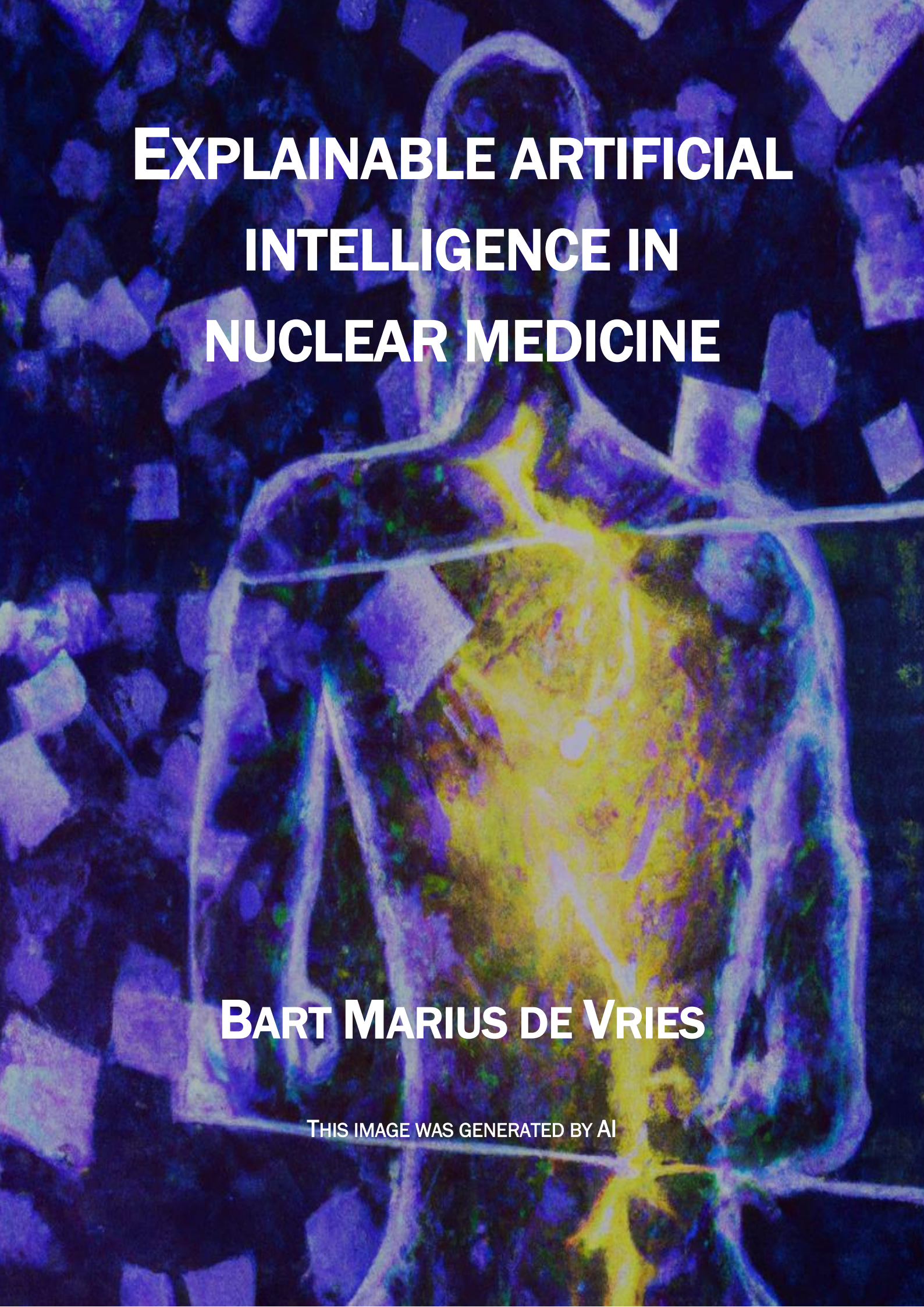
Explainable artificial intelligence in nuclear medicine

Bart Marius de Vries



DEVELOPMENT AND EVALUATION OF A MEDICAL-BASED EXPLAINABLE ARTIFICIAL INTELLIGENCE APPROACH TO PREDICT PROGRESSION FREE SURVIVAL IN PATIENTS WITH METASTATIC COLORECTAL CANCER USING PRE-TREATMENT ^{18}F-FDG PET

B.M. (Bart) de Vries

Student number : 5399009

11 April 2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. Radiology and Nuclear Medicine, Amsterdam UMC, location VUmc
Dept. Medical Oncology, Amsterdam UMC, location VUmc
November 2022 – April 2023

Thesis committee members:

Prof. dr. John van den Dobbelsteen	TU Delft	Chair
Prof. dr. Ronald Boellaard	Amsterdam UMC	Technical supervisor
Dr. Willemien Menke	Amsterdam UMC	Medical supervisor
Dr. Floris van Velden	LUMC	Technical supervisor (LDE)

Development and evaluation of a medical-based explainable artificial intelligence approach to predict progression free survival in patients with metastatic colorectal cancer using pre-treatment ^{18}F-FDG PET

Bart M. de Vries[1], Sophie Gerritse[2], Gerben J.C. Zwezerijnen[1], Sandeep S.V. Golla[1], Anne I.J. Arens[3], Floris H.P. van Velden[4], C. Willemien Menke-van der Houven van Oordt[5*], and Ronald Boellaard[1*]

[1]Cancer Center Amsterdam, Department of Radiology and Nuclear Medicine, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, The Netherlands

[2]Department of Oncology, Erasmus Medical Center Cancer Institute, Dr. Molewaterplein 40, Rotterdam, The Netherlands

[3]Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

[4]Department of Radiology, Leiden University Medical Center, Albinusdreef 2, Leiden, The Netherlands

[5]Cancer Center Amsterdam, Department of Oncology, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, The Netherlands

*These authors have contributed equally to this work.

First and corresponding author:

Bart M. de Vries - student

ORCID: 0000-0002-6421-8303

Department of Radiology & Nuclear Medicine

Amsterdam Universities Medical Center, Location VUmc

De Boelelaan 1117

1081 HV Amsterdam, The Netherlands

Email: b.devries1@amsterdamumc.nl

Abstract word count: 362

Manuscript word count: 6074

Running title: XAI for prognostication of mCRC using 18F-FDG PET

Manuscript category: Original Article

Abstract

Purpose: ^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) is used in the diagnostic process and management of patients with metastatic colorectal cancer (mCRC). Also, ^{18}F-FDG PET radiomic features have been found to hold prognostic value for clinical outcome in mCRC. However, no prognostic model has yet been developed to predict clinical outcome in mCRC using ^{18}F-FDG PET images. Computer-aided pattern recognition can be helpful in this process but needs to be validated. The aim of this work was to develop and evaluate a medical-based explainable artificial intelligence (XAI) framework for discriminating between dichotomous progression free survival (PFS) in patients with mCRC undergoing anti-epidermal growth factor receptor (anti-EGFR) monoclonal antibody (mAb) treatment using pre-treatment ^{18}F-FDG PET images.

Methods: We conducted an analysis of ^{18}F-FDG PET images, expressed in standardized uptake values (SUV), obtained from 80 patients with mCRC who were eligible for third-line treatment with an anti-EGFR mAb as part of the IMPACT study. A coronal 2.5D Convolutional Neural Network (CNN) was built to capture features of the ^{18}F-FDG PET images specific for the two patient groups and a medical-based XAI framework was developed to extract the ^{18}F-FDG PET features used by the CNN. The images were randomly divided into a training and a validation set (10-fold cross-validation). Performance of the CNN was evaluated based on the average area under the curve (AUC), accuracy, sensitivity and specificity from the cross-validation. A statistical analysis was performed to assess the predictive value of the ^{18}F-FDG PET features extracted by the XAI framework.

Results: The coronal 2.5D-CNN was able to discriminate between dichotomous PFS (median PFS: 152 days) in patients with mCRC undergoing anti-EGFR mAb treatment using pre-treatment ^{18}F-FDG PET images, with an average AUC of 0.95 ± 0.11 (SD), accuracy of 94% ± 12, sensitivity of 91% ± 21 and specificity of 94% ± 21 %. The XAI framework showed that especially low ^{18}F-FDG PET uptake volume features hold significant differences between the two patient groups.

Conclusion: The coronal 2.5D-CNN showed good performance to predict dichotomous PFS from pre-treatment ^{18}F-FDG PET images in patients with mCRC undergoing anti-EGFR mAb treatment. Low ^{18}F-FDG PET uptake volume features seem to have potential as IB in this patient cohort, but further validation is required.

Keywords: metastatic colorectal cancer (mCRC), anti-epidermal growth factor receptor (anti-EGFR), ^{18}F-FDG, convolutional neural network (CNN), explainable artificial intelligence (XAI)

Introduction

^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) is widely used as diagnostic tool in the management of patients with metastatic colorectal cancer (mCRC) [1, 2]. Also, ^{18}F-FDG PET showed potential as a prognostic tool to predict clinical outcome in patients with mCRC who underwent third-line palliative treatment with the anti-epidermal growth factor receptor (anti-EGFR) monoclonal antibody (mAb) cetuximab [3, 4]. The introduction of anti-EGFR mAb cetuximab and panitumumab improved clinical outcome in patients with a left-sided primary tumour, KRAS, NRAS and BRAF wild-type mCRC [5-7]. It is of interest to develop a prognostic model in patients with mCRC eligible for palliative treatment, which could aid in more patient specific treatment decision using pre-treatment ^{18}F-FDG PET.

