Deep Learning for Outcome Prediction in Cancer using Positron Emission Tomography Images

Amine Amyar

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Rouen Normandie

Apprentissage profond pour la prédiction de la réponse au traitement et la survie en cancérologie en imagerie TEP

Présentée et soutenue par
Amine AMYAR

Thèse soutenue le 31/08/2021 devant le jury composé de		
M. JOHN LEE	PROFESSEUR DES UNIVERSITES, Université Catholique de Louvain	Rapporteur du jury
Mme DIANA MATEUS	PROFESSEUR DES UNIVERSITES, ECOLE CENTRALE NANTES	Rapporteur du jury
M. ROMAIN MODZELEWSKI	INGENIEUR, Université de Rouen Normandie	Membre du jury
M. VINCENT MORARD	INGENIEUR,	Membre du jury
M. PIERRE VERA	PROFESSEUR DES UNIV - PRATICIEN HOSP., Université de Rouen Normandie	Membre du jury
M. DIMITRIS VISVIKIS	DIRECTEUR DE RECHERCHE, UNIVERSITE BRET. OCCIDENTALE UBO	Membre du jury
Mme SU RUAN	PROFESSEUR DES UNIVERSITES, Université de Rouen Normandie	Directeur de thèse

Thèse dirigée par SU RUAN, Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

UNIVERSITÉ DE ROUEN NORMANDIE

MIIS

litis

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de « l'Université De Rouen Normandie »

Deep Learning for Outcome Prediction in Cancer using Positron Emission Tomography Images

Présentée et soutenue par
Amine Amyar

Thèse soutenue publiquement le (date de soutenance) devant le jury composé de		
M. Dimitris Visvikis	Dr, Inserm UMR, UBO	Président du Jury
M. John Lee	Pr, Université catholique de Louvain	Rapporteur
Mme. Diana Mateus	Pr, Centrale Nantes	Rapporteur
M.Pierre Vera	PUPH, Centre Henri Becquerel, Rouen	Examinateur
M. Vincent Morard	PhD, General Electric Medical Systems	Examinateur
M. Baptiste Perrin	Directeur R&D, General Electric Medical Systems	Membre invité
Mme. Su Ruan	Pr, Université de Rouen	Directrice
M. Romain Modzelewski	PhD, Centre Henri Becquerel	Encadrant

Thèse dirigée par Su Ruan, laboratoire LITIS

UNIVERSITÉ DE ROUEN NORMANDIE

GE Healthcare

ed MIIS

litis

“The good physician treats the disease; the great physician treats the patient who has the disease.”

“It is much more important to know what sort of patient has a disease than what sort of disease a patient has.”

“Medicine is a science of uncertainty and art of probability.”

William Osler

“A very great deal more truth can become known than can be proven”

Richard Feynman

“No one knows what the right algorithm is, but it gives us hope that if we can discover some crude approximation of whatever this algorithm is and implement it on a computer, that can help us make a lot of progress.”

Andrew Ng

Summary

Precision medicine (also known as personalized medicine) has been proposed to customize healthcare for each patient, from medical diagnosis to treatment, impacting medical decisions and practices as well as current workflow. To meet this objective, patients are placed into different groups based on some relevant similarities to take a medical decision. Precision medicine primarily uses information about a person's clinical records, biological information including proteins (proteomics), genes (genomics), and, more recently, images (radiomics). In the case of cancer, information about the tumor is also incorporated to make a diagnosis, decide on the type of treatment, monitor disease progression or predict treatment response or prognosis. Precision medicine for cancer relies on the use of tumor markers to aid in diagnosis or targeted therapies to treat certain types of cancer.

Radiomics is a research field where images are used for their potential in precision medicine. It is defined as the analysis of a large number of extracted features from medical images such as *Computed Tomography* (CT), *Magnetic Resonance Imaging* (MRI) or *Positron Emission Tomography* (PET). These features are used to uncover disease characteristics that fail to be found or quantified by the naked eye. The first step in radiomic analysis in oncology is the lesion segmentation, which is the process of isolating a *Region Of Interest* (ROI) from other regions with contours. After segmentation, thousands of features can be extracted from the ROI, and then the most relevant ones are selected. Finally, a machine learning algorithm such as *Random Forest* (RF) or *Support Vector Machine* (SVM) is applied to identify the best relevant features that predict the outcome. This classical workflow is limited for several reasons : segmentation requires a highly trainable physician, is time consuming and the defined ground truth is physician subjective and prone to error (intra and inter observer variability). Secondly, the handcrafted features defined from the ROI are limited since they are heavily influenced by many factors like the used segmentation method. Therefore they fail when the ROI is altered.

Recently, deep learning has dramatically changed the field of computer vision, including image classification, object detection, and image segmentation. In the medical imaging field, various applications of deep learning have emerged in different areas, including pathology classification, risk stratification, treatment response prediction, lesion and organ segmentation. Thus, artificial intelligence in general and deep learning in particular can come in handy to develop *Computer Aided Diagnostic* (CAD) applications. However, deep learning approaches are well known for their data hungry nature, and annotated data are usually hard to obtain in the medical imaging field.

The goal of this thesis is to go beyond the current radiomic paradigm which requires manual extraction of characteristics and replace it by deep radiomics. In our new approach, features are learned along with the prediction of the outcome. To achieve this, we develop different *Deep Learning* (DL) algorithms to create end-to-end architectures that take an image as input, learn feature representation and outcome prediction.

The first method that we propose is to create a deep radiomics paradigm by exploring

a *Convolutional Neural Network* (CNN) due to its predictive power. We created an end-to-end prediction model based on a 3D CNN, called 3D RPET-NET, that jointly extracts features from a 3D PET image volume and predict the outcome of therapy. The obtained results outperform classical radiomic approaches.

As mentioned above, annotated data is a major issue in the medical imaging field, where only a small subset of annotated images are available. We propose a *Weakly Supervised Learning* (WSL) method to solve this problem. Our method allows to segment automatically the lesion for radiomic analysis, without segmentation ground truth and with only a weak annotation (class of the pathology and one voxel in the region of the tumor). The key step is to segment the tumor in 3D. Our segmentation method is composed of four steps : 1) calculate two *"Maximum Intensity Projection"* (MIP) images from 3D PET images of lung and esophageal cancers in two directions 2) classify the MIP images into different types of cancers 3) generate the class activation maps through a multitask learning approach with a weak prior knowledge 4) segment the 3D tumor region from the two 2D activation maps with a proposed new loss. Our proposed approach can obtain state of the art of prediction results with a very weak segmentation ground truth.

Recent studies have shown the potential of peritumoral regions on boosting the accuracy of outcome prediction. Thus, the association of the intratumoral and peritumotal regions provides richer information than one region for radiomic analysis. Therefore, we develop a new segmentation network that does not give the same ground truth as physicians do, but to find the regions that contribute the most in the outcome prediction. Our method is based on *Multi-Task Learning* (MTL) framework, which is a type of learning algorithm that aims to combine several pieces of information from different tasks in order to improve the model's performance and its ability to better generalise. The basic idea of MTL is that different tasks can share a representation of common characteristics, and thus train them jointly.

Our method jointly performs 4 tasks : image reconstruction, pathology classification, tumor segmentation and outcome prediction in a multi-task learning way. We show that the encoder can benefits from multiple tasks to extract meaningful and powerful features that boost radiomic performance, and that subsidiary tasks serve as an inductive bias so the learned model can generalize better. Our model was tested and validated for treatment response and survival in lung and esophageal cancers, outperforming single task learning methods. We show also that, by using a MTL approach, we can boost the performance of radiomics analysis thanks to the rich information extracted from intratumoral and peritumoral regions. The MTL architecture was also tested on a COVID-19 dataset with success.

Résumé

La médecine de précision (également appelée médecine personnalisée) a été proposée pour personnaliser les soins de santé pour chaque patient, du diagnostic médical au traitement, ce qui a un impact sur les décisions et les pratiques médicales ainsi que sur le flux de travail actuel. Pour atteindre cet objectif, les patients sont placés dans différents groupes en fonction de certaines similitudes pertinentes pour prendre une décision médicale. La médecine de précision utilise principalement des informations sur les dossiers cliniques d'une personne, des informations biologiques, notamment des protéines (protéomique), des gènes (génomique) et, plus récemment, des images (radiomique). Dans le cas du cancer, des informations sur la tumeur sont également incorporées pour établir un diagnostic, décider du type de traitement, suivre la progression de la maladie ou prédire la réponse au traitement ou le pronostic. La médecine de précision pour le cancer repose sur l'utilisation de marqueurs tumoraux pour faciliter le diagnostic ou de thérapies ciblées pour traiter certains types de cancer.

La radiomique est un domaine de recherche où les images sont utilisées pour leur potentiel dans la médecine de précision. Elle se définit comme l'analyse d'un grand nombre de caractéristiques extraites d'images médicales telles que les CT, MRI ou PET. Ces caractéristiques sont utilisées pour découvrir les caractéristiques de la maladie qui ne peuvent être trouvées ou quantifiées à l'œil nu. La première étape de l'analyse radiomique en oncologie est la segmentation de la lésion, qui consiste à isoler une ROI des autres régions à l'aide de contours. Après la segmentation, des milliers de caractéristiques peuvent être extraites de l'ROI, puis les plus pertinentes sont sélectionnées. Enfin, un algorithme d'apprentissage automatique tel que RF ou SVM est appliqué pour identifier les meilleures caractéristiques pertinentes qui prédisent le résultat. Ce flux de travail classique est limité pour plusieurs raisons : la segmentation nécessite un médecin hautement qualifié, elle est chronophage et la vérité terrain définie est subjective et sujette à erreur (variabilité intra et inter observateur). Deuxièmement, les caractéristiques artisanales définies à partir du ROI sont limitées car elles sont fortement influencées par de nombreux facteurs tels que la méthode de segmentation utilisée. Par conséquent, elles échouent lorsque le ROI est modifié.

Récemment, l'apprentissage profond a radicalement changé le domaine de la vision par ordinateur, notamment la classification des images, la détection des objets et la segmentation des images. Dans le domaine de l'imagerie médicale, diverses applications de l'apprentissage profond sont apparues dans différents domaines, notamment la classification des pathologies, la stratification des risques, la prédiction de la réponse au traitement, la segmentation des lésions et des organes. Ainsi, l'intelligence artificielle en général et l'apprentissage profond en particulier peuvent s'avérer utiles pour développer des applications CAD. Cependant, les approches d'apprentissage profond sont bien connues pour leur nature avide de données, et les données annotées sont généralement difficiles à obtenir dans le domaine de l'imagerie médicale.

L'objectif de cette thèse est de surpasser le paradigme actuel de la radiomique qui né-

cessite une extraction manuelle des caractéristiques et de le remplacer par la radiomique profonde. Dans notre nouvelle approche, les caractéristiques sont apprises en même temps que la prédiction du résultat. Pour y parvenir, nous développons différents algorithmes pour créer des architectures de bout en bout qui prennent une image en entrée, apprennent la représentation des caractéristiques et la prédiction des résultats.

La première méthode que nous proposons consiste à créer un paradigme de radiomique profonde en explorant un CNN en raison de son pouvoir prédictif. Nous avons créé un modèle de prédiction de bout en bout basé sur un CNN 3D, appelé 3D RPET-NET, qui extrait conjointement les caractéristiques à partir d'une image CNN en 3D et prédit le résultat du traitement. Les résultats obtenus surpassent les approches radiomiques classiques.

Comme mentionné ci-dessus, les données annotées constituent un problème majeur dans le domaine de l'imagerie médicale, où seul un petit sous-ensemble d'images annotées est disponible. Nous proposons une méthode WSL pour résoudre ce problème. Notre méthode permet de segmenter automatiquement la lésion pour l'analyse radiomique, sans vérité terrain pour la segmentation et avec seulement une faible annotation (classe de la pathologie et un voxel dans la région de la tumeur). L'étape clé est de segmenter la tumeur en 3D. Notre méthode de segmentation est composée de quatre étapes : 1) calculer deux images MIP à partir d'images PET 3D de cancers du poumon et de l'œsophage dans deux directions 2) classer les images MIP en différents types de cancers 3) générer les cartes d'activation de classe par une approche d'apprentissage multitâche avec une faible connaissance a priori 4) segmenter la région tumorale 3D à partir des deux cartes d'activation 2D avec une nouvelle fonction de perte. L'approche que nous proposons permet d'obtenir des résultats comparable à l'état de l'art pour la prédiction avec une vérité terrain très faible pour la segmentation.

Des études récentes ont montré le potentiel des régions péritumorales pour améliorer la précision de la prédiction de la réponse au traitement et la survie. Ainsi, l'association des régions intratumorale et péritumorale fournit des informations plus riches qu'une seule région pour l'analyse radiomique. Par conséquent, nous développons un nouveau réseau de segmentation qui ne donne pas la même vérité terrain que les médecins, mais qui permet de trouver les régions qui contribuent le plus à la prédiction. Notre méthode est basée sur l'apprentissage MTL, qui est un type d'algorithme d'apprentissage visant à combiner plusieurs éléments d'information provenant de différentes tâches afin d'améliorer les performances du modèle et sa capacité à mieux généraliser. L'idée de base du MTL est que différentes tâches peuvent partager une représentation de caractéristiques communes, et donc les entraîner conjointement.

Notre méthode réalise conjointement 4 tâches : la reconstruction de l'image, la classification de la pathologie, la segmentation de la tumeur et la prédiction de la réponse au traitement et la survie, dans le cadre d'un apprentissage multi-tâches. Nous montrons que l'encodeur peut bénéficier de tâches multiples pour extraire des caractéristiques significatives et puissantes qui améliorent la performance radiomique, et que les tâches subsidiaires servent de biais inductif pour que le modèle appris puisse mieux généraliser. Notre modèle a été testé et validé pour la réponse au traitement et la survie dans les cancers du poumon et de l'œsophage, surpassant les méthodes d'apprentissage à tâche unique. Nous montrons également qu'en utilisant une approche MTL, nous pouvons améliorer les performances de l'analyse radiomique grâce à la richesse des informations extraites des régions intratumorales et péritumorales. L'architecture MTL a également été testée avec succès sur un jeu de données COVID-19.

Remerciements

First of all, I would like to thank the members of the jury of this thesis for their interest in my work. I particularly thank Mrs. Diana Mateus and Mr. John LEE, for having accepted to report my thesis. I also thank Mr. Dimitris Visvikis for having done me the honor of chairing the jury.

There are many people I would like to thank for helping me during my stay at LITIS at the university of Rouen and General Electric Healthcare.

I wish to express my gratitude to Mrs. Su Ruan, as well as to Mr. Romain Modzlewski who have supervised this thesis. I thank them for their very instructive advice in the scientific field, and for their notable availability, favourable to the good progress of my thesis. I would also like to thank them for their precious help which allowed me to perfect the realization of this thesis.

I also thank the company General Electric Healthcare for having accompanied my thesis, starting with Mr. Baptiste Perrin for having accepted me within his team while leaving me a great freedom in my research. Of course, I also thank Vincent Morard for his help and his expertise concerning deep learning and medical imaging.

My gratitude goes then to all the members of the Henri Becquerel Center of Rouen with whom I had the chance to work during this thesis. I will start by thanking Mr Pierre VERA for his expertise in the medical field and his interest in all forms of research which have been a source of motivation for me. My gratitude also goes to all nuclear medicine team : Pierrick Gouel, Pierre Decazes, Sébastien Hapdey and the others who accompanied me during my thesis and who knew how to listen to me when obstacles were presented. My gratitude also goes to Mr. Simon BERNARD, associate professor at the University of Rouen, for his expertise and his relevant advice.

I thank all the members of the QuantIF team, who made these years of thesis pleasant

and friendly : Pauline, Alexandre, Zhou, Paul, Elyse, Thibaud and Mohammed. I hope to have the opportunity to meet you in the future.

For their unfailing support and regular encouragement, I thank my family infinitely for making me the person I am today and for always giving me everything, especially my parents Belkacem and Nadia, my brothers Yahia and Bilal and my sister Fairouz.

Pour Rassim

Table des matières

General Introduction **1**

1 Medical imaging as a diagnostic and prediction tool **5**

1.1 Cancer 6

1.2 FDG PET imaging : principle and characteristics 14

1.3 Conclusion 25

2 Radiomics and machine learning **27**

2.1 Introduction 28

2.2 Basic notions in Machine learning 29

2.3 Radiomics 41

2.4 Machine learning for radiomics 42

2.5 Deep learning for radiomics 47

2.6 Objectives of the thesis 48

3 Hand crafted methods vs deep radiomics **51**

3.1 Introduction 52

3.2 Material and methods 55

3.3 Experimentations 60

3.4 Validation methodology 60

3.5 Results 63

3.6 Discussion 64

3.7 Conclusion 66

4 Weakly supervised learning for outcome prediction **69**

4.1 Introduction . . . 70

4.2 Material and methods . . . 74

4.3 Experiments . . . 80

4.4 Evaluation Methodology . . . 83

4.5 Results . . . 83

4.6 Discussion & Conclusion . . . 85

5 Multitask learning for radiomics analysis **87**

5.1 Introduction . . . 88

5.2 Related Work . . . 90

5.3 Method . . . 94

5.4 Experiments . . . 99

5.5 Validation Methodology . . . 100

5.6 Results . . . 102

5.7 Discussion . . . 106

5.8 Conclusion . . . 108

Conclusion and Discussion **109**

5.9 Limitations . . . 112

5.10 Perspectives for future research . . . 114

Bibliography **118**

Table des figures **135**

List of tables **139**

Glossaire **141**

General Introduction

Cancers present a strong heterogeneity within and between patients, which occurs at different levels/scales : genetic, cellular, tissue, organ ... etc. It also evolves during the course of the disease and therapy [Marusyk et al. 2012]. This limits the use of invasive procedures such as biopsies, on which molecular and genetic analyses are carried out, but on the other hand gives enormous potential to non-invasive imaging techniques [Yip and Aerts 2016]. Over the past decade, the use and role of medical imaging in clinical oncology has increased dramatically. Recent advances in medical imaging allow the use of image analysis methods that go beyond the localization of organs and tumors and simple measurements of their size. Imaging therefore has great potential to guide treatment, monitor progress, predict disease progression and response to treatment.

"Radiomics" [Kumar et al. 2012, Lambin et al. 2012b], namely the computational analysis of medical images is recently used as a surrogate for the determination of complex image features. Until recently, evaluation of images has been limited to what the eye can see, but the complex interactions between tissues at the image level is a treasure-trove of information that can only be fully utilized with computational methods. Here, we propose to harness this information by applying deep learning methods to predict patient's outcome, study intra and inter-tumoral heterogeneity to better assess the underlying tumor changes that may be impacting prognosis and therapy response.

Many methodologies have been proposed for patient stratification and biomarker identification [Elefsinioti et al. 2016, Sorani et al. 2010], some of them building a joint latent variable model to simultaneous infer cluster assignments from multiple data types [Shen et al. 2009], or building networks of patients as a basis for data integration [Wang et al. 2014]. Though powerful, these approaches do not scale well to high-dimensionality data,

making the algorithms sensitive to a necessary initial feature pre-selection step. Despite several decades of research, predictive biomarkers are scarce, limited to the metastatic setting and are more effective at identifying non-responders than patients who may benefit from treatment. Here, we aim to provide novel insights into radiomic and therapy responsiveness by developing new prediction methods based on deep learning approach that uses multiple layers to progressively extract higher-level features from the raw input and predict the outcome in an end-to-end model.

In the last years, deep learning has seen large success for different applications such as image classification [Krizhevsky et al. 2012], object detection [Zhao et al. 2019], speech recognition [Hinton et al. 2012] and in various applications in the medical imaging field [Lee et al. 2017, Ravì et al. 2016]. Deep learning methods are data hungry, which presents a problem in the field of medical imaging where usually only few labeled examples are available. In practice, to deal with small dataset, different well known regularization techniques are used to avoid overfitting. For instance, dropout is a regularization technique commonly used in deep neural network architectures to prevent co-adaptation between neurons [Srivastava et al. 2014a]. The key idea is to randomly drop units (along with their connections) from the neural network during training with the goal of generating an exponential number of different "thinned" networks. Other mechanisms such as L_{p} parameter norm or early stopping are usually used to reduce the model complexity. Therefore, we believe that deep learning is a relevant approach for radiomic study despite these challenges.

The objective of this thesis is to investigate an end-to-end deep frameworks that can jointly extract rich features and predict patient's outcomes on a small dataset. This thesis has three main contributions. First, we propose to go beyond classical radiomics based on handcrafted features by using deep radiomics. Our approaches can jointly learn characteristics and predict outcomes. Second, we propose a weakly supervised learning approach to segment automatically the lesion and then predict the patient's outcome based on the segmentation result. Finally, instead of doing segmentation and prediction separately and also to solve the overfitting problem when training complex models, we propose an architecture that includes segmentation, classification and prediction through multi-task

learning.

The work presented in this thesis was carried out in Becquerel cancer center with the Quantif-Litis team of the university of Rouen, and General Electric Healthcare. This thesis was financed in part by National Association for Research and Technology (ANRT).

Outline of the thesis. The manuscript is composed of two background chapters followed by three chapters each presenting one of our contribution as mentioned above.

— Chapter 1 starts by a general definition of cancer and the different tools used for diagnosis, treatment and follow up. Then, we present the principle of fluorodeoxyglucose (FDG) PET imaging, as well as its medical interest in oncology. Finally, we describe the first-order, second-order, and higher-order statistical features derived from medical images and their contribution in oncology.

— Chapter 2 introduces several machine learning paradigms covering supervised learning, weakly supervised learning and multi-task learning, which are the core of this thesis. Then, we present a review of the literature presenting the concept of radiomic using machine learning and deep learning algorithms, as well as current limitations.

— Chapter 3 presents our first contribution, which consists in the development of a deep radiomics framewrok based on 3D CNN to predict the response to treatment for patients with esophageal cancer. Our proposed method relies on two strategies to boost the prediction power of a CNN : *(i)* Develop a 3D CNN to extract 3D PET image features and predict the outcome *(ii)* Study the role of the volume of interest on the accuracy of the 3D CNN and other methods by using isotropic margins around the tumor volume, so as to reveal intra and peritumoral influence on the outcome prediction. We show experimentally that our approach allow us to achieve the best results compared to state-of-the-art methods.

— Chapter 4 is devoted to our second contribution which falls into the scope of weakly supervised learning where only little information is available for tumor segmentation. Image segmentation in 3D requires a lot of data and high computing power. In addition, the tumor is sometimes too small and included in a large 3D volume. We propose a method based on the principle of interpretability of a classification

network to detect the lesion. The originality of our contribution comes from the fact that we train a CNN to classify images into lung or esophageal cancer, whose ground truths are easy to obtain compared to manual segmentation of 3D images. Class activation maps which represent the areas of the lesion can be obtained at the same time. Then, using the segmentation results, we perform radiomic analysis to predict patient's outcome for esophageal cancer and survival for lung cancer. We show experimentally that the proposed method achieve state-of-the-art results for both segmentation and prediction tasks.

— Chapter 5 is dedicated to the presentation of our third contribution. Instead of training a *Neural Network* (NN) to do a segmentation as the physician's delineation, we let the NN decides which are the most peritumoral and intratumoral informative regions that boost the prediction performance. This is done through a MTL approach where the NN learns jointly the segmentation of the lesion and the outcome prediction. Since CNN needs a large dataset to learn useful representation and generalize to an unseen data, training a CNN on a small dataset presents the risk of overfitting. In order to overcome this limitation, we propose an approach based on parameter sharing. By adding subsidiary tasks such image reconstruction [Zeng 2010] and pathology classification, we have experimentally shown that the shared encoder works well for all four tasks because the number of grounth truths is increased. We evaluated our method on lung and esophageal cancer datasets and showed that MTL can improve the performance over a single learning approach. We also validated our method on a COVID-19 CT dataset.

— In chapter 6, we conclude our work and give some perspectives to improve it.

Chapitre 1

Medical imaging as a diagnostic and prediction tool

Sommaire

1.1 Cancer 6

1.1.1 Definition 6

1.1.2 Cancer staging 7

1.1.3 TNM Stage 8

1.1.4 Treatment 10

1.1.5 The WHO, RECIST and PERCIST criteria 13

1.2 FDG PET imaging : principle and characteristics 14

1.2.1 Principle of PET imaging 14

1.2.2 Fluoro-2-deoxy-D-glucose (FDG) 17

1.2.3 Standardized Uptake Value (SUV) 18

1.2.4 Features in FDG PET Imaging 19

1.3 Conclusion 25

1.1 Cancer

1.1.1 Definition

A healthy organism has vital functions (breathing, circulation, digestion,...) in good condition and balanced : a phenomenon called homeostasis. In this situation a healthy cell can divide by mitosis and give two daughter cells, clone of the mother cell. It frequently happens that a cell undergoes changes either in the nucleus or in its morphology. In the case of a deleterious mutation, either the cell commits suicide by apoptosis, or the cell dies causing inflammation, which is called necrosis. As long as these mechanisms are active, then the organism can remain healthy. However, a mutant cell may acquire the characteristic of multiplying in an uncontrolled manner, endangering the vital balance of the organism (Figure 1.1).

This disease characterized by abnormal cell proliferation within a living organism is cancer. Indeed, these cells increase in number, both by their important mitotic capacity linked to a loss of control of the cell cycle, but also by an insensitivity to apoptosis. There is also an anomaly in DNA repair. As these cells accumulate, a tumor can form in the target organ. Then, some primary tumors may progress to a more global invasion of the body, by escape of tumor cells : this is called metastasis. When organs are affected, they gradually lose their functionality, ultimately leading to death.

The 2 main categories of cancer are solid cancers and liquid/blood cancers. Solid tu-

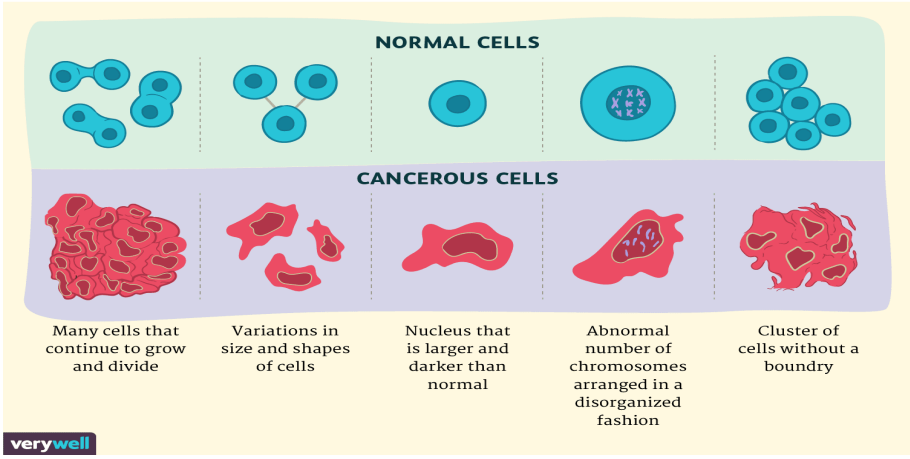


FIGURE 1.1 – Normal and cancerous cells : How Are they different. Source : verywellhealth

mors, such as carcinomas or sarcomas, are recognizable by a localized cluster of cells. They differ from blood cell cancers, such as leukemia or lymphoma, in which the cancer cells circulating in the blood or lymph are dispersed throughout the body.

In 2018, the number of new cancer cases in metropolitan France is estimated at 382,000. Between 2010 and 2018, the cancer incidence rate tends to stabilize in women; it is decreasing in men. The number of cancer deaths has been estimated at 157,400 (67,800 cancer deaths in women and 89,600 cancer deaths in men) [1]. Therefore, cancer care is a public health issue.

1.1.2 Cancer staging

For most of the cases where a person has cancer, physicians need to asses the stage of the disease. Staging is the process of determining how much cancer cells is within the body and its location. This step is crucial to determine the treatment. The stage of the cancer can also be used to predict the prognosis and the response to treatment. However, some cancers are not staged based on the spreading of the diseases, such as leukemia, which is a cancer of the blood cells and therefore the diseases have spread throughout the body by the time of the finding.

In order to asses the stage of the cancer, different techniques can be used. In many cases, the most reliable way to diagnose a person with cancer and to know the type of cancer it is, is by removing a small tissue called sample, and then analyze it under a microscope with the help of a pathologist. This operation is called biopsy. Blood tests can also be used to stage some type of cancer. Other exams such as endoscopy are sometimes used for the investigation and staging of cancer, for example, in the case of esophageal cancer. Staging can also be done using imaging tests such as CT, PET and MRI.

Generally, the stage of the cancer is determined at the time of diagnosis, however, this stage is usually updated later after the treatment and during the follow-up. When the staging is done based on physical exams such as medical imaging tests, endoscopy or biopsy before the treatment it is called clinical staging. When the stage is determined using a sample from a surgery given as first treatment it is called pathological stage. This stage

1. [Cancer In France /Edition 2019, National Institute of Cancer], November 2019, www.e-cancer.fr

may differ from the clinical stage, since it allows to determine more precisely the spread of the disease and may help also to predict treatment response and prognosis. In case of a recurrence, the staging is performed a second time in order to help guide decisions about the treatment. This is referred to as re-staging. An important thing to note here is that the new stage is added to the original stage, but it does not replace it. The first stage at the diagnosis level is the most important stage when performing statistics analysis or predicting the outcome.

1.1.3 TNM Stage

The most common and widely used system to asses the stage of solid cancer is the TNM stage. The overall stage in TNM is determined by investigating the 3 elements : tumor(T), node(N) and metastasis(M) as follow :

— T : the primary tumor

— N : if the cancer has spread to nearby lymph nodes

— M : if the cancer has spread to distant part of the body

This system provides physicians with important information about the size of the tumor, its location and whether or not it has spread. A letter or a number is assigned to each category to determine the spread of the disease. For the primary tumor category (T), the different sub-categories are :

— TX : no information about the tumor

— T0 : no evidence about the tumor

— T1 : the tumor invades the mucosa [2] or submucosa [3] (T1a and T1b)

— T2 : the tumor invades the muscularis [4]

— T3 : the tumor invades the adventitia [5]

— T4 : the tumor invades adjacent structures (other organs, ...)

For the lymph nodes category (N), the different sub-categories are :

2. A membrane that lines various cavities in the body and covers the surface of internal organs. Source : Wikipedia

3. A thin layer of tissue in various organs of the gastrointestinal, respiratory, and genitourinary tracts. Source : Wikipedia

4. Third layer of tissue in the colon.

5. The outer layer of fibrous connective tissue surrounding an organ

— NX : no information about the nearby lymph nodes

— N0 : no evidence of regional lymph node involvement

— N1 : spread to 1 or 2 neighboring lymph nodes

— N2 : spread to 3 to 6 neighboring lymph nodes

— N3 : spread to more than 7 neighboring lymph nodes

For the metastasis category (M), the different sub-categories are :

— M0 : no distance metastasis

— M1 : the cancer has metastasized, it has spread to another part of the body.

Once all these information are gathered and combined the TNM stage is defined (see table 1.1).

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

TABLEAU 1.1 – Regrouping of stages T (tumor), N (nodes) and M (metastatis) in a single stage TNM.

The cancer stage is a very important information that affect the treatment and also the patient's prognosis, along with the type of the cancer. The prognosis or survival rate is defined as the percentage of people with certain stage and type of cancer living after certain amount of time (usually 3- years), after being diagnosed. Survival rates are mainly based on the stage. There are indeed other factors that may affect the prognosis, such as the overall health of the patient, age and response to treatment. Finally, it should be noted that accurate cancer staging is difficult and complex due to the precision required to make accurate staging. Also, TNM system showed some limitations in the prediction of the response to treatment and survival in oncology [Huang and O'Sullivan 2017].

1.1.4 Treatment

Cancer is characterized by an inter- and intra-cellular heterogeneity. Due to this specificity, and the fact that different type of cancers are defined as different diseases, different type of treatments are proposed. The main classes of treatment are the surgery, chemotherapy, radiotherapy and immunotherapy. It can also be combined such as Chemoradiotherapy or surgery with Chemoradiotherapy.

Cancer surgery can be used for different purposes : to prevent, diagnose, stage and treatment. Surgery is used sometimes to diagnose cancer. When this procedure requires a surgery to take out a sample, it is called surgical biopsy. In the case of staging, it is done by examining the area around the tumor such as the lymph nodes and nearby organs, in order to determine how much the cancer has spread. Surgery for treatment is defined as the abduction of the whole or a part of the tumor. In the first case, it is called curative since it is given as the main treatment and the tumor is removed completely. In case where only a part of the tumor is removable, it can help other treatments to work better. In that case the operation is called debulking surgery.

Radiotherapy is the use of high doses of radiation to kill or damage DNA of cancer cells. It is used generally to treat some area of the body, and is called local since it treats or affects one part of the body. Cancer cells with irreparable DNA damage stop dividing or die. When the damaged cells die, they are destroyed and eliminated by the body. This process may take days or weeks, and even months before DNA is damaged enough for cancer cells to die. There are two main types of radiation : internal and external (Figure 1.2). The choice of the type of radiation depends on several factors such as the type of cancer, the size of the tumor, the tumor location and the overall health of the patient. Internal radiation therapy is a treatment where the source of radiation is put inside the body. The radiation source can be solid or dispersible. External radiation therapy is the most common used in radiotherapy. In this case, the source of radiation comes from a linear accelerator.

Chemotherapy is the use of drugs to treat a disease. In most of cases chemotherapy is used to imply drugs used for cancer treatment. It is considered as a systemic treatment,

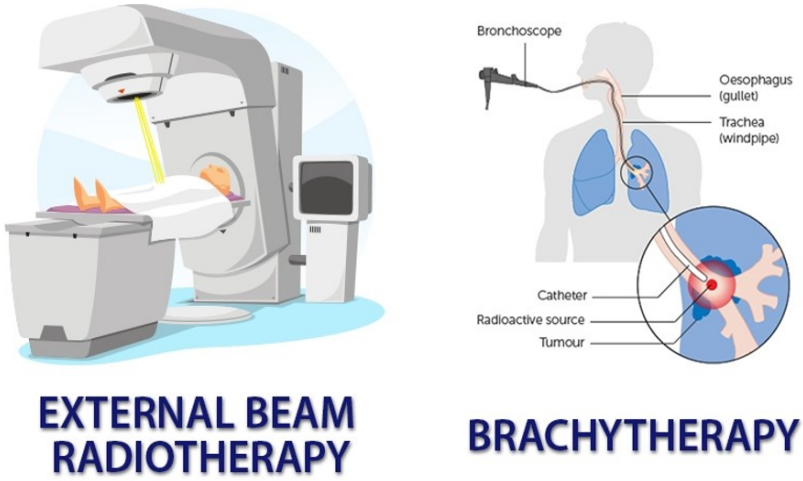


FIGURE 1.2 – Difference between external and internal (brachytherapy) radiation therapy. Source : Equicarehealth.

since the drugs travels through the body in the bloodstream to kill cancer cells. It allows to treat patients with metastasis that are far away from the primary tumor, which makes it different from surgery and radiotherapy. The oncologist decides on the doses, the way to administrate it, the frequency and the duration of treatment. Again, theses decisions are mainly based on the type and stage of the disease, with other factors such as the patient's age and overall health.

