Normandie Université

UNIVERSITÉ DE ROUEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Rouen Normandie**

Utilisation des réseaux de neurones profondsen imagerie multi-modalité pour la prédiction de la réponse au traitement et la survie cancérologie -Application au protocole RTEP7-IFCT 14.02

Présentée et soutenue par

MESBAH ZACHARIA

Thèse soutenue le 27/05/2026

devant le jury composé de :

M. THUREAU SEBASTIEN YVES	Professeur des Univ - Prat Hospitalier - URN - Université de Rouen Normandie	Directeur de thèse
M. MODZELEWSKI ROMAIN	Docteur - URN - Université de Rouen Normandie	Co-encadrant
M. HERAULT ROMAIN	Professeur des Universités - UCN - Université de Caen Normandie	Membre du jury
MME KHAMPHAN CATHERINE	Docteur - INSTITUT DU CANCER AVIGNON PROVENCE - SAINTE CATHERINE	Membre du jury
MME SCHICK ULRIKE	Professeur des Univ - Prat Hospitalier - Universite Bret. Occidentale Ubo	Membre du jury
MME ANTONI DELPHINE	Professeur des Univ - Prat Hospitalier - Université de Strasbourg	Rapporteur du jury
M. LUCIA FRANCOIS	Maître de Conf Univ - Prat Hospitalier - Université Catholique de l'Ouest Bretagne Nord	Rapporteur du jury

Thèse dirigée par **THUREAU SEBASTIEN YVES** (ANALYSE INTEGREE MULTIMODALE EN SANTE)

ED 497 NBISE
Normande de Biologie Intégrative, Santé, Environnement

AIMS
Analyse Intégrée Multimodale en Santé

Thèse effectuée au sein du Centre Henri Becquerel



Thèse CIFRE financée par Siemens Healthineers et l'ANRT





Remerciements

I would like to thank my thesis director Sébastien Thureau and my co-supervisor Romain Modzelewski for including and guiding me in this PhD. My gratitude also goes to Pierrick Gouel for his continued guidance and enlightening discussions, first during my internship, then during my PhD. Many thanks to Pierre Vera, the head of the AIMS lab and director of the Henri Becquerel Cancer Center for hosting me.

My colleagues Solène Perret, Léo Mottay and Dhruv Jain have helped me on some of the projects presented in this thesis. They contributed in many ways to the work presented below, through discussions, ideas sharing, implementations and presentation. A special thanks to Dhruv Jain for the collaboration on other projects which helped me gain expertise on topics relevant to my thesis.

I would like to acknowledge the work of the staff of Litis lab, nBISE doctoral school, AIMS lab and Rouen Université, Lynda Letêtu, Brigitte Diarra and Céline Breton. They made this thesis possible administratively. A special thanks to Mariana Rubio for overseeing my CSI, as well as Ulrike Schick, Joël Castelli and Julien Serrier for participating in it.

I would like to thank my colleagues at Centre Henri Becquerel for their daily participation, their advice and patience. In particular, thank you my fellow students, Léo Mottay, Solène Perret, Aghiles Kebaili, Tristan Loukianenko, Aya Terro, Tom Simon, Adrien Pommier, Mohammad Yamout and Zexin Ji. My gratitude also goes to the senior researchers at Centre Henri Becquerel for their insight on physics and medical problems: Pierre Decazes, Arnaud Dieudonné, Arthur Decazes and Pierre Bohn.

Many thanks to Matthieu Lepetit-Coiffé and Cédric Pasquier for helping me integrate at Siemens Healthineers.

Finally, I would like to thank my family and friends, for their support, patience and listening through the ups, downs, achievements and many setbacks which occurred during the last four and a half years.

Abstract

Multi-modality imaging has helped physicians to understand the human body and how it is affected by oncologic diseases. A growing number of imaging exams is performed each year and the practice of using these images for treatment has evolved towards personalized, precise care protocols.