Computer-aided pattern recognition algorithms have been developed to evaluate and identify disease specific PET patterns based on ^{18}F-FDG PET images in patients with mCRC [4, 8]. These studies used radiomic features extracted from tumour segmentations to predict clinical outcome, i.e. find tumour specific radiographic image characteristics (e.g. tumour intensity, shape, texture) for clinical outcome. Low metabolic tumour volume, heterogeneity and high sphericity on the ^{18}F-FDG PET showed an association with improved clinical outcome [3]. Despite yielding reasonable results, feature extraction depending on volume of interest (VOI) parcellation can be time-consuming and is observer dependent in case of manual delineation. Also, non-tumour patterns (e.g. ^{18}F-FDG uptake in liver tissue, bone-marrow and muscle) might be of prognostic value to predict clinical outcome [9]. Deep learning based algorithms, such as Convolutional Neural Networks (CNN), use the whole ^{18}F-FDG image to extract features and therefore may provide superior performance. Therefore, it is of interest whether a CNN could effectively be applied to predict progression free survival (PFS) in patients with mCRC undergoing anti-EGFR mAb treatment using pre-treatment ^{18}F-FDG PET.

However, in recent years, there has been an increasing concern about the black box nature of these CNNs [10]. Since CNNs are trained in an unsupervised way, many latent (i.e. not observable by humans) features can be extracted from images. Because of this complex nature, it is difficult to understand how CNNs extract features and come to a decision. Therefore, there is an increasing need for explainable artificial intelligence (XAI) methods to provide a post-hoc explanation of these trained CNNs. SHapley Additive exPlanations (SHAP) is a XAI method that shows great interest because of better consistency with human intuition than other XAI techniques [11, 12]. However, similar as most other post-hoc XAI methods, SHAP is developed/optimized for non-medical usage, prone to spatial variation in the input image and (still) provides moderately interpretable attribution images in medical imaging [13]. Therefore, there is high need for a XAI method that is able to accurately and robustly describe CNN feature extraction, but even so importantly provides medical-based explanation. Therefore, we propose a novel medical-based XAI method, which is a combination of current state-of-the-art XAI methods and methods used in medical imaging. The main aim is to discover new imaging biomarkers (IB), which have the potential to be used in routine management of patients with mCRC. In other words, this XAI method may provide new insight in disease characteristics, which alternatively can be used as an indicator of responses to anti-EGFR mAb cetuximab and panitumumab treatment. However, before this XAI method can be applied as an useful and trustworthy tool for either testing research hypotheses, or clinical decision-making, it must cross 'translational gaps', through performing and reporting technical and clinical validation.

In this proof of concept study, we retrospectively evaluate the performance of a CNN to predict dichotomous PFS in patients with mCRC undergoing anti-EGFR mAb cetuximab and panitumumab treatment using pre-treatment ^{18}F-FDG PET images. Also, we will develop a novel medical-based XAI method to develop potential IB as an indicator of responses to mAb cetuximab and panitumumab treatment in patients with mCRC.

Methods

Population

A total of 80 patients with mCRC from the IMPACT-CRC study [14] were included in this study and used for training the CNN (five subjects were excluded because of no available whole-body PET scan or no available injected dose activity). For the XAI analysis, 68 patients with mCRC were included (12 subjects were excluded because of no available/complete/misaligned low-dose computed tomography (ldCT) scan). The IMPACT-CRC is

a phase I-II multicentre image-guided dose escalation study (NCT02117466). Patients were eligible for inclusion in case of confirmed KRAS and NRAS wild-type adenocarcinoma of the colon or rectum with metastatic disease. During the clinical trial, evidence showed that patients harbouring a BRAF p.V600E mutation and right-sided mCRC might not respond to anti-EGFR mAb. Therefore, after December 2016, patients with this mutation and/or right-sided mCRC were no longer included in this trial [15]. Patients were 18 years or older, ECOG performance status ≤ 2, and an adequate renal and liver function. The study was performed at the Amsterdam University Medical Center location VUmc, Radboud University Medical Center, Erasmus University Medical Center, University Medical Center Groningen, Jeroen Bosch Medical Center, Antoni van Leeuwenhoek Medical Center, and Rijnstate Medical Center, the Netherlands. The central Medical Research Ethics Committee of the Amsterdam University Medical Center location VUmc approved the study. All patients gave written informed consent prior to any study procedure.

Data acquisition

Within two weeks before anti-EGFR mAb treatment, 60 min after tracer injection (3-4 MBq/kg), a ldCT scan was acquired for attenuation and scatter corrections, followed by a 20-min static ^{18}F-FDG PET scan according to the European Association of Nuclear Medicine guidelines using EARL-accredited PET scanners [16].

The ^{18}F-FDG PET images were converted to standardized uptake value (SUV) images, and rebinned to a matrix dimension of 160x160x500 and a voxel size of 4.0 mm in all three directions. The ldCT images were also rebinned to a matrix dimension of 160x160x500 and a voxel size of 4.0 mm in all three directions.

Clinical outcome

In this study, two patient groups were evaluated using dichotomous PFS (short-term and long-term PFS) based on the median PFS. The PFS was defined as the period starting from the date of the first treatment with an anti-EFGR mAb to the date of disease progression. Disease progression is defined as ≥ 20% increase of the sum of diameters of maximally 5 lesions [≤2 per organ, lesion diameter ≥ 10 mm (long axis) or ≥ 15 mm (short axis) for lymph nodes] on CT or MRI according to Reponse Evaluation Criteria in Solid Tumors (RECIST 1.1) [17].

Model architecture and hyper parameter tuning

CNNs are able to learn latent features from the ^{18}F-FDG PET images. Therefore, a CNN is proposed to find prognostic features from the ^{18}F-FDG PET images to support clinical prognostication in patients with mCRC. In

this study, because of the small dataset, instead of 3D-CNNs, coronal 2.5D-CNNs (Figure 1) to extract features from ^{18}F-FDG PET images are proposed [18]. Similar to the 3D-CNN, the 2.5D-CNN requires a input volume (3D), but instead of generating 3D feature images, it generates 2D features images. This could be described in terms of creating a maximum importance projection, which consists of projecting the voxel/feature with the highest importance (depicted by CNN) throughout every XZ (coronal) coordinate.

The CNNs consisted of convolution blocks, consisting each of one convolution layers, one Rectified Linear Unit (ReLU) activation layers, one batch normalization layers (first and third block), one max-pooling layer and one dropout. Hyper parameters (Table 1) were tuned to optimize model convergence (based on training accuracy) each fold. The classification layer consisted of a global average pooling (GAP) layer and a sigmoid dense layer for binary classification.

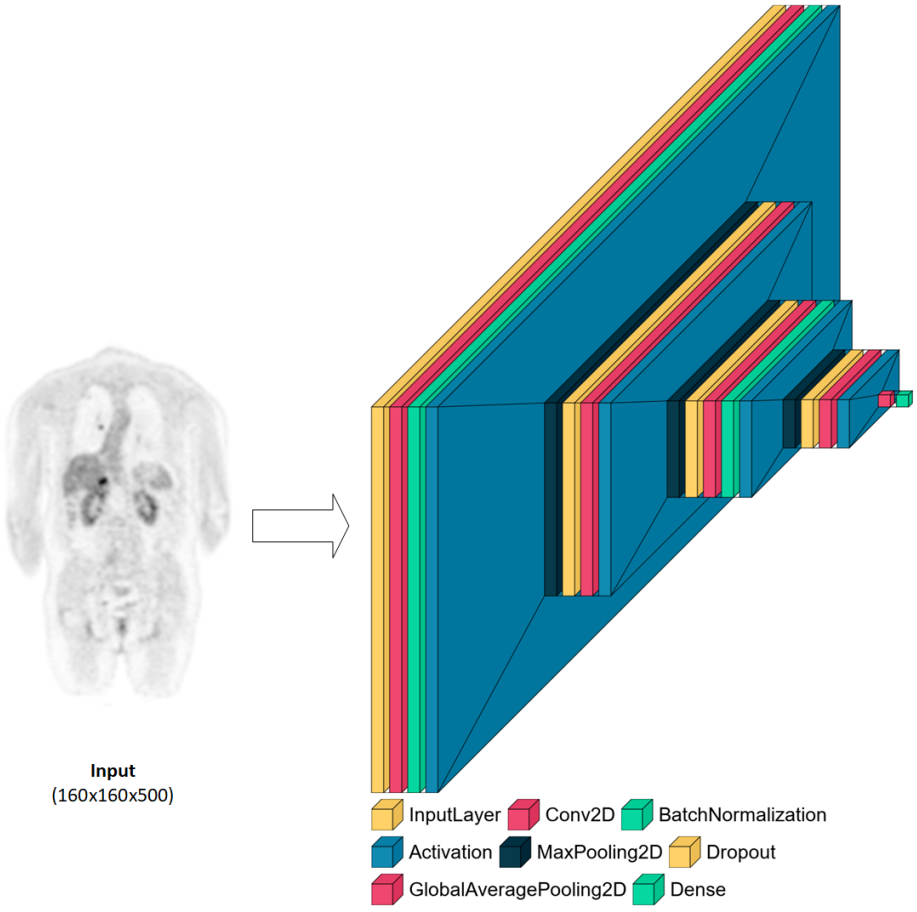


Figure 1: The architecture of a coronal 2.5D-CNN model used in this study.