Radiotherapy and chemotherapy are given to cure, to control or as a palliation treatment. In case where cure is not possible, they are given to control the disease and stop the growing and spreading of the tumor, with the hope to decrease its size. In that case, the cancer is treated as a chronic illness. When curing and controlling is not possible, radiotherapy and/or chemotherapy can be given to ease symptoms and relieve patient from pain or pressure caused by a tumor so the patient feels better. This is called palliative treatment or palliation.

Finally, immunotherapy is based on the use of a person's own immune system to target and kill cancer cells. This can be done using different strategies. The first one is stimulation, or boosting the immune system to work its hardest or smartest. The second one is to make substances similar to the immune system of a person and using them to improve the immune system works in order to kill cancer cells. Recently, immunotherapy showed very promising results for different type of cancers such as bladder cancer [Fuge et al. 2015],

brain cancer [Jackson et al. 2014], breast cancer [Emens 2018], cervical cancer [Tewari and Monk 2014], leukemia [Beyar-Katz and Gill 2018] and others [Couzin-Frankel 2013].

Cancer treatment is an area of ongoing research. Current criteria to choose a treatment are based on stage, tumor localisation and clinical information. However, these criteria have shown some limitations, and recent studies revealed the need of more accurate information for the choose of treatment [Lambin et al. 2012b]. These criteria should include personal information about the person such as genes (genomics), proteins (proteomics) and images (radiomics). The incorporation of these sources of information will help to accurately identify the stage of the disease but also to personalize the treatment, which is called precision or personalized medicine (see Figure 1.3).

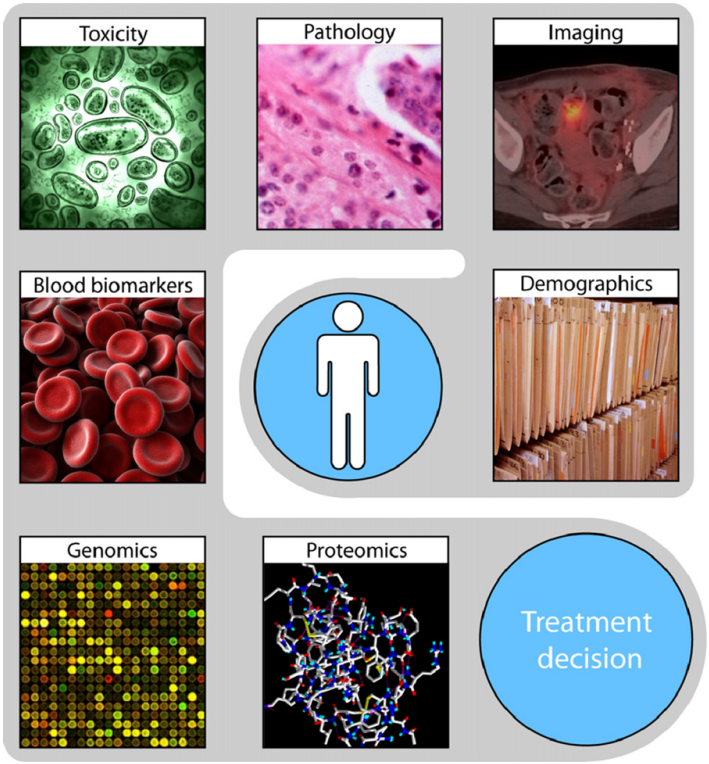


FIGURE 1.3 – Example of different information, including the radiomic information of the image contributing to a personalized treatment, according to [Lambin et al. 2012b]

During and at the end of the treatment the patient is monitored so that physicians could evaluate the efficiency of the treatment. Medical imaging plays a fundamental role in the follow up of the disease and re-staging of cancer.

1.1.5 The WHO, RECIST and PERCIST criteria

World Health Organization (WHO) introduced a standard measure to assess the response to treatment for solid tumors [Miller et al. 1981]. This standard categorize patients in 4 groups : Complete Response (CR), Partial Response (PR), Stable Disease (SD) and Progressive Disease (PD). This categorisation is based on several information including anatomical criteria for the evolution of the tumor at the end of treatment. However, this criteria have shown several limitations : it is not very robust to measurement bias and the maximum/minimum number of lesions to be considered in the evaluation is not specified. Also, the WHO criteria is subjective.

The response evaluation criteria in solid tumors (RECIST) [Eisenhauer et al. 2009, Therasse et al. 2005] was proposed to solve these issues. This criteria uses computed tomography scan to follow the development of the largest diameter of lesions over time. This measurement is made by assuming that the shape of the tumors is elliptical. A patient is considered to have CR if all lesions have disappeared and all affected lymph nodes are less than 10 mm in diameter. A patient is said as to have PR if the diameter of the lesions has decreased by 30% on average, PD if the diameter has increased by at least 20% and SD in other cases.

The PET response criteria in solid tumors (PERCIST) was introduced in 2009 by [Wahl et al. 2009]. The advantage of PERCIST is the use of metabolic quantitative information from a radiotracer binding intensity index corresponding to the average of the intensities within a 1 mL zone around the maximum intensity called "peak". This index is calculated and then compared between 2 successive exams, thus allowing the patients to be separated into 4 categories. A patient is considered to be in CR if all lesions have disappeared and the intensity of fixation on post-therapy imaging of the lesion is lower than that of a healthy reference zone (liver or aorta). A patient is defined as in PR if the variation of the "peak" between 2 examinations is -30% for the zone of highest intensity, PD if the "peak" is +30% and SD in other cases.

Given the great interest of PET in oncology [Lemarignier et al. 2014, Vera et al. 2014], from diagnostic to evaluation and follow-up [Ben-Haim and Ell 2008], we have chosen to

use PET modality with *2-[18]-Fluoro-2-desoxy-D-glucose* (FDG), to study treatment response and survival. In the following sections, we will present the principle of PET imaging with FDG, and the characteristics that can be derived from it. Next, we will show the value of PET imaging in oncology, and cover other advanced features that can be extracted from the images.

1.2 FDG PET imaging : principle and characteristics

1.2.1 Principle of PET imaging

PET (Figure 1.4) is one of the most widely used medical imaging techniques today to visualize the distribution of a tracer in an organism. Unlike anatomical imaging such as X-ray, CT, or MRI, functional imaging such as *Functional Magnetic Resonance Imaging* (FMRI) or PET allows the study of biochemical or physiological phenomena. In PET, the gamma-rays measured are emitted by annihilation of positrons released by an exogenous substance (tracer) in body. This is refereed to as imaging antimatter or annihilation of antimatter with matter [Morgan Jr and Hughes 1970]. The gamma-rays are emitted in all sort of directions. To look inside the set up of a PET scanner, it is composed of different elements such as the detector ring, a coincidence processing unit, the computer for image reconstruction, and the process placed at the heart of the object to be detected (see Figure 1.5).

In CT, absorption of the x-rays is the contrast generating parameter. The subject is irradiated with x-rays, and the different absorption allows to distinguish between the different tissues. While in PET, the absorption is, in principle, undesirable. Indeed, the absorption of gamma-rays in PET is considered as a nuisance effect. Figure 1.6 shows an unstable parent nucleus with in red the neutrons, and in blue, the positrons.

When this unstable nucleus decays, there is a conversion of a positron into a neutron. Thus, the charge changes by minus one. With a different charge and with the charge conservation law, this positive charge will be emitted from the nucleus in the form of a positron. This positron will then diffuses through the tissues, undergoes various interac-

FIGURE 1.4 – Discovery IQ PET/CT scanner image courtesy of GE Healthcare.

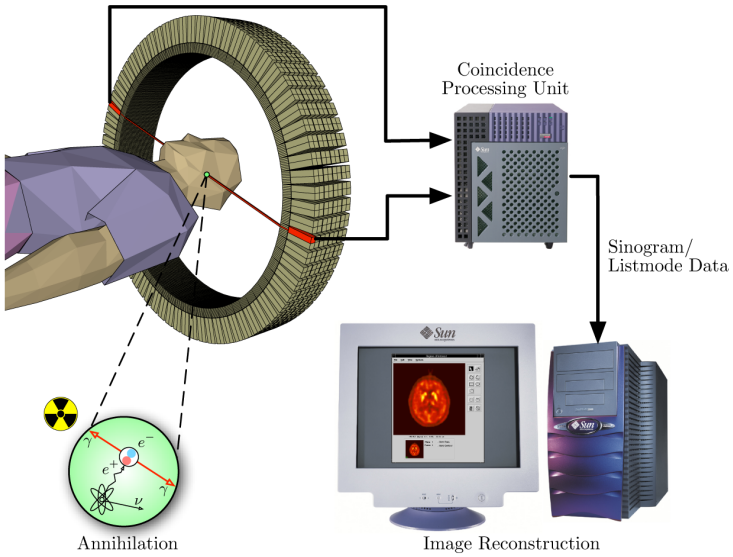


FIGURE 1.5 – Diagram of the PET scan acquisition process. Source : Wikipedia.

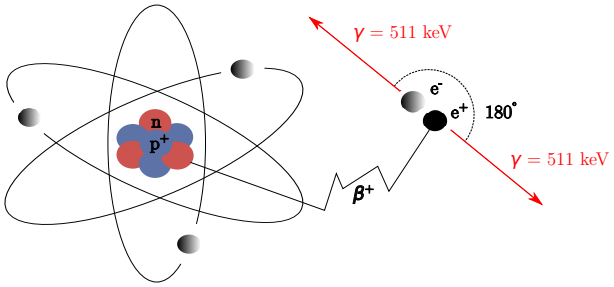


FIGURE 1.6 – After a short distance, the positron e^+ obtained by emission β^+ is annihilated with an electron e^- giving birth to two photons γ emitted in the same direction, in opposite direction at 180° from each other and with an energy of 511 keV each.

tions, and akin to the electrostatic Coulomb interaction. After this, positron has lost most of its kinetic energy, at some point, it will be attracted by a nearby electron. They combine, it is matter and antimatter, and when matter and antimatter meet, they are annihilated, so there is the emission of two gamma-rays. These gamma-rays travel at 3×10^8 meters per second. Thus, in 3 nanoseconds, the gamma-ray will have traversed 1 meter in the scanner. This essentially means that in order to determine the location of an annihilation event, the detector must detect two events that occurred simultaneously, i.e. two gamma-rays at the same time. This will indicate that at some point, there was an electron–positron annihilation. This is the basic principle of detection. However, in practice, the detection of the two gamma-rays does not happen to be always simultaneously. There are several cases :

— True coincidence (Figure 1.7a) two photons that are being sent off in opposite directions and detected simultaneously.

— Random coincidences (Figure 1.7b) coincidences where two positron electron annihilation processes happened simultaneously. Of these four photons that are being produced, two of them are lost.

— Scattered coincidence.(Figure 1.7c) when an annihilation occurs and one of the photons is being Compton scattered in the tissue, so it's being deviated by a certain angle.

— Multiple events (Figure 1.7d) when more then 2 events are detected simultaneously.

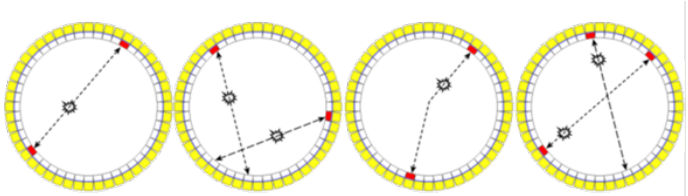


FIGURE 1.7 – Types of coincidences recorded by the detection system : (a) true coincidence, (b) diffuse coincidence, (c) fortuitous coincidence and (d) multiple coincidence.

Finally, standardization must be added to this process. Once all the events are detected, standardization must also be taken into account. Standardization is essentially a process that takes into account the imperfection of the scanner.

The detected signals are stored in a 2D matrix called sinogram. A sinogram allows to

describe the projection detected around the patient from a certain angle ϕ (Figure 1.8). Each line of the sinogram represents the number of events detected in all the parallel response lines forming the same angle with respect to the tomograph axis. Then the reconstruction of the 3D images is done using analytical or iterative methods. PET machines now-days are coupled with CT to add an anatomical information, which helps physicians in image interpretation, and also for attenuation correction [Kinahan et al. 1998]. More recently, MRI also can be coupled with PET images [Wagenknecht et al. 2013].

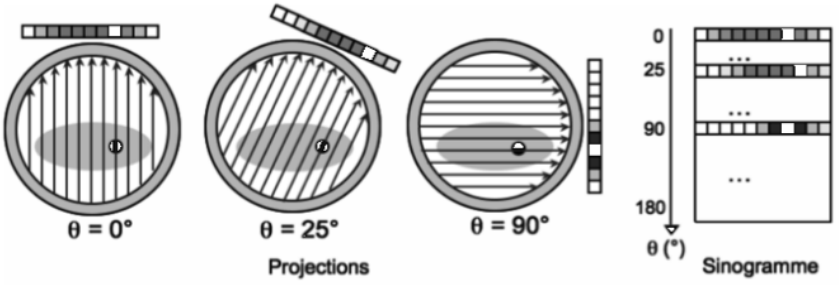


FIGURE 1.8 – Example of the creation of three lines of the sinogram according to the projection angle. Source : Tylski

For more details on how PET works as well as CT reconstruction, the reader can refer to [Das B K. and Das 2015].

1.2.2 Fluoro-2-deoxy-D-glucose (FDG)

The most widely used tracer for PET is FDG. It's the glucose where there's a fluorine attached at the two position (Figure 1.9). Instead of OH group, there's a fluorine attached. Fluorine 18 is an unstable nucleus that decays, and as it decays it emits a positron. The radioactive half-life is 110 minutes with an energy of 0.64 MeV.

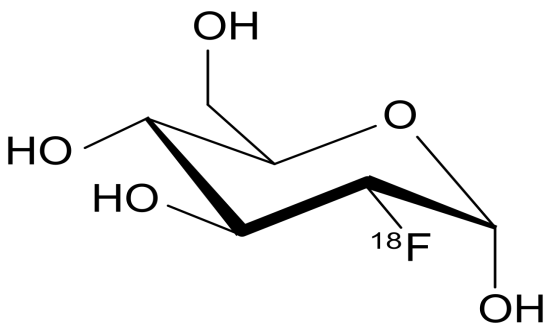
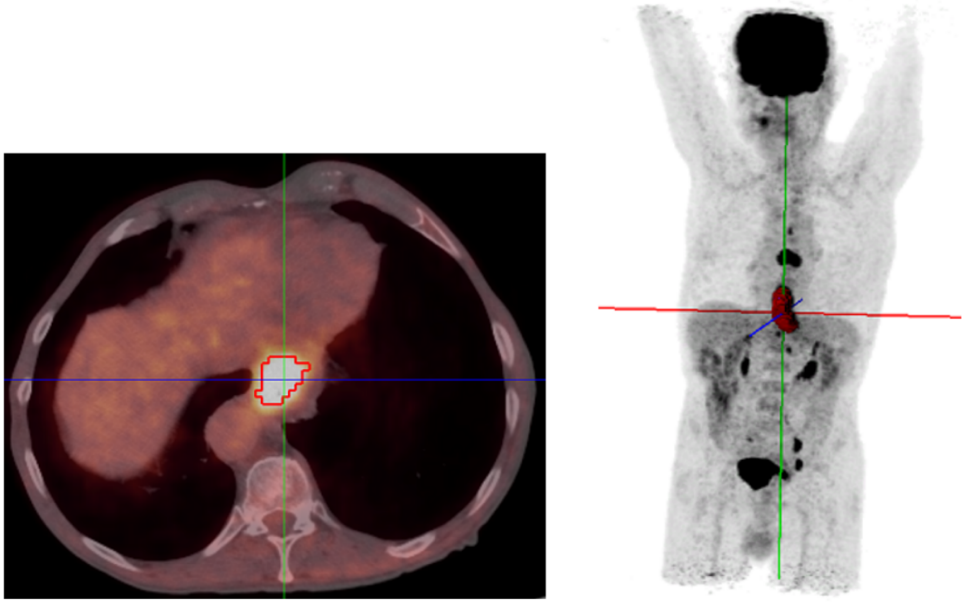


FIGURE 1.9 – Stereo skeletal formula of fluorodeoxyglucose (18F). Source : Wikipedia.

The accumulation of 18F-FDG visible on the PET images highlights the abnormally high carbohydrate metabolism of tumors. Thus, PET imaging is routinely used as an aid in the diagnosis and staging of many cancers. Due to its functional property, it is a powerful tool for diagnosing the malignancy of a lesion 4.8.



(a) (b)

FIGURE 1.10 – (a) Cross-section PET to FDG and (b) MIP of a patient with an esophageal cancer. The tumor appears stained on the MIP. Other organs with normal FDG fixation are : the brain due to its permanent activation, the kidneys and bladder for their filtration role.

1.2.3 Standardized Uptake Value (SUV)

"Standardized Uptake Value" (SUV) [Woodard et al. 1975] was introduced as a simple means to measure the absolute metabolic activity using PET. Tracer fixation in tissues depends in particular on the injected dose and the blood volume in which the activity is distributed. A simple way to normalize the measured uptake is therefore to apply the following formula 1.1 :

$$\mathrm{SUV_{BW}} = \frac{\text{Activity concentration}\left(\frac{\mathrm{kBq}}{\mathrm{mL}}\right)}{\text{Injected dose(kBq)/Body weight(g)}} \tag{1.1}$$

This weight-standardized definition is usually given without unit. The fixation expressed in Bq/mL, corresponds to the image quantified in an absolute way. So in summary, an SUV value equal to 1 means that the tracer concentration corresponds to the average

concentration in the patient. A value of 10 in a lesion means that 18F-FDG binding is 10 times greater than the uniform distribution of the tracer.

To interpret a fixation, it is necessary to normalize it by this quantity. It is a passage from relative quantification (recorded events / voxel) to absolute qunatification (Kbq/mL). The ability to measure the SUV value of a lesion has the advantage of normalizing the images so as to compare the intensity of the fixation for patients who have received different activities relative to their weight. It also makes it possible to evaluate the therapeutic response and to monitor the patient.

1.2.4 Features in FDG PET Imaging

Identifying new non-invasive approaches to predict a patient's response to treatment has the potential to significantly improve clinical outcome. As shown above, several studies have reported that the amount of FDG in initial PET images of the tumor can provide predictive power [Hatt et al. 2011, Javeri et al. 2009, Rizk et al. 2009]. In addition to the SUV_{max}, other measures were derived from the SUV 1.11 such as :

— SUV_{mean} : is the mean value of the SUV in a defined metabolic volume.

— SUV_{min} : represents the SUV in the voxel of activity with the lowest value.

— SUV_{peak} [Wahl et al. 2009] : is the average SUV of the voxels contained in a parallelepiped volume of interest with three voxels on each side and whose position is chosen such that the average SUV of the voxels contained in the volume of interest is as high as possible. Its center belongs to the segmented metabolic volume.

— TLG (Total Lesion Glycolysis) [Larson et al. 1999] : is defined as the product of the mean SUV with the metabolic tumor volume measured in the tumor.

— MV (Metabolic volume) : is defined as the total volume in cubic centimeters (cc) of the tumor.

Recently, a new approach is of increasing interest in PET imaging, namely the characterization of intra-tumor heterogeneity of radiotracer uptake. This approach consists in extracting image characteristics based on classical 1^{st} and 2^{nd} order statistical methods.

For instance, texture is very important in the analysis of images, this is due to its presence in the vast majority of images, thus a large number of methods for its analysis have

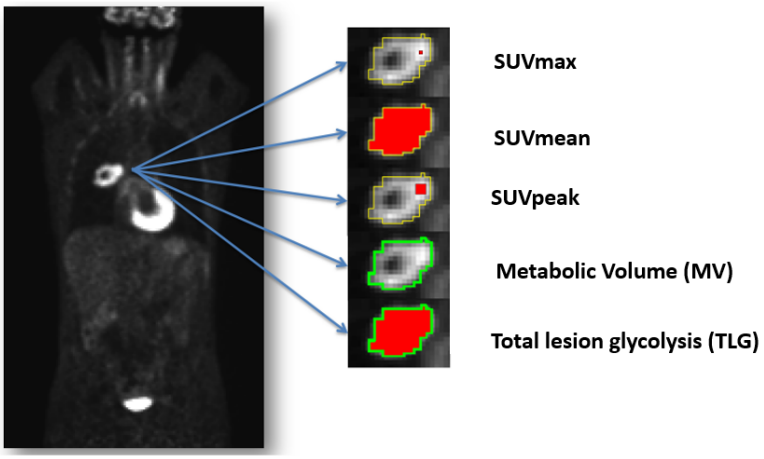


FIGURE 1.11 – Representation of SUV derivatives of a lesion. Source : Orlhac.

been developed. There are several definitions of texture in the field of image processing, the most common one defines texture as a region in an image, where a set of pixels that have a spatial relationship between them. Texture can appear in different symmetrical, recurring, dynamic forms ... etc.

Work on texture has given rise to several approaches to characterize it and thus, to recognize the texture in an image, we find in the literature two main approaches.

1.2.4.1 Structural approach

It is based on two main elements : the primitives used and the spatial relations that link them together. The methods of this approach are mainly based on signal processing, topography and geometry. Their strong point is that they can be used with classical segmentation methods such as Edge-Detection. Some methods consist in finding the texels (basic component of a texture) then using heuristics to find the positioning rules, other so-called syntactic methods use language theory to generate the texture by applying production rules, moreover a texture can be generated by several grammars. The results of this approach are more used in texture synthesis than in texture analysis.

1.2.4.2 Statistical approach

First-order statistics :

It is a method based on the distribution of pixels without taking into account the relationships between them. The means used to represent the distribution is the histogram, a

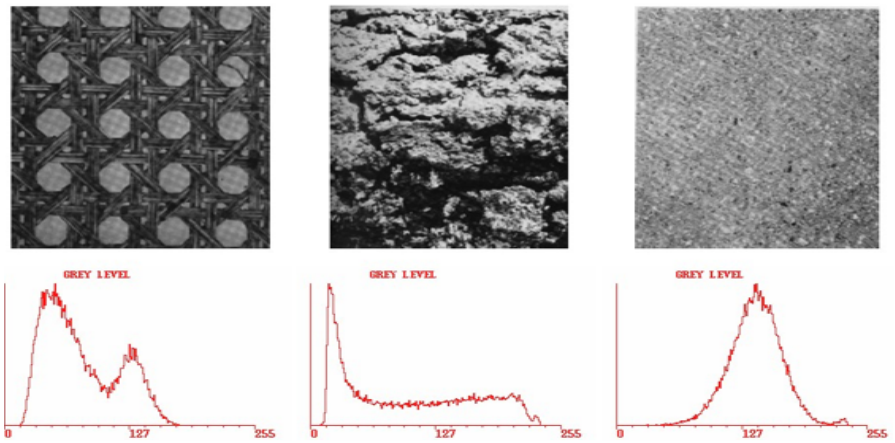


FIGURE 1.12 – Images and corresponding histograms

TABLEAU 1.2 – Different properties that can be calculated using the histogram.

Property	Formula
Mean	$\mu = \sum_{i}^{G-1} ih(i)$ (1.4)
Variance	$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 h(i)$ (1.5)
Skewness	$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 h(i)$ (1.6)
Kurtosis	$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 h(i) - 3$ (1.7)
Energy	$\mu_5 = \sum_{i}^{G-1} h(i)^2$ (1.8)
Entropy	$\mu_6 = \sum_{i}^{G-1} h(i) log_2 h(i)$ (1.9)

graph allowing to study the distribution of a variable, the X axis will represent the different gray level values, the Y axis will represent the number of occurrences 1.12.

The creation of a histogram is done with the following function :

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \sigma(f(x,y),i) \tag{1.2}$$

where σ is 1 when f(x,y) = i, 0 otherwise. Then, the final equation is the following :

$$H(i) = \frac{h(i)}{N * M} \tag{1.3}$$

After the calculation of the histogram, different properties can be calculated as shown in table 1.2.

Second-order statistics :

First order statistics are limited since there are no relation kept between the different voxels. For further statistical analysis, other methods have been proposed to account for this voxel relationship, such as textural analysis. The resulted characteristics are of second order or higher since relationships between neighboring elements 2 x 2 or more are kept. There are four main texture matrices proposed in the literature :

— Co-occurrence matrix (*"Gray Level Cooccurrence Matrix"* (GLCM)) [Haralick et al. 1973] : The co-occurrence matrix is a matrix of dimension N*N where N is the number of gray level values, each cell of the matrix C (i, j) represents the number of occurrences of pixels i and j according to a relation of distance and orientation. So we obtain a co-occurrence matrix for each distance and orientation (See figure **??**). For each matrix we can calculate 14 properties of which the most important are shown in table 1.3.

— *"Gray Level Difference Matrix"* (GLDM) [Amadasun and King 1989] : It describes differences in intensity between neighbors and contains statistics and contains statistics of a higher order than the previous matrix.

— (*"Gray Level Run Length Matrix"* (GLRLM)) [Galloway 1975] : characterize the length of ranges of the same intensity in a given direction.

— (*"Gray Level Size Zone Matrix"* (GLSZM)) [Thibault et al. 2009] : It gives the length of the zones having the same intensity in all directions simultaneously.

Table 1.3 summarizes the most important characteristics that can be obtained from these matrices.

The most used method nowadays for handcrafted feature extraction is based on first and second order analysis and texture analysis. More recently, to describe 18-FDG uptake heterogeneity in a lesion, other characteristics have been proposed. For instance, [Bundschuh et al. 2014] found that *Coefficient Of Variation* (VOC) is an important predictive factor for patients with rectal cancer. El Naqa et al. proposed the extraction of characteristics from the SUV-Volume Histogram [El Naqa et al. 2009], such as SUVx (minimum SUV for the highest x% SUV) and Vx (Percentage of volume with at least x% SUV). Moreover, in the same paper, El Naqa et al. found that features extracted from the GLCM [Hara-

TABLEAU 1.3 – Different properties that can be calculated using the Co-occurence matrix.

Property	Formula
Mean	mean = $\sum_x \sum_y p(x,y)$ (1.10)
Variance	var = $\sum_x \sum_y (i - mean)^2 p(x,y)$ (1.11)
Energy	energy = $\sum_x \sum_y (p(x,y))^2$ (1.12)
Contrast	contrast = $\sum_x \sum_y (p(x,y))^2 * p(x,y)$ (1.13)
Entropy	entropy = $\sum_x \sum_y p(x,y) log(p(x,y))$ (1.14)

TABLEAU 1.4 – Main statistical characteristics of 2nd and higher order

Matrices	**2nd order characteristics and more**
Matrices de cooccurrences (GLCM)	Variance, Energie, Entropy, Correlation, Dissimilarity, Contrast, Homogeneity, Moment differential inverse (IDM), "Cluster shade", "Cluster tendency"
Gray level difference matrix (GLDM)	"Coarseness", "Contrast", "Busyness", "Complexity", "Strength"
Matrices des longueurs de plages homogènes (GLRLM)	"Short Run Emphasis" SRE, "Long Run Emphasis" (LRE), "Low Gray level Run Emphasis" (LGRE), "High Gray-level Run Emphasis" (HGRE), "Short Run Low Gray-level Emphasis" (SRLGE), "Long Run Low Gray-level Emphasis" (LRLGE), "Short Run High Gray-level Emphasis" (SRHGE), "Long Run High Gray-level Emphasis" (LRHGE), "Run Percentage" (RPr), "Gray Level Non-Uniformity" (GLNUr), "Run Length Non-Uniformity" (RLNU)
Matrice des longueurs de zones homogènes (GLSZM)	"Short Zone Emphasis" (SZE), "Long Zone Emphasis" (LZE), "Low Gray level Zone Emphasis" (LGZE), "High Gray-level Zone Emphasis" (HGZE), "Short Zone Low Gray-level Emphasis (SZLGE), "Long Zone Low Gray-level Emphasis" (LZLGE), "Short Zone High Gray-level Emphasis" (SZHGE), "Long Zone HighGray-level Emphasis" (LZHGE), "Zone Percentage" (ZP), "Gray Level Non-Uniformity" (GLNUz), "Zone Length Non-Uniformity" (ZLNU)

lick et al. 1973], which characterizes intensity ratios between pairs of neighboring pixels, are among the most important features in cervical cancer prediction. Other texture matrices have also been proposed in the literature, such as the GLDM [Amadasun and King 1989] which characterizes the intensity difference between neighbors, the GLRLM [Galloway 1975] and the GLSZM [Thibault et al. 2009] characterizing the intensity size ranges in one direction or in all directions, respectively. In the end, it is possible to extract different characteristics per matrix, which leads to a large number to be processed. [Tixier et al. 2011a] studied the importance of having a large number of PET image features in esophageal cancer using *"Receiver Operating Characteristic"* (ROC) curves measuring *"Area Under ROC Curves"* (AUC). Among 38 characteristics, they found that GLCM characteristics (entropy, local contrast, correlation, second angular momentum, homogeneity and dissimilarity) and GLSZM (ZLNU, GLNUz) characteristics are relevant for predicting patient response to chemo-radiotherapy.

1.3 Conclusion

In this chapter, we presented the interest of the functional imaging PET with FDG for cancer care. We also covered the interest of this modality in therapeutic follow-up, as well as the prediction of treatment response and survival. The SUV_{max} is considered to be the first feature that allowed such prediction to be made prior to treatment. Numerous characteristics emerged later in an abundant literature, proposing features based on 1^{st} order, 2^{nd} order and higher order statistics.

Since many features can be extracted from the images, *Machine Learning* (ML) is the most relevant technique to take into account all the features together. In a classic scheme, a ML algorithm is first used to select the most relevant features for prediction, and then another or the same ML algorithm is applied to the selected features to predict patient outcome.

In the next chapter, we will draw up a state of the art in ML algorithms and the process of predicting patient survival using radiomic features. We will also discuss different ML algorithms, as well as different radiomics framework.

Chapitre 2

Radiomics and machine learning

Sommaire

2.1 Introduction 28

2.2 Basic notions in Machine learning 29

2.2.1 Principe of Learning 29

2.2.2 Capacity, Overfitting and Underfitting 32

2.2.3 Hyperparameters and Validation Sets 32

2.2.4 Cross-Validation 33

2.2.5 Weakly supervised learning 33

2.2.6 Multitask learning 34

2.2.7 Artificial Neural Networks 34

2.2.8 Multilayer perceptron (MLP) 36

2.2.9 Neural Network training 36

2.2.10 Convolutional Neural Networks 37

2.2.11 Interpretabilty 39

2.2.12 Conclusion 40

2.3 Radiomics 41

2.3.1 Concept and principle 41

2.4 Machine learning for radiomics 42

2.5 Deep learning for radiomics 47

2.6 Objectives of the thesis 48

> We are drowning in information and
> starving for knowledge.
>
> John Naisbitt

2.1 Introduction

Today we live in a world where data is available in immense quantities, to the point where it is becoming the new oil [Hirsch 2013]. Conventional methods for manipulating these vast amounts of data and mining knowledge are becoming very limited, hence the interest in developing and using new adapted methods. ML, a subclass of *Artificial Intelligence* (AI), is a paradigm in which the methods developed make use of this data to uncover a pattern in order to predict future data or outcomes. AI is defined as a program that mimic the human intelligence. ML is a branch of AI where designed algorithms improve their performances with experience. DL is a class of ML that uses *Artificial Neural Network* (ANN) with representation learning to progressively extract higher level features from the raw input (see Figure 2.1).

Two components are essential in machine learning : learning (training) and testing (inference). Learning requires the availability of a dataset in order to uncover a pattern, which will help later during the inference to make a decision on a new, unseen samples. Most of machine learning algorithms rely on handcrafted features instead of raw input. This process requires domain knowledge to extract manually meaningful features, and then passed to a ML algorithm in order to learn. For inference, the same features are extracted to be used by the algorithm. DL proposes to replace this framework with an end-to-end model that extract features and perform prediction at the same time. DL is composed of several layers : the first ones tend to learn low level representations while the latter ones high level features (Figure 2.7).

In this chapter, we will present first basic notion in machine learning, and then a state of the art works in radiomics with ML. Finally, we will show the research directions that we are going to carry out in this thesis.

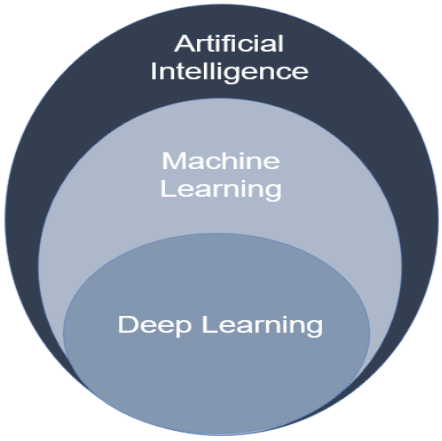


FIGURE 2.1 – Artificial intelligence, machine learning and deep learning.

2.2 Basic notions in Machine learning

2.2.1 Principe of Learning

Learning from data is what defines a machine learning algorithm. Mitchell provides a formal definition to learning : "A computer program is said to learn from Experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [Mitchell et al. 1997]. In this thesis, we will focus on the classification task with supervised learning experience, weakly supervised learning and multi-task learning.

The Task, *T* :

If you ask a person if a tree is in a picture, he can answer very quickly. If the same person is asked what a tree is, he may not answer with great accuracy to describe a tree. ML algorithms allow to solve this problem by learning from the images the features of the tree and decide on the type of tree, which is called learning from data. Thus, ML enables to address problems that are hard to be solved by a written fixed program. The task is what the ML algorithm will learn to do. For instance, if we want the algorithm to learn to predict the survival of a patient with lung cancer after treatment, the task is survival prediction.

The Performance, *P* :

Performance is a way to measure the efficiency of an algorithm. For the classification, we usually measure the accuracy of the learned model. Accuracy is the proportion of data that have been correctly classified. This measure provides a simple yet very useful information on how the model is doing on classification. However, sometimes one measure is limited. For instance, in a binary classification task, if 99% of the data belong to class 0 while only 1% belong to class 1, a simple model that attribute the class 0 to all data will achieve an accuracy of 99% while the class 1 is completely ignored. In this case, other measures could be added to measure the efficiency, such as sensitivity (Sens) and specificity (Spec) :

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

where TP is the true positives, FN is the false negatives. In our binary example, TP + FN is the number of data points classified as 1.

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.2}$$

where TN is the true negatives, FP is the false positives. In our binary example, TN + FP is the number of data points classified as 0.

In that case the accuracy (ACC) could be defined as :

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{2.3}$$

Similarly, we can calculate the error rate, which is the proportion of misclassified data. The objective of the learning algorithm is to minimize the error rate, called also the cost function. This cost function and performance are calculated on the training dataset, but in practice we are interested on how well the model is doing on a complete new unseen dataset, which is called a test set. The choice of the performance metric depends on the task to be learned.

The Experience, *E* :

In machine learning, different approaches can be used to learn from unstructured data (see Figure 2.2). Two main classes are usually presented : supervised learning and unsupervised learning. In a supervised learning approach, the goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs $D = (X_i, Y_i)_{i=1}{}^N$. D is the training set and N is the number of training examples. The outputs y could be categorical or nominal such as healthy or pathological patient, or a real-valued scalar such as survival time in days. In the first case it is called a classification problem, and in the second one it is known as regression problem.

The second type of machine learning is unsupervised learning. In this case, only inputs x are given, $D = (X_i)_{i=1}{}^N$, and the goal is to uncover a pattern or a relation between x's in the data. This approach is also refereed to as knowledge discovery. In this work we are interested mainly in the supervised learning approach.

