). Additionally, applying a threshold of 0.75 identified a greater number of features meeting the criteria, but the overall trend remained unchanged, with the difference being approximately the same at around 1%.

		DL	Non-DL
All sequences	**All Features**	**49.27** (369)	**48.86** (366)
	Shape Features	74.49 (73)	75.51 (74)
	First Order Features	53.97 (68)	50.79 (64)
	Higher Order Features	43.43 (228)	43.43 (228)
FLAIR	All Features	**63.55** (68)	**58.88** (63)
	Shape Features	78.57 (11)	78.57 (11)
	First Order Features	61.11 (11)	61.11 (11)
	Higher Order Features	61.33 (46)	54.66 (41)
2DT1	All Features	**39.25** (42)	**48.59** (52)
	Shape Features	50 (7)	50 (7)
	First Order Features	77,77 (14)	66,66 (12)
	Higher Order Features	28 (21)	44 (33)
2DT2	All Features	**58,88** (63)	**61,68** (66)
	Shape Features	50 (7)	57,14 (8)
	First Order Features	72,22 (13)	72,22 (13)
	Higher Order Features	57,33 (43)	60 (45)
3DT1	All Features	**14.95** (16)	**14.02** (15)
	Shape Features	92.86 (13)	92.86 (13)
	First Order Features	0	0
	Higher Order Features	4.00 (3)	2.66 (2)
3DT2	All Features	**61.68** (66)	**60.75** (65)
	Shape Features	92.86 (13)	92.86 (13)
	First Order Features	55.55 (10)	61.11 (11)
	Higher Order Features	57.33 (43)	54.66 (41)
Diffusion	All Features	**15.89** (17)	**8.41** (9)
	Shape Features	64,29 (9)	64,29 (9)
	First Order Features	11,11 (2)	0
	Higher Order Features	8,00 (6)	0
T2IP	All Features	**90,65** (97)	**89,72** (96)
	Shape Features	92,86 (13)	92,86 (13)
	First Order Features	100 (18)	94 (17)
	Higher Order Features	88 (66)	88 (66)

Table 4: Number of repeatable features defined as an OCCC>0.9 by feature category, sequence and reconstruction method. Data are presented as % (n)

Discussion

Our study aimed to evaluate the impact of a recent Deep Learning reconstruction method - increasingly integrated into routine radiologic practice - on radiomic feature stability - a critical factor for the clinical application of radiomics. DLIR offers the benefits of noise reduction and enhanced image quality but also impacts radiomic features, potentially altering their stability. This hypothesis has led to prior investigations in CT (17,18) but has never been investigated in MRI.

Our findings indicate that the repeatability of radiomic features is minimally affected by the reconstruction method, with only a 0.4% difference in repeatable features between DLIR and non-DLIR. In contrast, DLIR notably enhances the reproducibility of radiomic features, with an improvement of 9.75% more features deemed reproducible compared to non-DLIR reconstruction.

It was expected that higher-order features would be less reproducible than first-order ones (18,31), but it was not initially clear that reproducibility differences would arise between DLIR and non-DLIR. Our findings indicate that all features' classes are indeed more repeatable and reproducible with DLIR. Additionally, the influence of DLIR on the reproducibility of higher-order features is especially pronounced, with 12% more features considered reproducible compared to non-DLIR. Non-stationary and textured noise within images degrades overall quality and contributes to low reproducibility of radiomic features (18). DLIR reduces this noise (12,13), which might explain the observed improvements in feature reproducibility. The high sensitivity of higher-order features to noise variability likely accounts for the more substantial improvement in their reproducibility with DLIR.

In our study, shape features appear to be the most stable, with 74.49% in DL and 75.51% in non-DL of features having excellent repeatability, and 76.53% in both DL and non-DL having excellent reproducibility. These results are consistent with those found by Bernatz et al. (20). The difference in repeatability between reconstruction methods is negligible, with only a 1-feature difference, which is not surprising given that both methods are applied to the same acquisition (same k-space), and segmentation masks were identical. In our protocol, we studied T1 and T2 sequences in 2D and 3D, observing that overall repeatability and reproducibility rates of radiomic features slightly favor 3D for T2 sequences and clearly favor 2D for T1. However, the stability rates of shape features are significantly higher in 3D sequences. A possible explanation is that the improved spatial resolution in 3D makes non-shape radiomic features less stable due to increased detail and greater sensitivity to noise. These findings are consistent with those reported by Bianchini et al. (8).