To handle this large amount of data, research on artificial intelligence has been extended to medical images, to solve simplified versions of tasks from clinical practice. The contouring of structures, such as organs and lesions, in single or multi-modality imaging, has seen the most progress, through the collection of large annotated datasets and the publication of state-of-the-art segmentation tools. Automated and semi-automated contouring tools were the first successful applications of A.I. on medical images and were made possible thanks to community efforts.

This thesis presents how A.I. can be applied to accelerate the oncology workflow, using deep neural networks on multi-modality imaging for organs and lesions segmentation; and machine learning on imaging features to predict survival and support physician decision.

We developed an improved segmentation algorithm to **delineate organs-at-risk** in the head and neck area, on CT/MRI imaging. Our algorithm reached the third place in the HaN-Seg challenge, with a performance on par with human experts. This convincing result was obtained, in spite of the small size of the training dataset (42 patients), thanks to a specifically designed algorithm which relies on large scale pre-training and modified inference. Constraints on the runtime and inference hardware demonstrated the adaptability of SoTA segmentation methods based on deep neural networks to the conditions of clinical practice.

Focusing on nuclear imaging, we then tackled the task of **automated lesion segmentation on PET/CT** scans. For 3 years, we participated in the AutoPET challenge series and developed novel algorithms to improve over the SoTA and solve new problems inspired from clinical reality.

Our first algorithm achieved a higher performance than the baseline model trained with nnUNet, for automated lesion segmentation on FDG PET/CT scans.

The next year, we developed an algorithm capable of performing lesion segmentation on both PSMA and FDG PET/CT scans, using a tracer discriminator.

Finally, we adapted our algorithm to handle user input and perform semi-automated lesion segmentation. To feed user inputs to the DNN, we designed PET-aware lesion probability heatmaps which guided the segmentation.

Radiomic features extracted from the PET/CT contours can be used to forecast events, such as death or relapse. To improve the performance of machine learning models which use radiomic features as input, we developed **Radiomics Augmentation**.

This framework applies data augmentation transforms on the images and contours to create new synthetic samples. The goal of this framework is to reduce the over-fitting phenomenon, by filling holes in the data distribution, through the addition of modified but realistic samples.

Out of the 8 models we benchmarked, 7 showed an improved performance when using Radiomics Augmentation (from 0.2% to 8.6%) and 4 models exhibited a globally reduced over-fitting.

Going beyond lesion radiomics, we explored organ-level radiomics: **organomics**.

To this end, we developed and published Organ-omics, an automated organomics extraction pipeline, which relies on TotalSegmentator and PyRadiomics.

We applied Organ-omics to an internal dataset of NSCLC patients, which gathered PET/CT scans at 2 time-points for each patient. Using this dataset, we compared the predictive power of organomics, clinical data and lesion radiomics for overall survival.

We found that organomics largely outperformed clinical data ($p \ll 0.001$) and lesion radiomics ($p \ll 0.001$). On top of that, we showed that organomics extracted from CT images performed better than PET organomics ($p \ll 0.001$).

Through these contributions, we showed how A.I. could be applied to medical images in oncology. Deep Learning methods can generate accurate contours of organs and lesions from medical images. These contours can then be used to extract descriptive radiographic features such as radiomics or organomics. Radiomics and organomics can serve as inputs to machine learning models, to predict survival and treatment response.

The novel methods we have developed were made easily accessible, through the release of public tools.

Résumé

L'imagerie multi-modalité a ouvert la voie à une meilleure compréhension du corps humain, ainsi que de l'impact que les pathologies oncologiques ont sur celui-ci. Le nombre de ces examens a augmenté rapidement ces dernières années et les médecins les ont utilisés pour construire des protocoles de traitement adaptés à chaque patient.