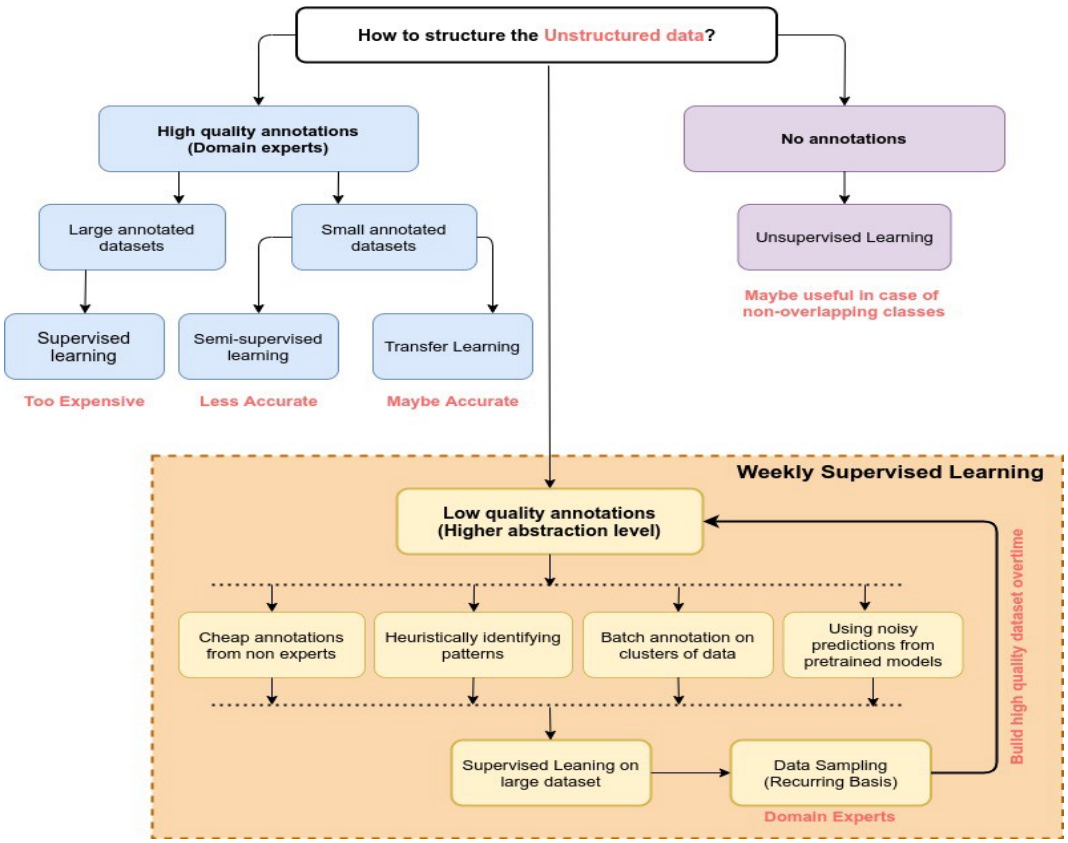


FIGURE 2.2 – Various categories of approaches for structuring the unstructured information. Source : towardsdatascience.

2.2.2 Capacity, Overfitting and Underfitting

The capacity of a model is defined as its generalization to perform well according to a performance metric P on a new unseen data set, the test set. The error rate on the training set is known as the training error, and on the test set as the generalization error or test error. Among the main differences between optimization and ML is that in ML not only we desire to minimize the training error but also the generalization error. Thus, finding a trade-off between the two measures is necessary, which is a challenging problem in ML known as underfitting and overfitting. The underfitting occurs when the model is not able to fit the training dataset, which results in a high training error. Overfitting occurs when the model does fit very well the training set, but the gap between the training error and the test error is very high. One of the main factors influencing underfitting and overfitting is the capacity of the model. It is defined as the set of functions the model can fit, called hypothesis space. Models with low capacity tend to underfit. Models with high capacity may overfit the training set, thus perform poorly on the test test. One way to address this problem is by limiting the hypothesis space that a model can explore, thus, reducing the capacity of the learning algorithm.

2.2.3 Hyperparameters and Validation Sets

As mentioned above, the capacity of the model influences the exploitation of a hypothesis space. This capacity in fact depends on a set of hyperparameter of the algorithm. In general, a hyperparameter is a parameter used to control the behavior of the algorithm. The values of hyperparameters are not adapted through learning (though a field called meta-learning in which a second algorithm learns the optimal hyperparameters for another algorithm does exist). If the hyperparameters are chosen based on the training set, the setting will be optimized to minimize the training error, resulting in overfitting. To solve this problem, another set called validation set which is constructed from the training set is used to evaluate the model. In practice, we split the training set into two subset, the first one for training ($\approx$ %80 of the data), and the second one for validation ($\approx$ %20). The validation set is used to estimate the generalization error during training, and allows to

update the hyperparameters accordingly. It should be noted here that the test set is used only to measure the performance of the final model after training and hyperparameters optimization.

2.2.4 Cross-Validation

The availability of a large database, especially in the medical field, is not always evident. Thus, dividing a small dataset into training and test may not be the best approach to measure the performance of a model. A small test set makes it hard to compare two algorithms due to the uncertainty around the test error. In that case, an alternative procedure called k-fold cross-validation can be applied. This procedure consists of dividing the dataset into k subsets called folds, where k-1 are used in training and one in test. The process is repeated k times. Different test set can then be chosen to measure the performance of the model when using the other k-1 folds for training. The final measure can be obtained as the mean of the k performances.

2.2.5 Weakly supervised learning

In a WSL, only few labels are available (see Figure 2.2). Theses labels are used to retrieve a signal that labels a large amount of data. ML algorithms are well known for their data hungry nature, although in many of real life problems, such as in medical imaging, having a sufficient quantity of labeled data may be difficult due to the need for an expert to label the data manually and such a task is time consuming. Three well known types of weak labels are usually presented :

— Imprecise or inexact labels : it is based on the definition of heuristics based on experts workflow to label the dataset, defining the expected distributions, or by imposing constraints on the training data [Cabannnes et al. 2020, Ratner et al. 2016, Zhou 2018].

— Inaccurate labels : based on non-experts to label the data, which results in a low quality annotations [Zhou 2018]

— Existing resources : such as knowledge bases or pre-trained models to label the

data for a certain task that may be helpful, but not perfectly suited for the given task [Ratner et al. 2019, Zhou 2018]

The advantage of this method is the possibility to increase the size of the database without worrying about labels. Indeed, in a supervised learning approach, the whole dataset should be labeled, which limits the use of large databases that are available but not annotated, as is the case in medical imaging.

2.2.6 Multitask learning

The standard method in machine learning is to learn one task at a time. Large problems are broken into small sub-problems that are learned separately and then recombined. MTL [Caruana 1997] is a type of learning algorithm that aims to combine several pieces of information from different tasks in order to improve the model's performance and its ability to better generalise [Zhang and Yang 2017]. The basic idea of MTL is that different tasks can share a representation of common characteristics [Zhang and Yang 2017], and thus train them jointly. The use of different data sets from different tasks allows learning an efficient representation of the common characteristics of all tasks, because all data sets are used to obtain it, even if each task has a small data set, thus improving the performance of each task.

2.2.7 Artificial Neural Networks

ANNs or NNs for simplicity, are a computing system inspired roughly by biological neural networks (see Figure 2.4). It consists of a number of units that receive an information, process it and send it to the next units. The simple component of a NN is perceptron [Rosenblatt 1958]. A perceptron represents a single neuron. It is a simple function with a linear parameter with respect to its input, as represented by the following formula :

$$f(x) = \phi(x \cdot w + b) = \phi(\sum_{i=1}^{D} x_i w_i + b) \tag{2.4}$$

where $x \in \mathbb{R}^D$ is an input vector, w is a vector of parameters known as weights. b is a scalar parameter known as bias. ϕ is an activation function. Given an input node, ϕ

outputs a value to decide if a neuron should contribute or not in the neural network, and on how should it contribute. The most popular activation functions used in NNs are nonlinear (see figure 2.3), thus the model could capture high representations using small nodes.

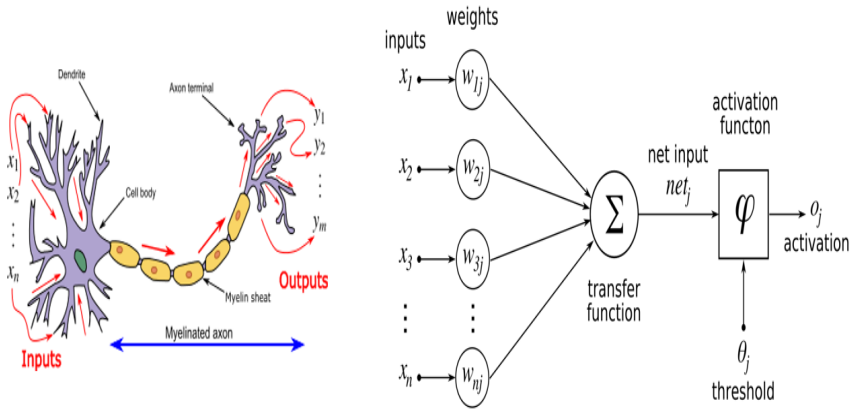


FIGURE 2.3 – Similarities between biological neuron (left) and artificial neuron (right). Source : Wikimedia.

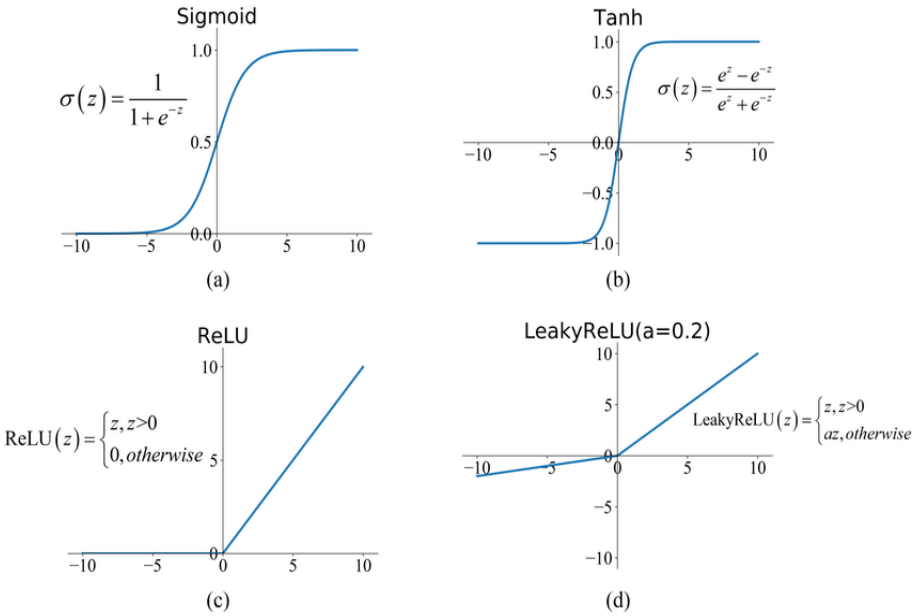


FIGURE 2.4 – Popular nonlinear activation functions used for NN training. Source : [Feng et al. 2019]

2.2.8 Multilayer perceptron (MLP)

Multi-Layer Perceptron (MLP) or deep feedforward networks are the core of deep learning models. A feedforward network can be represented as a function approximation where $y = f(x)$ and f is unknown. In classification, the goal of MLP is to estimate f using a labeled training set $D = (X_i, Y_i)_{i=1}^{N}$, and then to use this approximated function h to estimate $\hat{y}$ using inputs x. We name the approximated function h the hypothesis function. A MLP defines a mapping as $y = h(x, \theta)$, where θ are the parameters to learn that results in the best approximation function. It is composed by an input layer, followed by one or more hidden layers and an output layer (see figure 2.5).

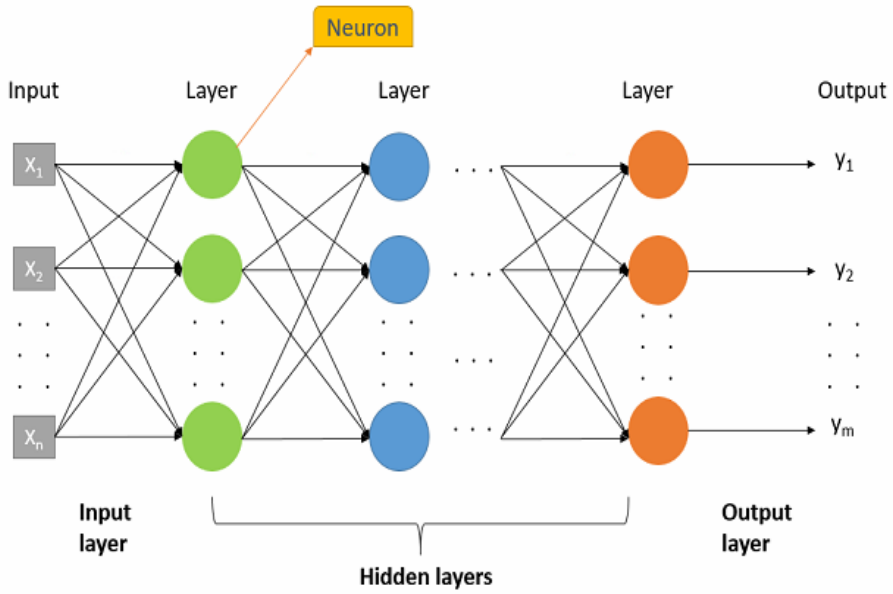


FIGURE 2.5 – An example of a multi-layer perceptron.

2.2.9 Neural Network training

The most widely used algorithm to train neural network is Back-propagation [Rumelhart et al. 1986]. After a feed-forward pass, the NN predicts an output $\hat{y}$ for each input x. The $\hat{y}_s$ are compared then to the expected output $\hat{y}$ via a cost function, which gives us an idea about the model performance. The error is propagated into the network from output to input layer via back propagation. During this process, the weights and the biases of the models are updated in order to minimize the cost (loss) function (see figure 2.6).

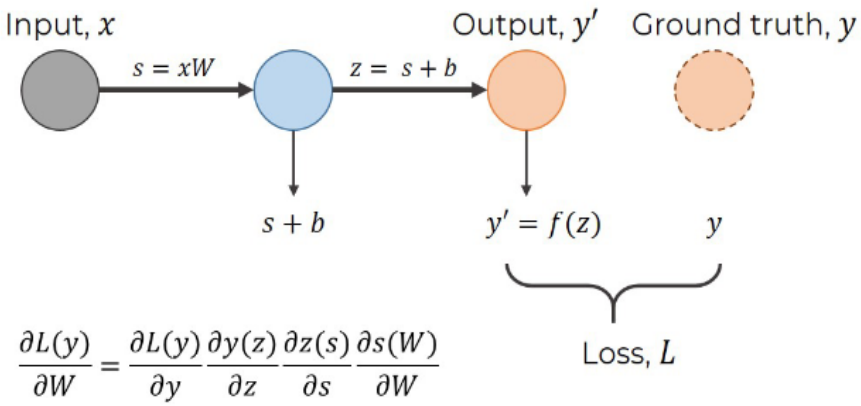


FIGURE 2.6 – An example of back propagation. By updating the weight w and the bias b through back propagation, the model minimize the loss L by approximating the ground truth y. Source : Javaid.

2.2.10 Convolutional Neural Networks

In 1980, Fukushima introduced a hierarchical multilayer neural network called neocognitron [Fukushima 1980], which is considered as the originial convolutional neural network CNN. Neocognitron was used in several applications such as handwritten character recognition and other pattern recognition tasks. The architecture was inspired from the work of Hubel & Wiesel [Hubel and Wiesel 1959], where they found two types of cells in the visual primary cortex. They have shown that first layers in the neural network tends to learn simple patterns, using simple cells, while advanced layers tend to learn more abstract patterns, using complex cells.

CNN was introduced and become known by LeCun [LeCun et al. 1998]. It is usually composed of convolutional layers followed by pooling layers, then a multi-layer perceptron. Convolutional layers are based on a convolution operation, where a kernel is used to convolve the image. A no linear activation function is applied then to the resulted image. This operation is usually followed by a pooling layer. Different methods exist for the pooling operation : max pooling which consists on keeping the maximum value within a region, or average pooling which return the average of values within a region. The size of the kernel for the convolution may differ from a layer to another by increasing the depth generally. At the end of the last convolutional layer, the raw input become small in width and high but bigger in depth. Then, a flatten operation that consists on putting the result

of the last convolutional layer into a 1 big dimensional tensor. A MLP is finally used to make a decision.

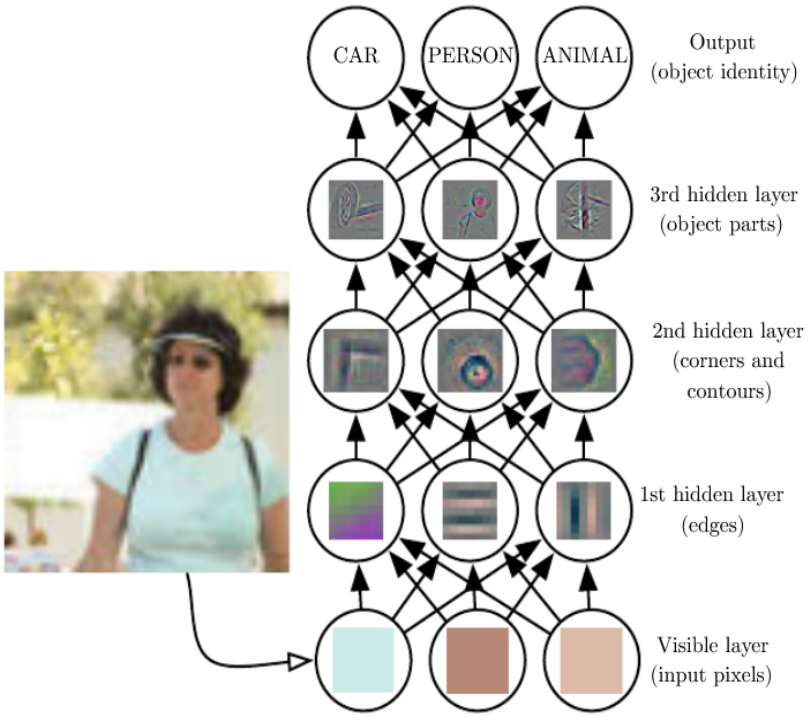


FIGURE 2.7 – Illustration of a deep learning model. Source : deeplearningbook [Goodfellow et al. 2016].

Convolution in neural networks comes with three important properties : sparse interactions (see figure 2.8 and figure 2.9), parameter sharing and equivariant representations. In a multi-layer perceptron, each unit of the actual layer is connected with every unit of the next layer. This process create separate parameters describing the interaction between each input unit and each output unit. Sparse interactions, called also sparse connectivity or sparse weights, refers to the small connectivity between a kernel and the input. This property allows the detection of small, meaningful features from an image, such as edges. Unlike the fully connected neural network, only a small kernel with tens of parameters is used, which results in a fewer parameters for the processing. This results in a large efficiency, since fully connected neural network is based on a matrix multiplication. Which means for an input n and an output m, the matrix multiplication requires m x n parameters. With a small k connectivity, only m x k parameters are required.

In a fully connected neural network, since each input unit is connected with each output unit, the connection is used only once. In a CNN, the same kernel is used for the whole

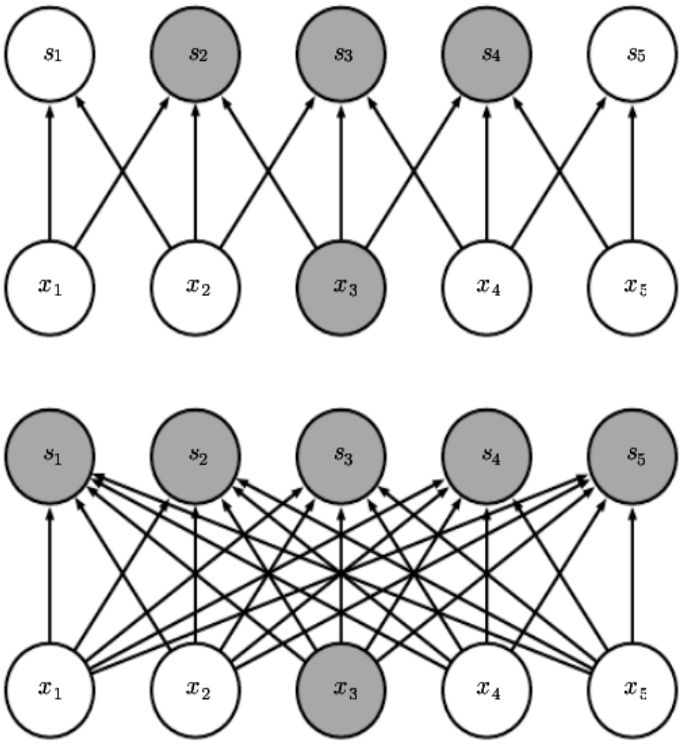


FIGURE 2.8 – Sparse connectivity (top) compared to fully connectivity (bellow). Using a kernel of size 3, only 3 S outputs are affected by input x3 (top). With a fully connectivity, all S outputs are affected by x3. Source : deeplearningbook [Goodfellow et al. 2016].

input, which results on sharing the parameters between the different input units.

2.2.11 Interpretabilty

Despite their success, deep learning models often function as black-boxes, and provide very little understanding about the inner workings. While opaqueness concerning machine behaviour might not be a problem in deterministic domains, in health care, model interpretability is crucial to build trust in the performance of a predictive system. To date no single method can provide a detailed human-understandable explanation of how a model makes a decision, however recent efforts in the field of interpretable artificial intelligence have produced various methods that can help bridge the gap between low-level features and phenotypic predictions. Perturbation-based approaches change parts of the input and observe the impact on the output of the network [Alipanahi et al. 2015, Zhou and Troyanskaya 2015]. Backpropagation-like methods, also known as saliency methods, use signals from gradients or output decomposition to infer a “saliency map” [Simonyan et al. 2013]. An alternative strategy is the Layer-wise Relevance Propagation (LRP) [Bach

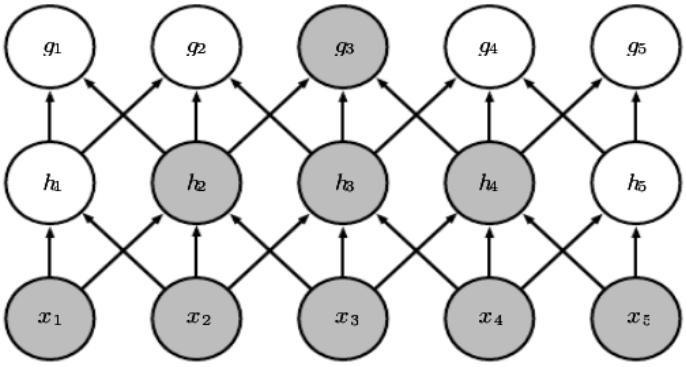


FIGURE 2.9 – The receptive field of the units in the deeper layers of a convolutional network is larger than the receptive field of the units in the shallow layers. This means that even though direct connections in a convolutional net are very sparse, units in the deeper layers can be indirectly connected to all or most of the input image. Source : deeplearningbook [Goodfellow et al. 2016].

et al. 2015]. Interpretable surrogate models aim to approximate a large, slow, but accurate model by a surrogate models a smaller, interpretable, yet still accurate model [Che et al. 2015, Hinton et al. 2015, Ribeiro et al. 2016]. Modifications have been proposed to Generative Adversarial Networks(GANs) to encourage the network to learn interpretable and meaningful representations [Chen et al. 2016]. Models with built-in explainability, such as attention mechanisms [Hendricks et al. 2016], can identify a posteriori the most informative features underlying a prediction.

2.2.12 Conclusion

In the first section we have covered the basics of ML with supervised learning approach, WSL and MTL. We have presented ANN with the basic component of a NN : perceptron. We covered then the MLP with CNN, showing how to train NN and preseting several hyperparameters that influences on the training and the performance of the NN.

In the next section, we introduce the concept of radiomics. We will show the differences between classical radiomics, based on handcrafted features with or without features selection strategy followed by ML, and deep radiomics, where the features are learned jointly with classification or prediction.

2.3 Radiomics

2.3.1 Concept and principle

Precision medicine is a reality in some tumor types [Arnedos et al. 2015]. It allows to separate patients based on some biomarkers to two categories : patients with good prognosis and patients with worse prognosis (see figure 2.10). Radiomics is a promising way towards precision medicine in oncology. It consists on the extraction of features from images to identify disease characteristics that help predict the outcome. In 2012, the concept of radiomics was introduced corresponding to the calculation of several dozens of features from medical images emerged [Kumar et al. 2012, Lambin et al. 2012a], extending the old notion of image quantification towards the design of predictive models based on selected features. Several reviews of the literature [Gardin et al. 2019, Yip and Aerts 2016] show the potential impact of radiomics in oncology for the prediction of the treatment response and patient survival. Several hundreds of quantitative handcrafted features can be extracted per lesion and image modality, related to the tumour volume, shape and textural properties.

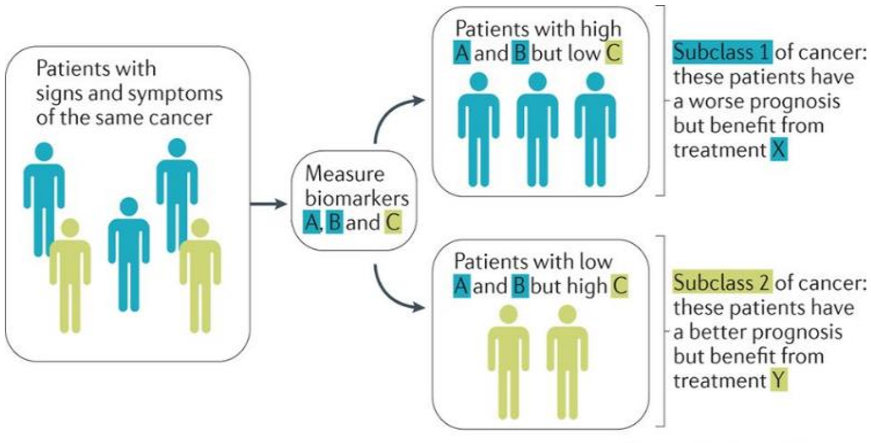


FIGURE 2.10 – Precision medicine allows to separate patients into different groups to personalize treatment. Source : [Vargas and Harris 2016].

The use of all the features extracted from the images does not enhance necessarily the performance, but may be responsible for the redundancy of the information and disrupt the model. [Orlhac et al. 2014] have shown that certain texture characteristics are highly correlated with MTV (Metabolic Total Volume) in three types of tumors. Similarly, [Tixier

et al. 2011a] have shown that GLRLM characteristics are highly correlated with GLSZM, therefore, they do not bring no additional information. Typically, conventional statistics are used to assign a degree of importance to each characteristic for prediction [Van De Wiele et al. 2013].

Because of the large number of characteristics to be studied and the non-linear relationship between them, standard mathematical tools such as linear regression, are not powerful enough. In this context, machine learning methods can be of great interest because of their ability to process a large number of characteristics and to capture a non-linear pattern, providing much better results than conventional statistics when analyzing several dozen characteristics [El Naqa et al. 2009]. The traditional classifiers generally used are the SVM and RF.

2.4 Machine learning for radiomics

The most used machine learning methods in radiomics are RF, SVM and MLP. Random forest or random decision forests [Breiman 2001] are an ensemble of multiple decision trees [Breiman et al. 1984]. Decision tree is a tree-like model where the population is divided in 2 progressively based on a feature so that it separate at best the 2 populations. Decision trees can be used for both classification and regression. RF makes use of handcrafted features, thus its performance is highly impacted by the features manually defined. SVM [Boser et al. 1992, Cortes and Vapnik 1995] constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used to separate different classes. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (called functional margin), since in general the larger the margin, the lower the generalization error of the classifier [Noble 2006].

To build models predicting treatment response or patient survival, Machine Learning (ML) approaches and Deep Learning (DL) have been used but their application to radiomics is still in its early stage. For instance, from a database of 65 patients with esophageal cancer treated using chemo-radiotherapy and 61 clinical and baseline FDG-PET features,

[Desbordes et al. 2017a] have shown the superiority of RF over SVM and conventional statistical analysis, using a single concatenated vector and several feature selection strategies. The best signature included both clinical and radiomic features.

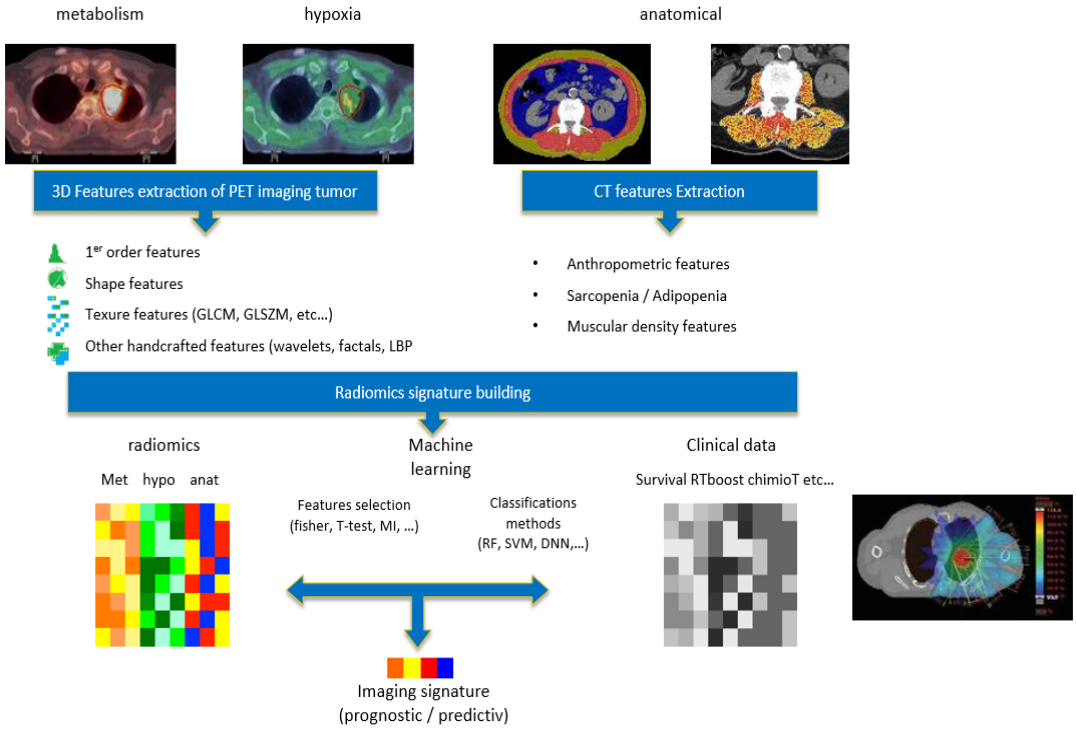


FIGURE 2.11 – Current workflow of radiomics based on machine learning and features selection strategy.

To date, most works have addressed this challenge by concatenating all together multiple groups of features in one single feature vector resulting in a high dimensional low sample size machine learning task [Parmar et al. 2015, Zhou et al. 2017]. Feature selection is the most commonly used method to reduce the dimension, either by using filter [Wu et al. 2016], wrapper [Farhidzadeh et al. 2016] or embedded [Wang et al. 2017c] methods. Though feature selection methods may be independent from the classifier, simple to implement and computationally fast, they may also filter some useful information for the classification task, whereas the objective of extracting a large number of features is precisely to bring additional information.

To summarize, machine learning for radiomics workflow is as follow (see figure 2.11 :

— The first step is the collection of the dataset and the definition of the ROI. Collecting a representative dataset is a challenge in the medical imaging field. In addition, the annotation of the dataset requires a highly trainable physician and is time

consuming. Thus, many radiomic studies do not include more than 50 patients in their studies [Balagurunathan et al. 2014, Cameron et al. 2015, Carneiro et al. 2017, Chung et al. 2015, Farhidzadeh et al. 2016].

— The second step consists of the extraction of features from the ROIs. Thousands of features can be extracted such as first order features, shape features,GLCM features,GLSZM features,GLRLM features,GLDM features and more [Van Griethuysen et al. 2017].

— Once the features are obtained, usually a feature selection strategy is applied to keep only a representative group of the whole set.

— Finally, a classical machine learning algorithm such as RF, SVM or MLP is applied to learn a radiomic signature that predict best the outcome [El Naqa et al. 2018, Jia et al. 2019, Shi et al. 2019, Wei et al. 2019].

— The obtained radiomic signature can be combined afterward with other data such as clinical data or other imaging modalities like CT or MRI.

In table 2.1 are referenced articles in the literature that address radiomic and machine learning in PET, for different type of cancers and using various ML algorithms and different purposes, from tumor diagnosis to outcome prediction. For instance, [Hyun et al. 2019] used machine learning based radiomics to successfully identify the histological subtypes of lung cancer (210 lung adenocarcinoma (ADC) from squamous cell carcinoma (186). [Cysouw et al. 2021] predicted metastatic disease or high-risk pathological tumor features in a prospective study. In [Li et al. 2019] radiomic with machine learning was used to detect bone marrow involvement in 41 patients with leukemia. They used random forest with 1826 achieving an accuracy, a sensitivity and a specificity of 88.6%, 87.5% and 89.5% respectively, outperforming visual analysis (accuracy = 62.5%, sensitivity = 73.7% and specificity = 68.6%). Other studies have shown the importance of incorporating peritumoral regions [Dou et al. 2018, Hao et al. 2018].

TABLEAU 2.1

Radiomic studies using machine learning algorithms for different types of cancer in PET.