Repeatability analysis by sequence reveals substantial heterogeneity, independent of the statistical method used. Notably, 2DT1 and 2DT2 features show greater repeatability with non-DLIR—a finding that contrasts with the general trend favoring DLIR. However, such variability across sequences aligns with observations from other research teams, where the stability of specific sequences varies across protocols. In contrast, reproducibility analysis presents a more consistent pattern, with DLIR enhancing reproducibility across all sequences. The diffusion sequence, in particular, demonstrates the greatest improvement, with 31.78% more features deemed reproducible under DLIR. This pronounced difference may be attributed to the diffusion sequence's sensitivity to artifacts and contrast variations.

Literature reviews reveal a lack of consensus regarding the statistical methods used to assess the repeatability and reproducibility of radiomics (6,32). Furthermore, certain studies fail to specify the exact type of ICC used or to provide details on its implementation. Variations in ICC definitions can also exist (27,33), though a recent review has tried to simplify these distinctions (34). In our study, we empirically tested the most commonly used methods in the literature for evaluating both repeatability

and reproducibility. All the statistical approaches applied confirmed the same trend, consistently ranking DLIR and non-DLIR reproducibility in the same order. Although the absolute counts of reproducible or repeatable features varied by method, this did not impact on our study's main goal which was a comparison between the two reconstruction techniques. A technical explanation of the differences between statistical methods is beyond the scope of our study. However, we were able to confirm that our results remained consistent regardless of the statistical approach chosen.

Consequently, it is essential to examine the percentages of reproducible (53.27%) and repeatable (49.27%) DLIR features in a critical manner, as these values, though improved compared to non DLIR, may still appear relatively low. Indeed, studies on the reproducibility and repeatability of MRI-based radiomic features using test-retest protocols remain limited. Furthermore, the existing studies often apply varying statistical tests and thresholds, which limits direct comparability with our results. For example, Baeßler et al. (9) reported an interobserver ICC >0.75 on the FLAIR sequence, a metric that is challenging to compare directly with our findings due to differences in study design, and the absence of description of the type of ICC. However, among our study results, certain findings are comparable to data in the literature. Using ICC (3,1) and OCCC with a 0.75 threshold, 79.97% of features met reproducibility criteria, and 73.43% met repeatability criteria, yielding similar outcomes to some publications (25,35).

The first limitation of our study is that using a phantom cannot fully replicate a clinical patient study. However, this methodological approach is widely used in methodological radiomic research for several reasons: first, conducting repeated, lengthy acquisitions on human subjects is hardly feasible; second, phantoms provide signal characteristics closely aligned with those observed in human tissues (28). Another limitation of our study is that we performed segmentation only once and with a single operator, which did not allow us to analyze the impact of segmentation variability on the stability of radiomic features across reconstruction methods. However, through repeated sequences and acquisitions, we introduced sufficient variability within the radiomic features to reveal the differences between the two reconstruction techniques. The final limitation of our study is the use of a single imaging device, which restricts the generalizability of our findings to DLIR implementations from other manufacturers.

In conclusion, our study demonstrates that MRI Deep Learning-based reconstruction enhances both the repeatability and reproducibility of radiomic features. This improvement emphasizes the suitability of MRI DLIR as a preferred alternative to conventional reconstruction techniques in MRI radiomic studies, strengthening its potential to advance clinical radiomics applications.

References

1. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017;14(12):749–762. doi: 10.1038/nrclinonc.2017.141.

2. Temperley HC, O'Sullivan NJ, Waters C, et al. Radiomics; Contemporary Applications in the Management of Anal Cancer; A Systematic Review. Am Surg. 2024;90(3):445–454. doi: 10.1177/00031348231216494.

3. Deng K, Chen T, Leng Z, et al. Radiomics as a tool for prognostic prediction in transarterial chemoembolization for hepatocellular carcinoma: a systematic review and meta-analysis. Radiol Med. 2024;129(8):1099–1117. doi: 10.1007/s11547-024-01840-9.

4. Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol. 2021;31(1):1–4. doi: 10.1007/s00330-020-07108-w.

5. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. doi: 10.1148/radiol.2020191145.

6. Teng X, Wang Y, Nicol AJ, et al. Enhancing the Clinical Utility of Radiomics: Addressing the Challenges of Repeatability and Reproducibility in CT and MRI. Diagnostics. 2024;14(16):1835. doi: 10.3390/diagnostics14161835.

7. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep. 2019;9(1):4800. doi: 10.1038/s41598-019-41344-5.

8. Bianchini L, Santinha J, Loução N, et al. A multicenter study on radiomic features from T 2 - weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. Magnetic Resonance in Med. 2021;85(3):1713–1726. doi: 10.1002/mrm.28521.

9. Baeßler B, Weiss K, Pinto dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Investigative Radiology. 2019;54(4):221–228. doi: 10.1097/RLI.0000000000000530.

10. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Physica Medica. 2019;61:44–51. doi: 10.1016/j.ejmp.2019.04.009.

11. Wichtmann BD, Harder FN, Weiss K, Schönberg SO, Attenberger UI, Alkadhi H. Influence of Image Processing on Radiomic Features From Magnetic Resonance Imaging. Investigative Radiology. 2023;58(3).

12. van Stiphout JA, Driessen J, Koetzier LR, et al. The effect of deep learning reconstruction on abdominal CT densitometry and image quality: a systematic review and meta-analysis. Eur Radiol. 2021;32(5):2921–2929. doi: 10.1007/s00330-021-08438-z.

13. Tsuboyama T, Onishi H, Nakamoto A, et al. Impact of Deep Learning Reconstruction Combined With a Sharpening Filter on Single-Shot Fast Spin-Echo T2-Weighted Magnetic Resonance Imaging of the Uterus. Invest Radiol. 2022;57(6):379–386. doi: 10.1097/RLI.0000000000000847.

14. Kanan A, Pereira B, Hordonneau C, et al. Deep learning CT reconstruction improves liver metastases detection. Insights Imaging. 2024;15(1):167. doi: 10.1186/s13244-024-01753-1.

15. Xue G, Liu H, Cai X, et al. Impact of deep learning image reconstruction algorithms on CT radiomic features in patients with liver tumors. Front Oncol. 2023;13:1167745. doi: 10.3389/fonc.2023.1167745.

16. Zhong J, Xia Y, Chen Y, et al. Deep learning image reconstruction algorithm reduces image noise while alters radiomics features in dual-energy CT in comparison with conventional iterative reconstruction algorithms: a phantom study. Eur Radiol. 2023;33(2):812–824. doi: 10.1007/s00330-022-09119-1.

17. Yang B, Chen X, Yuan S, Liu Y, Dai J, Men K. Deep learning improves image quality and radiomics reproducibility for high-speed four-dimensional computed tomography reconstruction. Radiotherapy and Oncology. 2022;170:184–189. doi: 10.1016/j.radonc.2022.02.034.

18. Michallek F, Genske U, Niehues SM, Hamm B, Jahnke P. Deep learning reconstruction improves radiomics feature stability and discriminative power in abdominal CT imaging: a phantom study. Eur Radiol. 2022;32(7):4587–4595. doi: 10.1007/s00330-022-08592-y.

19. Li H, Alves VV, Pednekar A, et al. Impact of Emerging Deep Learning–Based MR Image Reconstruction Algorithms on Abdominal MRI Radiomic Features. J Comput Assist Tomogr. 2024; doi: 10.1097/RCT.0000000000001648.

20. Bernatz S, Zhdanovich Y, Ackermann J, et al. Impact of rescanning and repositioning on radiomic features employing a multi-object phantom in magnetic resonance imaging. Sci Rep. 2021;11(1):14248. doi: 10.1038/s41598-021-93756-x.

21. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30(9):1323–1341. doi: 10.1016/j.mri.2012.05.001.

22. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21):e104–e107. doi: 10.1158/0008-5472.CAN-17-0339.

23. Li Y, Reyhan M, Zhang Y, et al. The impact of phantom design and material-dependence on repeatability and reproducibility of CT-based radiomics features. Med Phys. 2022;49(3):1648–1659. doi: 10.1002/mp.15491.

24. Muenzfeld H, Nowak C, Riedlberger S, et al. Intra-scanner repeatability of quantitative imaging features in a 3D printed semi-anthropomorphic CT phantom. European Journal of Radiology. 2021;141:109818. doi: 10.1016/j.ejrad.2021.109818.

25. Hertel A, Tharmaseelan H, Rotkopf LT, et al. Phantom-based radiomics feature test–retest stability analysis on photon-counting detector CT. Eur Radiol. 2023;33(7):4905–4914. doi: 10.1007/s00330-023-09460-z.