Pour traiter cette quantité grandissante d'images, la recherche en intelligence artificielle a été étendue à l'imagerie médicale. Le contourage de structures, telles que les organes et les lésions, dans des images simples ou multi-modalités est le domaine qui a connu le plus de progrès. Ce progrès a été rendu possible notamment par la récolte d'importants jeux de données annotés, ainsi que par l'émergence d'outils puissants, servant de *baseline* robuste. Des outils de segmentation automatique et semi-automatique, la première application réussie de l'I.A. sur des images médicales, ont été construits grâce aux efforts collectifs de la communauté.

Cette thèse présente comment l'I.A. peut être appliquée pour accélérer le travail des médecins en oncologie, en utilisant des réseaux de neurones profonds pour contourer les organes et les lésions; ainsi que de l'apprentissage machine sur les caractéristiques extraites de ces contours, pour assister la décision médicale.

Nous avons développé un algorithme capable de contourer les organes à risque de la tête et du cou sur des images TDM/IRM. Cet algorithme a obtenu la troisième place dans le challenge HaN-Seg, avec une performance au niveau de celle d'experts humains. Ce résultat a été obtenu en dépit de la taille réduite du jeu de données d'entraînement (42 patients), grâce à l'utilisation de pré-entraînement à grande échelle et une inférence adaptative. Des contraintes placées sur le temps d'exécution et le matériel d'inférence ont permis de démontrer que les algorithmes à l'état de l'art peuvent être adaptés pour se conformer aux conditions de la routine clinique.

Nous nous sommes par la suite concentrés sur l'imagerie nucléaire, avec en premier lieu le contourage automatique des lésions sur les scanners TEP/TDM. Durant 3 années consécutives, nous avons contribué aux challenges AutoPET en développant des algorithmes innovants, afin d'améliorer l'état de l'art et de l'adapter aux nouveaux problèmes proposés par les organisateurs, inspirés de la réalité de la pratique clinique. Notre premier algorithme a obtenu une performance meilleure que la *baseline*, un modèle entraîné avec nnUNet, pour le contourage de lésions sur TEP/TDM au FDG. Pour le second algorithme, nous avons étendu les capacités de segmentation aux TEP/TDM au PSMA, avec une détermination automatique du type de traceur utilisé. Enfin, nous avons ajouté la

capacité pour un utilisateur de guider l'algorithme, rendant le contourage semi-automatique, grâce à des cartes de probabilité innovantes, informées par la TEP.

Une fois les lésions contourées, des caractéristiques radiomiques peuvent être extraites de l'image pour prédire des évènements, tels que le décès et la rechute. Pour améliorer les performances des modèles d'apprentissage automatique qui utilisent les radiomiques comme entrées, nous avons développé *Radiomics Augmentation*. Ce framework crée de nouvelles données d'entraînement en appliquant des augmentations de données sur les images et les contours, avant d'extraire les radiomiques. L'objectif de ce framework est de réduire le sur-apprentissage afin d'améliorer les performances des modèles, en remplissant les parts de la distribution initialement absentes du jeu d'entraînement. Sur 8 modèles que nous avons évalués, 7 ont montré des performances améliorées (de +0.2% à +8.6%) et 4 ont montré une réduction du sur-apprentissage.

Pour aller au-delà des lésions, nous avons exploré les radiomiques des organes: les organomiques. Pour cela, nous avons développé et publié Organ-omics, un pipeline d'extraction des organomiques, qui repose sur TotalSegmentator et PyRadiomics. Nous avons appliqué Organ-omics à un jeu de données issu d'une étude clinique incluant des patients souffrant de CBNPC. Deux imageries TEP/TDM ont été récoltées pour chaque patient, avant et pendant le traitement par radiothérapie. En comparant les performances prédictives des données cliniques, radiomiques de lésions et des organomiques, nous avons trouvé que les organomiques étaient beaucoup plus performantes que les données cliniques ($p \ll 0.001$) et les radiomiques ($p \ll 0.001$) pour prédire la survie globale. De plus, les organomiques extraites du CT ont obtenu de meilleurs résultats que celles extraites du PET ($p \ll 0.001$).