Reference	Type of cancer	Nb of patients	Purpose of the study	Nb of features	Methods	Accuracy	AUC
[Hyun et al. 2019]	Lung	396	Distinguish lung adenocarcinoma (ADC) from squamous cell carcinoma (SCC)	44	logistic regression (LR) & ANN	0.769	0.859
[Cysouw et al. 2021]	Prostate	76	Predict metastatic disease or high-risk pathological tumor features	51	Random Forest	/	0.86-0.76
[Peng et al. 2019]	Nasopharyngeal carcinom	707	Predicting disease-free survival (DFS)	296	4 CNNs	/	0.722
[Toyama et al. 2020]	Pancreatic cancer	161	Prognosis	42	Random Forest	/	0.72
[Zhong et al. 2021]	Larynx and hypopha-rynx	72	Predict early disease progression	/	Random Forest	/	0.70
[Xie et al. 2020]	Head and neck	348	Prognosis	19	LR & SVM & RF & XGboost	/	0.72
[Du et al. 2020]	Nasopharyngeal Carcinoma	76	Local recurrence versus inflammation	478	k-nearest neighbo-rhood & SVM & RF	/	0.87
[Alongi et al. 2020]	Prostate	46	Prognosis	4867	/	0.66	/
[Ou et al. 2020]	breast	44	Breast carcinoma vs breast lymphoma	11	Linear discriminant analysis	0.808	0.806
[Ren et al. 2020]	Lung	315	ADC vs SCC	14	LASSO regression analysis	/	0.901
[Mu et al. 2019]	Lung	194	Prognosis	790	/	/	0.81

Reference	Type of cancer	Nb of patients	Purpose of the study	Nb of features	Methods	Accuracy	AUC
[Hao et al. 2018]	Lung	100	Predict distant failure	34	SVM	0.83	0.79
[Wang et al. 2017a]	Lymphoma	168	Classifying mediastinal lymph node metastasis	95	SVM & RF & Adaboost & ANN & CNN	0.81-0.85	0.87-0.92
[Nair et al. 2020]	Lung	50	Identify tumors with mutations	326	Logistic regression	0.87	0.71
[Li et al. 2018]	Lung	100	Predicting treatment response and OS	722	Clustering	0.64	/
[Li et al. 2019]	Leukemia	41	Bone marrow involvement detection	1826	Random Forest	0.88	/
[Papp et al. 2018]	Brain	70	Survival Prediction	56	Geometric probability covering algorithms	/	0.81
[Jeong et al. 2019]	Osteosarcoma	70	Treatment response prediction	/	SVM & RF & Gradient Boost	/	0.72-0.82
[Mi et al. 2015a]	lung & Esophageal	25 & 36	Treatment response prediction	79 & 29	SVM	100% & 0.94	
[Lian et al. 2016]	lung & Esophageal & Lymphoma	25 & 36 & 45	Recurrence or no-recurrence	52 & 29 & 27	Evidential K-Nearest-Neighbor	100% & 0.89 & 0.93	100% & 0.77 & 0.95
[Desbordes et al. 2017b]	Esophageal	65	Predicting treatment response and OS	58	Random Forest	0.82 & 0.80	0.82 & 0.75

2.5 Deep learning for radiomics

The use of DL algorithm is more recent [Peng et al. 2019]. While many studies have begun to explore the benefit of the analysis of texture to predict patient's outcome [Cook et al. 2013, El Naqa et al. 2009, Ha et al. 2014, Tixier et al. 2011a, Willaime et al. 2012], drawing a definitive conclusion is difficult because each study is based on different texture definitions and deploys different prediction models. In order to overcome this problem, deep machine learning methods such as convolutional neural networks have been used. They allow to extract features in a hierarchical way and to preserve the spatial relationship between the different slices. The strength of this method lies in the non-intervention in the manual extraction of the features which can cause a bias in the learning phase (see figure 2.12).

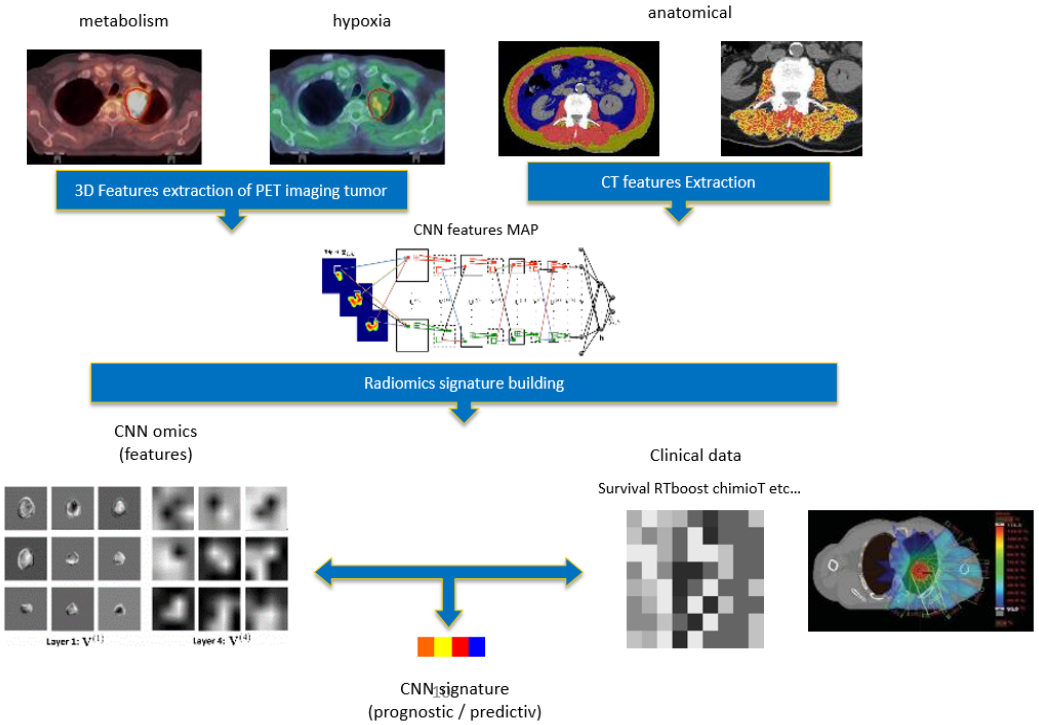


FIGURE 2.12 – Deep radiomics workflow. The model extract features and predict the outcome jointly.

[Ypsilantis et al. 2015] have proposed a hierarchical representation learned directly from PET images using two two-dimensional CNN architectures. The first called 1S-CNN, where the exam is separated on m slices representing the tumor and the its input is one slice at a time. A binary label is associated with each slice, 1 if the patient has responded to

treatment and 0 otherwise. Then the model is evaluated by a majority vote using all slices encompassing the region of interest (ROIs). The second architecture is a 3S-CNN where the input entry is the combination of 3 slices. For each exam with m slices, each set of three spatially adjacent slices is taken as an input resulting in a total of m-2 possible combinations. Figure 2.13 presents different slices for a patient X_i, where each slice presents a ROI for a specific tumor. The combination Z is done by selecting 2 adjacent slices.

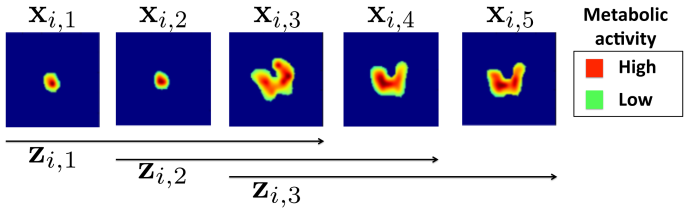


FIGURE 2.13 – ROIs of a specific tumor i after segmentation embedded into larger square background of standard size of 100 × 100 pixels. From [Ypsilantis et al. 2015]

[Wang et al. 2017b] compared CNNs with four traditional ML methods including RF, SVM, *Adaptive Boosting* (AB) and ANN. All methods were evaluated on 1397 lymphoma nodules from 168 patients. The study showed no significant difference between CNN and the other four methods for classifying non-small-cell lung carcinoma (NSCLC) mediastinal lymph node metastasis from PET/CT images. However, since CNNs do not require manual extraction of features like the other methods, it is considered more interesting to use them compared to the other methods.

2.6 Objectives of the thesis

Radiomics [Lambin et al. 2012b], can have a great clinical impact, since imaging is used in clinical routine all over the world. In PET imaging, there is a growing interest in identifying the features that characterize the spatial distribution and heterogeneity of 18F-FDG in a tumor [Hicks et al. 2004, Miller et al. 2003]. The dominant method for obtaining quantitative descriptors of spatial heterogeneities is based on texture analysis [Castellano et al. 2004]. These techniques encompass a large number of mathematical descriptors that can be used to evaluate the variation in intensity between voxels in a PET slice as well as in adjacent slices, in order to retrieve measures of intra-lesion heterogeneity.

Classical radiomic methods tend to use handcrafted methods to extract features, followed by a statistical method or machine learning algorithm such as RF SVM for the prediction. This workflow is based on two separated mechanism : extraction of features followed by learning the prediction of the outcome. Thus, it does not allow the model to learn useful data representation. Thus, in this thesis, we will study DL methods. The objective is to predict the cancer outcome from PET images. Because DL needs a lot of annotated data, it is not always available in medical imaging. To solve this problem, we propose to study weakly supervised learning and multitasking learning. In the next chapters, three proposed methods will be presented.

Chapitre 3

Hand crafted methods vs deep radiomics

Sommaire

3.1 Introduction 52

3.2 Material and methods 55

3.2.1 Database presentation 55

3.2.2 Image preprocessing 56

3.2.3 3D RPET-NET architecture 57

3.2.4 Implementation 59

3.3 Experimentations 60

3.4 Validation methodology 60

3.5 Results 63

3.6 Discussion 64

3.7 Conclusion 66

In this chapter, we conduct an exhaustive comparative study between classical radiomics, where the features are manually designed, and deep radiomics, where the model is extracts end-to-end feature maps. We propose deep learning approaches based on convolutional neural networks (CNNs) for outcome prediction from positron emission tomography (PET) images. In particular, two 2D CNNs and one 3D CNN are developed and evaluated to predict the outcome for patients with esophageal cancer on PET images. The results were compared to 3 state of the art classical radiomics algoithms : random forest without features selection (RF), random forest with genetic algorithm to select features (GARF), and random forest with feature importance (FIC).

3.1 Introduction

Predicting patient response to radio-chemotherapy (RCT) is a very promising field of research in personalized medicine. PET imaging with ^{18}F-FDG, which is a radioactive glucose analog, has mainly been used in radiomics analysis, but other radio-tracers have also been tested [Lu et al. 2016]. However, the roles of traditional imaging biomarkers such as SUVmax and metabolic tumor volume (MTV) have not been well established in esophageal cancer for therapy response [Kwee 2010]. Other biomarkers such as handcrafted texture features have been proposed [Tixier et al. 2011b] that are associated with standard statistics or advanced statistical classifiers [Desbordes et al. 2017c, Mi et al. 2015b].

The concept of radiomics is defined as the extraction of dozens of quantitative features from the image that could be incorporated in predictive models for patient management [Lambin et al. 2012a]. Many reports suggest that radiomic features extracted from baseline images can contribute to improving patient prognosis and prediction of treatment response in oncology [Avanzo et al. 2017]. Images can be obtained from computed tomography (CT) [Bogowicz et al. 2017], magnetic resonance imaging (MRI) [Nishioka et al. 2002] and positron emission tomography (PET) [Cook et al. 2014]. The visualization of glucose metabolism of tumor cells and other radiotracers in PET provides additional information to that obtained from anatomical imaging (CT or MRI). These so-called radiomic features are assumed to highlight some informative tissue characteristics, such as

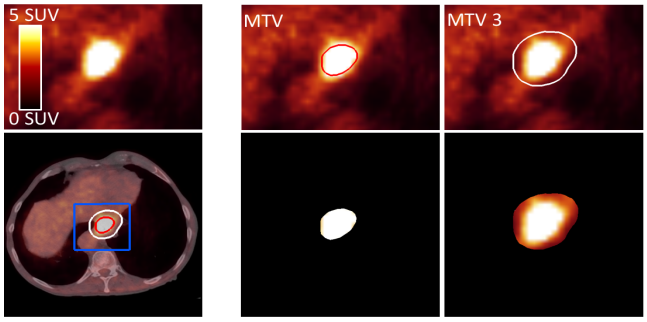


FIGURE 3.1 – Columns from left to right : Fused PET/CT slice, zoomed on the esophageal tumor seen on FDG-PET only. MTV (40% SUVmax thresholding) in red and MTV included in the cuboid. MTV3 (MTV + 3 cm isotropic margin) in white and MTV3 included in the cuboid.

heterogeneity in glucose metabolic activity, necrosis, etc. Numerous image features have been proposed in the literature [Gardin et al. 2019, Kumar et al. 2012, Sollini et al. 2017] based on the shape and size of the lesion, 1st order statistics, textural features, filter and model-based features, potentially leading to hundreds of image characteristics.

Several authors have used machine learning (ML) methods to build models for predicting treatment response or patient survival based on radiomic features, such as random forests (RFs) and support vector machines (SVMs) with or without a feature selection strategy [Desbordes et al. 2017c, Leger et al. 2017, Mi et al. 2015b]. The main drawback of these approaches is that they require an initial extraction of radiomic features using hand crafted methods, which usually results in a large number of features and cannot always find the most representative ones. in addition, handcrafted features are affected by some parameters [Hatt et al. 2017] such as noise, reconstruction, etc. and significantly by the contouring methods used.

CNNs have proven to be very powerful tools in computer vision for classifying images from different domains. CNN architectures for medical imaging have been introduced and usually containing fewer convolutional layers because of the small datasets [Frid-Adar et al. 2018]. Recently, a new paradigm in PET radiomic analysis has been proposed based on CNNs for predicting response to therapy [Ypsilantis et al. 2015]. It has been shown that deep learning architectures can outperform traditional ML methods in classification tasks.

CNNs were not fully studied in radiomics, especially in PET imaging. Some papers have investigated baseline PET analysis based on 2 Dimensional (2D) CNN architectures

[Wang et al. 2017b, Ypsilantis et al. 2015], but to our knowledge, there are no studies using 3D-CNN. These first two applications dealt with the prediction of the response to neoadjuvant chemotherapy in esophageal cancer [Ypsilantis et al. 2015] and the classification of mediastinal lymph node metastasis of non-small cell lung cancer (NSCLC) [Wang et al. 2017b].

In [Ypsilantis et al. 2015], Ypsilantis *et al.* proposed to learn a hierarchical representation directly from PET images in 107 patients with esophageal cancer using two CNN architectures. The first one, called 1S-CNN, corresponds to an architecture where the input is one slice. The process is repeated on each slice where the tumor is present. The spatial dependency between slices is not exploited in this architecture. For this reason, a second architecture was proposed where the input of the CNN is composed of 3 adjacent slices, called 3S-CNN. For each exam containing m slices, each set of three spatially adjacent slices is taken as input, leading to a total of m-2 possible combinations. This 3S-CNN better exploits the spatial relationship between slices but is limited to 3 slices. For both architectures, a post processing step is required to predict the response based on a majority vote process using all slices that include a tumor for a patient. This study has shown the superiority of these two deep learning methods compared to other ML methods, such as RF, SVM, gradient boosting, and logistic regression.

In [Wang et al. 2017b], Wang *et al.* used a centered axial slice and two others that were separated by 4 mm in two image modalities (PET and CT) to obtain a limited number of six slices for each tumor to make a prediction. They compared the performances of their CNN and four other methods including RF, SVM, adaptive boosting, and artificial neural network. The methods were evaluated to discriminate against benign and malignant lymph nodes (1397) in 168 patients. The study showed that there were no significant differences between the CNN and the best classical ML method for classifying mediastinal lymph node metastasis of NSCLC from PET/CT images. Nevertheless, Wang *et al.* concluded that CNNs are more convenient to use because the method does not require an initial feature extraction.

Radiotherapy planning is based on CT by delineating the gross tumor volume (GTV). This GTV can also be segmented using other image modalities, such as MR and PET

images. Segmentation of the tumor in PET imaging is usually performed using a fixed threshold value of 40% of the maximum standard uptake value (SUVmax) [Galavis et al. 2010], leading to the biological or metabolic target volume (BTV or MTV). Then, the radiation oncologist adds several margins that take into account the non-visible tumor infiltration (CTV : clinical tumor volume) as well as uncertainties in positioning and treatment to obtain the PTV (planning target volume) [Dubray et al. 2013]. The peritumoral part of the tumour is therefore a volume that is not neglected in the treatment. By analogy, taking into account the intratumoral and peritumoral regions in radiomics analysis is likely a strategy that can improve the results. At present, a few studies have tested this hypothesis in other modalities [Braman et al. 2017, Zhou et al. 2018a] but never with PET imaging.

Our goal is to develop a new 3D-CNN architecture, that we name 3D RPET-NET, to predict the response to treatment by learning from FDG-PET images of the tumor. Considering our small dataset, a four-layer 3D-CNN is proposed. Our study used a database of baseline FDG PET images of 97 patients treated by radio-chemotherapy (RCT) for esophageal cancer. The optimal hyperparameters of 3D RPET-NET and the influence of the learning volume (intratumoral volume with different peritumoral volumes) are investigated and will be reported in Results section. The performances of the model were compared to 1S-CNN and 3S-CNN [Ypsilantis et al. 2015], as well as to three RF methods [Desbordes et al. 2017c] considered as state-of-the-art radiomics classifiers.

3.2 Material and methods

3.2.1 Database presentation

In this study, 97 patients with one lesion that was histologically proven to be locally advanced esophageal cancer and eligible for RCT are included. All procedures performed in this study are conducted according to the principles expressed in the Declaration of Helsinki. The study was approved as a retrospective study by the Henri Becquerel Center Institutional Review Board (number 1506B). All patient information is de-identified and anonymized prior to analysis.

All patients underwent a FDG-PET/CT exam before treatment (baseline PET), at the initial stage. They were then treated by RCT, corresponding to an uninterrupted radiation therapy in the form of external radiation delivered by a 2-field technique of 2 Gy per fraction per day, 5 sessions per week, for a total of 50 Gy, as well as chemotherapy including platinum and 5-fluorouracil.

The PET/CT data were acquired on a Biograph® Sensation 16 Hi-Rez device (Siemens Medical Solutions, IL, USA). This device does not provide point spread function (PSF) modeling or time-of-flight (TOF) technology. Patients were required to fast for at least 6 hours before imaging. A total of 5 MBq/kg of FDG was injected after 20 min of rest. Sixty minutes later (±10 min), 6 to 8 bed positions per patient were acquired using a whole-body protocol (3 min per bed position). The PET images were reconstructed using Fourier rebinding (FORE) and attenuation-weighted ordered subset expectation maximization algorithms (AW-OSEM with 4 iterations and 8 subsets). The images were corrected for random coincidences, scatter, and attenuation. Finally, the FDG-PET images were smoothed with a Gaussian filter (full width at half maximum (FWHM) = 5 mm). The reconstructed image voxel size was $4.06 \times 4.06 \times 2.0$ mm^3.

For the determination of treatment response, the response assessment included clinical examination, CT, FDG-PET, and esophagoscopy with biopsies performed 1 month after the end of treatment. Patients were classified as showing a clinically complete response (CR, 56 patients) to RCT if no residual tumor was detected on the endoscopy (negative biopsies) and if no locoregional or distant disease were identified on CT or PET evaluation. Patients were classified as showing a non-complete response (NCR, 41 patients) if a residual tumor or locoregional or distant disease was detected or if death occurred.

3.2.2 Image preprocessing

Tumor images were spatially normalized by re-sampling all the dataset to an isotropic resolution of $2 \times 2 \times 2$ mm^3 using the k-nearest neighbor interpolation algorithm.

The metabolic tumor volume (MTV) was segmented by a physician who manually defined a cuboid volume around the lesion and used a fixed threshold value of 40% of the maximum standard uptake value (SUVmax) in the cuboid. To study the influence of the

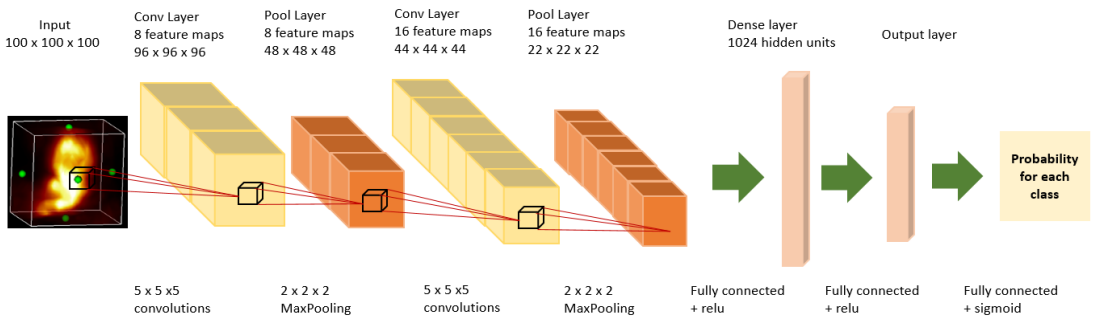


FIGURE 3.2 – 3D RPET-NET architecture composed by two 3D convolutional layers followed by 3D pooling layers and two dense layers.

volume of interest on the performances of 3D RPET-NET, several isotropic margins of 1, 2, 3 and 4 cm around MTV were also applied, leading to defining MTV1 to MTV4. In Fig. 1, an example of a PET/CT slice with two volumes of interest (MTV and MTV3) is shown.

Tumor gray level intensities were normalized to absolute SUV level between [0 30] and translated between [0 1] to be used in CNN architectures. The volumes of interest were included into a 3D empty cuboid of standard width, length and height of 100^3 voxels to learn tumoral radiomic features.

3.2.3 3D RPET-NET architecture

We have developed a CNN architecture based on two 3D convolutional layers and two fully connected layers, as shown in figure 3.2 for radiomic study. As we do not have a large amount of data and our architecture is in 3D, we take here only 4 layers. Each convolutional layer, denoted $C^{(m)}$, consists of $F^{(m)}$ feature maps, where m is the layer number (1 or 2). For the first layer, $C^{(1)}$, each feature map is obtained by convolving the volume of interest with a weight matrix $W_i^{(1)}$ to which a bias term $b_i^{(1)}$ is added, where i is the feature map number. Then, the output is processed by a non linear function f(x) called the activation function, where x is the input to a neuron, such as :

$$c_i^{(1)} = f(b_i^{(1)} + W_i^{(1)} * \mathrm{x}) \qquad \text{with } i = 1, ..., F^{(1)}. \tag{3.1}$$

Each element of a feature map, $c_i^{(1)}$, is obtained by convolving the input x with a 3D kernel. A large receptive field tends to better preserve the relationship between slices and

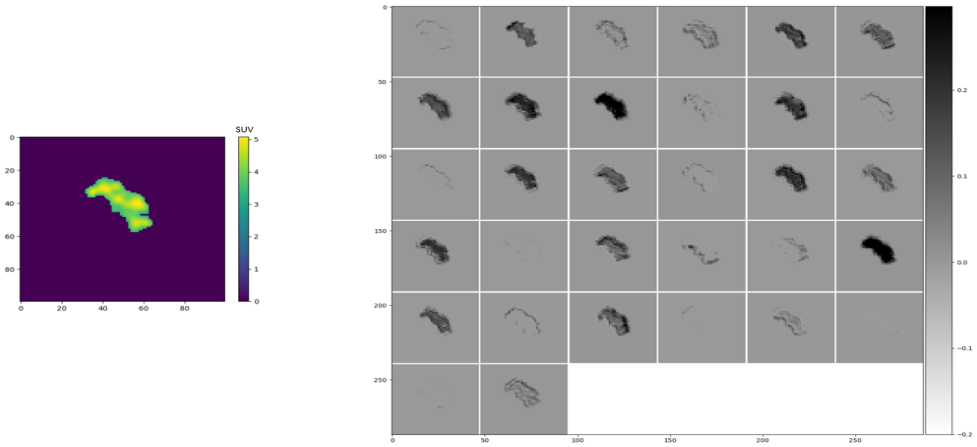


FIGURE 3.3 – Visualization of a 2D slice of a segmented tumor and the resulting 32 feature maps in the second convolutional layer of the 1S-CNN architecture.

the local 3D tumor information than a small one ($(5 \times 5 \times 5)$ *vs.* $(3 \times 3 \times 3)$). The $F^{(1)}$ weight matrices (one matrix per feature map) are learned by observing different positions of the input, leading to the extraction of the description of features. Thus, the weight parameters are shared for all tumor input sites, so that the layer has an equivariance property and is invariant to the input tumor transformations (such as translation and rotation). It also results in a sparse weight, which means that the kernel can detect small, but meaningful features, as shown in figure 3.3. For instance, it can be seen that some kernels are learning the tumor shape (e.g, feature maps [(1,1),(1,3),(2,4),(2,6)..etc.]), while others tend to focus on features within the tumor (e.g, feature maps [(1,2),(1,5),(2,2),(2,3)..etc.]).

Then, the output of this first convolutional layer is followed by a 3D pooling layer, to reduce the dimensionality of feature maps. The max-pooling operator is used as a stage detector to report the maximum value within each cuboid of size $(2 \times 2 \times 2)$ for all feature maps. The purpose of this operation is to down sample the feature maps by a factor of 2 along each direction (width, high, length) and to better generalize learning by selecting approximately invariant features. This invariance to local translation is very important in radiomics because tumors do not have a particular direction. The resulting feature maps are denoted $P^{(m)}$.

To extract high-level features from the low-level ones obtained in the initial layer, a second convolutional layer is added, followed by a pooling layer. This convolutional layer learns from the pooled feature maps of the first layer (see figure 3.2).

The parameters of the CNN consist of all the convolutional weights W, and the weight matrix Wh, denoted by θ. They are learned by minimizing the binary cross-entropy function :

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{3.2}$$

which is a special case of the multinomial cross-entropy loss function for m = 2 :

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log(\hat{y}_{ij}) \tag{3.3}$$

where n is the number of patients, y is the the ground truth : 1 if the patient responds to treatment, 0 otherwise.

In our experiments, the adaptive gradient algorithm optimizer (AdaDelta) is used with mini batches. At each update of weights using the AdaDelta algorithm, only one mini batch of training data was used, which is changed for each gradient calculation. Our CNN also incorporated L2 normalization of the weights and a dropout regularization of 50% to prevent the model from overfitting.

To find the best 3D RPET-NET we test different parameters. The network using the optimal parameters is 3D RPET-NETBest. The hyperparameters to be optimized include the number of 3D feature maps (we tested from 8 to 64 feature maps), the number of neurons (128, 256, 512, 1024, 2048 and 4096), as well as different receptive field sizes ($3\times3\times3$, $5\times5\times5$) and different sizes of mini-batches (2, 4, 8 and 16). We have evaluated several activation functions (relu, elu, selu and tanh), the numbers of 3D convolutional layers,3D pooling layers (2 to 5) and fully connected layers (2, 3, 4 and 5) to find the the best model.

3.2.4 Implementation

The implementation of 3D RPET-NET is conducted using the Keras library which is built on top of Theano and Tensorflow. We take advantage of graphical processing units (GPUs) to accelerate the algorithm. The CNNs training is performed on an NVIDIA Tesla 80 with 12 GB of memory.

3.3 Experimentations

Three experiments were performed to evaluate our 3D RPET-NET.

Experiment 1 : The first experiment consisted of tuning the optimal hyperparameters to find 3D RPET-NETBest based on MTV. Optimizing the hyperparameters was performed entirely on the training dataset.

Experiment 2 : The second experiment consisted of comparing our architecture with 2 other CNN methods proposed in the literature : 1S-CNN and 3S-CNN [Ypsilantis et al. 2015]. The same tuning process of 3D RPET-NET was performed to find the best 1S-CNN and 3S-CNN hyperparameters. This experiment was performed on test data.

We carried out a comparative study between our method and three RF-based methods : one method without any feature selection strategy, called RF, and two other RF methods proposed in the literature using a feature selection strategy. The first selection strategy, called GARF, uses a genetic algorithm based on random forest, and the second one called FIC, uses features important coefficient methods. For the details of these methods refer to [Desbordes et al. 2017c]. Briefly, 45 image features were extracted from PET images corresponding to first-order statistics (18), one feature of the lesion form, and textural features (26). Five hundred decision trees were built leading to the creation of the random forest classifiers.

Experiment 3 : The third experiment consisted of assessing the influence of the volume of interest on the performances of 3D RPET-NETBest, RF, GARF and FIC according to the size of the volume of interest.

3.4 Validation methodology

For the evaluation of our method, cross-validation (CV) was performed. We split the data into 2 groups to train and test the machine learning methods for each fold. One group was used for training the models (77 patients) and one group for testing (20 patients). Furthermore, for the CNN, the training samples were split into a dataset of 2 groups, a train

set (55 patients) and a validation set (20 patients), and a grid search was conducted to derive the optimal hyperparameters based on the validation set. For a fair comparison, different machine learning methods were trained and tested with the same fold, i.e, trained with the same training sets and tested with the same test sets. To keep the same ratio between the two classes CR and NCR, for each fold, the training set contained 44 CR patients and 33 NCR patients, and the testing set contained 12 CR and 8 NCR.

The performances of the methods were evaluated for each cross-validation, including sensitivity (Sens), specificity (Spec), accuracy (Acc), and area under the receiver operating characteristic (ROC) curve (AUC). For each curve, the definition of the thresholds was determined using the method proposed by Fawcett [Fawcett 2006], and the optimal cut-off point was defined using Youden's index.

A comparison between different methods was mainly performed based on the AUC values. Due to the 5-fold CV, 5 groups of performance values were calculated for each method; therefore, paired hypothesis tests of 5 samples were performed. The p values were calculated using Student t-test. To correct for multiple comparisons, we additionally adjusted p-values by the false-discovery-rate (FDR) procedure according to Benjamini-Hochberg [Benjamini and Hochberg 1995]. The null hypotheses were rejected at the level of $p < 0.05$ after correction.

	Method	VOI	Acc	Sens	Spec	AUC
Experiment 1						
	3D RPET-NETBest	**MTV**	**0.83±0.04**	**0.91±0.06**	**0.73±0.16**	**0.81±0.06**
	3D RPET-NET1	MTV	0.80±0.06	0.93±0.05	0.61±0.15	0.77 ±0.06
	3D RPET-NET2	MTV	0.76±0.04	0.87±0.12	0.62±0.19	0.75±0.05
Experiment 2						
	3D RPET-NET	**MTV**	**0.72±0.08**	**0.79±0.17**	**0.62±0.21**	**0.70±0.04**
	1S-CNN	MTV	0.69±0.06	0.79±0.15	0.57±0.24	0.65±0.08
	3S-CNN	MTV	0.67±0.08	0.73±0.19	0.60±0.20	0.67±0.08
	GARF	MTV	0.68±0.08	0.80±0.11	0.46±0.09	0.62±0.04
	FIC	MTV	0.65±0.07	0.78±0.21	0.46±0.38	0.61 ±0.16
	RF	MTV	0.65±0.04	0.65±0.18	0.53 ±0.18	0.59±0.04
Experiment 3						
	3D RPET-NET	MTV1	0.73±0.04	0.76±0.07	0.69±0.1	0.72±0.04
	GARF	MTV1	0.70±0.08	0.74±0.07	0.54±0.07	0.62±0.02
	FIC	MT1V	0.62±0.10	0.58±0.18	0.64±0.12	0.59 ±0.04
	RF	MTV1	0.62±0.09	0.62±0.08	0.61 ±0.07	0.59±0.03
	3D RPET-NET	**MTV2**	**0.75±0.03**	**0.76±0.45**	**0.74±0.15**	**0.74±0.02**
	GARF	MTV2	0.71±0.09	0.73±0.11	0.54±0.09	0.63±0.04
	FIC	MTV2	0.58±0.01	0.58±0.25	0.57±0.18	0.54 ±0.07
	RF	MTV2	0.62±0.11	0.56±0.20	0.65 ±0.12	0.59±0.05
	3D RPET-NET	MTV3	0.72±0.09	0.71±0.09	0.74±0.14	0.72±0.09
	GARF	MTV3	0.66±0.07	0.68±0.19	0.57±0.12	0.63±0.04
	FIC	MTV3	0.61±0.11	0.63±0.17	0.58±0.16	0.59 ±0.04
	RF	MTV3	0.62±0.14	0.66±0.17	0.55 ±0.20	0.59±0.04
	3D RPET-NET	MTV4	0.63±0.09	0.77±0.10	0.46±0.21	0.61±0.11
	GARF	MTV4	0.65±0.09	0.73±0.14	0.52±0.16	0.62±0.02
	FIC	MTV4	0.59±0.08	0.54±0.14	0.63±0.08	0.56 ±0.04
	RF	MTV4	0.60±0.13	0.66±0.12	0.56±0.05	0.58±0.04

TABLEAU 3.1 – Classification results : Each result corresponds to the average of five independent experiments and the standard deviation, using the training dataset (Experiment 1) or the test dataset (Experiment 2 and 3).

3.5 Results

The main results from the 3 experiments evaluated by accuracy, sensitivity, specificity and AUC of ROC curves are shown in table 1.

Experiment 1 : As shown in Fig.2, the best accuracy Acc=0.72 and AUC=0.70 were achieved by two 3D convolutions layers and two 3D pooling layers, followed by two fully connected layers with the following hyperparameters for the first 3D convolutional layer : eight 3D feature maps with a filter size of $5 \times 5 \times 5$ and a relu activation function. This operation is followed by 3D Max-pooling of size $2 \times 2 \times 2$. The second 3D convolutional layer corresponds to sixteen 3D feature maps of $5 \times 5 \times 5$ convolutions, followed again by a $2 \times 2 \times 2$ 3D pooling layer. Then, the last two layers are composed of fully connected layers of 1024 hidden neurons and finally 2 neurons for both classes.

In Experiment 1, the results of two other models show also interesting performances, with no significant difference from 3D RPET-NETBest. 3D RPET-NETBest and 3D RPET-NET1 differ by the activation function (relu *vs.* elu). 3D RPET-NETBest and 3D RPET-NET2 differ by the activation function (relu *vs.* elu) and the kernel size ((5 x 5 x 5) *vs.* (3 x 3 x 3)).

Experiment 2 : the best results obtained with 1S-CNN, 3S-CNN, RF, GARF and FIC are shown in table 3.1. The ROC curves of Experiment 2 are presented in Fig. 4.a.

The best results are obtained with 3D RPET-NETBest. 1S-CNN, seems to have lower performances (Acc=0.67±0.06, AUC=0.67±0.06), but the 1S-CNN ROC curve is not statistically significantly different from 3D RPET-NETBest (p=0.53) and 3S-CNN (p=0.48) ROC curves. For the RF classifiers, the best results are obtained with the GARF algorithm. The GARF ROC curve is not statistically significantly different from 1S-CNN (p=0.10) and 3S-CNN (p=0.058) ROC curves, while the 3D RPET-NETBest ROC curve obtains better results than the GARF ROC curve (p=0.028).

Experiment 3 : The results of Experiment 3 are given in table 3.1 and the comparisons of different AUC in figure 3.4.b. When studying the influence of the volume of interest, the best performances of 3D RPET-NETBest are obtained with MTV2 (Acc=0.75 and AUC=0.74). The performances of the 3D RPET-NETBest increase from no margin to a magin of 2 cm, and then decrease with higher margins (MTV3 and MTV4). Only 3D RPET-

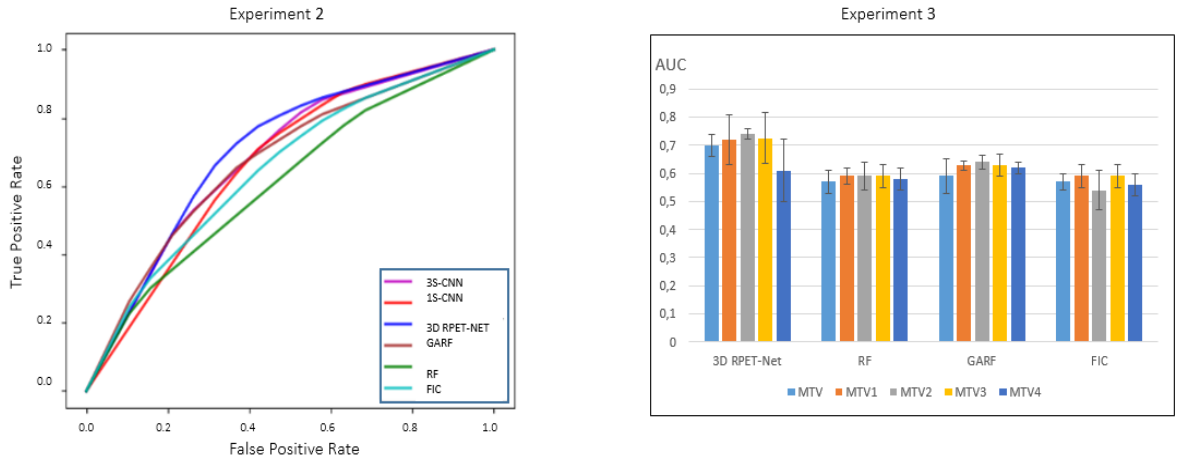


FIGURE 3.4 – a. On the left : ROC curve comparing the 6 classifiers (RF, GARF, FIC, 1S-CNN, 3S-CNN and 3D RPET-NET) with the best parameters on MTV. b. Right : Comparison of the four classifiers on different VOIs (MTVs). Error bars correspond to standard deviation.