26. Barnhart HX, Haber M, Song J. Overall Concordance Correlation Coefficient for Evaluating Agreement Among Multiple Observers. Biometrics. 2002;58(4):1020–1027. doi: 10.1111/j.0006-341X.2002.01020.x.

27. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–428. doi: 10.1037//0033-2909.86.2.420.

28. Bologna M, Tenconi C, Corino VDA, et al. Repeatability and reproducibility of MRI-radiomic features: A phantom experiment on a 1.5 T scanner. Medical Physics. 2023;50(2):750–762. doi: 10.1002/mp.16054.

29. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016;15(2):155–163. doi: 10.1016/j.jcm.2016.02.012.

30. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; https://www.R-project.org/.

31. Cattell R, Chen S, Huang C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art. 2019;2(1):19. doi: 10.1186/s42492-019-0025-6.

32. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics. 2018;102(4):1143–1158. doi: 10.1016/j.ijrobp.2018.05.053.

33. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30–46. doi: 10.1037/1082-989X.1.1.30.

34. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – A discussion and demonstration of basic features. Chiacchio F, editor. PLoS ONE. 2019;14(7):e0219854. doi: 10.1371/journal.pone.0219854.

35. Rai R, Holloway LC, Brink C, et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med Phys. 2020;47(7):3054–3063. doi: 10.1002/mp.14173.

Supplemental Material 1

Comparison of 4 methods and 2 thresholds in radiomic feature **reproducibility**.

MAIN CRITERIA: ICC(3,1)>0.9

		DL	Non-DL
All sequences	**All Features**	**53.27**(399)	**43.52** (326)
	Shape Features	76.53 (75)	76.53 (75)
	First Order Features	61.11 (77)	53.17 (67)
	Higher Order Features	47.04 (247)	35.04 (184)

ICC(2,1)>0.9

		DL	Non-DL
All sequences	**All Features**	**30.04** (225)	**24.57** (184)
	Shape Features	76.53 (75)	76.53 (75)
	First Order Features	31.75 (40)	27.78 (35)
	Higher Order Features	20.95 (110)	14.1 (74)

OCCC>0.9

		DL	Non-DL
All sequences	**All Features**	**28.84** (216)	**24.17** (181)
	Shape Features	76.53 (75)	76.53 (75)
	First Order Features	30.95 (39)	27.78 (35)
	Higher Order Features	19.43 (102)	13.52 (71)

OCCC>0.9 AND DR>0.9

		DL	Non-DL
All sequences	**All Features**	**28.84** (216)	**24.17** (181)
	Shape Features	76.53 (75)	76.53 (75)
	First Order Features	30.95 (39)	27.78 (35)
	Higher Order Features	19.43 (102)	13.52 (71)

ICC(3,1)>0.75

		DL	Non-DL
All sequences	**All Features**	**79.97** (599)	**78.24** (586)
	Shape Features	88.78 (87)	88.78 (87)
	First Order Features	92.06 (116)	89.68 (113)
	Higher Order Features	75.43 (396)	73.52 (386)

ICC(2,1)>0.75

		DL	Non-DL
All sequences	**All Features**	**65.95** (494)	**55.27** (414)
	Shape Features	85.71 (84)	86.73 (85)
	First Order Features	74.6 (94)	67.46 (85)
	Higher Order Features	60.19 (316)	46.48 (244)

OCCC>0.75

		DL	Non-DL
All sequences	**All Features**	**65.55** (491)	**54.34** (407)
	Shape Features	85.71 (84)	85.71 (84)
	First Order Features	74.6 (94)	65.87 (83)
	Higher Order Features	59.62 (313)	45.71 (240)

Supplemental Material 2

Comparison of 3 methods and 2 thresholds in radiomic feature **repeatability**.

MAIN CRITERIA: OCCC>0.9

		DL	Non-DL
All sequences	**All Features**	**49,27** (369)	**48,87** (366)
	Shape Features	74,49 (73)	75,51 (74)
	First Order Features	53,97 (68)	50,79 (64)
	Higher Order Features	43,43 (228)	43,43 (228)

ICC(1,1)>0.9

		DL	Non-DL
All sequences	**All Features**	**49,53** (371)	**49,27** (369)
	Shape Features	74,49 (73)	75,51 (74)
	First Order Features	53,97 (68)	51,59 (65)
	Higher Order Features	43,81 (230)	43,81 (230)