Grâce à ces travaux, nous avons démontré comment l'I.A. pouvait être appliquée à l'imagerie médicale en oncologie. Les méthodes d'apprentissage profond permettent de générer les contours des organes et des lésions à partir de ces images. Ces contours peuvent ensuite être utilisés pour extraire des caractéristiques radiographiques afin d'entraîner des modèles d'apprentissage automatique, pour prédire la survie et la réponse au traitement. Ainsi l'I.A. peut assister les décisions médicales.

Les nouvelles méthodes que nous avons développées sont facilement accessibles grâce à la mise à disposition d'outils publics.

List of Acronyms

A.I. Artificial Intelligence

CNN Convolutional Neural Network

CT Computed Tomography

CPU Central Processing Unit

DL Deep Learning

DSC Dice-Sorensen Similarity Coefficient

DNN Deep Neural Network

EHR Electronic Health Records

FOV Field Of View

GPT Generative Pre-trained Transformer

GT Ground Truth

IoU Intersection over Union

IP Internet Protocol

LLM Large Language Model

ML Machine Learning

MRI Magnetic Resonance Imaging

NLP Natural Language Processing

OAR Organs-At-Risk

OS Overall Survival

PET Positron Emission Tomography

RGB Red Green Blue

SoTA State of The Art

SSL Self-Supervised Learning

TOF Time-Of-Flight

TCP Transport Control Protocol

UUID Universally Unique Identifiers

ViT Visual Transformer

VLM Vision Language Model

XR X-rays

Contents

1 Introduction **1**

1.1 Using Artificial Intelligence in Medical Image Analysis . . . 2
1.2 Problem Statement . . . 3
1.3 Thesis structure . . . 4
1.4 Contributions and Publications . . . 4

2 A.I. for Medical Imaging **7**

2.1 Advances in Artificial Intelligence . . . 9
2.1.1 Artificial Intelligence, Machine Learning and Deep Learning . . . 9
2.1.2 The Origins of Deep Learning . . . 9
2.1.3 Transformers and Large Language Models . . . 10
2.2 Tasks & Paradigms for Artificial Intelligence . . . 11
2.2.1 Tasks Nomenclature for Artificial Intelligence . . . 11
2.2.2 Artificial Intelligence Training Paradigms . . . 15
2.3 Tools for Medical Image Analysis . . . 17
2.3.1 Medical Image Storage . . . 17
2.3.2 Medical Image Processing . . . 19
2.3.3 Tools for Machine learning with Medical Images . . . 20
2.3.4 Tools for Deep Learning with Medical Images . . . 21
2.4 Large Scale Oncology Imaging Datasets . . . 22
2.4.1 Limiting factors for large medical imaging datasets in oncology . . . 22
2.4.2 Examples of large datasets . . . 24
2.5 Conclusion . . . 26

3 Organ Segmentation in Medical Images **27**

3.1 Image Segmentation . . . 29
3.1.1 Semantic Segmentation vs Instance Segmentation . . . 29
3.1.2 Image segmentation using deep learning . . . 29
3.1.3 Losses and Metrics . . . 31
3.2 Medical Image Segmentation . . . 32
3.2.1 Specific Challenges with medical images . . . 32
3.2.2 Advances in automated medical images segmentation . . . 32
3.2.3 A Call for Rigorous Validation . . . 36
3.2.4 Advances in semi-automated medical images segmentation . . . 36

3.3 Head and Neck OARs Segmentation . . . 39
3.3.1 Context . . . 39
3.3.2 The HaN-Seg dataset and challenge . . . 42
3.3.3 Proposed Algorithm . . . 44
3.3.4 Results . . . 47
3.3.5 Discussion . . . 48
3.3.6 Conclusion . . . 49