NETBest performances on MTV2 are statistically significantly better than those on MTV4 (p=0.04). The same trend is observed with RF classifiers.

3.6 Discussion

We have developed an end-to-end 3D convolutional neural network (3D PET-NET) based on PET images. We have also evaluated 5 other methods from the literature [Ypsilantis et al. 2015] [Desbordes et al. 2017c]. For each CNN, the search for the best architecture is achieved by using a validation procedure to tune hyperparameters, such as the number of feature maps and the size of filters.

Apart from the numerous advantages of CNNs (avoiding handcrafted feature design and feature selection), it is now well known that convolutional architectures build high level representations of the input signals. They typically extract low level features such as textures of edge detectors in the low layers and accumulate these information to form higher level features in the last layers. Low level features are generally rather generic and can be exploited through transfer learning [Belharbi et al. 2017]. Higher level features are more domain-specific and depend upon the application. A neural network is often considered as a “black box”, but CNN layers provide interpretability through the feature maps that highlight the activation of each kernel within the input signal. Therefore, we think

that CNN features are likely to be related to classical hancrafted radiomic features (see figure 3.3).

PET imaging suffers from low resolution and high noise, leading to challenges in PET radiomics [Hatt et al. 2017]. However, neural networks provide a robust mechanism to avoid encoding the noise in the data such as 'early-Stopping' and 'dropout' which provide better generalization [Srivastava et al. 2014b].

Unlike Ypsilantis et al. in [Ypsilantis et al. 2015] who claimed that the use of a 3D ROI as direct input of the CNN is infeasible because every tumor has a different shape and size, we show that englobing the tumor into a 3D cuboid of standard width, length and height allows the benefit of the spatial relationship between slices using a large 3D receptive field to be realized. Our assumption is that a neural network architecture able to capture patterns of FDG uptake that occur within the whole lesion may detect imaging features that are more relevant to predict treatment response than each slice individually or 3 adjacent slices. Under this assumption, we propose an architecture that initially fuses the spatial information across intra-slices images. 3D RPET-NETBest is composed of only 2 convolutional layers. A higher number of convolutional layers were tested, without conclusive results. The small number of patients in our database (without artificial data augmentation) is a limiting factor not only for the development of a deeper network but also for radiomic analysis in general. Indeed, the current trend is in favour of the use of a network with an increasing number of convolutional layers (very deep neural network). This is only possible on large image databases (e.g., ImageNet [Russakovsky et al. 2015], containing now more than 14 million images, 30 high level categories and 20K subcategories) that are not currently available in medical imaging. It is possible to artificially increase the number of data. However since, learning takes place on a tumor inside a black box, this solution leads to overfitting.

To ensure a fair comparison between the different methods, the database was divided into 3 groups of 57 patients for the training, 20 for the validation and 20 for the test before any operation. Every CNN and RF classifier used the same folds to obtain an exact comparison between methodologies.

There are several segmentation methods available for PET imaging. Many automa-

tic frameworks have been proposed during the last decade [Foster et al. 2014], but few of them are used/available in clinical routine. The simple threshold is still mainly used but with different values depending on pathologies [Dewalle-Vignion et al. 2012]. A segmentation of the MTV can be accurately performed with a 40% threshold value because esophageal cancer can be considered as a massive non moving tumor [Kawakami et al. 2015] and it has been proven that this segmentation is highly correlated with a manual segmentation [Lambin et al. 2012a].

We have shown that isotropic dilation of MTV tends to increase the performances of RPET-NET 3D. When the margin around the MTV is too large (>2 cm) the network performances decrease. When the MTV is increased by a margin which is too large, the volume of interest can include parts of metabolically active organs that are likely to interfere with the CNN analysis. Our results suggest that between 3 cm and 4 cm of the peritumoral volume, the relevant information to predict treatment response decreases, is responsible for a drop in the model's performance. Adding a peritumoral volume to the radiomic analysis has already been tested in MRI [Braman et al. 2017] but never in PET imaging. These initial results must be confirmed on other types of cancer. Moreover, the influences of the initial volume of interest and the segmentation methods require further study.

3.7 Conclusion

The analysis of PET tumor images with a 3D CNN architecture (3D-RPET-NET) shows very promising results in the prediction of treatment response in esophageal cancer. 3D-RPET-NET outperformed 2D CNN architectures, as well as the traditional radiomics approach (such as RF classifiers). Moreover, since the CNN does not take hand-crafted features as input, it eliminates the need for feature selection, making the entire process much more convenient and less prone to user bias. In addition, we have shown that the best volume to be used for PET radiomic prediction is the metabolic tumor volume with an isotopic margin of 2 cm. This peritumoral region seems to contain information that is potentially relevant to building better prediction algorithms since currently approaches are based only on the quantification of the intratumoral region alone.

Even though our CNN-based method can give good results, it needs to know segmented tumor regions. However, manual segmentation of the tumor in 3D is a very tedious and time consuming task. To solve this problem, we propose a weakly supervised learning approach which will be presented in the next section.

Article Details :

— **3D RPET-NET : DEVELOPMENT OF A 3D PET IMAGING CONVOLUTIONAL NEURALNETWORK FOR RADIOMICS ANALYSIS AND OUTCOME PREDICTION.** Amine Amyar, Su Ruan, Isabelle Gardin, Clément Chatelain, Pierre Decazes, Romain Modzelewsk. IEEE Transactions on Radiation and Plasma Medical Sciences, 3(2), pp.225-231.

Other Related Publications :

— **Radiomics-net : Convolutional neural networks on FDG PET images for predicting cancer treatment response.** Amine Amyar, Su Ruan, Isabelle Gardin, Romain Herault, Chatelain Clement, Pierre Decazes, Romain Modzelewski. Journal of Nuclear Medicine, 59(supplement 1), pp.324-324.

— **Prédiction de la réponse au traitement du cancer de l'œsophage en boostant l'analyse (deep) radiomique avec des caractéristiques cliniques et anthropométriques.** Amine Amyar, Su Ruan, Pierre Decazes, Isabelle Gardin, Romain Modzelewski. Médecine Nucléaire, 44(2), p.106.

— **Radiomics-net : analyse Deep-radiomics des images TEP FDG pour prédire la réponse au traitement du cancer.** Amine Amyar, Su Ruan, Pierre Decazes, Isabelle Gardin, Romain Modzelewski.Médecine Nucléaire, 44(2), p.105.

Chapitre 4

Weakly supervised learning for outcome prediction

Sommaire

4.1 Introduction 70

4.2 Material and methods 74

4.2.1 Main idea 74

4.2.2 Maximum Intensity Projection 75

4.2.3 New Design of Class Activation Map 76

4.2.4 Segmentation 80

4.2.5 Prediction 80

4.3 Experiments 80

4.3.1 Dataset 80

4.3.2 Setup 81

4.3.3 Implementation 83

4.4 Evaluation Methodology 83

4.5 Results 83

4.6 Discussion & Conclusion 85

After demonstrating the effectiveness and usefulness of deep learning in predicting patient's outcome, this chapter introduces a novel weakly supervised learning approach to segment the lesions in order to conduct a radiomic analysis. By the concern of lack of annotations necessary for a supervised learning, we propose here a method which does not require the ground truth of segmented tumor, but only the classes of the tumors. We propose to use explainable deep learning techniques in the classification decision to detect the tumor under prior knowledge. We transform the classification neural network to the tumor detection and segmentation tasks. The results are compared to supervised learning approach for tumor segmentation, and with radiomic based on manual segmentation for outcome prediction.

4.1 Introduction

To better appreciate the volume of interest in oncological radiotherapy and also the biological component of a tumor, radiomics is proposed as a field of study that makes use of images [Gillies et al. 2016]. Radiomics allows from an initial PET exam the prediction of the survival of a patient and the response to radio-chemotherapy treatment, and therefore to help to personalize treatment [Amyar et al. 2019b, Lian et al. 2016]. The first step in a radiomics analysis is to localize tumor region for which radiomcis features can be extracted. Manual segmentation is tedious and time consuming, especially in 3D.

Deep learning is a very promising tool for the automatic detection of lesions in PET images, but due to their data-hungry nature, they require very large amounts of annotated images, they are usually not available in medical imaging field. Most of segmentation methods use large annotated databases, however, annotating pixel-level tumor requires highly trainable physicians and they don't have a lot of time to do manual segmentation, especially in 3D. Moreover, physicians annotations can be subjective. In contrast, image-level labels indicating the presence of a lesion, or the type of cancer when they make the diagnosis are easy for the physicians and can be quickly obtained. Therefore, we propose an approach based on a weakly supervised learning (WSL), where image-level information is used to train a classifier based on CNN to predict the class label in a supervised

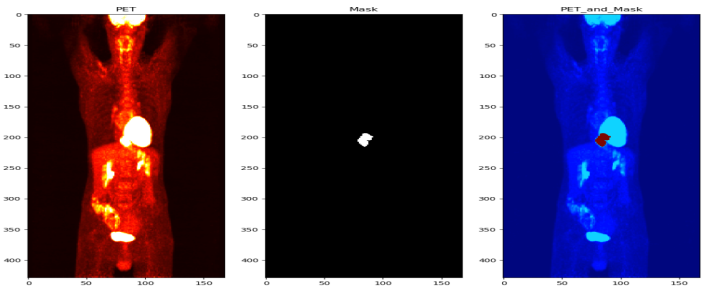


FIGURE 4.1 – An example of a PET image with oesophagus cancer on the left in which the tumor is barely visible, and the same image on the left in which the localisation of the tumor is shown in red color. It is not straightforward to learn the difference between tumor fixation and a normal fixation in a PET image.

learning way. Thanks to the explainability of neural networks [Zeiler and Fergus 2014] when making its decision, tumor pixels can be detected in an unsupervised way. Using only image-level labels to segment pixel-level image remains unexplored in PET images. To achieve this end, our strategy is to try to interpret how a neural network makes a classification decision.

The work on explanability and interpretability of neural network decision-making is an ongoing area of research. CNNs have yielded impressive results for a wide range of visual recognition tasks [Girshick et al. 2014, Krizhevsky et al. 2012], especially in medical imaging [Hannun et al. 2019, Rajpurkar et al. 2017]. As autonomous machines and black-box algorithms begin making decisions previously entrusted to humans, it becomes necessary for these mechanisms to explain themselves [Gilpin et al. 2018]. Many approaches for understanding and visualizing CNN have been developed in the literature [Mahendran and Vedaldi 2015, Zeiler and Fergus 2014, Zhou et al. 2014]. For instance, Zeiler et al [Zeiler and Fergus 2014] use deconvolutional networks to visualize the activation. The deep feature maps can be aggregated to extract class-aware visual evidence [Zhou et al. 2016]. However, when using fully connected layers for classification, the ability to locate objects in convolutional layers is lost. Different studies tried to solve this problem by using a fully convolutional neural networks (FCNs) such as Network in Network (NN) [Lin et al. 2013] and GoogLeNet [Szegedy et al. 2015]. Typically, conventional CNNs are first converted to FCNs to produce class response maps in a single forward pass. Although image-level class labels indicate only the existence of objects classes, they can be used to meaningful indices for image segmentation, called Class Attention Maps (CAMs) [Amyar et al. 2019a,

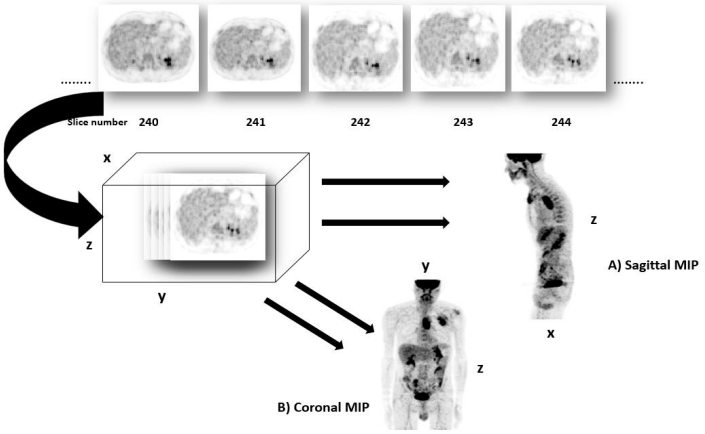


FIGURE 4.2 – Maximum intensity projection (MIP) of PET exam. A) projection in Sagittal. B) Projection in Coronal

Selvaraju et al. 2017, Zhou et al. 2016]. These class response maps can indicate discriminant regions in the image that make a CNN take a decision. However, it can not distinguish the different objects present in the image, which therefore makes precise segmentation difficult at the pixel level [He et al. 2017]. Different works have shown that although a CNN is trained to classify images, it can be used to localize objects at the same time [11, 12]. Zhou et al [10] demonstrated that CNNs can recognize objects while being trained for scene recognition, and that the same network can perform both image recognition and object localization in a single training. They have shown that convolutional units of different CNNs layers can behave as object detectors despite the lack of object labels.

[Ahn et al. 2019] presents an approach for instance segmentation using only image-level class as label. They trained an image classifier model, and by identifying seed areas of object from attention maps, a pseudo instance segmentation labels are generated, then, propagated to discover the entire object areas with precise boundaries. Zhou et al. reported that local maximums in a class response map correspond to strong visual cues residing inside each instance [Zhou et al. 2018b]. They create a novel architecture based on peak class response for instance segmentation using only the image-level label. First, a peak from a class response map is stimulated, then, back-propagated and mapped to highly informative regions of each object instance, such as instance boundaries. Works on WSL in medical imaging field play an important role due to the lack of annotations. However, the works on outcome prediction are limited. We propose here a WSL based on

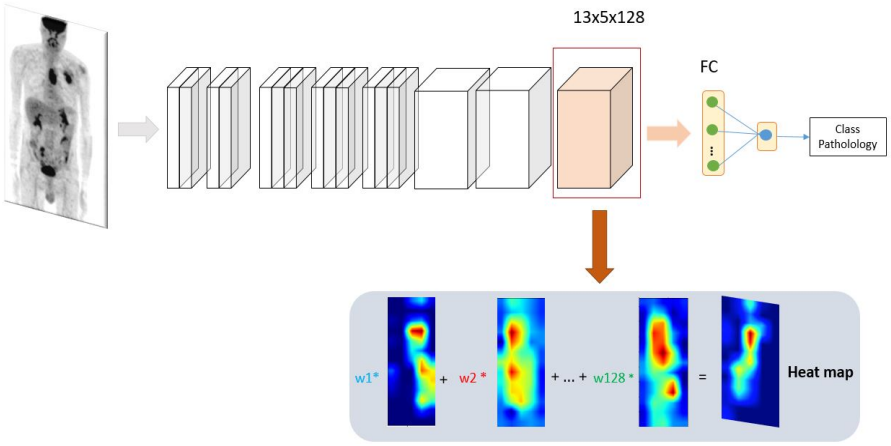


FIGURE 4.3 – The proposed architecture. The MIP is used to predict the pathology. A heat map is generated from the last conolutional layer

class attention maps to segment tumor region in PET images.

Once the tumor region is known, the outcome prediction can be performed on the tumor region. To this end, machine learning based methods are commonly used such as random forests (RFs) and support vector machines (SVMs) with or without a feature selection strategy [Cameron et al. 2015, Desbordes et al. 2017c, Leger et al. 2017]. The main disadvantage of these classical approaches is the need of an initial extraction of radiomic features using hand crafted methods, which usually yields a large number of features. In addition, hand-crafted features are affected by some parameters [Hatt et al. 2017] such as noise, reconstruction, etc. and significantly by the contouring methods used. Recent studies aim to develop classifiers based CNN which can automatically extract image features [Hosny et al. 2018, Zhou et al. 2018a]. Our work presented in the previous chapter has proved its effectiveness by comparing it to other methods [Amyar et al. 2019b]. In this work, we therefore use it to predict the response to treatment and the survival of patients.

Due to low PET image resolution, class attention maps cannot be directly used as supervision for pixel segmentation since they cannot distinguish between physiological fluorodeoxyglucose uptakes (normal fixation/ no tumor) and pathological uptakes (tumor), see Fig. 1. The main concern of this method for processing PET images is the difficulty of identifying only the tumor region, because certain other regions of the image can also be identified as participating in the classification decision due to their strong visual information. To resolve this problem, we integrate a weak prior knowledge. In this work,

we tackle the challenging problem of training CNNs with image-level labels for pixel-level tumor segmentation. We show that, by using a CNN architecture and its class attention maps integrated with a weak prior knowledge, we can transform the classification task into tumor localization in PET images. In this chapter, we present a new method for learning segmentation at the pixel level with class labels (at the image level) using the class response map, to make a CNN capable of segmenting the tumor at the pixel level but without pixel labels. To date, using only image-level labels to segment pixel-level image remains unexplored in PET images.

In addition, we propose to fully identify and locate tumor in 3D PET images from only two 2D MIP images with face and profile views, which allow to enormously reduce the complexity of the architecture and the learning time.

The chapter is organized as follow : in section 2, we describe our weakly supervised model model, explaining the CAM and the new loss function introduced. Section 3 presents the experimental studies. In section 4, we show the results of our work. Section 5 is for discussion and conclusion.

4.2 Material and methods

4.2.1 Main idea

Our method consists of two stages : segmentation of the tumor region and prediction of the treatment outcomes. The core of our method is to develop a new method to generate class activation maps to locate the tumor region. We propose a new loss function to improve the generation of class activation maps, and therefore to locate the tumor more precisely. First, to make class activation maps more relevant, we introduce prior knowledge. For each patient data we randomly define a point at the approximate center of the tumor, which can be achieved easily compared to the delineation of the tumor contours in 3D PET images. Then, we define a new loss function based on two loss terms : the accuracy to classify the type of tumor, and the distance between the generated class activation maps at the current iteration and the central point. To that end, an 8 layers CNN

is created to learn image-level labels and to generate an improved class activation maps to locate the tumors. After each feed forward of a mini batch of 8 images, a probability of belonging to a tumor class is obtained and then a binary cross-entropy loss function is calculated, noted $Lclass$, which is the first term. A A class activation map is generated for each image and a distance between the CAM and the central point in the tumor is then calculated, noted ($Ldistance$) which is the second loss term for tumor localization. Finally, the back-propagation is performed in respect to both $Lclass$ and $Ldistance$ to update the weights.

4.2.2 Maximum Intensity Projection

Maximum intensity projection (MIP) is a 2D image that represents 3D image for fast interpretation in clinical applications [Prokop et al. 1997]. Our idea is to use MIP to deal with 3D images, allowing in one hand to greatly reduce the complexity of the networks and avoids over-fitting due to the small size of the medical image data set, and on the other hand to keep useful 3D information for classification. Two MIPs calculated from opposite points of view are symmetrical images because they are rendered by orthographic projection. MIP imaging is used routinely by physicians in interpreting PET images. It can be used for the detection of lung nodules in lung cancer screening programs for example. MIP enhances the 3D nature of these nodules, making them stand out from pulmonary bronchi and vasculature [Valencia et al. 2006]. Considering the advantages of this technique which is also faster in terms of calculation, we can use it for the classification of images, to classify the different pathology such as lung cancer or esophageal cancer. However, the radiomics features obtained from MIP images are not rich enough to predict the outcome of treatment and survival, due to the loss of depth information (the third dimension). To obtain a 3D tumor region, we propose to use both sagittal and coronal MIPs. The intersection of these two orthogonal views allows us to define the region of interest in 3D, as shown in Fig. 2. Our strategy is to use 2D images to find 3D tumor region, which can speed up the tumor localisation in 3D. Indeed, instead of generating a 3D activation response map whose corresponding 3D network is time consuming and difficult to train with limited resources, we only design two 2D classifiers to generate two class activation

maps.

4.2.3 New Design of Class Activation Map

Interpreting machine learning models is a key element towards making it easier for physicians to embrace these methods. To interpret a convolutional neural network, we can produce class activation maps to detect the zones in images that contribute the most to the network classification decision. In this work, the classification involves classifying the PET images into two classes : the esophagus class in which the esophagus tumor is present in the images ; and the lung class in which the lung tumor is present. It is a key step in our method, since it will be used to recover the entire tumor area in a PET image. When a CNN, typically having a series of layers, classifies an MIP image, its first layers capture low-level features while later layers capture higher-level visual information that is relevant to the classification task. The last convolutional layer is flattened, and then passed to a fully connected layers to provide a certain probability of belonging to the oesophagus class or lung one, see Fig 3.

In a CNN based classifier, once the features are flattened, the spatial information is lost. Therefore, if we want to visualize locations of the features , we have to visualize the features with their locations before the flattening. We thus take the feature maps of the last convolutional layer to generate class activation map. However, these feature maps are much smaller in size than the input size. Typically, the width and high of a class activation map are 1/33 of that of the input image and the number of feature maps is the same as the output of the last layer (128). We note the total number of feature maps in the last layer by D. To go from these feature maps with size of 13 x 5 to a heat maps over the whole image, we need to unpack these feature maps. Let f^{i} be the ith feature map. For each feature map f^{i}, a weight w is associated to it, where i=1...D. Then, a pre-heat maps is obtained by adding each feature map multiplied by its weight as in 4.1 :

$$pre\_hmap = \sum_{i=1}^{D} [w^{\mathrm{i}} f^{\mathrm{i}}] \tag{4.1}$$

Each feature maps contains 13×5 elements (65 in total), where ${f_{\mathrm{j,z}}}^{\mathrm{i}}$ is (j,z) element

of the ith feature map, where j=1..13 and z=1..5. To obtain the wights w for each of these feature maps, we calculate the influence of $f^{i}_{j,z}$ on the output $\hat{y}$, by computing the partial derivative of $\hat{y}$ with respect to each feature in f_{i}, such as :

$$I = \frac{\partial \hat{y}}{\partial f^{i}_{j,z}} \tag{4.2}$$

Then, w^{i} is calculated by taking the average of the feature influences at each j, z position as in 4.3 :

$$w^{i} = \frac{1}{N} \sum_{j=1}^{J} \sum_{z=1}^{Z} \frac{\partial \hat{y}}{\partial f^{i}_{j,z}} \tag{4.3}$$

where N is the number of elements in the feature map, J is the width and Z is the height. Finally, we keep only features with positive influence. We apply ReLU function to keep only positive values. The heat map is finally obtained by :

$$h\_map = ReLU(\sum_{i=1}^{D} [w^{i} f^{i}]) \tag{4.4}$$

where $ReLU(X)$ is defined as :

$$Relu = max(0, X) \tag{4.5}$$

Because the heat map is generated at a low resolution of 13×5, we interpolate it to adapt it to the size of the MIP images. In our application, two different types of cancer, corresponding to two classes : lung cancer and oesophagus c ancer are considered. Let C denote the class $\in$ {lung, oesophagus}. From 4.1 and **??** we have :

$$w^{i}_{C} = \frac{1}{N} \sum_{j=1}^{J} \sum_{z=1}^{Z} \frac{\partial \hat{y^{C}}}{\partial f^{i}_{j,z}} \tag{4.6}$$

$$h\_map^{C} = ReLU(\sum_{i=1}^{D} [w_{C}{}^{i} f^{i}]) \tag{4.7}$$

The obtained heat maps will be used afterwards to calculate a new loss function in the classification step(see next section).

We introduce this novel loss function to prevent heat maps from further resolution drop. A large loss indicates that the current representation of the networks does not ac-

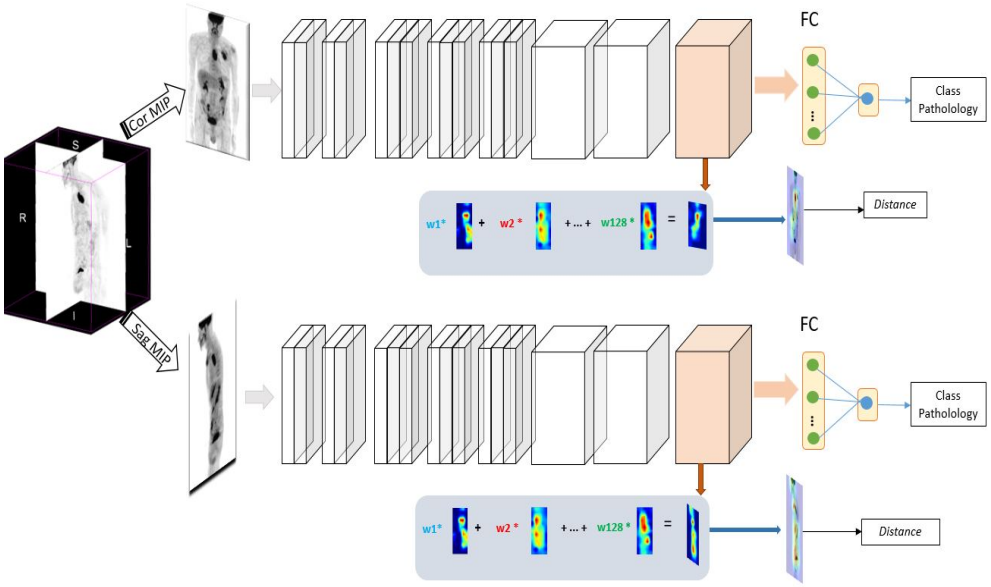


FIGURE 4.4 – Our proposed architecture. The neural network learns to classify the type of cancer from two 2D MIP images (sagittal and coronal). The generated heatmap is back-propagated and corrected to identify accurately tumor regions.

curately capture the lesion's visual patterns, and it is therefore necessary to provide an additional mechanism for self-improvement through back-propagation. The resulting architecture 4.4 is a novel convolutional neural network with an attention feedback, having an improved localisation capability.

4.2.3.1 Classification

A CNN consisting of a two Dense layers with 128 and 64 neurons respectively, is used in our classification step. The resulting set of feature maps, encloses the entire spatial local information, as well as the hierarchical representation of the input. Each feature map is flattened out, and all the elements are collected into a single vector V of dimension K, providing the input for a fully connected hidden layer, called h, consisting of H units. The activation of the $i^{(th)}$ unit of the h hidden layer is given by :

$$h_i = g(b_i + W h_i * V) \qquad \text{with } i = 1, ..., H. \tag{4.8}$$

A dropout of 0.5 and the activation function elu are used for learning. The last layer is a Dense layer with one neuron for image classification using a sigmoid activation. The

binary cross entropy is used as the loss function (Lclass) for classification :

$$Lclass = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{4.9}$$

where n is the number of patients, y is the cancer lung label (binary, 1 if the patient has lung cancer, 0 if it is oesophagus cancer) and $\hat{y}_{ij} \in (0,1) : \sum_j \hat{y}_{ij} = 1 \ \forall i,j$ is the prediction of a lung cancer presence.

4.2.3.2 Distance constraint using prior knowledge

As shown above, the class activation map depends on the derivation of feature maps. Since the patients' bodies have different widths, the areas where there is the contour of the body can also make the class activation map meaningful. To deal with this problem, we propose to use prior knowledge to construct a distance constraint. We assume that when CNN classifies the images, the decision is focused on the tumor region. This means that the class activation map must include the tumor region. Based on this prior knowledge, we randomly select a point approximately in the center of the tumor. Therefore, we define a distance constraint as our second loss term of classification.

The distance between the selected point p and the points in the generated heat maps, defined as follow :

$$Ldistance = \sqrt{\sum_{i=1}^{m} |q_i - p|} \tag{4.10}$$

where q_i notes a point i and m is the number of points in a heatmap. This second loss function makes it possible to correct the errors of the heat maps generated through the distance constraint. In fact, instead of focusing on the discriminating regions, which may include information other than the location of the tumor for classification, the heat map is regularized with the distance constraint to emphasize the region of the tumor and at the same time keep a good classification (see Fig 5).

The global loss function (loss glob) for the 2 tasks is defined by :

$$loss\_glob = Lclass + \alpha Ldistance \tag{4.11}$$

where α is a constant weight coefficient. We take alpha = 1 in our study. As the class activation map has low resolution, it does not accurately capture visual patterns of the lesion, and it is therefore necessary to provide an additional mechanism for self-improvement by backpropagation. The resulting architecture (see Figure 4.4) is a novel convolutional neural network with attention feedback based on the proposed loss function. This can greatly improve the locating ability.

4.2.4 Segmentation

Once we obtain the heat maps for sagittal and coronal MIP views, we retrieve the lesions mask on the 3D image. Sagittal MIP allows to retrieve y and z axis, and coronal MIP the x and z axis. Combining the 3 coordinates finally results in the 3D volume of the tumor, see figure 4.6.

4.2.5 Prediction

Once we obtain the 3D tumor region, we conduct a radiomics analysis to predict patient survival and treatment outcome. We use 3d-rpet-net [Amyar et al. 2019b], a CNN classifier based on two 3D convolutional layers and two fully connected layers to conduct radiomics analysis (see figure 4.7). The same model is applied on both 3D volumes manually segmented by a physician and automatically segmented by our method in order to compare their performances.

4.3 Experiments

4.3.1 Dataset

Patients underwent a whole body FDG PET/CT, at the initial stage of the pathology and before any treatment. The PET/CT data were acquired on the same device, and with the same acquisition and reconstruction procedure used in routine care, and presented in the above chapter. The reconstructed exam voxel size was $4.06 \times 4.06 \times 2.0$ mm^3 and were spatially normalized by re-sampling all the dataset to an isotropic resolution of $2 \times 2 \times 2$

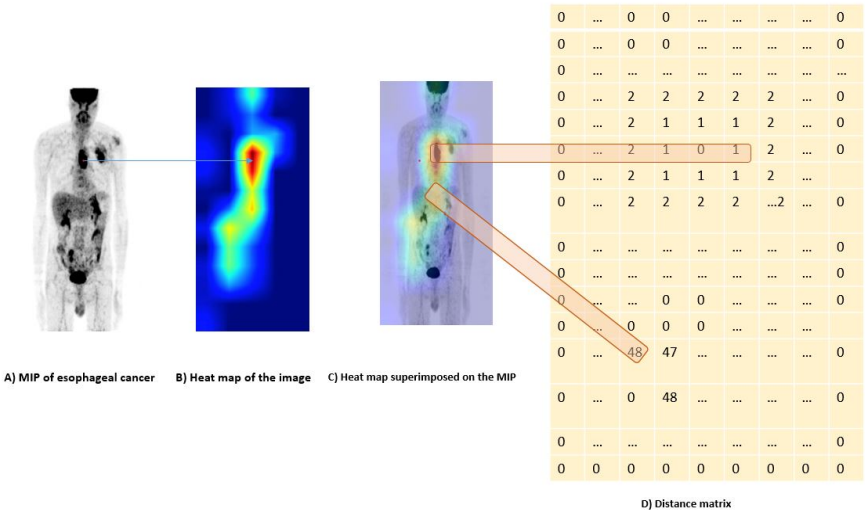


FIGURE 4.5 – Distance matrix between p at the center of the tumor and the points qi generated by the heat map. A) is a Coronal MIP for a patient with esophageal cancer. A point p is randomly defined at the tumor region. B) is the heat map generated using our proposed model. C) shows the overelay of the MIP and the heat map. D) is the distance matrix showing the distance between the points qi generated by the heat map and the point p.

TABLEAU 4.1 – Results for 3D segmentation. WPk : without prior knowledge. CAM : class activation map.

	Method	Dice
Oesophagus cancer	U-NET	0.42±0.16
	CAMsWPK	0.53±0.17
	Ours	**0.73±0.09**
Lung cancer	U-NET	0.57±0.19
	CAMsWPK	0.63±0.14
	Ours	**0.77±0.07**

mm^3 using the k-nearest neighbor interpolation algorithm.

4.3.2 Setup

We firstly generated maximum intensity projection (MIP) for coronal view and for sagittal view. MIP is a 2D image that summarizes 3D images for fast interpretation. Tumor gray level intensities were normalized to have SUV level between [0 30] and then translated between [0 1] to be used in CNN architecture. The neural network is trained to classify the type of cancer : oesophagus vs lung cancers. For each mini-batch, CAMs are generated, backpropagated and corrected via a distance function (3), to differentiate tumor regions from normal regions. Then, the two resulted corrected CAMs, for face and profile view are combined to retrieve the 3D tumor.

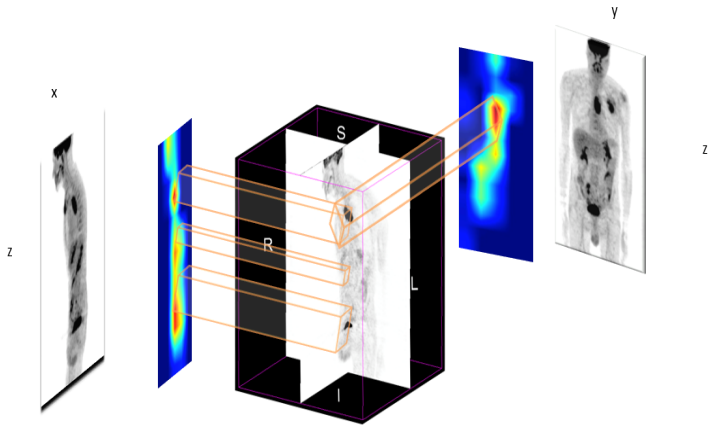


FIGURE 4.6 – Segmentation : the 3D tumor region from the two 2D heat maps. Coronal heat map allows to retrieve y and z axis, while sagittal heat map return x and z axis. The tumor is selected by the intersection of the two heat maps.

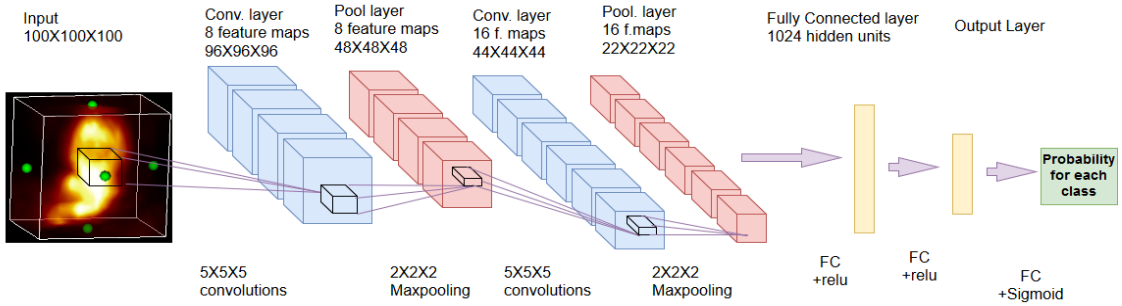


FIGURE 4.7 – 3D RPET-NET architecture composed by two 3D convolutional layers followed by 3D pooling layers and two dense layers.

Two experiments are conducted to evaluate our model.

Experiment 1 : The first experiment consisted of segmenting the lesions on the 3D PET images for patients with oesophagus cancer and lung cancer, using only 2D MIPs. The results were compared to the state of the art method U-NET [Ronneberger et al. 2015], which is commonly used in medical imaging for fully supervised segmentation, and CAMs without prior knowledge.