ICC(2,1)>0.9

		DL	Non-DL
All sequences	**All Features**	**49,53** (371)	**49,27** (369)
	Shape Features	74,49 (73)	75,51 (74)
	First Order Features	53,97 (68)	51,59 (65)
	Higher Order Features	43,81 (230)	43,81 (230)

OCCC>0.75

		DL	Non DL
All sequences	**All Features**	**73,43** (550)	**72,1** (540)
	Shape Features	82,65 (81)	83,67 (82)
	First Order Features	83,33 (105)	78,57 (99)
	Higher Order Features	69,33 (364)	68,38 (359)

ICC(1,1)>0.75

		DL	Non DL
All sequences	**All Features**	**73,43** (550)	**72,36** (542)
	Shape Features	74,49 (73)	75,51 (74)
	First Order Features	53,97 (68)	50,79 (64)
	Higher Order Features	43,43 (228)	43,43 (228)

ICC(2,1)>0.75

		DL	Non DL
All sequences	**All Features**	**73,43** (550)	**72,36** (542)
	Shape Features	82,65 (81)	83,67 (82)
	First Order Features	83,33 (105)	78,57 (99)
	Higher Order Features	69,33 (364)	68,76 (361)

Conclusion

Dans ce travail, nous avons pu apporter une preuve de l'intérêt d'une innovation technologique vendue par un constructeur de scanner, en montrant que cette nouvelle reconstruction améliorait la détection des métastases hépatiques de nos patients. Nous avons ensuite proposé deux modèles de radiomique, répondant à deux questions cliniques – l'évaluation des muscles axiaux dans les néoplasies ORL et le pronostic des patients traités par immunothérapie pour un mélanome métastatique. Nous avons enfin montré que la nouvelle reconstruction d'images d'IRM vendue par un constructeur permettait d'améliorer la stabilité des paramètres de radiomique. Cette dernière étude permet de valider l'utilisation de ce type de reconstruction dans les études de radiomique.

A travers ce travail, nous avons abordé une des façons dont l'apprentissage profond peut améliorer la prise en charge de nos patients. Nous avons abordé aussi l'impact d'une technique d'analyse d'images, la radiomique, en montrant ses apports mais en soulignant aussi ses limites. L'avenir des études qui développent de nouveaux modèles de la radiomique sera nécessairement celui d'études moins nombreuses mais à la méthodologie plus rigoureuse, impérativement multicentriques pour augmenter les tailles d'effectif et assurer une validation externe. L'avenir plus immédiat passe aussi par la validation externe de modèles de radiomique déjà publiés, ce qui une première direction de la suite de nos travaux de recherche. Cet avenir de la radiomique passe enfin par la poursuite d'études de méthodologie de la radiomique, ce qui sera une seconde direction de la suite de nos travaux de recherche.

Il peut se poser la question de la pertinence de maintenir des études de radiomique au moment du développement de l'apprentissage profond. Cette étape d'extraction de paramètres décrivant les propriétés des images qu'est la radiomique peut en effet paraître superflue, puisqu'en théorie, il doit être possible de créer un réseau de neurones qui va être capable de capter ce type d'information. La solution optimale sera peut être plutôt de combiner les deux techniques. Cette approche existe déjà, en cherchant par exemple à comparer des paramètres de radiomique extraits de façon classique à des paramètres extraits de réseaux de neurones profonds (deep features) ou bien en essayant d'utiliser les réseaux de neurones pour pallier aux problèmes de reproductibilité de la radiomique. C'est l'utilisation combinée de l'apprentissage profond et de la radiomique qui sera une troisième direction de la suite de nos travaux de recherche.

Bibliographie

1. Snoeckx A, Franck C, Silva M, Prokop M, Schaefer-Prokop C, Revel M-P. The radiologist's role in lung cancer screening. Transl Lung Cancer Res. 2021;10(5):2356–2367. doi: 10.21037/tlcr-20-924.

2. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. The Lancet Oncology. 2017;18(3):e143–e152.

3. Cheson BD, Ansell S, Schwartz L, et al. Refinement of the Lugano Classification lymphoma response criteria in the era of immunomodulatory therapy. Blood. 2016;128(21):2489–2496. doi: 10.1182/blood-2016-05-718528.

4. Chartrand G, Cheng PM, Vorontsov E, et al. Deep Learning: A Primer for Radiologists. RadioGraphics. 2017;37(7):2113–2131. doi: 10.1148/rg.2017170077.