Table 1: (Hyper) parameters used to optimize model convergence each fold.	
(Hyper) parameter	Value(s)
Convolution block(s)	1/2/3/4
Number of units/filters	16/32
Kernel size(s)	3x3/6x6/9x9
Pool size(s)	2x2/4x4/6x6
Batch size(s)	4/8/16
Learning rate	0.0001
Dropout	0.2

The proposed CNNs were implemented in the Keras library in Python (version 3.6), using TensorFlow as backend. For weights optimization, an Adam optimizer was used with a fixed low learning rate of 1×10^{-5}, and binary cross entropy as loss function.

Model performance

A stratified 10-fold cross-validation was used to evaluate the performance of each CNN model its average validation area under the curve (AUC), accuracy, sensitivity and specificity. Also, an receiver operating characteristic (ROC) curve will be provided. No external test dataset was used due to the small size of the dataset.

Explainable Artificial Intelligence (XAI)

Current application of XAI in medical imaging shows the inability to provide accurate and/or clear explanation of features used by the CNN for outcome prediction [13]. These XAI methods are not developed/optimized for medical imaging and therefore are not (yet) useful to provide reliable IB which can be used for management of (oncology) patients. In this study we propose a novel medical-based XAI method, which combines current knowledge of XAI methods with knowledge used in medical imaging. We therefore utilize strengths of both fields to provide a XAI method, which is both technically as clinically easy to understand, accurate and robust.

Our XAI method is based on SHAP, a state-of-the-art post-hoc perturbation XAI algorithm, which showed consistency with human explanations [11, 19]. SHAP decomposes the output prediction of the CNN on the input image by propagating the contribution of all neurons in the CNN to every feature present in the input image. It

compares the activation of each neuron to an estimated reference activation (background/neutral images) and assigns attribution scores to each voxel according to its difference. However, SHAP is prone to spatial variation in the input images because of this intra-voxel comparison and obtaining reliable neutral reference images is difficult to achieve, which makes technical and clinical validation difficult.

A neutral image/value means that it should not contain any prognostic value for one of the patient groups. Therefore, instead of a (non-neutral) image-based reference activation, we use a SUV value of 0 (SUV_0) as a neutral value to assess the negative/positive attribution of the ^{18}F-FDG PET image features utilized by the CNN. This way the XAI method is not affected by spatial variation between the images, and provides a value that should not contain any prognostic value. Attribution scores are assigned to each voxel according to the difference between the benchmark/original CNN probability and the new probability after replacing/perturbing the ^{18}F-FDG PET voxels with SUV_0.

Data perturbation is performed using a three XAI-resolution medical-based XAI framework (Figure 2). First, 104 structures/organs are segmented from the ldCT using nnU-Net, an open-source (ld-)CT segmentation software package in Slicer3D [20]. However, nnU-Net does not provide segmentation of the mesentery, which is a common metastatic site in mCRC, and also major muscles are not included; muscle mass have shown prognostic value for adverse outcomes in cancer patients [21]. Therefore, to ensure coverage of the mesentery and muscles, an additional segment is created (called soft-tissue) from the remaining tissue by subtracting the 104 segments from the ^{18}F-FDG PET images.

These 105 segments are used as initial masks for the ^{18}F-FDG PET images (XAI-resolution 1). Subsequently, $SUV_{[n-n+1]}$ masks (start: SUV_0, end: SUV_{max}, interval: SUV_1) are created per organ, to intra-subject assess the contribution of $SUV_{[n-n+1]}$ per organ (XAI-resolution 2). Also, these two XAI-resolutions function as a preselection for the last (highest) XAI-resolution. The last XAI-resolution consists of systematic perturbation of the positive contributing ^{18}F-FDG PET voxels (obtained from the previous two XAI-resolutions) per organ using a 1x1x1-4x4x4 (SUV_0) voxel matrix to obtain more fine-grained attribution images (XAI-resolution 3). The idea behind this three XAI-resolution approach, is that direct systematic perturbation (without preselection) using a similar kernel would result in exceptionally long computation time. Also, this approach provides a clear way to perturbate the ^{18}F-FDG PET images.

Similar to SHAP, we propose a XAI method that incorporates both local as (semi-)global attribution features. First, we assess the ^{18}F-FDG PET images per organ to achieve local explanation of prognostic features. However, cancer cells can spread to other parts in the body, in mCRC most often to the liver, but also to the lungs and mesentery and therefore requires a XAI method that also incorporates interaction between different parts of the body. This global explanation requires coalitions (i.e. combinations of features) to obtain the interrelationships of ^{18}F-FDG PET image features utilized by the CNN. For this only the positive contributing ^{18}F-FDG PET voxels (obtained from the local XAI method) are used and again are assessed per organ. However, assessing all possible coalitions is not feasible because of the high computation time and therefore we estimate the global interaction by only using single organ, two organs, and all but one organ coalitions. This way you can assess the target organ attribution on its own, in combination with one other organ, and in absence of the target organ.

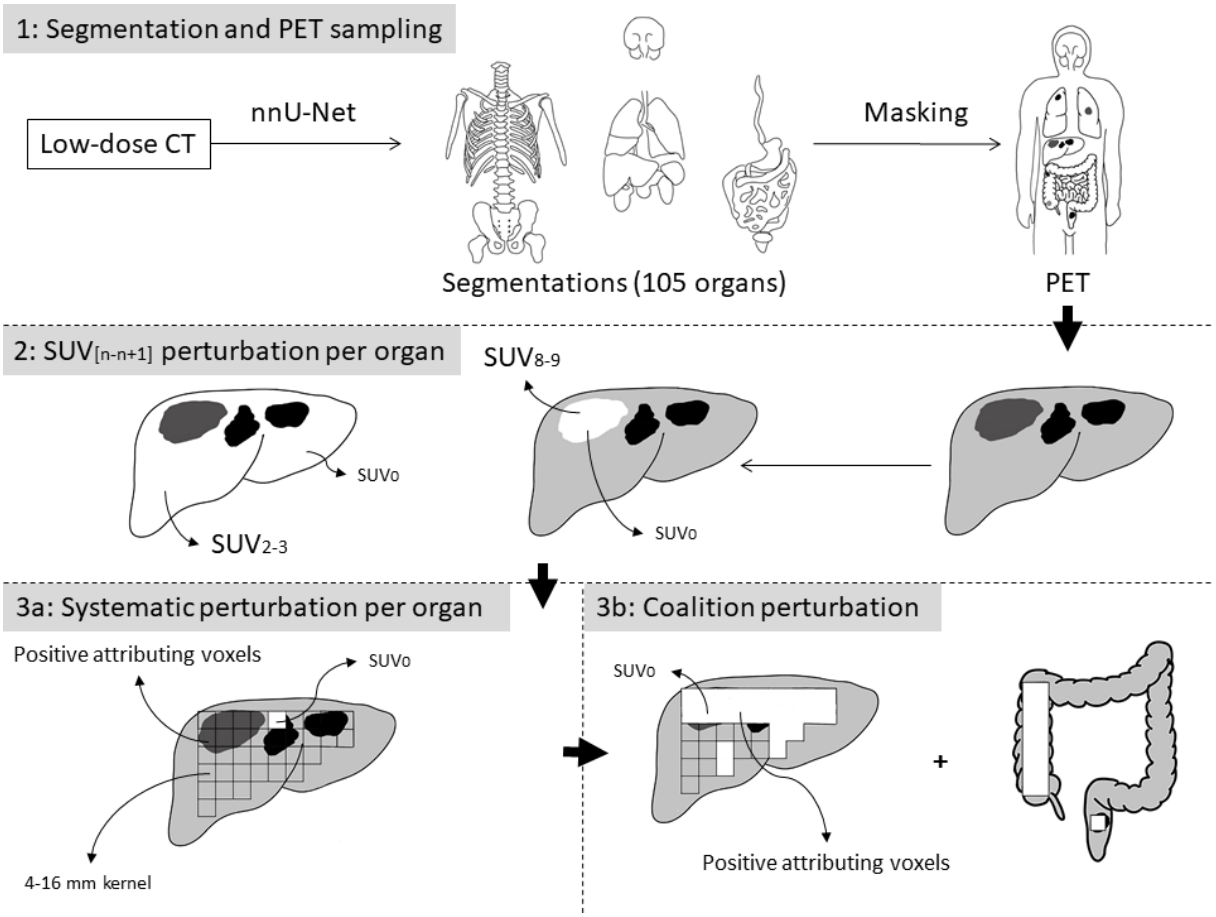


Figure 2: The proposed three XAI-resolution medical-based XAI framework. In XAI-resolution 1 segmentations mask are derived from the ldCT using nnU-Net to mask/sample each organ from the ^{18}F-FDG PET image. Subsequently in XAI-resolution 2, per organ the $SUV_{[n\ -n+1]}$ is perturbated using SUV_0. This functions as a preselection for XAI-resolution 3, where systematic perturbation using a SUV_0 voxel matrix is performed (3a). To

obtain (semi-)global attribution features, coalition perturbation (3b) using the from XAI-resolution 3a obtained positive attributing ^{18}F-FDG PET voxels is performed.