4 PET/CT Lesion Segmentation 51
4.1 Context . . . 53
4.1.1 PET/CT imaging . . . 53
4.1.2 PET Quantification: Standardized Uptake Value . . . 55
4.1.3 PET Acquisition methods . . . 55
4.1.4 Multimodal PET/CT acquisition . . . 55
4.1.5 PET/CT Analysis . . . 56
4.2 AutoPET: data and challenge history . . . 57
4.2.1 Datasets . . . 57
4.2.2 Challenges . . . 57
4.3 Our methods . . . 59
4.3.1 AutoPET II . . . 59
4.3.2 AutoPET III . . . 60
4.3.3 AutoPET IV . . . 62
4.4 Discussion . . . 66
4.4.1 The Winning Solutions from Each Year . . . 66
4.4.2 Insights from AutoPET . . . 67
4.5 Conclusion . . . 68

5 Data Augmentation for Radiomics 71
5.1 Radiomics: high-dimensional medical imaging descriptors . . . 73
5.1.1 Medical imaging descriptors . . . 73
5.1.2 Radiomics: high dimensional descriptors . . . 73
5.1.3 Extracting Radiomics . . . 74
5.2 Survival Prediction using Radiomics . . . 75
5.2.1 Feature Selection . . . 76
5.2.2 Machine learning for survival prediction . . . 76
5.2.3 Metrics for Survival Prediction Evaluation . . . 77
5.3 Radiomics Augmentation against overfitting . . . 80
5.3.1 Context . . . 80
5.3.2 Material and methods . . . 81
5.3.3 Results . . . 87
5.3.4 Discussion . . . 90
5.3.5 Conclusion . . . 91

6 Organomics: Radiographic information from organs 93
6.1 The Organ-omics pipeline: Overview . . . 95
6.2 Detailed description . . . 95
6.2.1 Structure and Configuration . . . 95
6.2.2 Organs Segmentation . . . 97
6.2.3 Radiomics Extraction . . . 97
6.3 Applications and Limitations . . . 98
6.3.1 Applications . . . 98
6.3.2 Limitations and Future Improvements . . . 99
6.4 Application of organomics to survival prediction of NSCLC patients: The RTEP7-IFCT 1402 protocol . . . 100
6.4.1 Introduction . . . 100
6.4.2 Material and Methods . . . 101
6.4.3 Results . . . 106
6.4.4 Discussion . . . 109
6.4.5 Conclusion . . . 111

7 Conclusion and perspectives 113
7.1 Conclusion . . . 113
7.2 Perspectives . . . 114
7.2.1 Building foundational medical image segmentation models . . . 114
7.2.2 Improving the state of survival prediction . . . 115
7.2.3 Quantifying the whole human body . . . 116
7.3 Closing Remarks . . . 117

List of Figures 123

List of Tables 127

Appendix 1 129

Bibliography 137

Chapter 1

Introduction

Contents

1.1 Using Artificial Intelligence in Medical Image Analysis 2
1.2 Problem Statement 3
1.3 Thesis structure 4
1.4 Contributions and Publications 4

1.1 Using Artificial Intelligence in Medical Image Analysis

Over the years, the number of imaging exams performed increased consistently. For Computed Tomography (CT) scans, a report projected a 4% growth in the number of scans performed annually[1]. This increase in the volume of imaging is powered by the development of screening programs[2],[3], aiming at detecting diseases, such as lung cancer, early enough to maximize patient chances at survival, while reducing the need for heavier and more costly treatment options. For example, an earlier detection of cancerous lung nodules enables the use of minimally invasive surgery, such as video-assisted thoracoscopy[4], while stage III lung cancer is often inoperable and requires chemotherapy, which carries more adverse effects.

Along with this increase in the number of imaging scans, the nature of these scans has changed over time, with improvements brought to the imaging technologies themselves. Taking the example of Positron Emission Tomography (PET) scans, technological improvements such as TOF PET increased the image resolution and shortened the image acquisition time, which in turn increased the amount of information stored in a single scan volume.