Experiment 2 : The second experiment consists of radiomics analysis. We predict the treatment survival for oesophagus cancer, and patient's survival for lung cancer. The response to treatment was evaluated 3 months after the end of treatment, and the overall survival (OS) used for the prognostic study was estimated at 3 years after the end of the treatment.

4.3.3 Implementation

The model was implemented using python with pytroch deep learning library, and trained for 2 days on nvidia p6000 quadro GPU with 24gb.

4.4 Evaluation Methodology

We divide the dataset into 3 groups : training, validation and test. For a fair comparison, all the methods were trained, validated and tested with the same group of data. The performance of the models were evaluated using the dice coefficient for the segmentation task, and the accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the ROC curve (AUC) for the classification, such as :

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.12}$$

where TP is the true positives, FN is the false negatives and TP + FN is the number of patients classified positively.

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.13}$$

where TN is the true negatives, FP is the false positives and TN + FP is the number of patients classified negatively.

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{4.14}$$

4.5 Results

Table 1 shows results for of tumor segmentation for both oesophagus and lung cancers. Different methods were compared to our proposed model with : U-NET using fully supervised learning, and CAMs without prior knowledge.

Table 2 shows results of radiomic analysis, for the prediction of patient's treatment, 3 months after the end of radiochemotherapy for oesophagus cancer, and the prediction

TABLEAU 4.2 – Results for radiomics analysis. WPk : without prior knowledge. Ms : manual segmentation

	Method	Accuracy	Sensibility	Specificity	AUC
Oesophagus cancer	CAMsWPK	0.57±0.03	0.61±0.28	0.56±0.24	0.53±0.26
	MS	**0.72±0.08**	0.79±0.17	**0.62±0.21**	**0.70±0.04**
	Ours	0.69±0.04	**0.80±0.14**	0.59±0.26	0.67±0.08
Lung cancer	CAMsWPK	0.61±0.07	0.59±0.21	0.57±0.15	0.55±0.24
	MS	**0.68±0.17**	**0.72±0.09**	0.54±0.07	**0.61±0.03**
	Ours	0.65±0.05	0.65±0.18	**0.58±0.15**	0.59±0.04

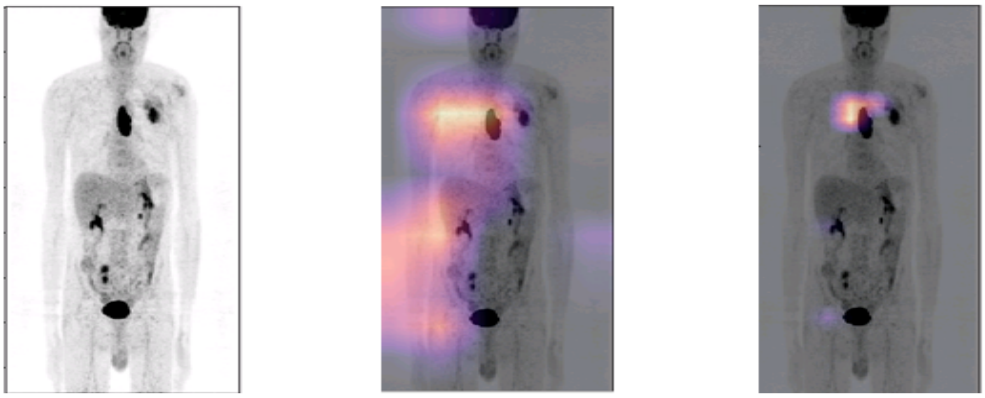


FIGURE 4.8 – Comparison between different models. From left to right : PET exam, CAMs without prior knowledge, ours

of survival for patients with lung cancer.

Fig 8 shows a CAM of one patient. We can see the improvement with the distance constraint.

All the methods were compared based on the ability to detect accurately the tumor and to conduct a radiomics analysis. The performances are measured by accuracy, sensibility, specificity and the area under the ROC curve. The results were obtained using a 5 fold cross-validation. Best results for segmentation were obtained using our proposed model for both lung and oesophagus cancer. For radiomics, 3d-rpet-Net with manual segmentation was not statistically significantly different from our model (p=0.59) for oesophagus and (p=0.63) for lung. Our model tend to have a better sensibility for oesophagus and a better specificity for lung cancer with no significant differences. This means that our method, which does not need ground truths, can obtain similar results as using manual segmentation. This is very encouraging for the automatic radiomics analysis.

4.6 Discussion & Conclusion

In this study, a new weakly supervised learning model was developed to localize lung and oesophagus tumors in PET images. It utilizes two fundamental components : a new class activation map to locate the tumor and a new loss function to improve localisation precision. The model could detect tumors with better accuracy compared to fully supervised models such as U-NET, or classical CAMs, see Fig 6. Our model outperformed other methods in terms of the dice index. As for radiomics analysuis, 3d-rpet-net with manual segmentation is showing slightly better results than our model in radiomics analysis. However, it is based in manual pixel-level annotations of tumor, which requires a physician expert and also is time consuming.

By detecting the tumor with 2D MIP images for face and profile views, we can obtain x,y and z coordinates to segment the 3D image. The segmentation in the 3D images were used to conduct a radiomics analysis with state-of-the-art results. This simple and yet powerful technique, can be integrated in future workflow/software dedicated to automatic analysis of PET exams to conduct radiomics analysis.

Article Details :

— **Weakly Supervised Tumor Detection in PET Using Class Response for Treatment Outcome Prediction.** Amine Amyar, Romain Modzelewski, Pierre Vera, Vincent Morard, Su Ruan. Under review.

Other Related Publications :

— **Contribution of class activation map on WB PET deep features for primary tumour classification.** Amine Amyar, Pierre Decazes, Su Ruan, Romain Modzelewski. Journal of Nuclear Medicine, 60(supplement 1), pp.1212-1212.

— **Contribution des cartes d'activation de classe des réseaux de neurones profonds pour la classification des tumeurs primaires en TEP-FDG.** Amine Amyar, Su Ruan, Pierre Decazes, Romain Modzelewski. Médecine Nucléaire, 44(2), p.133.

Chapitre 5

Multitask learning for radiomics analysis

Sommaire

5.1 Introduction 88

5.2 Related Work 90

5.3 Method 94

5.3.1 Model description 94

5.3.2 Dataset 98

5.3.3 Implementation 99

5.4 Experiments 99

5.5 Validation Methodology 100

5.6 Results 102

5.7 Discussion 106

5.8 Conclusion 108

In the previous chapter we showed the interest of weakly supervised learning to detect automatically the lesions, in order to perform a radiomic analysis afterward. In this chapter, we introduce the multi-task learning (MTL) framework, where neural networks is trained to conduct several tasks in the same time. In this MTL approach we are interested of learning segmentation not to be exactly as the physician ground truth, but to let the neural network decides which are the regions that contribute the most in the outcome prediction.

5.1 Introduction

Radiomic is a field of study where images have great potential for precision and personalized medicine [Aerts et al. 2014, Lambin et al. 2012a]. It is defined as the extraction of a large number of features from medical images such as computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography (PET) [Kumar et al. 2012]. These features are used to uncover disease characteristics that fail to be found or quantified by the naked eye. The first step in radiomic analysis in oncology is the lesion segmentation (see figure 5.1). This task requires a highly trainable physician, is time consuming and the ground truth defined is physician subjective. Recently, deep learning showed very promising results in image classification [Ciregan et al. 2012], object detection [Szegedy et al. 2013], and image segmentation [Badrinarayanan et al. 2017]. In the medical imaging field, various applications have emerged in different areas, including pathology classification [Janowczyk and Madabhushi 2016], treatment response prediction [Amyar et al. 2018], lesions segmentation [Kamnitsas et al. 2017] and organs at risk segmentation [Trullo et al. 2017]. Thus, artificial intelligence in general and deep learning in particular can come in handy to develop computer aided diagnostic applications (CAD). However, deep learning approaches are well known for their data hungry nature, and annotated data are usually hard to obtain in the medical imaging field. Recent works tried to tackle this problem with a weakly supervised learning strategy to segment the lesions, and then predict the outcome [Amyar et al. 2020]. This approach showed very promising results outperforming state of the art supervised approaches such as U-Net [Ronneber-

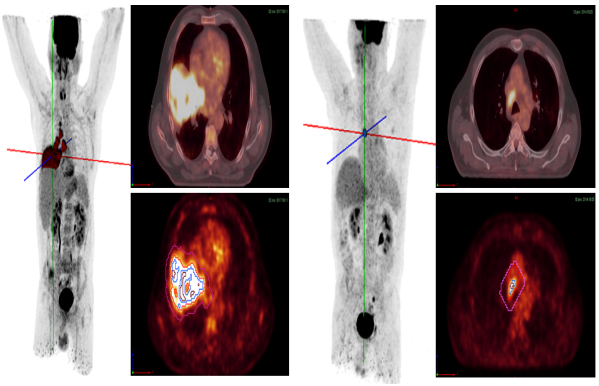


FIGURE 5.1 – Columns from left to right : Fused PET/CT slice, zoomed on the lung tumor (left) and esophageal tumor (right) seen on FDG-PET only. Metabolic Tumor Volume MTV (40% SUVmax thresholding) in red. MTV3 (MTV + 3 cm isotropic margin) include the tumor and peritumoral region.

ger et al. 2015] for image segmentation, and comparable results with supervised learning [Amyar et al. 2019b] for radiomics analysis. However, the drawback of this method is the two stage segmentation-outcome prediction. In addition, recent studies have shown the potential of peritumoral regions on boosting the accuracy of outcome prediction [Braman et al. 2019, Dou et al. 2018, Prasanna et al. 2017]. Thus, the association of the intratumoral and peritumotal regions provides rich information for radiomic analysis [Braman et al. 2017, Hu et al. 2020].

The standard method in machine learning is to learn one task at a time. Large problems are broken into small subproblems that are learned separately and then recombined. Multi-task learning (MTL) [Caruana 1997] is a type of learning algorithm that aims to combine several pieces of information from different tasks in order to improve the model's performance and its ability to better generalise [Zhang and Yang 2017]. The basic idea of MTL is that different tasks can share a representation of common characteristics [Zhang and Yang 2017], and thus train them jointly. The use of different data sets from different tasks allows learning an efficient representation of the common characteristics of all tasks, because all data sets are used to obtain it, even if each task has a small data set, thus improving the performance of each task.

Contribution : In this work, we tackle the challenging problem of training a neural network to classify the pathology, segment the lesion, reconstruct the image and predict the outcome based on the segmentation results. We believe that the global information

in the entity image volume describing the relationship between the tumor and other organs is also useful as the characteristics of the tumor. We show that, by using a multi-task learning approach, we can boost the performance of radiomics analysis while extracting rich information of intratumoral and peritumoral regions. More specifically, we present a new method of learning segmentation not to be exactly as the physician ground truth, but to let the neural network decides which are the regions that contribute the most in the outcome prediction. Our main contributions are summarized as follows :

1. Our proposed architecture is the first to use jointly global features extracted from entire image and local features from tumor regions to predict the outcome in a radiomics study.

2. We design a new multi-tasking learning network to jointly segment the tumor on a 3D PET image and predict the outcome, which is simultaneously associated with two subsidiary tasks, classification and reconstruction. The last two tasks are added to make the features more relevant and also to serve as an inductive bias to better generalize.

3. We utilize a multi-scale feature extraction so that the model can predict the outcome from tumor and tumor neighborhoods features, and also global features at the encoder level.

4. We conduct extensive validation strategy with multiple ablation experiments, comparison with state of the art methods in both supervised and multi-task learning.

5.2 Related Work

In previous studies, several methods for segmentation of the region of interest and joint classification have been proposed. For instance, Yang et al. [Yang et al. 2017] created a multi-task deep neural network for skin lesion analysis, in order to solve different tasks

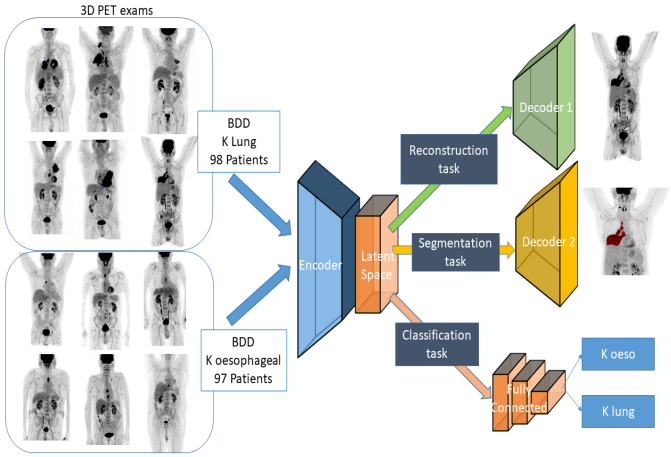


FIGURE 5.2 – Hard parameter sharing for multi-task learning in deep neural networks used in our proposed architecture.

simultaneously such as lesion segmentation and two independent binary lesion classifications. The MTL model improved learning efficiency and prediction accuracy for each task, in comparison to single task models. They achieved an average Jaccard score of 0.724 for lesion segmentation, while the average values of the area under the receiver operating characteristic curve (AUC) on two lesion classifications are 0.880 and 0.972, respectively. The model consists of a common encoder for the 3 tasks based on GoogleNet [Szegedy et al. 2015], one decoder for segmentation and two fully connected branches for classification. In [Asgari et al. 2019] Asgari et al. proposed a multi-class segmentation as multi-task learning for drusen segmentation in retinal optical coherence tomography. The model is based on a multi-decoder architecture that tackles drusen segmentation as a multi-task problem. Instead of training a multi-class model for two classes segmentation, they used one decoder per target class and an extra task for the area between the layers. They used connections between each class-specific branch and the additional decoder to increase the regularization effect of this surrogate task. The model was validated on a dataset of 366 images. They achieved a mean dice of 0.73 compared to 0.68 with multi-class U-Net or 0.66 with a binary U-Net.

In [Thome et al. 2019] Thome et al. proposed a multi-task classification and segmentation model for cancer diagnosis in mammography. The architecture is based on a fully convolutional networks (FCN) [Long et al. 2015]. The model was evaluated on the DDSM database [Heath et al. 2000] with cancer classification and pixel segmentation with five

classes. They showed that the model could learn shared representations that are beneficial for both tasks when trained in MTL approach compared to STL. The model achieved a mean dice of 38.28% and an AUC of 84.02% compared to a mean dice of 34.98% and an AUC of 81.37% for STL. In [He et al. 2020] He et al. used a multi-task learning approach for the segmentation of organs at risk with label dependence. They used a MTL to accurately determine the contour of organs at risk in CT images. They used an encoder-decoder framework for two tasks. The main task is the segmentation of organs, while the secondary task is the multi-label classification of organs.

While previous studies showed the advantage of using MTL compared to single task U-Net for image segmentation, recent works have shown the benefit of using U-Net, V-NET [Milletari et al. 2016] or Faster-RCNN [Ren et al. 2016] as the backbone network. In [Playout et al. 2018], Playout et al. proposed an extension to U-Net architecture relying on multi-task learning with one common encoder, and two decoders to jointly detect and segment red and bright retinal lesions which are essential biomarkers of diabetic retinopathy. At the encoder level, they used residual connections at every scale, mixed pooling for spatial compression and large kernels for convolutions at the lowest scale. Segmentation results are refined with conditional random fields (CRF) and the model is trained with Kappa-based function loss. They achieved a sensitivity of 66,9% and a specificity of 99,8% on a public dataset.

In [Vesal et al. 2018a] Vesal et al. proposed a multi-task framework for Skin Lesion Detection and Segmentation. The model is based on Faster-RCNN to generate bounding boxes for lesion localization in each image, and "SkinNet" [Vesal et al. 2018b], which is a modified version of U-Net. The model was trained and evaluated on ISBI 2017 challenge and the PH2 datasets, outperforming other STL methods in terms of Dice coefficients (0.93), Jaccard index (0.88), accuracy (0.96) and sensitivity (0.95), across five-fold cross validation experiments. In [Zhou et al. 2020] Zhou et al. used an MTL framework for segmentation and classification of tumors in 3D automated breast ultrasound images. The main motivation behind their work is the correlation between tumor classification and segmentation, therefore learning these two tasks jointly may improve the outcomes of both tasks. The framework is based on an encoder-decoder network for segmentation

and a light-weight multi-scale network for classification, with VNet as the backbone.

These above methods cannot be directly applied to 3D PET images to jointly segment the lesion, classify the pathology and predict the outcome. For instance, the tumor boundaries in PET images for esophageal cancer are not well defined, and sometimes hard to separate from other normal fixation (no tumor). In addition, peritumoral which is defined as the pathology around the tumor is an important information that can boost the prediction accuracy, but it is not taken into account with previous and classical approaches. Finally, due to the variation in size of the tumors, a mutli-scale approach could be a benefice to capture small features as well as investigating bigger ones. In this work, we take advantage of previous proposed methods and propose a new architecture for radiomics analysis. The main tasks are outcome prediction and lesion segmentation, and the secondary tasks are image reconstruction and pathology classification. We propose a multi-scale feature learning for the outcome prediction, by jointly predicting on the local features and global ones.

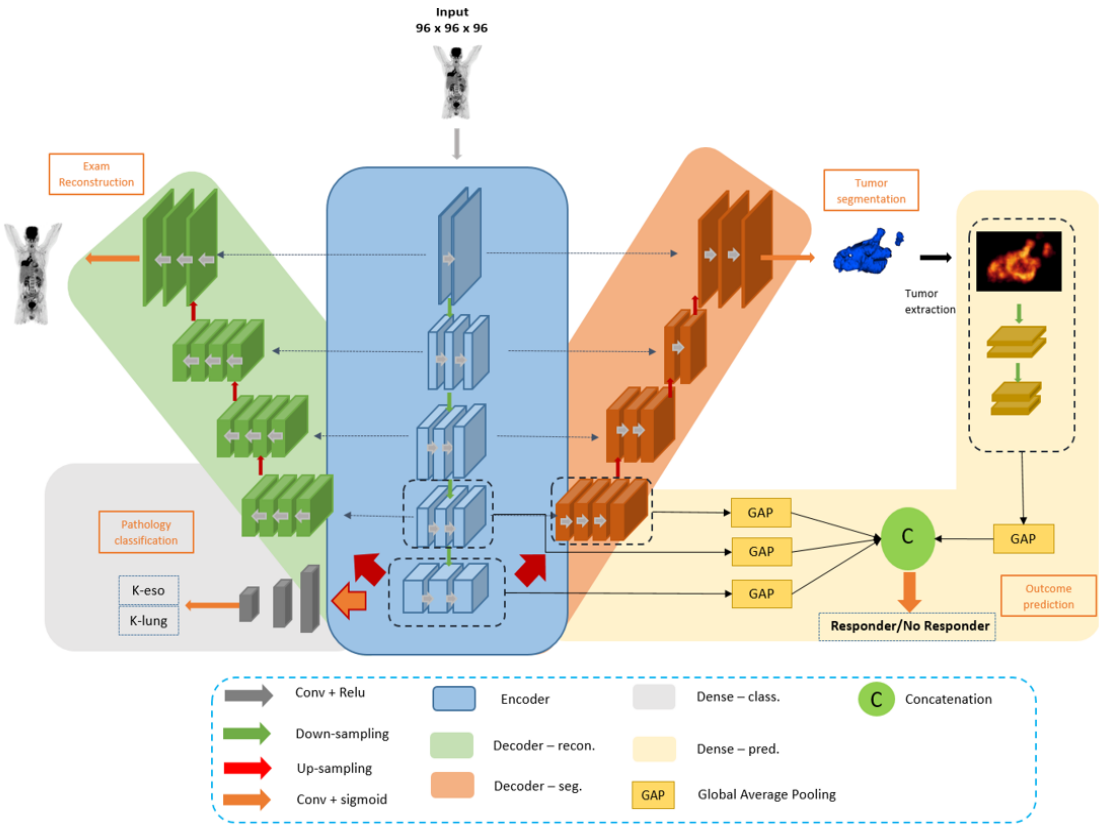


FIGURE 5.3 – Our proposed architecture, composed of an encoder and two decoders for image reconstruction and tumor segmentation. A fully connected layers are added for classification (Oesophageal vs lung cancer), and a multi-scale outcome prediction.

5.3 Method

We propose a new multi-task learning algorithm to improve generalization by leveraging the domain-specific information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better. Two major strategies are used when training a MTL algorithm, hard parameters sharing [Caruana 1997] or soft parameters sharing [Ruder 2017]. Hard parameter sharing is the most commonly used approach to MTL in neural networks and greatly reduces the risk of overfitting [Ruder 2017], see figure 5.2. It is generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers. In this work we utilize hard parameters sharing due to its great performance and wide utilization.

The reconstruction and pathology classification are extra tasks that serve as an inductive bias. The power of MTL framework lay in the fact that it is able to determine how tasks are related without being given an explicit training signal for task relatedness. We make the assumption that since the four outputs share common hidden layers, it is possible for internal representations that arise in the hidden layer for one task are used by other tasks.

5.3.1 Model description

We propose a new architecture called W-Net to jointly segment the lesion, classify the pathology, reconstruct the image and predict the outcome. The proposed network is shown in figure 5.3. We use U-Net as the backbone due to its great performance in 3D medical image segmentation. The W-Net architecture consists of four parts : (i) a common encoding part, (ii) a decoding part for reconstruction, iii) a decoding part for the segmentation and (iii) skip connections between them, which form a W, see figure 5.3. To that we add a multi-layer perceptron (MLP) for the classification task, and a convolutional neural network for the outcome prediction based on the segmentation result. Finally, we use multi-scale approach to feed global features to the CNN, in order to predict make a prediction on both global features and tumor ones. To summarize, many classic image classification networks use transfer learning [Pan and Yang 2009] to extract high level features

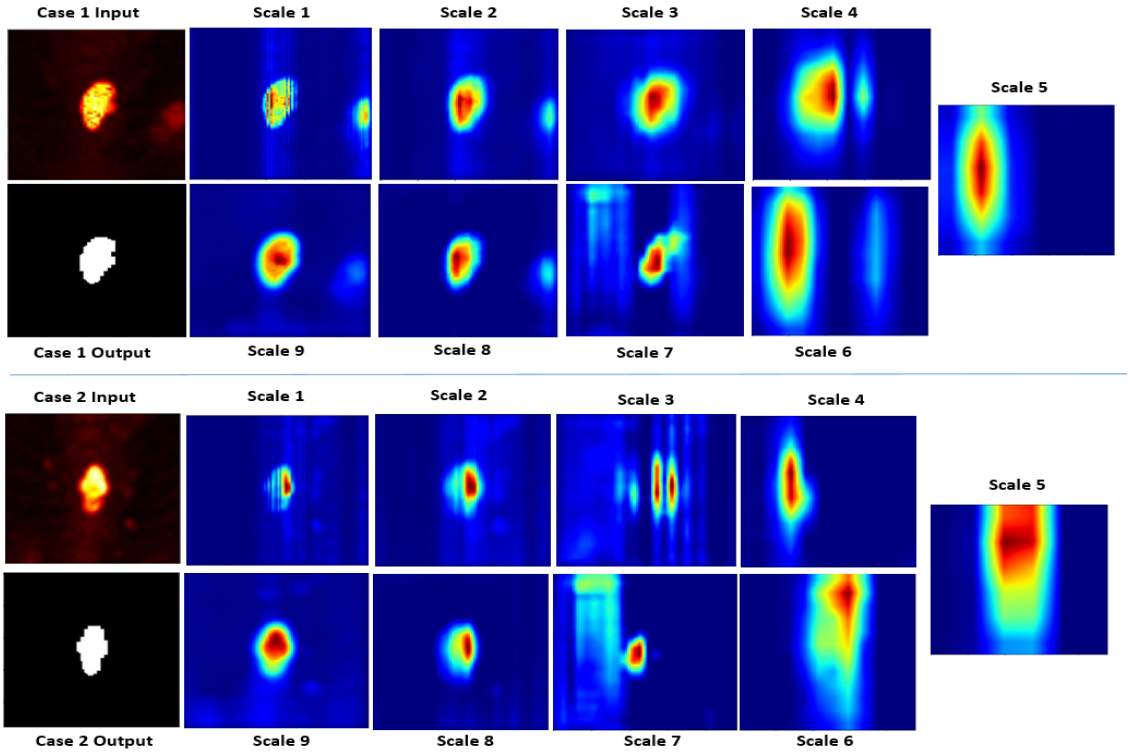


FIGURE 5.4 – Heatmap from different scales with two different input.

from CNN models, such as VGG16 [Simonyan and Zisserman 2014] or ResNET [He et al. 2016]. Motivated by this, we use the same encoder for lesion segmentation and pathology classification to to extract common features. We add reconstruction task as a secondary task so that the neural network can extract meaningful features about PET images.

5.3.1.1 Encoder-Decoder

The encoder is used to obtain the disentangled feature representation. It is a 10-layers 3D convolutional neural networks with convolution filters of $3 \times 3 \times 3$ and a maxpooling of $2 \times 2 \times 2$ after each 2 convolutional layers and a skip connection. The number of feature maps increases from 64 for the 2 first layers to 1024 for the last ones. We use $relu$ activation function and a Dropout of 0.5 after the last convolutional layer. The structure of the 2 decoders is the same, with upsampling to return to the original image size followed by convolutional layers to reduce the number of features by a factor of 2. These features are concatenated with the ones from the corresponding level of the encoder.

5.3.1.2 Multi-scale Feature Extraction

For the outcome prediction, we take advantage of both local features and global features. Local features are extracted from the segmentation result, while the global features are obtained from the common encoder. To benefit from features of different scales, we designed a multi-scale feature concatenation model for the radiomics task, as shown in figure 5.4. We concatenate feature maps from Level 3 to 5 in the encoder with the convolutional network in the outcome prediction. As a strong tool to evaluate and analyze the decision made by the neural network, we visualize heatmaps at different levels of the encoder and the decoder for the segmentation. To visualize the heatmaps, we use Grad-cam technique [Selvaraju et al. 2017] to produce visual explanation at each scale. We can observe that scale 4, 5 and 6 extract rich features at the tumor level and beyond, including peritumotal regions and other important fixations. To incorporate this information at the prediction level we design a multi-scale feature concatenation model by fusing feature maps from scale 4, 5 and 6 with the tumor features. We use a channel-wise global average pooling (GAP) as in [Zhou et al. 2020] to reduce the complexity in training time and to keep also important features, since it is more robust to spatial translation.

5.3.1.3 The reconstruction task Task1

We trained the model with a linear activation for the output and a mean squared error for the loss function (Lrecon) and used accuracy as the metric :

$$Lrecon = \frac{1}{n}\sum_{t=1}^{n}(y\_true - y\_predict)^2 \tag{5.1}$$

where y_true is the true label and y_predict is the predicted label.

5.3.1.4 The segmentation task Task2

We used the same architecture as the reconstruction except for the activation function for the output, which is a sigmoid. The loss function is based on the dice coefficient loss (Lseg) which is considered as the metric :

$$dice\_coef = \frac{2 * |X \cap Y| + \epsilon}{|X| + |Y| + \epsilon} \tag{5.2}$$

$$Lseg = -dice\_coef \tag{5.3}$$

where ϵ is the the smoothing factor and used to avoid a division by zero.

5.3.1.5 The classification task Task3

The resulting set of feature maps, encloses the entire spatial local information, as well as the hierarchical representation of the input. Then, each feature map is flattened out, and all the elements are collected into a single vector V of dimension K, providing the input for a fully connected hidden layer, called h, consisting of H units. The activation of the $i^{(th)}$ unit of the h hidden layer is given by :

$$h_{\mathrm{i}} = f(b_{\mathrm{i}} + W h_{\mathrm{i}} * V) \qquad \text{with } i = 1, ..., H. \tag{5.4}$$

In details, the output of the encoder is a tensor of mini_batch x 32 x 32 x 1024 to which we add a convolutional layer followed by a maxpooling, and then a flatten operation to convert the data to a mono-dimensional tensor to perform the classification. The multi-layer perceptron consist of a two Dense layer with 128 and 64 neurons respectively, with a dropout of 0.5 and the activation function elu. The last layer is a Dense layer with three neurons for image classification using a sigmoid activation and a binary cross entropy is used as the loss function (Lclass) :

$$Lclass = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{5.5}$$

which is a special case of the multinomial cross-entropy loss function for m = 2 :

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log(\hat{y}_{ij}) \tag{5.6}$$

where n is the number of patients and y is the class label (esophageal cancer, lung cancer).

5.3.1.6 The prediction task Task4

The prediction branch is connected to three layers from the encoder and segmentation decoder, in order to incorporate global features, in addition to tumor features extracted from the segmentation result. It is composed of 2 convolutional layers with 64 feature maps each followed by a maxpooling and 2 other convolutional layers with 128 feature maps each. Then, we apply a global average pooling to concatenate tumor based features (local features) with encoder-decoder global features in a multi-scale. Finally, three fully connected layers are used for the prediction with 128, 128 and 1 neurons respectively. The loss function is the binary cross-entropy and the performance metric is the accuracy :

$$Lpredict = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{5.7}$$

where n is the number of patients and y is the outcome.

5.3.1.7 multi-task Loss Function

We use reconstruction task to learn more meaningful features of PET exams, and outcome prediction task so that the network will focus attention on the most discriminator regions for the segmentation task so that. In our experiments, the Adam optimizer [Kingma and Ba 2014] algorithm was used with a mini batches of 4 and a learning rate of 0.0001. The global loss function (loss glob) for the 4 tasks is defined by :

$$loss\_glob = \alpha Lrecon + \beta Lseg + \omega Lclass + \lambda predict \tag{5.8}$$

where $\alpha = \beta$ and ω= 1-($\lambda + 2 \times \alpha$). Our model was trained for 1500 epochs with an early stopping of 70.

5.3.2 Dataset

Our experiments were run on 195 PET image volumes with lung (98) and oesophagus (97) cancer, from Henri Becquerel Center, Rouen, France. All patients underwent whole body FDG PET with a CT (baseline PET), at the initial stage of the pathology and before

any treatment. The PET/CT data were acquired on the same device, and with the same acquisition and reconstruction procedure used in routine care. The reconstructed exam voxel size was $4.06 \times 4.06 \times 2.0$ mm^3 and were spatially normalized by re-sampling all the dataset to an isotropic resolution of $2 \times 2 \times 2$ mm^3 using the k-nearest neighbor interpolation algorithm. We split the data into 2 groups to train and test the deep learning methods. One group was used for training the models (77 Oeso and 78 Lung) and one locked group for testing (40 patients). Furthermore, for the CNN, the training samples were split into 2 groups, a train set (57 Oeso and 58 Lung) and a validation set (40 patients).

5.3.3 Implementation

All models were implemented using python and keras deep learning library, with tensorflow as backend, and trained on nvidia p6000 quadro gpu with 24gb. Some tested state of the art models were developed using pytorch library.

5.4 Experiments

We compare the performance of single task learning (STL—learning just one task at a time) and multi-task learning. We present an empirical test that rules out these mechanisms and thus ensures that the benefit from MTL is due to the information in the extra tasks.

Experiment 1 : The first experiment consists of the optimization of the network by testing the different combination of tasks. The models developed include single task models, 2 and 3 tasks models, and all tasks models. Reconstruction and pathology classification are secondary tasks, thus they are combined either with segmentation or outcome prediction or both of them. Also, outcome prediction with and without global or local features were evaluated. In total, 15 models were developed for the outcome prediction.

Experiment 2 : The second experiment is to evaluate the performance of the best model with state of the art methods for image segmentation such as U-Net, V-NET and a weakly supervised multi-task approach [Amyar et al. 2020]. The WSL-MTL model uses a priori knowledge by defining two points in two 2D maximum intensity projection (MIP)

images for coronal and sagittal views. The model learn to classify the two MIPs into lung and esophageal cancers, and by generating a class activation map (CAM), it calculate a distance between the CAM generated and the two points defined, and learn to minimize the distance between the CAMs and the 2 points in a multi-task learning approach. Finally, the corrected CAMs for sagittal and coronal views are used to retrieve the tumor in the 3D space.

Experiment 3 : The third experiment is to compare our models with state of the art methods for image classification and outcome prediction. We use : Alexnet [Krizhevsky et al. 2017], VGG-16 [Simonyan and Zisserman 2014], ResNET50 [He et al. 2016], 169-layer DenseNet [Huang et al. 2017] and InceptionV3 [Szegedy et al. 2016]. We compare also our results with deep radiomics such as 3D RPET-NET [Amyar et al. 2019b] and a 6 layers 3D convolutional neural network.

Experiment 4 : In experiment 4 we study the effects of the hyperparameter λ on the multitask learning. We have tested different values : 0.1, 0.3, 0.5, 0.7 and 0.9.

Experiment 5 : Finally, we compare our proposed method with state of the art multi-task methods, including [Zhou et al. 2021],[Wang et al. 2018], [Qu et al. 2019] and [Chen et al. 2018]. We extended 2D networks to 3D. In order to incorporate both local and larger contextual information, we employ a multi-scale feature extraction for outcome prediction.

5.5 Validation Methodology

The performances of the models were evaluated using the dice coefficient for the segmentation task, and the area under the ROC curve (AUC) and the accuracy (Acc), for both classification and prediction.

TABLEAU 5.1 – Results of experiment 1 : segmentation, classification and prediction results from different scenarios, for esophageal and lung cancers. Task1 : reconstruction, Task2 : segmentation, Task3 : pathology classification, Task4 : outcome prediction.

				Seg.	**Class.**		**Pred.**	
bf Type of cancer	**Tasks**	**Global features**	**Tumor features**	**Dice_coef**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
Esoph.	Task1 & Task4	✓	x	/	/	/	0.60	0.60
	Task1 & Task3 & Task4	✓	x	/	0.95	0.94	0.65	0.68
	Task1 & Task2 & Task4	✓	x	0.77	/	/	0.65	0.64
	Task1 & Task2 & Task4	x	✓	0.76	/	/	0.65	0.63
	Task1 & Task2 & Task4	✓	✓	0.74	/	/	0.70	0.63
	Task2 & Task4	✓	x	0.73	/	/	0.70	0.71
	Task2 & Task4	x	✓	0.70	/	/	0.70	0.74
	Task2 & Task4	✓	✓	**0.79**	/	/	0.73	0.72
	Task2 & Task3 & Task4	✓	x	0.71	0.94	0.93	0.72	0.70
	Task2 & Task3 & Task4	x	✓	0.69	0.91	0.92	0.70	0.71
	Task2 & Task3 & Task4	✓	✓	0.71	0.93	0.91	0.75	0.73
	Task3 & Task4	x	x	/	**0.98**	**0.97**	0.60	0.59
	Task1 & Task2 & Task3 & Task4	✓	x	0.73	0.96	0.95	0.70	0.67
	Task1 & Task2 & Task3 & Task4	x	✓	0.75	0.97	0.95	0.76	0.74
	Task1 & Task2 & Task3 & Task4	✓	✓	0.73	0.97	0.94	**0.79**	**0.77**
Lung	Task1 & Task4	✓	x	/	/	/	0.49	0.51
	Task1 & Task3 & Task4	✓	x	/	0.95	0.94	0.59	0.56
	Task1 & Task2 & Task4	✓	x	0.84	/	/	0.60	0.58
	Task1 & Task2 & Task4	x	✓	0.84	/	/	0.64	0.60
	Task1 & Task2 & Task4	✓	✓	0.83	/	/	0.65	0.62
	Task2 & Task4	✓	x	**0.86**	/	/	0.65	0.57
	Task2 & Task4	x	✓	0.81	/	/	0.67	0.65
	Task2 & Task4	✓	✓	0.82	/	/	0.67	0.66
	Task2 & Task3 & Task4	✓	x	0.80	0.94	0.93	0.68	0.66
	Task2 & Task3 & Task4	x	✓	0.76	0.91	0.92	0.68	0.65
	Task2 & Task3 & Task4	✓	✓	0.79	0.93	0.91	0.69	0.67
	Task3 & Task4	x	x	/	**0.98**	**0.97**	0.70	0.65
	Task1 & Task2 & Task3 & Task4	✓	x	0.81	0.96	0.95	0.70	0.62
	Task1 & Task2 & Task3 & Task4	x	✓	0.83	0.97	0.95	**0.71**	0.69
	Task1 & Task2 & Task3 & Task4	✓	✓	0.82	0.97	0.94	0.70	**0.71**

5.6 Results

The main results of the five experiments are shown in Tables 1 to 5. The neural network was trained for 1500 epochs with an early stopping of 70.