The area over the perturbation curve (AOPC) was used to assess the ability of the XAI method to obtain prognostic features from the ^{18}F-FDG PET images used by the CNN for each patient group [22]. The idea behind this is that important features from the attribution image should correspond with important features from the ^{18}F-FDG PET image. So the more the benchmark CNN probability decreases by perturbation, the better an attribution method is capable to identify relevant input features resulting in a high AOPC.

Quantitative XAI

In addition to the attribution images, first-order features of the attributing ^{18}F-FDG PET voxels (based on the XAI framework) are extracted: SUV_{max}, SUV_{peak}, SUV_{mean}, SUV_{median} and area under the cumulative SUV-volume histogram (AUC-CSH: a quantitative index of tumour uptake heterogeneity) for each organ [23]. These first-order features are acquired using the attribution image as sampling mask/VOI for each organ. In a similar way, the volume of attributing ^{18}F-FDG PET voxels is assessed using $SUV_{[n-n+1]}$ (start: SUV_0, end: SUV_{max}, interval: SUV_1), $SUV_{>4}$, $SUV_{>6}$ and $SUV_{>8}$ (last three approximate tumour uptake) for each organ. The $SUV_{>4}$, $SUV_{>6}$ and $SUV_{>8}$ are used, since comparison using $SUV_{[n-n+1]}$ may be negatively impacted by the high inter-subject ^{18}F-FDG PET uptake heterogeneity, especially in tumours. Subsequently, a student T-test is used to obtain significant ($p < 0.05$) different XAI derived PET features between the two patient groups. For visualisation purposes, only the colon and rectum (primary tumour site), and the most common metastatic sites in mCRC (liver, lungs and soft-tissue, i.e. mesentery and muscles) are presented in the result section [24].

In addition, first-order (SUV_{max}, SUV_{peak}, SUV_{mean}, SUV_{median} and AUC-CSH) and volume features are also extracted from corresponding non-XAI derived ^{18}F-FDG PET voxels for each organ for comparison. This comparison demonstrates the possible added value of XAI compared to non-XAI derived PET features for patient group discrimination, with lower p-values indicating better capability to discriminate between the patient groups. It also functions as quality control for the XAI derived PET features; similar patterns between the XAI and non-XAI derived PET features strengthens the reliability of the XAI-method. Boxplots are created for the significant XAI and corresponding non-XAI derived PET features.

Qualitative XAI

From both the attribution (XAI) images as the ^{18}F-FDG PET images, average images of the two patient groups were created for more pragmatic visual comparison between the patient groups. First, all the ^{18}F-FDG PET images are aligned to a fixed (average/healthy ^{18}F-FDG PET image: average patient size, no tumour bulk, etc.) image using a rigid and a non-rigid registration algorithm based on the co-registration framework Dipy [25]. The fixed image is smoothed with a 4 mm full-width-half-maximum Gaussian kernel to suppress noise. The non-rigid registration is used to transform extreme differences between the moving and fixed scans, e.g. arms up and down, and patient size. Subsequently, per voxel the average of the patient group is taken. Although the co-registration has major effect on quantification reliability, the average attribution images provide a more pragmatic visualisation than the individual images.

In addition, ten (5/5) random ^{18}F-FDG PET were visually assessed as a quality control of the obtained quantitative features. In other words, we assessed whether the (significant) quantitative features were also visible on the ^{18}F-FDG PET images.

XAI permutation test

Random perturbation of the ^{18}F-FDG PET images was performed through randomly adding SUV_0 to the images (where SUV is not zero) using a similar 4x4x4 (SUV_0) voxel matrix to assess the significance ($p < 0.05$) of the XAI method. This was done for each image using the same amount of perturbation as used by the proposed XAI method. This permutation test shows whether the XAI method uses non-specific/specific features and therefore is similar/better than random perturbation.

Results

Model performance

In Table 2, patient demographics and clinical data are presented. Demographic and clinical data were all not significant ($p < 0.05$) different between the two patient groups (except for left and right-sided mCRC and BRAF mutation). Based on the best hyper parameters from each fold, the CNN showed the ability to predict dichotomous PFS of patients with mCRC using pre-treatment ^{18}F-FDG PET image, with an average validation AUC of 0.95 ± 0.11, accuracy of 94 ± 12%, sensitivity of 91 ± 21% and specificity of 94 ± 11% (Figure 3). The high

variance in hyper parameter performance, and the high standard deviations, indicate that the performance of the models highly relies on the data used for training (and validation).

Table 2: Demographic and clinical data of the patients included in this study.

	Part 1 (n = 34)		**Part 2 (n = 46)***		**All patients (n = 80)**	
Variables	**No. (%)**	**Median (range)**	**No. (%)**	**Median (range)**	**No. (%)**	**Median (range)**
Age - Years		64 (50 - 82)		69 (31 - 82)		65 (31 - 82)
Sex						
Male	25 (75)		36 (78)		61 (76)	
Female	9 (25)		10 (22)		19 (24)	
WHO status						
0	9 (27)		11 (28)		20 (25)	
1	22 (65)		24 (62)		46 (58)	
2	3 (9)		4 (10)		7 (9)	
Unknown	0 (0)		7 (15)		7 (9)	
Prior treatment						
Capecitabine/5-FU	34 (100)		46 (100)		80 (100)	
Oxaliplatin	34 (100)		42 (91)		76 (95)	
Irinotecan	30 (88)		29 (63)		59 (74)	
Bevacizumab	23 (68)		27 (59)		50 (63)	
BRAF mutation*						
No	30 (88)		43 (93)		73 (91)	

Yes	4 (12)		3 (7)		7 (88)	
Side*						
Left-sided	25 (75)		40 (87)		65 (81)	
Right-sided	9 (25)		6 (13)		15 (19)	
Months since metastatic disease		24 (5-65)		24 (6-48)		24 (5-65)
Baseline laboratory results						
LDH (U/L)		299 (148-1636)		328 (145-2628)		309 (145-2628)
CEA (µg/L)		59 (5-28386)		69 (2-1697)		63 (2-28386)
Plasma Glucose (mmol/L)		5.5 (4.1-11)		5.7 (5–6.4)		5.5 (4.1–11)
PFS		138 (31-622)		158 (29-350)		152 (29-622)
Dichotomous PFS (152 days)	20/14 (59/41)		20/26 (43/57)		40/40 (50/50)	

*Significant different ($p < 0.05$) between dichotomous PFS patient groups.

WHO: World Health Organization; LDH: Lacto dehydrogenase; CEA: Carcinoembryonic antigen; PFS: Progression free survival

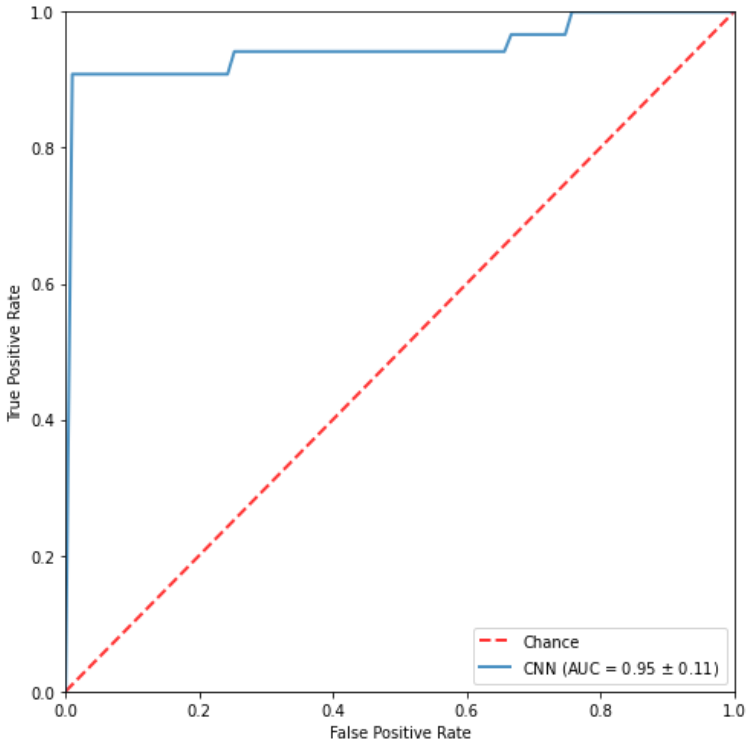


Figure 3: Mean cross-validated ROC curves of the 2.5D-CNNs using ^{18}F-FDG PET images.

Explainable Artificial Intelligence (XAI)

An average AOPC of 83.15 ± 29.59 (standard deviation) and 90.51 ± 25.64 was obtained for the short- and long-term patient groups, respectively. The high standard deviation is because of four images (2/2) which were not able to be described/assessed by the XAI method.