Better resolution for each scan and more individual scans, translates in terms of computer science to more signal and more data. These characteristics were the ideal ground to start using artificial intelligence methods. Precisely, machine learning and deep learning methods can be used on this data to train prediction models for different patient outcomes or to obtain useful information from the scans.

The medical use-cases for these prediction models can be varied, like classifying patients as healthy or sick, determining if a nodule or tumor is benign or malignant, obtaining organ contours before planning external or internal radiotherapy treatment, or even estimating the risk of an adverse event (death, treatment side-effect, relapse) occurring after a given time. For most of these use cases, training a prediction model requires accumulating pairs consisting of the imaging data on one hand (PET, CT, MRI imaging volumes) and the clinical outcome (malignancy status, survival time) or medical decision (chosen treatment, organs delineation) on the other hand. Using these data pairs, prediction models can be trained to learn patterns and rely on them to extrapolate on new, unseen data, thus becoming a predictive tool which can support the decision of physicians, radiologists and oncology specialists.

In the long run, these predictive models could make the work of medical specialists more efficient, by streamlining tedious, repetitive and time-consuming tasks such as contouring healthy organs and target volumes in radiotherapy planning CT scans; or detecting, localizing and assessing the malignancy of lung nodules in a CT scan acquired during a screening program. The time saved by physicians could then be used to drive further progress on other research activities, such as research for improved treatment. This is summarized in figure 1.1.

Freeing up medical time could also allow doctors to dedicate more time to patient care. If coupled with adequate funding policies, this could result either in an increase in the time spent with each patient, improving the quality of listening and care, or in the possibility of treating a larger number of patients in a given time frame. The latter is particularly important, given the growing number of patients in oncology: WHO is projecting new cancer cases to grow past 35 million in 2050[5].

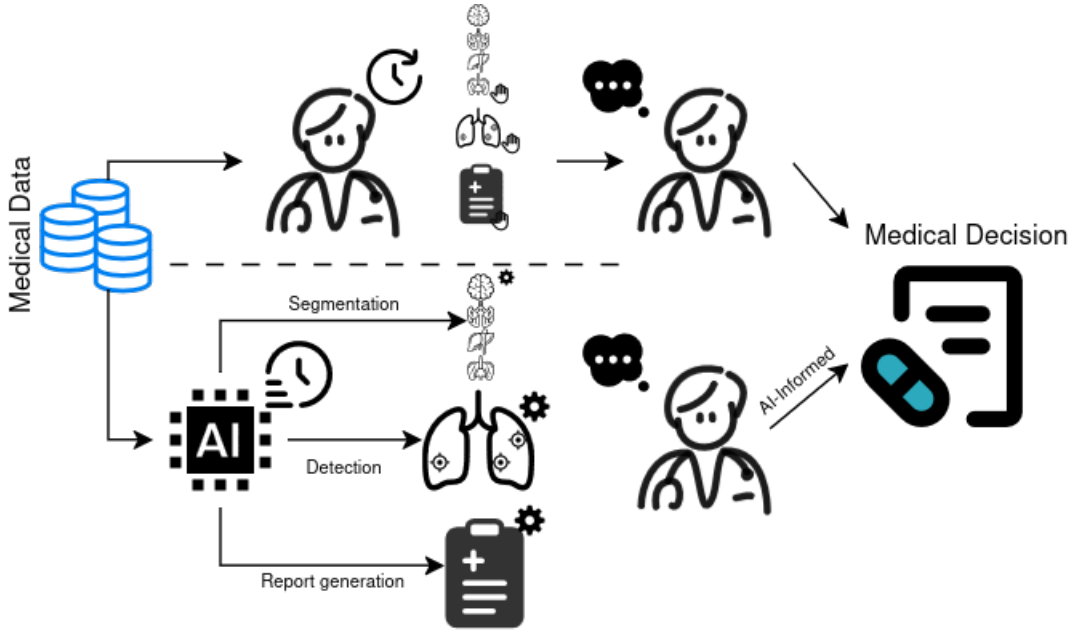

Figure 1.1: A structural improvement can be achieved using A.I. to help doctors with tedious, repetitive and time-consuming tasks. Some tasks, such as lesion detection, organ segmentation or even imaging report generation can be automated, saving the physician's time for the more complex medical decisions.