5. Ma J, Zhang Y, Gu S, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. Lancet Digit Health. 2024;6(11):e815–e826. doi: 10.1016/S2589-7500(24)00154-7.

6. Wang C, Huang Y, Liu C, et al. Diagnosis of Clinically Significant Portal Hypertension Using CT- and MRI-based Vascular Model. Radiology. 2023;307(2):e221648. doi: 10.1148/radiol.221648.

7. Wei J, Song X, Wei X, et al. Knowledge-Augmented Deep Learning for Segmenting and Detecting Cerebral Aneurysms With CT Angiography: A Multicenter Study. Radiology. 2024;312(2):e233197. doi: 10.1148/radiol.233197.

8. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392(10162):2388–2396. doi: 10.1016/S0140-6736(18)31645-3.

9. Cai JC, Nakai H, Kuanar S, et al. Fully Automated Deep Learning Model to Detect Clinically Significant Prostate Cancer at MRI. Radiology. 2024;312(2):e232635. doi: 10.1148/radiol.232635.

10. Hu B, Shi Z, Lu L, et al. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. Lancet Digit Health. 2024;6(4):e261–e271. doi: 10.1016/S2589-7500(23)00268-6.

11. Jiang X, Zhao H, Saldanha OL, et al. An MRI Deep Learning Model Predicts Outcome in Rectal Cancer. Radiology. 2023;307(5):e222223. doi: 10.1148/radiol.222223.

12. Kim H, Jin KN, Yoo S-J, et al. Deep Learning for Estimating Lung Capacity on Chest Radiographs Predicts Survival in Idiopathic Pulmonary Fibrosis. Radiology. 2023;306(3):e220292. doi: 10.1148/radiol.220292.

13. Lim DSW, Makmur A, Zhu L, et al. Improved Productivity Using Deep Learning-assisted Reporting for Lumbar Spine MRI. Radiology. 2022;305(1):160–166. doi: 10.1148/radiol.220076.

14. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. Moy L, editor. Radiology. 2024;311(1):e232714. doi: 10.1148/radiol.232714.

15. Lin DJ, Johnson PM, Knoll F, Lui YW. Artificial Intelligence for MR Image Reconstruction: An Overview for Clinicians. J Magn Reson Imaging. 2021;53(4):1015–1028. doi: 10.1002/jmri.27078.

16. Kiryu S, Akai H, Yasaka K, et al. Clinical Impact of Deep Learning Reconstruction in MRI. RadioGraphics. 2023;43(6):e220133. doi: 10.1148/rg.220133.

17. Lebel RM. Performance characterization of a novel deep learning-based MR image reconstruction pipeline. arXiv; 2020. http://arxiv.org/abs/2008.06559. Accessed October 28, 2024.

18. Tajima T, Akai H, Sugawara H, et al. Breath-hold 3D magnetic resonance cholangiopancreatography at 1.5 T using a deep learning-based noise-reduction approach: Comparison with the conventional respiratory-triggered technique. European Journal of Radiology. 2021;144:109994. doi: 10.1016/j.ejrad.2021.109994.

19. Tsuboyama T, Onishi H, Nakamoto A, et al. Impact of Deep Learning Reconstruction Combined With a Sharpening Filter on Single-Shot Fast Spin-Echo T2-Weighted Magnetic Resonance Imaging of the Uterus. Invest Radiol. 2022;57(6):379–386. doi: 10.1097/RLI.0000000000000847.

20. Rastogi A, Brugnara G, Foltyn-Dumitru M, et al. Deep-learning-based reconstruction of undersampled MRI to reduce scan times: a multicentre, retrospective, cohort study. The Lancet Oncology. 2024;25(3):400–410. doi: 10.1016/S1470-2045(23)00641-1.

21. Geyer LL, Schoepf UJ, Meinel FG, et al. State of the Art: Iterative CT Reconstruction Techniques. Radiology. Radiological Society of North America; 2015;276(2):339–357. doi: 10.1148/radiol.2015132766.

22. Nagayama Y, Sakabe D, Goto M, et al. Deep Learning-based Reconstruction for Lower-Dose Pediatric CT: Technical Principles, Image Characteristics, and Clinical Implementations. Radiographics. 2021;41(7):1936–1953. doi: 10.1148/rg.2021210105.

23. Akagi M, Nakamura Y, Higaki T, et al. Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. Eur Radiol. 2019;29(11):6163–6171. doi: 10.1007/s00330-019-06170-3.