Quantitative analysis

The ^{18}F-FDG PET first-order and volume features are presented in Table 3 and 4, respectively. The first-order features do not differ significantly between the patient groups, except for soft-tissue. However, this difference is because of ^{18}F-FDG accumulation at the radiotracer injection site.

The volume features, however, show significant differences. For the colon and rectum, a higher prevalence (amount of voxels) of SUV_{1-2} is positively associated with long-term PFS (PFS ≥ 152 days) (Figure 4). Similar results are observed for other organs: SUV_{2-4} for the liver, SUV_{1-2}, for the lungs and SUV_{1-3} for the soft-tissue (Figure 4). Especially the ^{18}F-FDG PET volume features of the liver show high significance (p-value: 2.29×10^{-8}). Interestingly, the corresponding non-XAI derived ^{18}F-FDG PET volume features of the liver show high significance as well (p-value: 0.0021). Also, SUV_{6-7}, $SUV_{>4}$ and $SUV_{>6}$ for the lungs, and SUV_{9-11}, SUV_{15-16}, $SUV_{>4}$, and $SUV_{>6}$ for the soft-

tissue are significantly different, but the prevalence (amount of voxels) is low. Furthermore, the total volume of colon and rectum seems to be significantly larger (13625 ± 6358 (PFS < 152 days) vs 17098 ± 6718 (PFS ≥ 152 days) voxels) of the patients with long-term PFS, however the segmentation of this organ can be highly impacted by the amount of defecation and/or air present. Additionally, most (92%) of the significant XAI-derived ^{18}F-FDG PET features have lower p-values compared to non-XAI derived features.

Table 3: XAI derived ^{18}F-FDG PET first-order features.

Organ	Feature	XAI p-value	Non-XAI p-value
Colon and rectum	SUV_{max}	0.31	0.14
	SUV_{peak}	0.20	0.095
	SUV_{mean}	0.12	0.43
	SUV_{median}	0.096	0.40
	AUC-CSH	0.43	0.14
Liver	SUV_{max}	0.18	0.31
	SUV_{peak}	0.24	0.35
	SUV_{mean}	0.39	0.23
	SUV_{median}	0.37	0.11
	AUC-CSH	0.16	0.34
Lungs	SUV_{max}	0.21	0.46
	SUV_{peak}	0.11	0.32
	SUV_{mean}	0.27	0.056
	SUV_{median}	0.48	0.19
	AUC-CSH	0.23	0.41
Soft-tissue	SUV_{max}	0.12	**0.045**
	SUV_{peak}	0.15	0.061
	SUV_{mean}	0.26	0.33
	SUV_{median}	0.22	0.33

	AUC-CSH	0.095	**0.045**

Table 4: XAI derived ^{18}F-FDG PET volume features with corresponding non-XAI derived ^{18}F-FDG PET volume features of the colon and rectum, liver, lungs and soft-tissue.

Organ	**SUV min**	**SUV max**	**XAI p-value**	**Non-XAI p-value**
Colon and rectum	1	2	**2.00×10^{-6}**	**0.0032**
	4	SUV_{max}	0.46	0.46
	6	SUV_{max}	0.43	0.44
	8	SUV_{max}	0.45	0.31
	Volume		**0.019**	
Liver	2	3	**2.29×10^{-8}**	**0.0021**
	3	4	**0.026**	0.15
	4	SUV_{max}	0.44	0.24
	6	SUV_{max}	0.43	0.20
	8	SUV_{max}	0.45	0.27
	Volume		0.24	
Lungs	1	2	**0.0023**	0.070
	7	8	**0.019**	**0.039**
	4	SUV_{max}	0.062	**0.046**
	6	SUV_{max}	**0.038**	**0.023**
	8	SUV_{max}	0.050	**0.027**
	Volume		0.41	
Soft-tissue	1	2	**8.4×10^{-6}**	**0.027**
	2	3	**0.0012**	0.13
	9	10	**0.042**	0.24
	10	11	**0.044**	0.26
	15	16	**0.044**	0.28

	4	SUV_{max}	**0.00036**	**0.016**
	6	SUV_{max}	**0.021**	0.084
	8	SUV_{max}	0.33	**0.018**
	Volume		0.052	

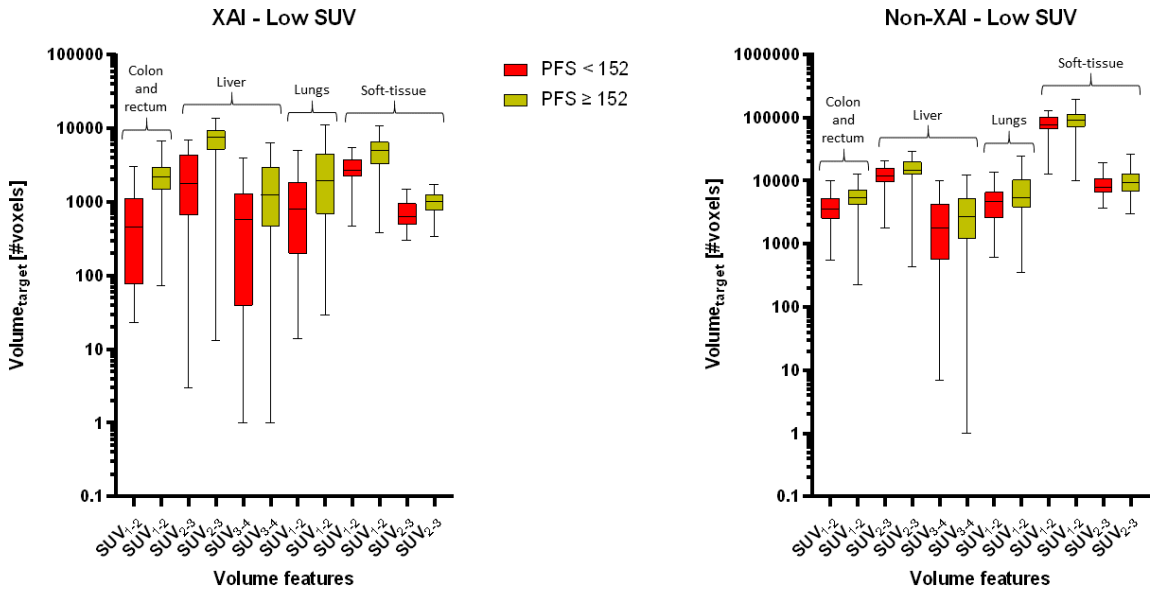


Figure 4: Boxplots of the significant XAI derived low ^{18}F-FDG PET uptake volume features of the two patient groups with corresponding non-XAI derived ^{18}F-FDG PET volume features of the colon and rectum, liver, lungs and soft-tissue.

Qualitative analysis

An attribution image for each ^{18}F-FDG PET image was created (Figure 5). Also, average attribution and ^{18}F-FDG PET images were created for each patient group (Figure 6). It can be seen that especially the liver, the colon, the brain, the kidneys and the rectum showed high attribution. In addition, in the average attribution images of the long-term PFS patient group, more homogenous attribution is seen in the liver, which corresponds with the high prevalence of $SUV_{2\text{-}4}$ in this patient group. For the short-term PFS patient group, more heterogeneous attribution was seen in the liver, which corresponds to the more heterogeneous ^{18}F-FDG PET uptake present in this patient group.

Visual comparison of ten ^{18}F-FDG PET images showed that the patient group with longer PFS has a higher prevalence of $SUV_{2\text{-}4}$ (significant), a similar prevalence of $SUV_{>4}$ and a lower prevalence of $SUV_{1\text{-}2}$ (non-significant) in the liver (Figure 7). Visual comparison of the other three organs (colon and rectum, lungs and mesentery) did not show clear difference between the two patient groups, however, this is due to the non-rigid and heterogeneous characteristics of these organs.

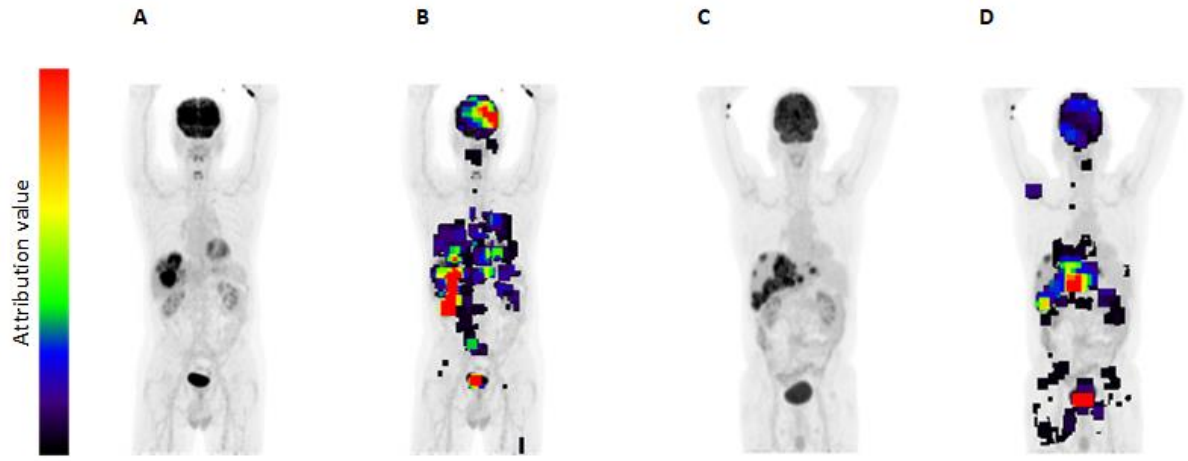


Figure 5: Coronal ^{18}F-FDG PET images (A: 132 days PFS; C: 230 days PFS) and corresponding attribution images (B and D) of two patients with mCRC.