On top of that, the ability of these models to retrieve and process information from large medical images could be used to obtain imaging features relevant to a patient's health and the progress of their disease; a second model could then summarize the imaging and clinical characteristics into a meaningful patient-level risk score, to assist oncology specialists and physicians to choose the best treatment for each patient, individually.

1.2 Problem Statement

Deep neural networks have improved their capacity to a level where they can satisfyingly perform many tasks, either on their own, or in collaboration with a human user. Physicians and radiologists in oncology could benefit from the automated processing of medical images, as it would free their time for more treatment-oriented activities. However, while the applications of DNNs seem to extend to most domains, they are still making mistakes, which cannot be accepted for critical tasks. Indeed, since patients' lives are at stake, this automated processing has to be reliable, to ensure safe and broad adoption by medical experts.

Machine learning has also made important progress in the last decade, with an ability to aggregate information from several sources and use them to support decisions taken by humans. Unlike DNNs, ML models cannot handle huge amounts of information, but can be trained on much less data and tend to be more explainable. Their usage in oncology is often to choose patients, for a treatment, or additional screening, based on a risk score.

Our research aims to discuss the following question:

How can we apply deep neural networks to accelerate the oncology workflow, particularly organs and lesions segmentation; and use machine learning on these segmentations to support physician decision by predicting patient survival and treatment response?

1.3 Thesis structure

The next chapters in this thesis are structured as follows:

Chapter 2 gives background elements on Artificial Intelligence (A.I.) and its uses in medical imaging and oncology. We will present nomenclature for the different A.I. tasks and useful tools to analyze and learn from medical images.

Chapter 3 presents the state-of-the-art and existing works in medical image segmentation. The different model architectures and frameworks used by the medical imaging community for this task are presented, compared and discussed. Our work on head and neck Organs-At-Risk (OAR) segmentation is also presented, as an example of automated segmentation performing on par with human annotation.

Chapter 4 goes over the work accomplished to advance automated and semi-automated PET/CT segmentation. During our participation in the AutoPET challenges, we explored the state-of-the-art automated and semi-automated segmentation algorithms. We then developed improved methods for automated lesion segmentation on PSMA and FDG PET/CT scans, as well as a novel algorithm for semi-automated lesion segmentation.

Chapter 5 describes a framework applying data augmentations to radiomic features extraction. Data augmentations are widely used in deep learning, mainly to combat overfitting, thanks to their capacity to create slightly different, yet realistic data samples. We showed that applying data augmentations to the images and contours before radiomics extraction improves survival prediction performance, specifically on the HECKTOR dataset.

Chapter 6 introduces our most important contribution: Organ-omics. Building on very recent, groundbreaking work[6][7], we implemented and made publicly available a tool for organ-based radiomic features extraction. We then applied Organ-omics to data from the RTEP7 trial and demonstrated its superiority to using only lesion radiomics for OS prediction.

1.4 Contributions and Publications

The main contributions presented in this thesis are:

- **Automated head and neck OARs segmentation for CT/Magnetic Resonance Imaging (MRI):** A fully automated segmentation algorithm, relying on a DNN, trained with nnUNet and using the HaN-Seg dataset. The algorithm, including dynamic inference depth which controls the algorithm's run-time, has been packaged in a Docker container for out-of-

the-box, cross-platform usability. This algorithm is available publicly on a GitHub repository: `https://github.com/Zhack47/HaNSeg-QuantIF`