Experiment 1 : As shown in Table 5.1, the best results for outcome prediction were obtained by the combination of the four tasks with multi-scale, and with tumor and global features. It achieved an accuracy of 0.79 and AUC of 0.77 for esophageal cancer outperforming 14 other scenarios which are composed of several combination of different tasks, with and without multi-scale and with and without tumor features. For lung cancer, our proposed model achieved an accuracy of 0.70 and AUC of 0.71 in multi-scale, and an accuracy of 0.71 and AUC of 0.69 when using only tumor features. Using only reconstruction and prediction resulted in a poor performance for both lung and esophageal cancers. For the segmentation, the best results were achieved by the combination of segmentation and prediction in multi-scale for esophageal cancer (dice coefficient = 0.79), and using only global features for lung cancer (dice coefficient = 0.86). Our proposed model achieved a dice score of 0.73 in multi-scale and 0.75 when using only tumor features for esophageal cancer, and a dice score of 0.82 in multi-scale and 0.83 when using only tumor features for lung cancer. The combination of the reconstruction, segmentation and prediction also resulted in good results for segmentation, when using only global features, only tumor features and in multi-scale : 0.77, 0.76 and 0.74 for esophageal cancer and 0.84, 0.84, 0.83 for lung cancer respectively. When using the classification task in addition to segmentation and prediction the performance on segmentation decreases : 0.71, 0.69 and 0.71 for esophageal cancer and 0.80, 0.76, 0.79 for lung cancer. This can be explained due to the fact that the reconstruction task helps in the extraction of rich meaningful features that contribute in the segmentation better than the classification task. For the classification, the best results were achieved with the combination of the classification and prediction tasks, without reconstruction and segmentation : accuracy = 0.98 and AUC = 0.97. Our proposed model achieved an accuracy of 0.97 and an AUC of 0.94 and 0.95 with multi-scale and only tumor features respectively. Since the goal of our study is to focus on the prediction task, the performance of the other 2 tasks (segmentation and classification) can be a lit-

	Method	Dice coef
Esophageal cancer	U-NET (Task2)	0.69
	V-NET	0.69
	WSL-MTL	0.73
	Ours	**0.73**
Lung cancer	U-NET (Task2)	0.80
	V-NET	0.77
	WSL-MTL	**0.85**
	Ours	0.82

TABLEAU 5.2 – Experiment 2 : Segmentation results for esophageal and lung cancer compared to the state of the art methods. WSL : weakly supervised learning model developed in [Amyar et al. 2020].

		Classification		**Prediction**	
	Method	Accuracy	AUC	Accuracy	AUC
Esophageal cancer	AlexNet	0.74	0.73	0.54	0.52
	VGG-16	0.79	0.77	0.53	0.51
	VGG-19	0.78	0.78	0.55	0.53
	ResNet50	**0.97**	**0.97**	0.62	0.63
	169-layers DenseNet	0.95	0.94	0.63	0.61
	InceptionV3	0.93	0.92	0.61	0.69
	3d-rpet-net	/	/	0.72	0.70
	6 layers CNN	0.80	0.81	0.69	0.68
	Ours	**0.97**	0.94	**0.79**	**0.77**
Lung cancer	AlexNet	0.74	0.73	0.51	0.49
	VGG-16	0.79	0.77	0.50	0.51
	VGG-19	0.78	0.78	0.51	0.51
	ResNet50	**0.97**	**0.97**	0.59	0.57
	169-layers DenseNet	0.95	0.94	0.61	0.60
	InceptionV3	0.93	0.92	0.57	0.59
	3d-rpet-net	/	/	0.68	0.61
	6 layers CNN	0.80	0.81	0.63	0.60
	Ours	**0.97**	0.94	**0.70**	**0.71**

TABLEAU 5.3 – Experiment 3 : classification and outcome prediction results compared to state of the art methods for esophageal and lung cancers.

tle higher when using only segmentation and prediction or classification and prediction, but not for the prediction. This is due to the fact that to improve the performance of the prediction task, the model tends to find the most informative and discriminating region in the image that allows this improvement. This results in the extraction of intratumoral and peritumoral tumor regions, which may differ from segmentation ground truth but improve the prediction. The combination of segmentation and prediction resulted in an accuracy of 0.70, 0.70 and 0.73 and an AUC of 0.71, 0.74 and 0.72 for esophageal cancer for global features, tumor features and multi-scale respectively, and an accuracy of 0.65, 0.65 and 0.67 and an AUC of 0.57, 0.65 and 0.66 for lung cancer. The combination of the classification and prediction resulted in an accuracy of 0.60 and an AUC of 0.59 for esophageal cancer and an accuracy of 0.70 and an AUC of 0.65 for lung cancer.

Experiment 2 : In Table 5.2, segmentation results for three other state of the art methods are reported and compared to our proposed model, for esophageal and lung cancers. The 3 models are : U-Net, which represents the task Task2 for the segmentation since it was used as the backbone in our model, V-Net and a weakly supervised multi-task learning (WSL-MTL) model for tumor segmentation. Our model achieved the best results with the WSL-MTL for esophageal cancer (dice coefficient = 0.73), and slightly worse then the WSL-MTL for lung cancer (dice coefficient = 0.82 and 0.85 respectively), since the WSL-MTL was trained to do the segmentation as a primary objective. Our model was better then single task (Task2) U-Net and V-Net : 0.69 and 0.69 for esophageal cancer and 0.80 and 0.77 for lung cancer. These results show that our model can correctly find the tumor regions from which local tumor features can well be extracted.

Experiment 3 : Table 5.3 shows the results of the third experiment. We compared our method with state-of-the-art deep learning models for image classification and prediction. Our proposed model outperformed other methods for the prediction task for both esophageal and lung cancers. ResNet50 had slightly better results for the classification (accuracy 0.97 and AUC 0.97) but very poor results the the prediction : accuracy = 0.62 and AUC = 0.63 for eophageal cancer and accuracy = 0.59 and AUC = 0.57 for lung cancer. AlexNet, VGG-16 and VGG-19 have not shown promising results.

Experiment 4 : In Table 5.4 the the influence of λ on the performance of our model

	λ	Dice coef	Accuracy	AUC	Accuracy	AUC
Esophageal cancer	0.1	**0.74**	0.94	0.94	0.70	0.69
	0.3	0.73	**0.97**	**0.94**	**0.79**	**0.77**
	0.5	0.71	0.95	0.94	0.77	0.76
	0.7	0.71	0.93	0.94	0.70	0.71
	0.9	0.69	0.89	0.88	0.73	0.71
Lung cancer	0.1	0.80	0.94	0.94	0.65	0.64
	0.3	0.82	**0.97**	**0.94**	0.70	**0.71**
	0.5	**0.85**	0.95	0.94	**0.71**	0.70
	0.7	0.80	0.93	0.94	0.69	0.70
	0.9	0.78	0.89	0.88	0.71	0.70

TABLEAU 5.4 – Experiment 4 : The effects of λ on the multi-task learning.

	Method	Dice coef	Accuracy	AUC	Accuracy	AUC
Esophageal cancer	[Zhou et al. 2021]	**0.75**	0.96	0.94	0.73	0.71
	[Chen et al. 2018]	0.68	0.91	0.90	0.68	0.66
	[Qu et al. 2019]	0.69	0.92	0.90	0.70	0.69
	Ours	0.73	**0.97**	**0.94**	**0.79**	**0.77**
Lung cancer	[Zhou et al. 2021]	0.77	0.96	0.94	0.67	0.65
	[Chen et al. 2018]	0.73	0.91	0.90	0.63	0.60
	[Qu et al. 2019]	0.80	0.92	0.90	0.68	0.65
	Ours	**0.82**	**0.97**	**0.94**	**0.70**	**0.71**

TABLEAU 5.5 – Experiment 5 : A quantitative comparison between our model and state of the art multi-task methods.

is reported. We achieved the best results for tumor classification and outcome prediction with $\lambda = 0.3$. When lowering the value of λ the model achieves slightly better result for the segmentation for esophageal cancer (0.74) but a worse prediction result for both esophageal and lung (accuracy = 0.70 and AUC = 0.69, accuracy = 0.65 and AUC = 0.64). For $\lambda = 0.5$ our model achieves comparable results for both pathologies : accuracy = 0.77 and AUC = 0.76 for esophageal cancer and accuracy = 0.71 and AUC = 0.70 for lung cancer, with a better dice coefficient (0.85) for the segmentation of lung tumors. Increasing λ does not result in an improvement of the prediction task, it decreases the performance of the segmentation and classification tasks and also the prediction : accuracy = 0.73, AUC = 0.71 for outcome predication with esophageal cancer and accuracy = 0.71 and AUC = 0.70 for lung cancer, accuracy = 0.89 and AUC = 0.88 for classification, and a dice coefficient = 0.69 and 0.78 for esophageal and lung cancers respectively for segmentation.

Experiment 5 : Table 5.5 reports the results of three state-of-the-art methods for multi-task learning for segmentation and classification. Our proposed model achieves the best results for both esophageal and lung cancers for the prediction and classification task, where [Zhou et al. 2021] achieves a slight improvement on the segmentation task for esophageal cancer (dice coefficient = 0.75), and comparable results for the classification (accuracy = 0.96 and AUC = 0.94).

5.7 Discussion

We have developed a new deep learning multi-task model to jointly identify esophageal and lung tumors, segment the tumor regions of interest and predict patient's outcome. Our architecture is general, which means that it can be used for other segmentation-classification-prediction applications. We have also compared our method with several state-of-the-art algorithms such as U-NET, V-NET and WSL-MTL for tumor segmentation, methods for image classification and prediction, and for multi-task learning such as [Chen et al. 2018, Qu et al. 2019, Zhou et al. 2021]. To show the performance of our method, we tested the different combinations of different tasks, as well as using only global features or only tumor features and a multi-scale regrouping tumor and global features.

We have added the reconstruction task in order to leverage useful information contained in multiple related tasks to improve both segmentation and prediction performances.

Multi-task learning can handle small data problems well, although each task can have a relatively small data set. In contrast to conventional radiomics, where only one pathology is studied at a time, multi-task learning allows to study different cancer types at the same time, thus, to increase the size of the dataset and help the model to learn meaningful features from PET images so that help to improve the prediction.

We have added global image features through a multi-scale by using a global average pooling and then concatenated with tumor features to predict the outcome. Having both global and local features helps to improve the performance of the model compared to using only tumor features as in classical radiomic. Although the segmentation performance drops a little when combining the 4 tasks compared to segmentation-prediction alone, the most important task in our study is the prediction, hence we let the model decides which is the most important region in the image that increases the prediction performance, resulting in encompassing intratumoral and peritumoral regions. Since dice coefficient measures the intersection between the ground truth and the segmentation result, it can drop a little its score. The segmented tumor region may not be exactly the same as the ground truth, but it may be more relevant for prediction. In our study, the dice coefficient is used to ensure that the result of the segmentation is anatomically correct, not to be perfect.

One of the main advantages of our proposed method relies in the fact that once the learning is finished, we no longer need segmentation ground truth to do radiomics. The model requires only the PET images as input, thus to avoid the tedious segmentation task for physicians. Also, the architecture is general. The model can be modified easily to add other cancer types to do radiomics without changing the architecture, just the classification branch.

5.8 Conclusion

In this chapter, we proposed a multi-task learning approach to predict patients outcome from PET images and segment the regions of interest simultaneously. Our method can improve prediction results even if we have only several small datasets. thanks to learning tasks in parallel while using a shared representation. Therefore, what is learned for each task can help other tasks be learned better. We show also that subsidiary tasks serve as an inductive bias so that the model can generalize better. Our model was tested and evaluated for treatment response and survival in lung and esophageal cancers, outperforming single task learning methods and state-of-the-art multi-task learning methods. In the future, we will add other cancers to validate our framework and develop an attention mechanism to combine the different features.

Article Details :

— **Multi-Task Multi-Scale Learning For Outcome Prediction in 3D PET Images.** Amine Amyar, Romain Modzelewski, Pierre Vera, Vincent Morard, Su Ruan. Under review.

Other Related Publications :

— **Multi-task Deep Learning Based CT Imaging Analysis For COVID-19 pneumonia : Classification and Segmentation.** Amine Amyar, Romain Modzelewski, Hua Li, Su Ruan. Computers in Biology and Medicine, 126, p.104037.

Conclusion and Perspectives

In this thesis, we have investigated deep learning (DL) to design new frameworks for radiomics. The main objective of this thesis was to exploit the potential of DL to automatically segment the lesions, extract deep radiomic features and predict patient's outcome in order to propose robust and reproducible models. Concretely, different CNNs architectures were proposed and evaluated using PET images mainly for esophageal and lung cancers.

We summarize our contributions as follows :

Deep radiomics

We propose a four-layer 3D-CNN. To find the best 3D RPET-NET, convolutional neural networks were used as a backbone to develop new architectures for outcome prediction. Classical methods based on random forest and handcrafted features such as : random forest without feature selection (RF), random forest with features importance (FIC) and random forest with genetic algorithm (GARF) were used as baseline for comparison. RF makes use of handcrafted features, thus its performance is highly impacted by the features manually defined. On the other hand, CNNs tend to learn representative features while making a decision in an end-to-end framework. Early layers of a CNN extract low level features, and latter layers high level ones. These rich features are then fed to a fully connected layers (FC) for classification or regression.

The main algorithms proposed in the literature for deep radiomic are based on 2 Dimensional (2D) CNN architectures. This approach requires to process each slice separately, therefore the spatial relationship is lost. Also, the final prediction for the whole 3D volume requires a majority vote, which add the need to find the best threshold to separate

accurately the 2 populations with good and bad outcome. To solve this issue, we englobed the tumor into a 3D cuboid of standard width, length and height. This method allows to take advantage of the spatial relationship between slices. Our assumption is that a neural network architecture able to capture patterns of FDG uptake that occur within the whole lesion in 3D may detect more relevant imaging features that are more relevant to predict treatment response than each slice individually or 3 adjacent slices. The influence of the learning volume (intratumoral volume with different peritumoral volumes) was also investigated. The peritumoral part of the tumor is therefore a volume that is not neglected in the treatment. By analogy, taking into account the intratumoral and peritumoral regions in radiomics analysis is likely a strategy that can improve the results.

To find a good compromise between network complexity and performance as well as our small dataset, we proposed a four-layer 3D-CNN. To find the best 3D RPET-NET, called 3D RPET-NETBest, with optimized hyperparameters. The hyperparameters optimized include the number of 3D feature maps (we tested from 8 to 64 feature maps), the number of neurons (128, 256, 512, 1024, 2048 and 4096), as well as different receptive field sizes ($3 \times 3 \times 3$, $5 \times 5 \times 5$) and different sizes of mini-batches (2, 4, 8 and 16). Several (4) expressions of f(x), the activation function, were also evaluated (relu, elu, selu and tanh). Several numbers of 3D convolutional layers and 3D pooling layers (2 to 5) and fully connected layers (2, 3, 4 and 5) were evaluated. Having achieved the best performance with 3D CNN, we used it for all the further studies for outcome prediction.

Image segmentation

The first step in radiomic analysis is the segmentation of the lesion. This process could be automated using computer-aided detection (CAD) tools. State of the art U-NET have shown very good performances for image segmentation in different fields. In medical imaging, it is usually used as a backbone for tumor or organs segmentation. The main drawback of U-NET is the need of a large dataset to work efficiently. Fully labeled dataset is very hard to obtain in the medical imaging field due to several reasons such as : protection of the patient's privacy, establishment of a specific protocol for data recovery, the need of an expert physician for data labeling ... etc. Moreover, physician's labels are

subjective and prone to error.

One possible solution is the use of a weakly supervised learning (WSL) approach. In order to make use of weakly labeled data, we transform a standard CNN used initially for classification, to perform a segmentation. This is done by the interpretation of the decision of the CNN using grad-cam, a method to visualize heatmaps developed by. In particular, two maximum intensity images (MIP) of the 3D PET scan are calculated for sagittal and coronal view. Then, a CNN is trained to classify the image for each view to lung or esophageal cancer, while generating a heatmap. This heatmap is used to retrieve the whole tumor volume. For each patient, we define at random a point in the center of the tumor, which is considered as a prior knowledge. Then, we define a loss function based on both the distance between the heatmap generated at the current iteration and the central point, and on the accuracy of the tumor type classification. We showed that training a neural network with weakly annotated data allows to achieve state of the art in both segmentation and outcome prediction.

Multi-task based deep learning for segmentation and prediction

Fully supervised learning approaches for tumor segmentation rely on large annotated dataset, and their performances depend on the ground truth (GT) defined manually by the physicians, thus the need of a well-defined GT.

To find a good compromise between segmentation accuracy and outcome prediction performance, we train a neural network to segment the tumor not to be precisely as the physician defined ground truth, but to maximize outcome prediction, thus the segmentation include peritumoral and intratumoral regions that contribute most to the prediction. To ensure that the neural network does not take into account noises we ensure that the predicted segmentation is not very far away from the GT. This is done through a multi-task learning (MTL) approach where the NN is trained to segment the lesion and predict the outcome on the segmentation result simultaneously. In single task two step learning process, the NN is first trained to segment the lesion and the parameters of the model are freezed after training. Then a second NN is trained to predict the outcome from the results of the first NN. MTL allows to train a single NN to perform both tasks in the same

time, hence each task can help the other. Other subsidiary tasks, can be added to extract meaningful features and to serve also as an inductive bias to generalize better. We proposed a MTL framework where the NN is trained to classify the pathology, segment the lesion, reconstruct the image and predict the outcome based on the segmentation results. In addition, we used global patient features with tumor and tumor regions features to predict the outcome in a radiomic study through a multi-scale design. We conducted extensive validation strategy with multiple ablation experiments, comparison with state of the art methods in both supervised and multi-task learning. Furthermore, our MTL approach was validated on a COVID-19 database to classify and segment COVID-19 pneumonia lesions. Our approach outperformed state of the art methods, showing the interest of combining jointly different tasks to improve both segmentation and classification performances. Moreover, adding a third task such as image reconstruction, the encoder can extract meaningful feature representation which help the other tasks (classification and segmentation) to improve even more their performances.

5.9 Limitations

The potential of deep radiomics has been demonstrated in this thesis. However, there are several important limitations that we would like to address here.

Quantity and quality of data

Deep learning methods are data hungry, and their performance rely heavily on the quality of the dataset used for training. Therefore, it is very challenging to train a NN on a small dataset with the aim to generalize very well to an unseen dataset. Two main problems may occur when training a NN on a small dataset : the NN fail to learn, which results on a poor performance on the training dataset, this phenomena is called underfitting. This problem is generally due to the low capacity of the model, and can be solved by increasing the number of parameters of the NN to learn better. This will results in a NN with a large number of parameters that may have easily millions, therefore when learning from a small dataset it will lead to an overfitting. DL algorithms tend to overfit when

no big dataset is available for training, thus one must be vigilant to include some regularization techniques. Regularization is the process that allows a NN to look for useful representation from the training dataset while being careful not to memorize its distribution to prevent overfitting. Different regularization methods are proposed in the literature, but the most common one is based on representation learning such as dropout, semi-supervised learning and manifold learning. These techniques may improve the results.

In order to improve model generalization, different approaches were used in our work. In order to reduce model complexity we used L_{p} parameter norm, training with early stopping, dropout and with other methods that do not alter the complexity of the model such as parameter sharing in the MTL framework. The MTL algorithm can improve generalization by leveraging the domain-specific information contained in the training signals of related tasks as an inductive bias.

A major limitation in advancing the field of precision medicine research is the ability to integrate data from a variety of different sources in order to improve patients classification, which arises the need for approach that focuses on establishing new methods for the computational analysis and integration of multi-modal data. More importantly, classifiers based on a single-data modality might ignore key biological features from other available data sources hat might be highly predictive of a patient's clinical status. In this work, only image information is used to train DL models. Predicting patient's outcome is a very hard question to answer. While it is true that medical imaging shows a promising results to tackle this challenges, other information is also highly relevant to add complementary value such as clinical notes, genomics and other imaging modalities such as CT scan or MRI. Training the NN with multi-modality to incorporate different modalities and with multi-view to integrate other relevant information can lead to an improved framework with better results.

5.10 Perspectives for future research

Multi-view and multimodal learning

Multi-view data are very common in the medical imaging field. One patient may have several exams such as PET scan, MRI or CT scan. Each modality provides rich and complementary information for other modalities, for example CT scan offers more anatomic detail while PET image gives functional information such as metabolism. These two modalities can be integrated into the same framework in order to extract rich features from both images to boost the NN performance. In particular, a multi-view NN can be trained using both modalities as raw input, and instead of using a single description about the patient and the lesion with one modality, using both of them provide more accurate and complete description. In addition, more information can be added in this framework, coming from heterogeneous sources : such as clinical data, genomics, proteomics or some pertinent handcrafted features. One of the promising direction in the precision medicine is the integration of all the available data types (genomics, proteomics and images) in a AI-driven multi-modal classifier that will be trained to predict patient's outcome, as well as other relevant clinical variables, such as staging, disease-free survival, etc.

Interpretabilty

Despite their success, deep learning models often function as black-boxes, and provide very little understanding about the inner workings. While opaqueness concerning machine behaviour might not be a problem in deterministic domains, in health care, model interpretability is crucial to build trust in the performance of a predictive system. To date no single method can provide a detailed human-understandable explanation of how a model makes a decision, however recent efforts in the field of interpretable artificial intelligence have produced various methods that can help bridge the gap between low-level features and phenotypic predictions. Perturbation-based approaches change parts of the input and observe the impact on the output of the network [Alipanahi et al. 2015, Zhou and Troyanskaya 2015]. Backpropagation-like methods, also known as saliency methods,

use signals from gradients or output decomposition to infer a “saliency map” [Simonyan et al. 2013]. An alternative strategy is the Layer-wise Relevance Propagation (LRP) [Bach et al. 2015]. Interpretable surrogate models aim to approximate a large, slow, but accurate model by a surrogate models a smaller, interpretable, yet still accurate model [Che et al. 2015, Hinton et al. 2015, Ribeiro et al. 2016]. Generative models : Modifications have been proposed to Generative Adversarial Networks(GANs) to encourage the network to learn interpretable and meaningful representations [Chen et al. 2016]. Models with built-in explainability, such as attention mechanisms [Hendricks et al. 2016], can identify a posteriori the most informative features underlying a prediction.

While a lot of progress has been achieved, many of these methods have been developed for specific types of data, and their application to medical imaging data is not always trivial. Besides, many of the previous approaches exploit heuristic ideas that work in very specific data types and models, limiting its generalization. Indeed, interpretabilty methods are, to some extent, black-boxes themselves and we have no consensus on which methods to use. In trying to understand a black-box, we have inadvertently created another. We currently have little understanding of human factors when it comes to accepting AI predictions in the clinic. Interpretability is important, but we have to always compare to the gold standard here : human physicians. In many cases, their decisions are not interpretable and object to a large inter and intra physician variability.

Data annotation

There has been a growing number of medical data annotation services over the past few years. These offer a network of experts to label the data for AI development. It is no surprise that almost 60% of ML work involves preparing data for models. Such services handle this bottleneck allowing startups to focus on the AI development and clinical integration aspects of their business. As healthcare data is silo-ed, the tools needed to curate each silo differ widely. As such, it is not entirely apparent if these services can offer a one-size-fits-all solution. As data cleanliness is paramount, the expertise of these "outsourced" annotators often comes into question.

Some of these services go one step further by offering access to curated medical images.

It is unclear how successful this model will be as it relies heavily on the success of medical AI startups. There are also issues around the exclusivity of data, and whether data can be re-purposed across multiple use cases. startups can have greater long-term benefits if they control and operate their own data curation pipelines and network of annotators.

Clinical perspectives

The increasing availability of omics datasets has opened new ways to characterize, categorize cancers and guide therapeutic interventions. Even though significant efforts have been made recently, many of them have shown limited clinical applicability and precise biomarkers that can inform clinicians of expected prognosis and offer the most beneficial treatment, while reducing unnecessary morbidity, are still needed. Furthermore, since cohort sizes are often limited in size, many studies employ single-locus analysis strategies that require the pre-selection of features in order to increase the statistical power of the study. Consequently, these strategies limit their search to a few known a priori candidate modifications and do not take full advantage of high-throughput datasets. From a clinicians' and radiologists' perspective all these efforts are only accepted, if it does not get more complicated. The irony is that while technology is designed to help productivity, it actually can add more work and complexity. The best AI application is the one that is invisible, that is seamlessly integrated into the workflows.

Three important concepts in clinical trials today that may help AI based application : Central reads, diversity, and real-world evidence / real-world data.

Central reads : Trials often rely on site-based reads where a radiologist on staff that given day will review a patient's images. The same radiologist may not even read all other time-points for that patient. This can cause a very high level of variability and bias in the data. Central reads ensure all data is read in a controlled environment complying with protocols and using the same software. AI can play a major role in this highly operational environment without the nuances of patient management.

Diversity : It is no surprise that trial populations are heavily skewed in terms of race, background, and even gender. The FDA has recently issued guidance related to this topic. Providers that have more diverse patient populations can now be seen as having "valuable

data" and are more likely to be selected to participate in trials.

RWE / RWD : As trials test the efficacy of drugs (a controlled experiment) as opposed to efficiency (in the real world), pharmaceutical companies are being asked to collect data about how their drug is performing in the wild. That is the "evidence" part. The collection of real-world data allows them to best identify what drugs need to be developed.

Bibliographie

Aerts H J, Velazquez E R, Leijenaar R T, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 2014 ; 5(1) :1–9. 88

Ahn J, Cho S, and Kwak S. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019 ; pages 2209–2218. 72

Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 2015 ;33(8) :831–838. 39, 114

Alongi P, Laudicella R, Stefano A, et al. Choline pet/ct features to predict survival outcome in high risk prostate cancer restaging : a preliminary machine-learning radiomics study. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging : Official Publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of*, 2020 ;. 45

Amadasun M and King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern*, 1989 ;19(5) :1264–1273. doi :10.1109/21.44046. 22, 24

Amyar A, Decazes P, Ruan S, et al. Contribution of class activation map on wb pet deep features for primary tumour classification. *Journal of Nuclear Medicine*, 2019a ;60(supplement 1) :1212–1212. 71

Amyar A, Modzelewski R, Vera P, et al. Weakly supervised pet tumor detection usingclass response. *arXiv preprint arXiv :200308337*, 2020 ;. 88, 99, 103, 139

Amyar A, Ruan S, Gardin I, et al. Radiomics-net : Convolutional neural networks on fdg pet images for predicting cancer treatment response. *Journal of Nuclear Medicine*, 2018 ; 59(supplement 1) :324–324. 88

Amyar A, Ruan S, Gardin I, et al. 3-d rpet-net : development of a 3-d pet imaging convolutional neural network for radiomics analysis and outcome prediction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2019b ;3(2) :225–231. 70, 73, 80, 89, 100

Arnedos M, Vicier C, Loi S, et al. Precision medicine for metastatic breast cancer—limitations and solutions. *Nature Reviews Clinical Oncology*, 2015 ;12(12) :693. 41

Asgari R, Orlando J I, Waldstein S, et al. Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019 ; pages 192–200. 91

Avanzo M, Stancanello J, and El Naqa I. Beyond imaging : the promise of radiomics. *Physica Medica*, 2017 ;38 :122–139. 52

Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015 ;10(7) :e0130140. 39, 115

Badrinarayanan V, Kendall A, and Cipolla R. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2017 ;39(12) :2481–2495. 88

Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and prognosis of quantitative features extracted from ct images. *Translational oncology*, 2014 ;7(1) :72–87. 44

Belharbi S, Chatelain C, Hérault R, et al. Spotting l3 slice in ct scans using deep convolutional network and transfer learning. *Computers in biology and medicine*, 2017 ;87 :95–103. 64

Ben-Haim S and Ell P. 18F-FDG PET and PET/CT in the Evaluation of Cancer Treatment Response. *J Nucl Med*, 2008 ;50(1) :88–99. doi :10.2967/jnumed.108.054205. 13

Benjamini Y and Hochberg Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal statistical society : series B (Methodological)*, 1995 ;57(1) :289–300. 61

Beyar-Katz O and Gill S. Novel approaches to acute myeloid leukemia immunotherapy. *Clinical Cancer Research*, 2018 ;24(22) :5502–5515. 12

Bogowicz M, Riesterer O, Ikenberg K, et al. Computed tomography radiomics predicts hpv status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. *International Journal of Radiation Oncology* Biology* Physics*, 2017 ;99(4) :921–928. 52

Boser B E, Guyon I M, and Vapnik V N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. 1992 ; pages 144–152. 42

Braman N, Prasanna P, Whitney J, et al. Association of peritumoral radiomics with tumor biology and pathologic response to preoperative targeted therapy for her2 (erbb2)–positive breast cancer. *JAMA network open*, 2019 ;2(4) :e192561–e192561. 89

Braman N M, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. *Breast Cancer Research*, 2017 ;19(1) :1–14. 55, 66, 89

Breiman L. Random forests. *Machine learning*, 2001 ;45(1) :5–32. 42

Breiman L, Friedman J, Olshen R, et al. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. 42

Bundschuh R A, Dinges J, Neumann L, et al. Textural parameters of tumor heterogeneity in 18f-fdg pet/ct for therapy response assessment and prognosis in patients with locally advanced rectal cancer. *Journal of Nuclear Medicine*, 2014 ;55(6) :891–897. 22

Cabannnes V, Rudi A, and Bach F. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*. PMLR, 2020 ; pages 1230–1239. 33

Cameron A, Khalvati F, Haider M A, et al. Maps : a quantitative radiomics approach for prostate cancer detection. *IEEE Transactions on Biomedical Engineering*, 2015 ; 63(6) :1145–1156. 44, 73

Carneiro G, Oakden-Rayner L, Bradley A P, et al. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017 ; pages 130–134. 44

Caruana R. Multitask learning. *Machine learning*, 1997 ;28(1) :41–75. 34, 89, 94

Castellano G, Bonilha L, Li L, et al. Texture analysis of medical images. *Clinical radiology*, 2004 ;59(12) :1061–1069. 48

Che Z, Purushotham S, Khemani R, et al. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv :151203542*, 2015 ;. 40, 115

Chen C, Bai W, and Rueckert D. Multi-task learning for left atrial segmentation on ge-mri. In *International workshop on statistical atlases and computational models of the heart*. Springer, 2018 ; pages 292–301. 100, 105, 106

Chen X, Duan Y, Houthooft R, et al. Infogan : Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv :160603657*, 2016 ;. 40, 115

Chung A G, Shafiee M J, Kumar D, et al. Discovery radiomics for multi-parametric mri prostate cancer detection. *arXiv preprint arXiv :150900111*, 2015 ;. 44

Ciregan D, Meier U, and Schmidhuber J. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012 ; pages 3642–3649. 88

Cook G J, Siddique M, Taylor B P, et al. Radiomics in pet : principles and applications. *Clinical and Translational Imaging*, 2014 ;2(3) :269–276. 52

Cook G J R, Yip C, Siddique M, et al. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non–Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy ? *J Nucl Med*, 2013 ;54(1) :19–26. doi :10.2967/jnumed.112.107375. 47

Cortes C and Vapnik V. Support-vector networks. *Machine learning*, 1995 ;20(3) :273–297. 42

Couzin-Frankel J. Cancer immunotherapy. 2013. 12

Cysouw M C, Jansen B H, van de Brug T, et al. Machine learning-based analysis of [18 f] dcfpyl pet radiomics for risk stratification in primary prostate cancer. *European journal of nuclear medicine and molecular imaging*, 2021 ;48(2) :340–349. 44, 45

Das B K and Das B K. Positron Emission Tomography- A Guide for Clinicians. Springer edition, 2015 ;. 17

Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS One*, 2017a ;12(3) :e0173208. doi :10.1371/journal.pone.0173208. 43

Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial fdg-pet features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS One*, 2017b ;12(3) :e0173208. 46

Desbordes P, Su R, Romain M, et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Computerized Medical Imaging and Graphics*, 2017c ;60 :42–49. 52, 53, 55, 60, 64, 73

Dewalle-Vignion A S, Yeni N, Petyt G, et al. Evaluation of pet volume segmentation methods : comparisons with expert manual delineations. *Nuclear medicine communications*, 2012 ;33(1) :34–42. 66

Dou T H, Coroller T P, van Griethuysen J J, et al. Peritumoral radiomics features predict distant metastasis in locally advanced nsclc. *PloS one*, 2018 ;13(11) :e0206108. 44, 89

Du D, Feng H, Lv W, et al. Machine learning methods for optimal radiomics-based differentiation between recurrence and inflammation : application to nasopharyngeal carcinoma post-therapy pet/ct images. *Molecular imaging and biology*, 2020 ;22(3) :730–738. 45

Dubray B, Thureau S, Nkhali L, et al. Nuclear imaging and target volumes for radiotherapy. *MEDECINE NUCLEAIRE-IMAGERIE FONCTIONNELLE ET METABOLIQUE*, 2013 ; 37(5) :198–202. 55

Eisenhauer E A, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours : Revised RECIST guideline (version 1.1). *Eur J Cancer*, 2009 ;45(2) :228–247. doi :10.1016/j.ejca.2008.10.026. 13

El Naqa I, Brock K, Yu Y, et al. On the fuzziness of machine learning, neural networks, and artificial intelligence in radiation oncology. *International journal of radiation oncology, biology, physics*, 2018 ;100(1) :1–4. 44

El Naqa I, Grigsby P W, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit*, 2009 ;42(6) :1162–1171. doi : 10.1016/j.patcog.2008.08.011. 22, 42, 47

Elefsinioti A, Bellaire T, Wang A, et al. Key factors for successful data integration in biomarker research. *Nature Reviews Drug Discovery*, 2016 ;15(6) :369–370. 1

Emens L A. Breast cancer immunotherapy : facts and hopes. *Clinical Cancer Research*, 2018 ;24(3) :511–520. 12

Farhidzadeh H, Kim J Y, Scott J G, et al. Classification of progression free survival with nasopharyngeal carcinoma tumors. In *Medical Imaging 2016 : Computer-Aided Diagnosis*, volume 9785. International Society for Optics and Photonics, 2016 ; page 97851I. 43, 44

Fawcett T. An introduction to roc analysis. *Pattern recognition letters*, 2006 ;27(8) :861–874. 61