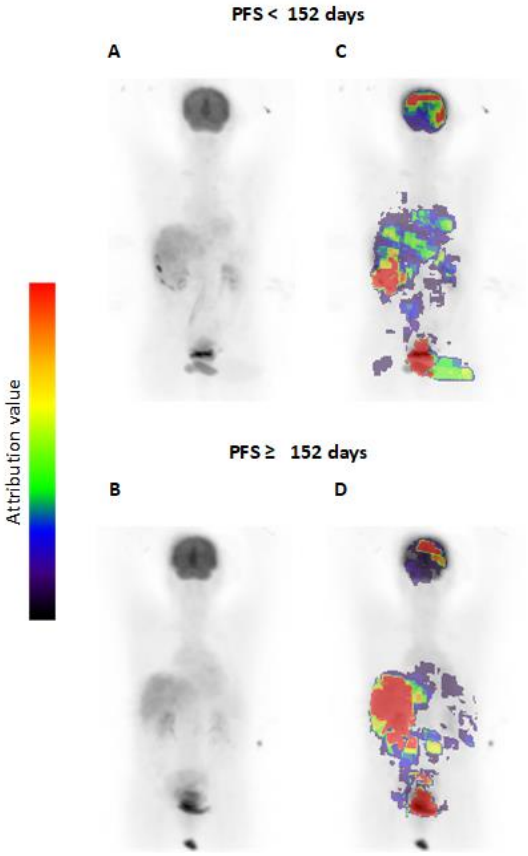


Figure 6: Average ^{18}F-FDG PET images (A and B) and corresponding average attribution images (C and D).

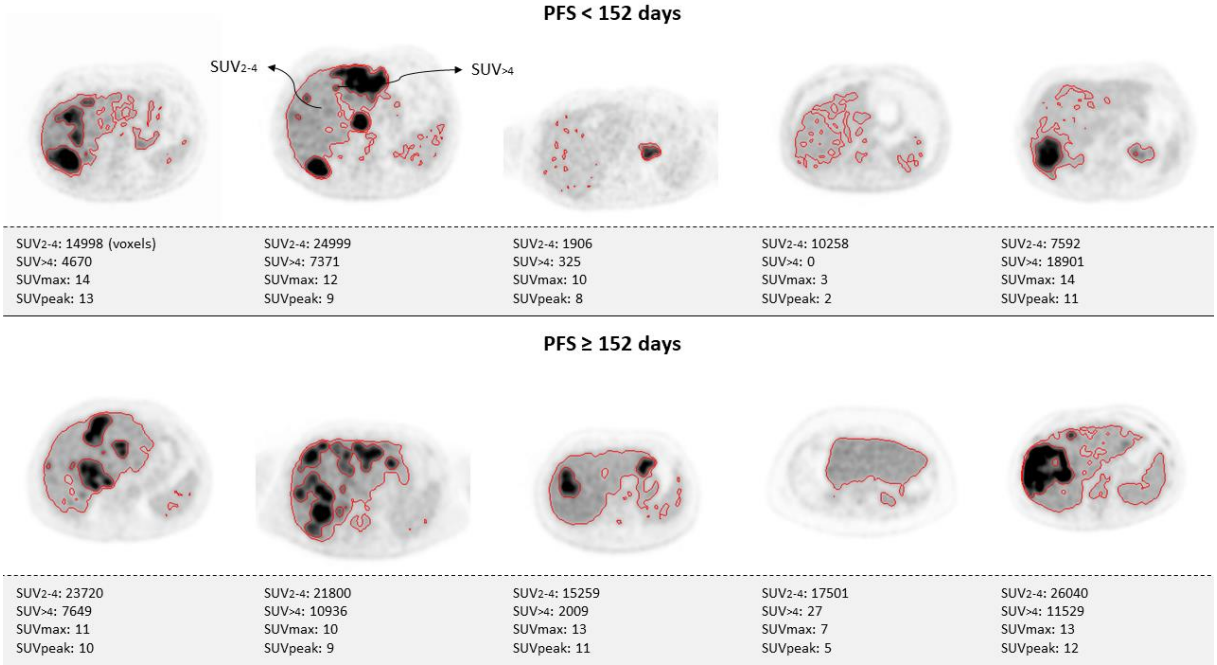


Figure 7: Qualitative assessment of the liver as a quality control of the obtained (significant) ^{18}F-FDG PET features of ten (5/5) mCRC patients. The red-line represents the voxels within $SUV_{2\text{-}4}$ (within the whole image).

XAI permutation analysis

Permutation (random perturbation) of the ^{18}F-FDG PET images using a similar amount of perturbation as the XAI method showed a significant lower AOPC of -0.0015 ± 0.0209 and 0.005275 ± 0.0296 for the short- and long-term PFS patient groups, respectively.

Discussion

A coronal 2.5D-CNN was developed to predict dichotomous PFS in patients with mCRC undergoing anti-EGFR mAb cetuximab and panitumumab treatment using pre-treatment ^{18}F-FDG PET scans. The CNN was able to discriminate between short- and long-term PFS with relative high performance. Also, a medical-based XAI method was developed to provide specific (explainable) ^{18}F-FDG PET features used by the CNN.

In a previous study performed by van Helden et al., radiomic features on pre-treatment ^{18}F-FDG PET were assessed as potential IB for response and survival in patients with mCRC [4]. However, this method may ignore certain abnormalities/features in the tumour (because of predefined radiomic features), and potential relevant features outside of the tumour lesions are not captured using the pre-defined VOIs, limiting the representativeness power of these extracted radiomics. For this reason, we proposed a deep learning framework based on a CNN, which can extract features from the entire image unsupervised and could therefore have a better representation of the actual data of (clinical) interest/potential than predefined radiomic features.

In this study, a coronal 2.5D-CNN over a 3D-CNN was preferred. The ability to learn interslice context (3D) comes with high computation costs, but more importantly, it is more prone to overfitting, especially in this small patient cohort [18]. Overfitting is a problematic challenge because of the limited amount of training data compared with the large number of learnable features. The high disease heterogeneity present in mCRC [26] makes development of a robust and accurate CNN even more difficult. Although we obtained a high average performance (AUC, accuracy, sensitivity and specificity), the high standard deviation suggests that the performance of the models highly depends on the training (and validation) data used for model development, i.e. is impacted by the small dataset and the high disease heterogeneity. Therefore, further validation should be performed using external test data to assess the model's robustness.

In addition to the ^{18}F-FDG PET images, other clinical data, such as (diagnostic) CT images, laboratory tests, medical history, demographics, etc. could also aid in developing more robust and accurate AI models. Current AI models

for nuclear medicine and radiology applications almost only consider pixel/voxel information [27]. However, in clinical practice, these additional data help healthcare professionals to interpret image findings more accurately [27-29]. Therefore, we hypothesize that adding additional medical data could improve model performance.

The ability of a CNN to learn features unsupervised comes at the cost of high complexity and extensive amount of learnable features. Yet, to be able to understand why a CNN makes a certain decision not only helps to improve the model, but also increases confidence of healthcare professional in utilization of these data-driven models [30]. Therefore, in this study we proposed the use of a medical-based XAI method to increase interpretability of the CNN. To the best of our knowledge, we are the first that developed a medical-based XAI method to mitigate the gap between these CNN models and healthcare professionals in medical imaging. For this we combined state-of-the-art XAI knowledge, an anatomical atlas acquired from the ldCT image using nnU-Net, and we (pre-)sampled the ^{18}F-FDG PET SUV image. This combination strengthens the usability of this XAI method. However, implementation of such XAI method in the clinical workflow still requires considerable improvement and validation.

First, misalignment between the ldCT and the ^{18}F-FDG PET image (due to breathing and/or movement of the patient) could results in incorrect interpretation of the ^{18}F-FDG PET image. In this study minor misalignment was acceptable, since most disease characteristics were present in the more bulky organs (e.g. liver). However, PET uptake can also be more subtle and more sensitive to misalignment as is the case in, for example, patients with Multiple Myeloma bone disease [31]. Therefore, for disease agnostic application of this XAI method, better alignment between the ldCT and the ^{18}F-FDG PET image should be pursued, such as respiratory-gating and post-scan image co-registration Also, incomplete segmentation of the structures using nnU-Net could impact the interpretation power of the method. Another potential problem is the partial volume effect as consequence of limited spatial resolution of PET and ldCT. For small organs (e.g. ribs) or organs with thin walls (e.g. colon), the limited spatial resolution can result in considerable quantitative underestimation and/or mis-segmentation as can be the case for the colon and rectum.