24. Nakamura Y, Higaki T, Tatsugami F, et al. Deep Learning–based CT Image Reconstruction: Initial Evaluation Targeting Hypovascular Hepatic Metastases. Radiology: Artificial Intelligence. 2019;1(6):e180011. doi: 10.1148/ryai.2019180011.

25. Jensen CT, Gupta S, Saleh MM, et al. Reduced-Dose Deep Learning Reconstruction for Abdominal CT of Liver Metastases. Radiology. 2022;303(1):90–98. doi: 10.1148/radiol.211838.

26. Park HJ, Choi S-Y, Lee JE, et al. Deep learning image reconstruction algorithm for abdominal multidetector CT at different tube voltages: assessment of image quality and radiation dose in a phantom study. Eur Radiol. 2022;32(6):3974–3984. doi: 10.1007/s00330-021-08459-8.

27. Brady SL, Trout AT, Somasundaram E, Anton CG, Li Y, Dillman JR. Improving Image Quality and Reducing Radiation Dose for Pediatric CT by Using Deep Learning Reconstruction. Radiology. 2020;202317. doi: 10.1148/radiol.2020202317.

28. Higaki T, Nakamura Y, Zhou J, et al. Deep Learning Reconstruction at CT: Phantom Study of the Image Characteristics. Acad Radiol. 2020;27(1):82–87. doi: 10.1016/j.acra.2019.09.008.

29. Racine D, Becce F, Viry A, et al. Task-based characterization of a deep learning image reconstruction and comparison with filtered back-projection and a partial model-based iterative reconstruction in abdominal CT: A phantom study. Phys Med. 2020;76:28–37. doi: 10.1016/j.ejmp.2020.06.004.

30. Hsieh J, Liu E, Nett B, Tang J, Thibault J-B, Sahney S. A new era of image reconstruction: TrueFidelity™. . https://www.gehealthcare.com/-/jssmedia/040dd213fa89463287155151fdb01922.pdf. Accessed November 20, 2023.

31. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–446. doi: 10.1016/j.ejca.2011.11.036.

32. Guerrisi A, Russillo M, Loi E, et al. Exploring CT Texture Parameters as Predictive and Response Imaging Biomarkers of Survival in Patients With Metastatic Melanoma Treated With PD-1 Inhibitor Nivolumab: A Pilot Study Using a Delta-Radiomics Approach. Front Oncol. 2021;11:704607. doi: 10.3389/fonc.2021.704607.

33. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563–577. doi: 10.1148/radiol.2015151169.

34. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. RadioGraphics. 2017;37(5):1483–1503. doi: 10.1148/rg.2017170056.

35. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. J Nucl Med. 2020;61(4):488–495. doi: 10.2967/jnumed.118.222893.

36. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. doi: 10.1148/radiol.2020191145.

37. Yin P, Mao N, Zhao C, et al. Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. European Radiology. 2019;29(4):1841–1847. doi: 10.1007/s00330-018-5730-6.

38. Mulé S, Thiefin G, Costentin C, et al. Advanced Hepatocellular Carcinoma: Pretreatment Contrast-enhanced CT Texture Parameters as Predictive Biomarkers of Survival in Patients Treated with Sorafenib. Radiology. 2018;288(2):445–455. doi: 10.1148/radiol.2018171320.

39. Miles KA, Ganeshan B, Griffiths MR, Young RCD, Chatwin CR. Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival. Radiology. 2009;250(2):444–452. doi: 10.1148/radiol.2502071879.

40. Jia L-L, Zhao J-X, Zhao L-P, Tian J-H, Huang G. Current status and quality of radiomic studies for predicting KRAS mutations in colorectal cancer patients: A systematic review and meta-analysis. Eur J Radiol. 2023;158:110640. doi: 10.1016/j.ejrad.2022.110640.

41. Chen J, Chen A, Yang S, Liu J, Xie C, Jiang H. Accuracy of machine learning in preoperative identification of genetic mutation status in lung cancer: A systematic review and meta-analysis. Radiother Oncol. 2024;196:110325. doi: 10.1016/j.radonc.2024.110325.

42. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology. 2019;291(1):53–59. doi: 10.1148/radiol.2019182023.

43. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407–415. doi: 10.1148/radiol.2018172361.

44. Orlhac F. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol. 2021;9.

45. Image processing — IBSI 0.0.1dev documentation. . https://ibsi.readthedocs.io/en/latest/02_Image_processing.html#recommendations. Accessed October 29, 2024.

46. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19(2):143–150. doi: 10.1109/42.836373.

47. Dewey BE, Zhao C, Reinhold JC, et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magnetic Resonance Imaging. 2019;64:160–170. doi: 10.1016/j.mri.2019.05.041.

48. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging. 2010;29(6):1310–1320. doi: 10.1109/TMI.2010.2046908.

49. Dai X, Lei Y, Liu Y, et al. Intensity non-uniformity correction in MR imaging using residual cycle generative adversarial network. Phys Med Biol. 2020;65(21):215025. doi: 10.1088/1361-6560/abb31f.

50. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21):e104–e107. doi: 10.1158/0008-5472.CAN-17-0339.

51. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A. MaZda--a software package for image texture analysis. Comput Methods Programs Biomed. 2009;94(1):66–76. doi: 10.1016/j.cmpb.2008.08.005.

52. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res. 2018;78(16):4786–4789. doi: 10.1158/0008-5472.CAN-18-0125.

53. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42(3):1341–1353. doi: 10.1118/1.4908210.

54. Zheng Y, Li J, Liu S, et al. MRI-Based radiomics nomogram for differentiation of benign and malignant lesions of the parotid gland. Eur Radiol. 2020; doi: 10.1007/s00330-020-07483-4.

55. Da-ano R, Masson I, Lucia F, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep. 2020;10(1):10248. doi: 10.1038/s41598-020-66110-w.

56. Da-ano R, Lucia F, Masson I, et al. A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. Bianconi F, editor. PLoS ONE. 2021;16(7):e0253653. doi: 10.1371/journal.pone.0253653.

57. Carré A, Battistella E, Niyoteka S, Sun R, Deutsch E, Robert C. AutoComBat: a generic method for harmonizing MRI-based radiomic features. Sci Rep. 2022;12(1):12762. doi: 10.1038/s41598-022-16609-1.

58. Balagurunathan Y, Kumar V, Gu Y, et al. Test–Retest Reproducibility Analysis of Lung CT Image Features. J Digit Imaging. 2014;27(6):805–823. doi: 10.1007/s10278-014-9716-x.

59. Bernadach M, Lapeyre M, Dillies A-F, et al. Predictive factors of toxicity of TPF induction chemotherapy for locally advanced head and neck cancers. BMC Cancer. 2021;21(1):360. doi: 10.1186/s12885-021-08128-5.

60. Lere-Chevaleyre A, Bernadach M, Lambert C, et al. Toxicity of induction chemotherapy in head and neck cancer: The central role of skeletal muscle mass. Head Neck. 2021; doi: 10.1002/hed.26954.

61. Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol. 2021;31(1):1–4. doi: 10.1007/s00330-020-07108-w.

62. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019;46(13):2638–2655. doi: 10.1007/s00259-019-04391-8.

63. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30(1):523–536. doi: 10.1007/s00330-019-06360-z.

64. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. Radiology: Artificial Intelligence. 2023;e220232. doi: 10.1148/ryai.220232.

65. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv; 2019. doi: 10.48550/arXiv.1612.07003.

66. Santinha J, Pinto Dos Santos D, Laqua F, et al. ESR Essentials: radiomics—practice recommendations by the European Society of Medical Imaging Informatics. Eur Radiol. 2024; doi: 10.1007/s00330-024-11093-9.

67. Xue G, Liu H, Cai X, et al. Impact of deep learning image reconstruction algorithms on CT radiomic features in patients with liver tumors. Front Oncol. 2023;13:1167745. doi: 10.3389/fonc.2023.1167745.

68. Zhong J, Xia Y, Chen Y, et al. Deep learning image reconstruction algorithm reduces image noise while alters radiomics features in dual-energy CT in comparison with conventional iterative reconstruction algorithms: a phantom study. Eur Radiol. 2023;33(2):812–824. doi: 10.1007/s00330-022-09119-1.

69. Michallek F, Genske U, Niehues SM, Hamm B, Jahnke P. Deep learning reconstruction improves radiomics feature stability and discriminative power in abdominal CT imaging: a phantom study. Eur Radiol. 2022;32(7):4587–4595. doi: 10.1007/s00330-022-08592-y.

70. Li H, Alves VV, Pednekar A, et al. Impact of Emerging Deep Learning–Based MR Image Reconstruction Algorithms on Abdominal MRI Radiomic Features. J Comput Assist Tomogr. 2024; doi: 10.1097/RCT.0000000000001648.