Feng J, He X, Teng Q, et al. Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. *Physical Review E*, 2019 ; 100(3) :033308. 35, 135

Foster B, Bagci U, Mansoor A, et al. A review on segmentation of positron emission tomography images. *Computers in biology and medicine*, 2014 ;50 :76–96. 66

Frid-Adar M, Diamant I, Klang E, et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 2018 ; 321 :321–331. 53

Fuge O, Vasdev N, Allchorne P, et al. Immunotherapy for bladder cancer. *Research and reports in urology*, 2015 ;7 :65. 11

Fukushima K. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*, 1980 ;36(4) :193–202. doi :10.1007/BF00344251. 37

Galavis P E, Hollensen C, Jallow N, et al. Variability of textural features in fdg pet images due to different acquisition modes and reconstruction parameters. *Acta oncologica*, 2010 ;49(7) :1012–1016. 55

Galloway M M. Texture analysis using gray level run lengths. *Comput Graph Image Process*, 1975 ;4(2) :172–179. doi :http://dx.doi.org/10.1016/S0146-664X(75)80008-6. 22, 24

Gardin I, Grégoire V, Gibon D, et al. Radiomics : principles and radiotherapy applications. *Critical reviews in oncology/hematology*, 2019 ;138 :44–50. 41, 53

Gillies R J, Kinahan P E, and Hricak H. Radiomics : images are more than pictures, they are data. *Radiology*, 2016 ;278(2) :563–577. 70

Gilpin L H, Bau D, Yuan B Z, et al. Explaining explanations : An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018 ; pages 80–89. 71

Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014 ; pages 580–587. 71

Goodfellow I, Bengio Y, Courville A, et al. *Deep learning*, volume 1. MIT press Cambridge, 2016. 38, 39, 40, 135, 136

Ha S, Lee H Y, and Kim S E. Prediction of response to neoadjuvant chemotherapy in patients with breast cancer using texture analysis of 18f-fdg pet/ct. *Journal of Nuclear Medicine*, 2014 ;55(supplement 1) :623–623. 47

Hannun A Y, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 2019 ;25(1) :65. 71

Hao H, Zhou Z, Li S, et al. Shell feature : a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer. *Physics in Medicine & Biology*, 2018 ;63(9) :095007. 44, 46

Haralick R M, Shanmugam K, and Dinstein I. Textural features for image classification. 1973. doi :10.1109/TSMC.1973.4309314. 22

Hatt M, Cheze Le Rest C, Albarghach N, et al. PET functional volume delineation : A robustness and repeatability study. *Eur J Nucl Med Mol Imaging*, 2011 ;38(4) :663–672. doi :10.1007/s00259-010-1688-6. 19

Hatt M, Tixier F, Pierce L, et al. Characterization of pet/ct images using texture analysis : the past, the present... any future ? *European journal of nuclear medicine and molecular imaging*, 2017 ;44(1) :151–165. 53, 65, 73

He K, Gkioxari G, Dollár P, et al. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2017 ; pages 2961–2969. 72

He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016 ; pages 770–778. 95, 100

He T, Hu J, Song Y, et al. Multi-task learning for the segmentation of organs at risk with label dependence. *Medical Image Analysis*, 2020 ;61 :101666. 92

Heath M, Bowyer K, Kopans D, et al. Pkegelmeyer. the digital database for screening mammography. In *The Proceedings of the 5th International Workshop on Digital Mammography. Madison, WI, USA : Medical Physics Publishing*. 2000 ; . 91

Hendricks L A, Akata Z, Rohrbach M, et al. Generating visual explanations. In *European Conference on Computer Vision*. Springer, 2016 ; pages 3–19. 40, 115

Hicks R J, Mac Manus M P, Matthews J P, et al. Early fdg-pet imaging after radical radiotherapy for non–small-cell lung cancer : inflammatory changes in normal tissues correlate with tumor response and do not confound therapeutic response evaluation. *International Journal of Radiation Oncology* Biology* Physics*, 2004 ;60(2) :412–418. 48

Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal processing magazine*, 2012 ;29(6) :82–97. 2

Hinton G, Vinyals O, and Dean J. Distilling the knowledge in a neural network. *arXiv preprint arXiv :150302531*, 2015 ;. 40, 115

Hirsch D D. The glass house effect : Big data, the new oil, and the power of analogy. *Me L Rev*, 2013 ;66 :373. 28

Hosny A, Parmar C, Coroller T P, et al. Deep learning for lung cancer prognostication : A retrospective multi-cohort radiomics study. *PLoS medicine*, 2018 ;15(11) :e1002711. 73

Hu Y, Xie C, Yang H, et al. Assessment of intratumoral and peritumoral computed tomography radiomics for predicting pathological complete response to neoadjuvant chemoradiation in patients with esophageal squamous cell carcinoma. *JAMA network open*, 2020 ;3(9) :e2015927–e2015927. 89

Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017 ; pages 4700–4708. 100

Huang S H and O'Sullivan B. Overview of the 8th edition tnm classification for head and neck cancer. *Current treatment options in oncology*, 2017 ;18(7) :1–13. 9

Hubel D H and Wiesel T N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 1959 ;148(3) :574–591. 37

Hyun S H, Ahn M S, Koh Y W, et al. A machine-learning approach using pet-based radiomics to predict the histological subtypes of lung cancer. *Clinical nuclear medicine*, 2019 ;44(12) :956–960. 44, 45

Jackson C M, Lim M, and Drake C G. Immunotherapy for brain cancer : recent progress and future promise. *Clinical Cancer Research*, 2014 ;20(14) :3651–3659. 12

Janowczyk A and Madabhushi A. Deep learning for digital pathology image analysis : A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 2016 ; 7. 88

Javeri H, Xiao L, Rohren E, et al. Influence of the baseline 18f-fluoro-2-deoxy-d-glucose positron emission tomography results on survival and pathologic response in patients with gastroesophageal cancer undergoing chemoradiation. *Cancer*, 2009 ;115(3) :624–630. 19

Jeong S Y, Kim W, Byun B H, et al. Prediction of chemotherapy response of osteosarcoma using baseline 18f-fdg textural features machine learning approaches with pca. *Contrast media & molecular imaging*, 2019 ;2019. 46

Jia T Y, Xiong J F, Li X Y, et al. Identifying egfr mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. *European radiology*, 2019 ;29(9) :4742–4750. 44

Kamnitsas K, Ledig C, Newcombe V F, et al. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 2017 ;36 :61–78. 88

Kawakami W, Takemura A, Yokoyama K, et al. The use of positron emission tomography/computed tomography imaging in radiation therapy : a phantom study for setting internal target volume of biological target volume. *Radiation Oncology*, 2015 ;10(1) :1. 66

Kinahan P E, Townsend D, Beyer T, et al. Attenuation correction for a combined 3d pet/ct scanner. *Medical physics*, 1998 ;25(10) :2046–2053. 17

Kingma D P and Ba J. Adam : A method for stochastic optimization. *arXiv preprint arXiv :14126980*, 2014 ;. 98

Krizhevsky A, Sutskever I, and Hinton G E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 2012 ; pages 1097–1105. 2, 71

Krizhevsky A, Sutskever I, and Hinton G E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017 ;60(6) :84–90. 100

Kumar V, Gu Y, Basu S, et al. Radiomics : the process and the challenges. *Magnetic resonance imaging*, 2012 ;30(9) :1234–1248. 1, 41, 53, 88

Kwee R M. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of 18f fdg pet : a systematic review. *Radiology*, 2010 ;254(3) :707–717. 52

Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics : extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 2012a ;48(4) :441–446. 41, 52, 66, 88

Lambin P, Rios-Velazquez E, Leijenaar R T H, et al. Radiomics : Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 2012b ; 48(4) :441–446. doi :10.1016/j.ejca.2011.11.036. 1, 12, 48, 135

Larson S, Erdi Y, Akhurst T, et al. Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging The Visual Response Score and the Change in Total Lesion Glycolysis. *Clin Positron Imaging*, 1999 ; 2(3) :159–171. 19

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*, 1998 ;86(11) :2278–2323. doi :10.1109/5.726791. 37

Lee J G, Jun S, Cho Y W, et al. Deep learning in medical imaging : general overview. *Korean journal of radiology*, 2017 ;18(4) :570. 2

Leger S, Zwanenburg A, Pilz K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific reports*, 2017 ; 7(1) :1–11. 53, 73

Lemarignier C, Di Fiore F, Marre C, et al. Pretreatment metabolic tumour volume is predictive of disease-free survival and overall survival in patients with oesophageal squamous cell carcinoma. *Eur J Nucl Med Mol Imaging*, 2014 ;i(2014) :2008–2016. doi : 10.1007/s00259-014-2839-y. 13

Li H, Galperin-Aizenberg M, Pryma D, et al. Unsupervised machine learning of radiomic features for predicting treatment response and overall survival of early stage non-small cell lung cancer patients treated with stereotactic body radiation therapy. *Radiotherapy and Oncology*, 2018 ;129(2) :218–226. 46

Li H, Xu C, Xin B, et al. 18f-fdg pet/ct radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia. *Theranostics*, 2019 ;9(16) :4730. 44, 46

Lian C, Ruan S, Denœux T, et al. Selecting radiomic features from fdg-pet images for cancer treatment outcome prediction. *Medical image analysis*, 2016 ;32 :257–268. 46, 70

Lin M, Chen Q, and Yan S. Network in network. *arXiv preprint arXiv :13124400*, 2013 ;. 71

Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015 ; pages 3431–3440. 91

Lu L, Lv W, Jiang J, et al. Robustness of radiomic features in [11 c] choline and [18 f] fdg pet/ct imaging of nasopharyngeal carcinoma : Impact of segmentation and discretization. *Molecular Imaging and Biology*, 2016 ;18(6) :935–945. 52

Mahendran A and Vedaldi A. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015 ; pages 5188–5196. 71

Marusyk A, Almendro V, and Polyak K. Intra-tumour heterogeneity : a looking glass for cancer ? *Nat Rev Cancer*, 2012 ;12(5) :323–334. doi :10.1038/nrc3261. 1

Mi H, Petitjean C, Dubray B, et al. Robust feature selection to predict tumor treatment outcome. *Artif Intell Med*, 2015a ;64(3) :195–204. doi :10.1016/j.artmed.2015.07.002. 46

Mi H, Petitjean C, Dubray B, et al. Robust feature selection to predict tumor treatment outcome. *Artificial intelligence in medicine*, 2015b ;64(3) :195–204. 52, 53

Miller a B, Hoogstraten B, Staquet M, et al. Reporting results of cancer treatment. *Cancer*, 1981 ;47(1) :207–214. doi :10.1002/1097-0142(19810101)47:1<207::AID-CNCR2820470134>3.0.CO;2-6. 13

Miller T R, Pinkus E, Dehdashti F, et al. Improved prognostic value of 18F-FDG PET using a simple visual analysis of tumor characteristics in patients with cervical cancer. *J Nucl Med*, 2003 ;44(2) :192–197. 48

Milletari F, Navab N, and Ahmadi S A. V-net : Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016 ; pages 565–571. 92

Mitchell T M et al. Machine learning. 1997 ;. 29

Morgan Jr D L and Hughes V W. Atomic processes involved in matter-antimatter annihilation. *Physical Review D*, 1970 ;2(8) :1389. 14

Mu W, Tunali I, Gray J E, et al. Radiomics of 18 f-fdg pet/ct images predicts clinical benefit of advanced nsclc patients to checkpoint blockade immunotherapy. *European journal of nuclear medicine and molecular imaging*, 2019 ;pages 1–15. 45

Nair J K R, Saeed U A, McDougall C C, et al. Radiogenomic models using machine learning techniques to predict egfr mutations in non-small cell lung cancer. *Canadian Association of Radiologists Journal*, 2020 ;page 0846537119899526. 46

Nishioka T, Shiga T, Shirato H, et al. Image fusion between 18fdg-pet and mri/ct for radiotherapy planning of oropharyngeal and nasopharyngeal carcinomas. *International Journal of Radiation Oncology* Biology* Physics*, 2002 ;53(4) :1051–1057. 52

Noble W S. What is a support vector machine ? *Nature biotechnology*, 2006 ;24(12) :1565–1567. 42

Orlhac F, Soussan M, Maisonobe J A, et al. Tumor Texture Analysis in 18F-FDG PET : Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. *J Nucl Med*, 2014 ;55(3) :414–422. doi : 10.2967/jnumed.113.129858. 41

Ou X, Zhang J, Wang J, et al. Radiomics based on 18f-fdg pet/ct could differentiate breast carcinoma from breast lymphoma using machine-learning approach : A preliminary study. *Cancer medicine*, 2020 ;9(2) :496–506. 45

Pan S J and Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009 ;22(10) :1345–1359. 94

Papp L, Pötsch N, Grahovac M, et al. Glioma survival prediction with combined analysis of in vivo 11c-met pet features, ex vivo features, and patient features by supervised machine learning. *Journal of Nuclear Medicine*, 2018 ;59(6) :892–899. 46

Parmar C, Grossmann P, Bussink J, et al. Machine Learning methods for Quantitative Radiomic Biomarkers (Supplement). *Sci Rep*, 2015 ;5(1) :13087. doi :10.1038/srep13087. 43

Peng H, Dong D, Fang M J, et al. Prognostic value of deep learning pet/ct-based radiomics : potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. *Clinical Cancer Research*, 2019 ;25(14) :4271–4279. 45, 47

Playout C, Duval R, and Cheriet F. A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018 ; pages 101–108. 92

Prasanna P, Patel J, Partovi S, et al. Radiomic features from the peritumoral brain parenchyma on treatment-naive multi-parametric mr imaging predict long versus short-term survival in glioblastoma multiforme : preliminary findings. *European radiology*, 2017 ; 27(10) :4188–4197. 89

Prokop M, Shin H O, Schanz A, et al. Use of maximum intensity projections in ct angiography : a basic review. *Radiographics*, 1997 ;17(2) :433–451. 75

Qu H, Riedlinger G, Wu P, et al. Joint segmentation and fine-grained classification of nuclei in histopathology images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019 ; pages 900–904. 100, 105, 106

Rajpurkar P, Irvin J, Zhu K, et al. Chexnet : Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv :171105225*, 2017 ;. 71

Ratner A, Bach S, Varma P, et al. Weak supervision : the new programming paradigm for machine learning. *Hazy Research Available via https ://dawn cs stanford edu//2017/07/16/weak-supervision/ Accessed*, 2019 ;pages 05–09. 34

Ratner A, De Sa C, Wu S, et al. Data programming : Creating large training sets, quickly. *Advances in neural information processing systems*, 2016 ;29 :3567. 33

Ravì D, Wong C, Deligianni F, et al. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 2016 ;21(1) :4–21. 2

Ren C, Zhang J, Qi M, et al. Machine learning based on clinico-biological features integrated 18 f-fdg pet/ct radiomics for distinguishing squamous cell carcinoma from adenocarcinoma of lung. *European Journal of Nuclear Medicine and Molecular Imaging*, 2020 ;pages 1–12. 45

Ren S, He K, Girshick R, et al. Faster r-cnn : Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 2016 ;39(6) :1137–1149. 92

Ribeiro M T, Singh S, and Guestrin C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016 ; pages 1135–1144. 40, 115

Rizk N P, Tang L, Adusumilli P S, et al. Predictive value of initial PET-SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. *J Thorac Oncol*, 2009 ;4(7) :875–879. doi :10.1097/JTO.0b013e3181a8cebf. 19

Ronneberger O, Fischer P, and Brox T. U-net : Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015 ; pages 234–241. 82, 88

Rosenblatt F. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 1958 ;65(6) :386. 34

Ruder S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv :170605098*, 2017 ;. 94

Rumelhart D E, Hinton G E, and Williams R J. Learning representations by back-propagating errors. *nature*, 1986 ;323(6088) :533–536. 36

Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015 ;115(3) :211–252. 65

Selvaraju R R, Cogswell M, Das A, et al. Grad-cam : Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 2017 ; pages 618–626. 72, 96

Shen R, Olshen A B, and Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009 ;25(22) :2906–2912. 1

Shi Z, Zhang C, Welch M, et al. Ct-based radiomics predicting hpv status in head and neck squamous cell carcinoma. In *Radiotherapy and Oncology*, volume 133. ELSEVIER IRELAND LTD ELSEVIER HOUSE, BROOKVALE PLAZA, EAST PARK SHANNON, CO . . . , 2019 ; pages S515–S515. 44

Simonyan K, Vedaldi A, and Zisserman A. Deep inside convolutional networks : Visualising image classification models and saliency maps. *arXiv preprint arXiv :13126034*, 2013 ;. 39, 115

Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :14091556*, 2014 ;. 95, 100

Sollini M, Cozzi L, Antunovic L, et al. Pet radiomics in nsclc : state of the art and a proposal for harmonization of methodology. *Scientific reports*, 2017 ;7(1) :358. 53

Sorani M D, Ortmann W A, Bierwagen E P, et al. Clinical and biological data integration for biomarker discovery. *Drug discovery today*, 2010 ;15(17-18) :741–748. 1

Srivastava N, Hinton G, Krizhevsky A, et al. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*, 2014a ;15 :1929–1958. doi :10.1214/12-AOS1000. 2

Srivastava N, Hinton G, Krizhevsky A, et al. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014b ;15(1) :1929–1958. 65

Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015 ; pages 1–9. 71, 91

Szegedy C, Toshev A, and Erhan D. Deep neural networks for object detection. In *Advances in neural information processing systems*. 2013 ; pages 2553–2561. 88

Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016 ; pages 2818–2826. 100

Tewari K S and Monk B J. New strategies in advanced cervical cancer : from angiogenesis blockade to immunotherapy. *Clinical Cancer Research*, 2014 ;20(21) :5349–5358. 12

Therasse P, Arbuck S G, Eisenhauer E A, et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *Clin Trials*, 2005 ;12(3) :16–27. 13

Thibault G, Fertil B, Navarro C, et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. *Pattern Recognit Inf Process*, 2009 ;pages 140–145. doi :Artn1357002\rDoi10.1142/S0218001413570024. 22, 24

Thome N, Bernard S, Bismuth V, et al. Multitask classification and segmentation for cancer diagnosis in mammography. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*. 2019 ; . 91

Tixier F, Cheze-Le Rest C, Hatt M, et al. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. *J Nucl Med*, 2011a ;52(3) :369–378. doi : 10.2967/jnumed.110.082404. 24, 41, 47

Tixier F, Le Rest C C, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine*, 2011b ;52(3) :369–378. 52

Toyama Y, Hotta M, Motoi F, et al. Prognostic value of fdg-pet radiomics with machine learning in pancreatic cancer. *Scientific reports*, 2020 ;10(1) :1–8. 45

Trullo R, Petitjean C, Nie D, et al. Joint segmentation of multiple thoracic organs in ct images with two collaborative deep architectures. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 21–29. Springer, 2017 ;. 88

Valencia R, Denecke T, Lehmkuhl L, et al. Value of axial and coronal maximum intensity projection (mip) images in the detection of pulmonary nodules by multislice spiral ct : comparison with axial 1-mm and 5-mm slices. *European radiology*, 2006 ;16(2) :325–332. 75

Van De Wiele C, Kruse V, Smeets P, et al. Predictive and prognostic value of metabolic tumour volume and total lesion glycolysis in solid tumours. *Eur J Nucl Med Mol Imaging*, 2013 ;40(2) :290–301. doi :10.1007/s00259-012-2280-z. 42

Van Griethuysen J J, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 2017;77(21) :e104–e107. 44

Vargas A J and Harris C C. Biomarker development in the precision medicine era : lung cancer as a case study. *Nature Reviews Cancer*, 2016;16(8) :525. 41, 136

Vera P, Mezzani-Saillard S, Edet-Sanson A, et al. FDG PET during radiochemotherapy is predictive of outcome at 1 year in non-small-cell lung cancer patients : a prospective multicentre study (RTEP2). *Eur J Nucl Med Mol Imaging*, 2014;41 :1057–1065. doi : 10.1007/s00259-014-2687-9. 13

Vesal S, Patil S M, Ravikumar N, et al. A multi-task framework for skin lesion detection and segmentation. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 285–293. Springer, 2018a;. 92

Vesal S, Ravikumar N, and Maier A. Skinnet : A deep learning framework for skin lesion segmentation. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. IEEE, 2018b ; pages 1–3. 92

Wagenknecht G, Kaiser H J, Mottaghy F M, et al. Mri for attenuation correction in pet : methods and challenges. *Magnetic resonance materials in physics, biology and medicine*, 2013;26(1) :99–113. 17

Wahl R L, Jacene H, Kasamon Y, et al. From RECIST to PERCIST : Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med*, 2009;50 Suppl 1(5) :122S—-50S. doi :10.2967/jnumed.108.057307. 13, 19

Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 2014;11(3) :333. 1

Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 f-fdg pet/ct images. *EJNMMI research*, 2017a;7(1) :1–11. 46

Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI Res*, 2017b;7(1) :11. doi :10.1186/s13550-017-0260-9. 48, 54

Wang J, Wu C J, Bao M L, et al. Machine learning-based analysis of mr radiomics can help to improve the diagnostic performance of pi-rads v2 in clinically relevant prostate cancer. *European radiology*, 2017c;27(10) :4082–4090. 43

Wang P, Patel V M, and Hacihaliloglu I. Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided cnn. In *International conference on medical image computing and computer-assisted intervention*. Springer, 2018 ; pages 134–142. 100

Wei L, Rosen B, Vallières M, et al. Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling. *Physics and Imaging in Radiation Oncology*, 2019;10 :49–54. 44

Willaime J, Turkheimer F, Kenny L, et al. Image descriptors of intra-tumor proliferative heterogeneity predict chemotherapy response in breast tumors. *Journal of Nuclear Medicine*, 2012;53(supplement 1) :387–387. 47

Woodard H Q, Bigler R E, Freed B, et al. Expression of tissue isotope distribution. *Journal of Nuclear Medicine*, 1975;16(10) :958–959. 18

Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Frontiers in oncology*, 2016;6 :71. 43

Xie C, Du R, Ho J W, et al. Effect of machine learning re-sampling techniques for imbalanced datasets in 18 f-fdg pet-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *European journal of nuclear medicine and molecular imaging*, 2020;47(12) :2826–2835. 45

Yang X, Zeng Z, Yeo S Y, et al. A novel multi-task deep learning model for skin lesion segmentation and classification. *arXiv preprint arXiv :170301025*, 2017;. 90

Yip S S F and Aerts H J W L. Applications and limitations of radiomics. *Phys Med Biol*, 2016;61(13) :R150—-R166. doi :10.1088/0031-9155/61/13/R150. 1, 41

Ypsilantis P P, Siddique M, Sohn H M, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. *PLoS One*, 2015; 10(9) :e0137036. doi :10.1371/journal.pone.0137036. 47, 48, 53, 54, 55, 60, 64, 65, 136

Zeiler M D and Fergus R. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 2014; pages 818–833. 71

Zeng G L. *Medical image reconstruction : a conceptual tutorial*. Springer, 2010. 4

Zhang Y and Yang Q. A survey on multi-task learning. *arXiv preprint arXiv :170708114*, 2017;. 34, 89

Zhao Z Q, Zheng P, Xu S t, et al. Object detection with deep learning : A review. *IEEE transactions on neural networks and learning systems*, 2019;30(11) :3212–3232. 2

Zhong J, Frood R, Brown P, et al. Machine learning-based fdg pet-ct radiomics for outcome prediction in larynx and hypopharynx squamous cell carcinoma. *Clinical Radiology*, 2021;76(1) :78–e9. 45

Zhou B, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv :14126856*, 2014;. 71

Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016; pages 2921–2929. 71, 72

Zhou H, Vallières M, Bai H X, et al. Mri features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro-oncology*, 2017;19(6) :862–870. 43

Zhou J and Troyanskaya O G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 2015;12(10) :931–934. 39, 114

Zhou Y, Chen H, Li Y, et al. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, 2020 ;page 101918. 92, 96

Zhou Y, Chen H, Li Y, et al. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, 2021 ; 70 :101918. 100, 105, 106

Zhou Y, Xu J, Liu Q, et al. A radiomics approach with cnn for shear-wave elastography breast tumor classification. *IEEE Transactions on Biomedical Engineering*, 2018a ; 65(9) :1935–1942. 55, 73

Zhou Y, Zhu Y, Ye Q, et al. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018b ; pages 3791–3800. 72

Zhou Z H. A brief introduction to weakly supervised learning. *National science review*, 2018 ;5(1) :44–53. 33, 34

List of figures

1.1 Normal and cancerous cells : How Are they different. Source : verywellhealth 6
1.2 Difference between external and internal (brachytherapy) radiation therapy. Source : Equicarehealth. . . . 11
1.3 Example of different information, including the radiomic information of the image contributing to a personalized treatment, according to [Lambin et al. 2012b] . . . 12
1.4 Discovery IQ PET/CT scanner image courtesy of GE Healthcare. . . . 15
1.5 Diagram of the PET scan acquisition process. Source : Wikipedia. . . . 15
1.6 After a short distance, the positron e^+ obtained by emission β^+ is annihilated with an electron e^- giving birth to two photons γ emitted in the same direction, in opposite direction at 180° from each other and with an energy of 511 keV each. . . . 15
1.7 Types of coincidences recorded by the detection system : (a) true coincidence, (b) diffuse coincidence, (c) fortuitous coincidence and (d) multiple coincidence. . . . 16
1.8 Example of the creation of three lines of the sinogram according to the projection angle. Source : Tylski . . . 17
1.9 Stereo skeletal formula of fluorodeoxyglucose (18F). Source : Wikipedia. . . . 17
1.10 (a) Cross-section PET to FDG and (b) MIP of a patient with an esophageal cancer. The tumor appears stained on the MIP. Other organs with normal FDG fixation are : the brain due to its permanent activation, the kidneys and bladder for their filtration role. . . . 18
1.11 Representation of SUV derivatives of a lesion. Source : Orlhac. . . . 20
1.12 Images and corresponding histograms . . . 21

2.1 Artificial intelligence, machine learning and deep learning. . . . 29
2.2 Various categories of approaches for structuring the unstructured information. Source : towardsdatascience. . . . 31
2.3 Similarities between biological neuron (left) and artificial neuron (right). Source : Wikimedia. . . . 35
2.4 Popular nonlinear activation functions used for NN training. Source : [Feng et al. 2019] . . . 35
2.5 An example of a multi-layer perceptron. . . . 36
2.6 An example of back propagation. By updating the weight w and the bias b through back propagation, the model minimize the loss L by approximating the ground truth y. Source : Javaid. . . . 37
2.7 Illustration of a deep learning model. Source : deeplearningbook [Goodfellow et al. 2016]. . . . 38

2.8 Sparse connectivity (top) compared to fully connectivity (bellow). Using a kernel of size 3, only 3 S outputs are affected by input x3 (top). With a fully connectivity, all S outputs are affected by x3. Source : deeplearningbook [Goodfellow et al. 2016]. 39
2.9 The receptive field of the units in the deeper layers of a convolutional network is larger than the receptive field of the units in the shallow layers. This means that even though direct connections in a convolutional net are very sparse, units in the deeper layers can be indirectly connected to all or most of the input image. Source : deeplearningbook [Goodfellow et al. 2016]. . . . 40
2.10 Precision medicine allows to separate patients into different groups to personalize treatment. Source : [Vargas and Harris 2016]. 41
2.11 Current workflow of radiomics based on machine learning and features selection strategy. 43
2.12 Deep radiomics workflow. The model extract features and predict the outcome jointly. 47
2.13 ROIs of a specific tumor i after segmentation embedded into larger square background of standard size of 100 × 100 pixels. From [Ypsilantis et al. 2015] 48

3.1 Columns from left to right : Fused PET/CT slice, zoomed on the esophageal tumor seen on FDG-PET only. MTV (40% SUVmax thresholding) in red and MTV included in the cuboid. MTV3 (MTV + 3 cm isotropic margin) in white and MTV3 included in the cuboid. 53
3.2 3D RPET-NET architecture composed by two 3D convolutional layers followed by 3D pooling layers and two dense layers. 57
3.3 Visualization of a 2D slice of a segmented tumor and the resulting 32 feature maps in the second convolutional layer of the 1S-CNN architecture. 58
3.4 a. On the left : ROC curve comparing the 6 classifiers (RF, GARF, FIC, 1S-CNN, 3S-CNN and 3D RPET-NET) with the best parameters on MTV. b. Right : Comparison of the four classifiers on different VOIs (MTVs). Error bars correspond to standard deviation. 64

4.1 An example of a PET image with oesophagus cancer on the left in which the tumor is barely visible, and the same image on the left in which the localisation of the tumor is shown in red color. It is not straightforward to learn the difference between tumor fixation and a normal fixation in a PET image. . . 71
4.2 Maximum intensity projection (MIP) of PET exam. A) projection in Sagittal. B) Projection in Coronal 72
4.3 The proposed architecture. The MIP is used to predict the pathology. A heat map is generated from the last conolutional layer 73
4.4 Our proposed architecture. The neural network learns to classify the type of cancer from two 2D MIP images (sagittal and coronal). The generated heatmap is back-propagated and corrected to identify accurately tumor regions. 78
4.5 Distance matrix between p at the center of the tumor and the points qi generated by the heat map. A) is a Coronal MIP for a patient with esophageal cancer. A point p is randomly defined at the tumor region. B) is the heat map generated using our proposed model. C) shows the overelay of the MIP and the heat map. D) is the distance matrix showing the distance between the points qi generated by the heat map and the point p. 81

4.6 Segmentation : the 3D tumor region from the two 2D heat maps. Coronal heat map allows to retrieve y and z axis, while sagittal heat map return x and z axis. The tumor is selected by the intersection of the two heat maps. 82
4.7 3D RPET-NET architecture composed by two 3D convolutional layers followed by 3D pooling layers and two dense layers. 82
4.8 Comparison between different models. From left to right : PET exam, CAMs without prior knowledge, ours 84

5.1 Columns from left to right : Fused PET/CT slice, zoomed on the lung tumor (left) and esophageal tumor (right) seen on FDG-PET only. Metabolic Tumor Volume MTV (40% SUVmax thresholding) in red. MTV3 (MTV + 3 cm isotropic margin) include the tumor and peritumoral region. 89
5.2 Hard parameter sharing for multi-task learning in deep neural networks used in our proposed architecture. 91
5.3 Our proposed architecture, composed of an encoder and two decoders for image reconstruction and tumor segmentation. A fully connected layers are added for classification (Oesophageal vs lung cancer), and a multi-scale outcome prediction. 93
5.4 Heatmap from different scales with two different input. 95

List of tables

1.1 Regrouping of stages T (tumor), N (nodes) and M (metastatis) in a single stage TNM. . . . 9
1.2 Different properties that can be calculated using the histogram. . . . 21
1.3 Different properties that can be calculated using the Co-occurence matrix. . 23
1.4 Main statistical characteristics of 2nd and higher order . . . 23

2.1 . . . 45

3.1 Classification results : Each result corresponds to the average of five independent experiments and the standard deviation, using the training dataset (Experiment 1) or the test dataset (Experiment 2 and 3). . . . 62

4.1 Results for 3D segmentation. WPk : without prior knowledge. CAM : class activation map. . . . 81
4.2 Results for radiomics analysis. WPk : without prior knowledge. Ms : manual segmentation . . . 84

5.1 Results of experiment 1 : segmentation, classification and prediction results from different scenarios, for esophageal and lung cancers. Task1 : reconstruction, Task2 : segmentation, Task3 : pathology classification, Task4 : outcome prediction. . . . 101
5.2 Experiment 2 : Segmentation results for esophageal and lung cancer compared to the state of the art methods. WSL : weakly supervised learning model developed in [Amyar et al. 2020]. . . . 103
5.3 Experiment 3 : classification and outcome prediction results compared to state of the art methods for esophageal and lung cancers. . . . 103
5.4 Experiment 4 : The effects of λ on the multi-task learning. . . . 105
5.5 Experiment 5 : A quantitative comparison between our model and state of the art multi-task methods. . . . 105

Glossaire

AB *Adaptive Boosting*. 48

AI *Artificial Intelligence*. 28

ANN *Artificial Neural Network*. 28, 34, 40, 48

AUC *"Area Under ROC Curves"*. 24

CAD *Computer Aided Diagnostic*. V, VII

CNN *Convolutional Neural Network*. VI, VIII, 3, 4, 37, 38, 40, 47, 48, 70

CT *Computed Tomography*. V, VII, 4, 7, 14, 17, 44, 48

DL *Deep Learning*. V, 28, 49

FDG *2-[18]-Fluoro-2-desoxy-D-glucose*. 14, 17, 18, 25, 135

FMRI *Functional Magnetic Resonance Imaging*. 14

GLCM *"Gray Level Cooccurrence Matrix"*. 22–24, 44

GLDM *"Gray Level Difference Matrix"*. 22–24, 44

GLNUr *"Gray Level Non-Uniformity"*. 23

GLNUz *"Gray Level Non-Uniformity"*. 23, 24

GLRLM *"Gray Level Run Length Matrix"*. 22–24, 42, 44

GLSZM *"Gray Level Size Zone Matrix"*. 22–24, 42, 44

HGRE *"High Gray-level Run Emphasis"*. 23

HGZE *"High Gray-level Zone Emphasis"*. 23

IDM *Moment différentiel inverse*. 23

LGRE *"Low Gray level Run Emphasis"*. 23

LGZE *"Low Gray level Zone Emphasis"*. 23

LRE *"Long Run Emphasis"*. 23

LRHGE *"Long Run High Gray-level Emphasis"*. 23

LRLGE *"Long Run Low Gray-level Emphasis"*. 23

LZE *"Long Zone Emphasis"*. 23

LZHGE *"Long Zone High Gray-level Emphasis"*. 23

LZLGE *"Long Zone Low Gray-level Emphasis"*. 23

MIP *"Maximum Intensity Projection"*. VI, VIII, 18, 135

ML *Machine Learning*. 25, 28, 29, 32, 33, 40, 44, 48

MLP *Multi-Layer Perceptron*. 36, 38, 40, 42, 44

MRI *Magnetic Resonance Imaging*. V, VII, 7, 14, 17, 44

MTL *Multi-Task Learning*. VI, VIII, 4, 34, 40

NN *Neural Network*. 4, 34–36, 40, 135

PET *Positron Emission Tomography*. V–VIII, 3, 7, 13, 14, 17–19, 24, 25, 44, 45, 47, 48, 70, 73, 135

RF *Random Forest*. V, VII, 42–44, 48, 49

RLNU *"Run Length Non-Uniformity"*. 23

ROC *"Receiver Operating Characteristic"*. 24

ROI *Region Of Interest*. V, VII, 43, 44

RPr *"Run Percentage"*. 23

SRE *"Short Run Emphasis"*. 23

SRHGE *"Short Run High Gray-level Emphasis"*. 23

SRLGE *"Short Run Low Gray-level Emphasis"*. 23

SUV *"Standardized Uptake Value"*. 18, 19

SVM *Support Vector Machine*. V, VII, 42–44, 48, 49

SZE *"Short Zone Emphasis"*. 23

SZHGE *"Short Zone High Gray-level Emphasis"*. 23

SZLGE *"Short Zone Low Gray-level Emphasis"*. 23

VOC *Coefficient Of Variation*. 22

WSL *Weakly Supervised Learning*. VI, VIII, 33, 40, 72

ZLNU *"Zone Length Non-Uniformity"*. 23, 24

ZP *"Zone Percentage"*. 23