Another consideration of this study, is the use of SUV_0 perturbation. Ideally you would prefer a perturbation value, which is not prognostic for both patient groups but also follows imaging modality specific characteristics. In other words, you would want to perturbate the ^{18}F-FDG PET image with a value that is not abnormal for that specific VOI; SUV_0 results in holes in the structures, which is clinically not feasible (with the exception of necrosis).

Therefore, you could speculate about the usability of SUV_0 perturbation in clinical terms. Also, SUV_0 perturbation results in steep gradients in the ^{18}F-FDG PET image, which may have prognostic value in ^{18}F-FDG PET image as well. Yet, from the significant lower AOPCS of the permutation test you could conclude that these steep gradients do not have much impact. Also, from both the quantitative as the qualitative analysis it can be seen that the XAI-derived (using SUV_0 perturbation) and non-XAI derived ^{18}F-FDG PET show similar patterns. In addition, zero perturbation is also current practise in widely used (although non-medical) XAI methods such as RISE [32], LIME [33] and systematic occlusion mapping [22].

An alternative approach is perturbation of the ^{18}F-FDG PET features using a physiological ^{18}F-FDG PET uptake value per organ. Although this provides more clinical realistic perturbation in structures with more homogeneous ^{18}F-FDG PET uptake (e.g. liver), in structures with heterogeneous ^{18}F-FDG PET uptake (e.g. colon) this will results in abnormal PET uptake as well. Also, this may overestimate the prevalence of the healthy organ tissue, which may be problematic as well. In addition, physiological ^{18}F-FDG PET uptake perturbation values also do not guarantee to hold no prognostic value for a specific patient group (SUV_0 more and less does). Furthermore, this approach cannot assess physiological ^{18}F-FDG PET features (which may have prognostic value as well) and therefore cannot provide a complete explanation of the ^{18}F-FDG PET features used by the CNN. In future research, additional steps have to be taken to develop perturbation values, which do not hold prognostic value and also provide clinically realistic images. But for now, SUV_0 seems to provide the most reliable and versatile perturbation value in ^{18}F-FDG PET imaging.

The quantitative analysis showed that first-order ^{18}F-FDG PET features and volume features of high ^{18}F-FDG PET uptake ($SUV_{>4}$; representing tumour uptake) in the liver, colon and rectum are not significant different between the two patient groups. Van Helden et al. conducted a study using similar data and found no significant differences in the tumours using first-order features as well [4]. For the lungs and soft-tissue, significant difference was seen between the patient groups. However, because of low prevalence of these lesions in this cohort, these results are not generalizable.

Interestingly, especially low ^{18}F-FDG PET uptake ($SUV_{<4}$) volume features showed significant difference between the patient groups. In multiple organs higher volume of low ^{18}F-FDG PET uptake is associated with long-term PFS (≥ 152 days). Clear example of this, is the significant higher volume of low ^{18}F-FDG PET uptake in the liver ($SUV_{2–4}$) in patients with longer PFS. It therefore seems that higher volume of $SUV_{2–4}$ in the liver has beneficial impact

on outcome in patient with mCRC undergoing anti-EGFR mAb cetuximab and panitumumab therapy. Total liver volume depends on patient size, liver diseases, previous treatment (e.g. resection, radiofrequency ablation, radiotherapy, etc.) of liver metastasis and presence of liver metastasis. The total liver volume was not significant different between the two patient groups, suggesting that the ratio between healthy liver tissue and tumour bulk and/or diseased tissue in the liver may be different. In other words, the short-term PFS patient group may have insufficient healthy liver tissue in comparison to the long-term PFS patient group. Similar observations were also observed for the colon and rectum, the lungs, and soft-tissue. However, not much is known about the influence of low ^{18}F-FDG PET uptake on the effectiveness of cancer therapy, making the legitimacy and application of these observations (currently) questionable.

Hypoalbuminemia is known to decrease ^{18}F-FDG uptake in the liver, and low serum albumin levels have shown to correlate with increased hepatocellular carcinoma aggressiveness [34, 35]. The short-term PFS patient group may therefore be impacted by lower serum albumin levels, because of the higher disease aggressiveness seen in this patient group. However, although this may be the case for some patients, a significant lower ^{18}F-FDG PET uptake in the liver is not seen for the short-term PFS patient group. This indicates that, although hypoalbuminemia could partly explain the significant lower $SUV_{2\text{-}4}$ prevalence in this patient group, it is the heterogeneous intra-patient group ^{18}F-FDG PET uptake that results in this inter-patient group difference. Although these observations are (currently) not completely supported by relevant literature, these observations may contain relevant information about biological processes important for this patient cohort and therapy. Therefore, we advocate for further assessment of these features in future studies.

Future studies could compare the biodistribution of cetuximab and panitumumab between the patient groups. A previous study performed by van Helden et al. did not show a relation between tumour PET uptake and therapy response using ^{89}Zr-cetuximab PET imaging in patients with mCRC [14]. However, in this study only the relation between SUV of the tumour lesions and therapy response were assessed. It would also be of interest to assess the relation of non-tumour features to therapy response in ^{89}Zr-cetuximab PET imaging. This may provide use with information about the possible biodistribution difference of cetuximab and panitumumab between the two patient groups and how it relates to ^{18}F-FDG PET uptake seen in this study.

In this study we only focussed on the primary tumour and the most common metastatic sites (liver, lungs, mesentery). However, other organs may hold prognostic value as well. Bone marrow may for instance also play

a role in disease prognostication; high bone marrow ^{18}F-FDG PET uptake has shown to be related to advanced staging in colorectal cancer [36]. Also, abnormal brain metabolic patterns observed in patients with colorectal cancer in a preliminary study performed by Jie Ma et al. suggest that cerebral metabolic patterns may be associated with disease burden in colorectal cancer [37]. Therefore, in future studies further assessment should be conducted to obtain a complete overview of the organs and ^{18}F-FDG PET features associated with disease aggressiveness in patients with mCRC.

Further quality assessment of the XAI method should be performed using other (but similar in performance) CNN weights derived from the cross-validation. It could be the case that current used weights are biased towards specific ^{18}F-FDG PET features (although unlikely because of similar patterns seen in XAI and non-XAI derived features). By comparing the ^{18}F-FDG PET features extracted from the different CNN weights, the reliability/robustness of the current derived ^{18}F-FDG PET features can further be assessed. Also, four ^{18}F-FDG PET images were not able to be described by the proposed XAI method. Possible causes of this is the (possible) incompleteness of current XAI method, for example, SUV_0 perturbation, coalitions which were not assessed and the spatial resolution of the XAI framework (not able to find (very) fine-grained details).

First steps are made to understand the behaviour of this CNN model. However, for successful implementation we still need to invest more in the development of clinical realistic and neutral perturbation values. This is not only of interest to get a better understanding of the behaviour of these CNNs, but also to investigate the legitimacy and usability of the XAI method in this field [38]. The obtained results require extensive further assessment, but because of the extensive quality control performed in this study, we believe that these results may hold valuable information to support clinical prognostication for treatment decision in this cohort.

Conclusion

A coronal 2.5D-CNN to classify dichotomous PFS from pre-treatment ^{18}F-FDG PET in patients with mCRC undergoing anti-EGFR mAb cetuximab and panitumumab treatment was successfully constructed and trained. From the medical-based XAI framework, low ^{18}F-FDG PET uptake volume features from multiple organs are promising to hold prognostic value for treatment with anti-EGFR mAb. However, extensive validation should be performed to develop IB which can be used clinically for patients with mCRC eligible for anti-EGFR mAb cetuximab and panitumumab treatment.

References

[1] R. Kochhar, S. Liong, and P. Manoharan, "The role of FDG PET/CT in patients with colorectal cancer metastases," (in eng), *Cancer Biomark,* vol. 7, no. 4, pp. 235-48, 2010, doi: 10.3233/cbm-2010-0201.

[2] D. Vriens, L. F. de Geus-Oei, W. T. van der Graaf, and W. J. Oyen, "Tailoring therapy in colorectal cancer by PET-CT," (in eng), *Q J Nucl Med Mol Imaging,* vol. 53, no. 2, pp. 224-44, Apr 2009.

[3] E. J. van Helden *et al.*, "Early 18F-FDG PET/CT Evaluation Shows Heterogeneous Metabolic Responses to Anti-EGFR Therapy in Patients with Metastatic Colorectal Cancer," (in eng), *PloS one,* vol. 11, no. 5, p. e0155178, 2016, doi: 10.1371/journal.pone.0155178.

[4] E. J. van Helden *et al.*, "Radiomics analysis of pre-treatment [(18)F]FDG PET/CT for patients with metastatic colorectal cancer undergoing palliative systemic treatment," (in eng), *Eur J Nucl Med Mol Imaging,* vol. 45, no. 13, pp. 2307-2317, Dec 2018, doi: 10.1007/s00259-018-4100-6.

[5] A. Lièvre *et al.*, "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer," (in eng), *Cancer Res,* vol. 66, no. 8, pp. 3992-5, Apr 15 2006, doi: 10.1158/0008-5472.Can-06-0191.

[6] J. C. Jones *et al.*, "(Non-V600) BRAF Mutations Define a Clinically Distinct Molecular Subtype of Metastatic Colorectal Cancer," (in eng), *J Clin Oncol,* vol. 35, no. 23, pp. 2624-2630, Aug 10 2017, doi: 10.1200/jco.2016.71.4394.

[7] F. Loupakis *et al.*, "Primary tumor location as a prognostic factor in metastatic colorectal cancer," (in eng), *J Natl Cancer Inst,* vol. 107, no. 3, Mar 2015, doi: 10.1093/jnci/dju427.

[8] P. Alongi *et al.*, "Artificial Intelligence Applications on Restaging [18F]FDG PET/CT in Metastatic Colorectal Cancer: A Preliminary Report of Morpho-Functional Radiomics Classification for Prediction of Disease Outcome," *Applied Sciences,* vol. 12, p. 2941, 03/13 2022, doi: 10.3390/app12062941.

[9] L. Wei *et al.*, "A deep survival interpretable radiomics model of hepatocellular carcinoma patients," (in eng), *Phys Med,* vol. 82, pp. 295-305, Feb 2021, doi: 10.1016/j.ejmp.2021.02.013.

[10] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis,* vol. 79, p. 102470, 2022/07/01/ 2022, doi: https://doi.org/10.1016/j.media.2022.102470.

[11] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, doi: https://doi.org/10.48550/arXiv.1705.07874.

[12] B. H. van der Velden, M. H. Janse, M. A. Ragusi, C. E. Loo, and K. G. Gilhuijs, "Volumetric breast density estimation on MRI using explainable deep learning regression," *Scientific Reports,* vol. 10, no. 1, pp. 1-9, 2020, doi: https://doi.org/10.1038/s41598-020-75167-6.

[13] G. J. C. Z. Bart M. de Vries, George L. Burchell, Floris H.P. van Velden, C. Willemien Menke-van der Houven van Oordt and Ronald Boellaard, "Explainable Artificial Intelligence (XAI) in radiology and nuclear medicine: A literature review," ed. Frontiers in Medicine, 2023.

[14] E. J. van Helden *et al.*, "[89Zr]Zr-cetuximab PET/CT as biomarker for cetuximab monotherapy in patients with RAS wild-type advanced colorectal cancer," *European Journal of Nuclear Medicine and Molecular Imaging,* vol. 47, no. 4, pp. 849-859, 2020/04/01 2020, doi: 10.1007/s00259-019-04555-6.

[15] "Systemische therapie bij niet lokaal behandelbare metastasen." https://richtlijnendatabase.nl/richtlijn/colorectaal_carcinoom_crc/gemetastaseerd_colorectaalcarcinoom_crc/systemische_therapie_bij_niet_lokaal_behandelbare_metastasen_bij_crc.html (accessed.

[16] R. Boellaard *et al.*, "FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0," (in eng), *Eur J Nucl Med Mol Imaging,* vol. 37, no. 1, pp. 181-200, Jan 2010, doi: 10.1007/s00259-009-1297-4.

[17] L. H. Schwartz *et al.*, "RECIST 1.1-Update and clarification: From the RECIST committee," (in eng), *Eur J Cancer,* vol. 62, pp. 132-7, Jul 2016, doi: 10.1016/j.ejca.2016.03.081.

[18] B. M. de Vries *et al.*, "Classification of negative and positive 18F-florbetapir brain PET studies in subjective cognitive decline patients using a convolutional neural network," *European Journal of Nuclear Medicine and Molecular Imaging,* vol. 48, no. 3, pp. 721-728, 2021/03/01 2021, doi: 10.1007/s00259-020-05006-3.

[19] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," *CoRR,* vol. abs/1704.02685, / 2017. [Online]. Available: http://arxiv.org/abs/1704.02685.

[20] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," (in eng), *Nat Methods,* vol. 18, no. 2, pp. 203-211, Feb 2021, doi: 10.1038/s41592-020-01008-z.

[21] H. N. Rier, A. Jager, S. Sleijfer, A. B. Maier, and M. D. Levin, "The Prevalence and Prognostic Value of Low Muscle Mass in Cancer Patients: A Review of the Literature," (in eng), *Oncologist,* vol. 21, no. 11, pp. 1396-1409, Nov 2016, doi: 10.1634/theoncologist.2016-0066.

[22] A. Lopatina, S. Ropele, R. Sibgatulin, J. R. Reichenbach, and D. Güllmar, "Investigation of Deep-Learning-Driven Identification of Multiple Sclerosis Patients Based on Susceptibility-Weighted Images Using Relevance Analysis," (in eng), *Frontiers in neuroscience,* vol. 14, p. 609468, 2020 2020, doi: doi:10.3389/fnins.2020.609468.

[23] F. H. van Velden *et al.*, "Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies," (in eng), *Eur J Nucl Med Mol Imaging,* vol. 38, no. 9, pp. 1636-47, Sep 2011, doi: 10.1007/s00259-011-1845-6.

[24] M. Riihimäki, A. Hemminki, J. Sundquist, and K. Hemminki, "Patterns of metastasis in colon and rectal cancer," (in eng), *Sci Rep,* vol. 6, p. 29765, Jul 15 2016, doi: 10.1038/srep29765.

[25] E. Garyfallidis *et al.*, "Dipy, a library for the analysis of diffusion MRI data," (in English), *Frontiers in neuroinformatics,* Methods vol. 8, 2014-February-21 2014, doi: 10.3389/fninf.2014.00008.

[26] C. Molinari, G. Marisi, A. Passardi, L. Matteucci, G. De Maio, and P. Ulivi, "Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine?," (in eng), *Int J Mol Sci,* vol. 19, no. 12, Nov 23 2018, doi: 10.3390/ijms19123733.

[27] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *npj Digital Medicine,* vol. 3, no. 1, p. 136, 2020/10/16 2020, doi: 10.1038/s41746-020-00341-z.

[28] M. D. Cohen, "Accuracy of information on imaging requisitions: does it matter?," (in eng), *J Am Coll Radiol,* vol. 4, no. 9, pp. 617-21, Sep 2007, doi: 10.1016/j.jacr.2007.02.003.

[29] A. Leslie, A. J. Jones, and P. R. Goddard, "The influence of clinical information on the reporting of CT by radiologists," (in eng), *Br J Radiol,* vol. 73, no. 874, pp. 1052-5, Oct 2000, doi: 10.1259/bjr.73.874.11271897.

[30] P. Papadimitroulas *et al.*, "Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization," *Physica Medica,* vol. 83, pp. 108-121, 2021/03/01/ 2021, doi: https://doi.org/10.1016/j.ejmp.2021.03.009.

[31] J. C. Regelink, G. J. C. Zwezerijnen, R. J. W. Groen, P. G. Raijmakers, and S. Zweegman, "In vivo (18) F-fluoride-PET imaging reveals pronounced heterogeneity in bone formation in multiple myeloma patients," (in eng), *Br J Haematol,* vol. 200, no. 6, pp. 755-758, Mar 2023, doi: 10.1111/bjh.18588.

[32] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421,* 2018.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016, doi: https://doi.org/10.48550/arXiv.1602.04938.
[34] Y. Otomi *et al.*, "A reduced liver (18)F-FDG uptake may be related to hypoalbuminemia in patients with malnutrition," (in eng), *Ann Nucl Med,* vol. 33, no. 9, pp. 689-696, Sep 2019, doi: 10.1007/s12149-019-01377-2.
[35] B. I. Carr and V. Guerra, "Serum Albumin Levels in Relation to Tumor Parameters in Hepatocellular Carcinoma Patients," *The International Journal of Biological Markers,* vol. 32, no. 4, pp. 391-396, 2017, doi: 10.5301/ijbm.5000300.
[36] F. Tustumi *et al.*, "Prognostic Value of Bone Marrow Uptake Using 18F-FDG PET/CT Scans in Solid Neoplasms," (in eng), *J Imaging,* vol. 8, no. 11, Oct 31 2022, doi: 10.3390/jimaging8110297.
[37] J. Ma *et al.*, "Cerebral Metabolic Analysis of Patients With Colorectal Cancer and Chronic Enteritis: Inquiry Into Gut-Brain Crosstalk," (in eng), *Frontiers in neuroscience,* vol. 16, p. 822891, 2022, doi: 10.3389/fnins.2022.822891.
[38] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *ArXiv,* vol. abs/2006.11371, 2020.
[39] G. S. Collins *et al.*, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence," (in eng), *BMJ Open,* vol. 11, no. 7, p. e048008, Jul 9 2021, doi: 10.1136/bmjopen-2020-048008.

Statements & Declarations

TRIPOD-AI

The authors declare that this study followed (if applicable) the protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence [39].

Funding

This work was funded by KWF - Alpe d'Huez [2012–5565]. The funders had no role in study design, data collection and analysis or preparationof the manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Bart M. de Vries, Sophie Gerritse, Floris H.P. van Velden, Ronald Boellaard, and Willemien C.

Menke-van der Houven van Oordt. The first draft of the manuscript was written by Bart M. de Vries and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request

Institutional Review Board Statement

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board (or Ethics Committee) of Amsterdam UMC, location VUmc. (Date: April 21, 2014/No. NCT02117466).

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.