- **A collection of PET/CT segmentation models:** built during the AutoPET II-III-IV challenges, three algorithms are made publicly available. All of these algorithms are provided, alongside a Docker building script, in a separate GitHub repository
  - **AutoPET II**: ^{18}F-FDG PET/CT automated segmentation algorithm available at `https://github.com/Zhack47/autopetii`
  - **AutoPET III**: multi-tracer PET/CT automated segmentation algorithm. This model offers segmentation for PET/CT scans using PSMA-based and FDG-based tracers, without requiring prior information on the tracer used. Available at `https://github.com/Zhack47/AutopetIII_CHB`
  - **AutoPET IV**: semi-automated, multi-tracer PET/CT segmentation algorithm, which uses expert input to guide the underlying DNN in its segmentation. A novel heatmap-based, PET-informed click encoding method is proposed, along with dynamic inference depth, to ensure the algorithm finishes before a given time. Available at `https://github.com/Zhack47/autopetiv`
- **Radiomics-augmentation, a framework to apply data augmentations before radiomic features extraction:** a complete pipeline is implemented to extract augmented radiomic features from medical images. This pipeline performs the image and mask augmentations as well as the radiomic features extraction and storage, for any radiomics-based prediction task. Data augmentations aim at obtaining better generalization in downstream tasks, by filling the gaps in data distribution
- **Organ-omics, a framework to extract information from healthy organs:** a complete, automated pipeline which uses TotalSegmentator to extract organ contours from CT scans and PyRadiomics to compute radiomic features for each organ. These organ radiomics (organomics) contain more information than lesion radiomics alone, as we showed it improves the performance of prediction models for survival prediction of NSCLC patients. This framework was implemented and packaged in Python, available publicly at `https://github.com/Zhack47/Organ-omics/`.

Publications

- Podobnik, G., Ibragimov, B., Tappeiner, E., Lee, C., Kim, J. S., **Mesbah, Z.**, Modzelewski, R., Ma, Y., Yang, F., Rudecki, M., Wodzinski, M., Peterlin, P., Strojan, P., & Vrtovec, T. (2024). HaN-Seg: The head and neck organ-at-risk CT and MR segmentation challenge. Radiotherapy and Oncology, 198, 110410. https://doi.org/10.1016/j.radonc.2024.110410[8] **URL:** `https://www.sciencedirect.com/science/article/pii/S1361841524001786`
- Dexl, J., Gatidis, S., Fruh, M., Jeblick, K., Mittermeier, A., Stüber, A. T., Schachtner, B., Topalis, J., Fabritius, M. P., Gu, S., Murugesan, G. K., VanOss, J., Ye, J., He, J., Alloula, A., Papież, B. W., **Mesbah, Z.**, Modzelewski, R., Hadlich, M., . . . Kustner, T. (2025). AutoPET Challenge on Fully Automated lesion Segmentation in Oncologic PET/CT

Imaging, Part 2: Domain Generalization. Journal of Nuclear Medicine, jnumed.125.270260. https://doi.org/10.2967/jnumed.125.270260[9] **URL:**`https://jnm.snmjournals.org/content/jnumed/early/2025/12/30/jnumed.125.270260.full.pdf`

- **Mesbah, Z.**, Mottay, L., Modzelewski, R., Decazes, P., Hapdey, S., Ruan, S., & Thureau, S. (2024). AutoPET III: The tracer frontier. What frontier? arXiv.org[10] **URL:**`https://arxiv.org/pdf/2410.02807`

- **Mesbah, Z.**, Perret, S., Mottay, L., Univ Rouen Normandie F-76000 Rouen, France, AIMS-QuantIF, & Siemens Healthineers. (2025). QuantIF: Semi-automatic PET/CT segmentation across tracers. In QuantIF: Semi-automatic PET/CT Segmentation Across Tracers. autopet.org[11] **URL:** `https://www.autopet.org/docu/autopetiv/Task1/3_Zhack_AutoPETIV_GoodFormat.pdf`

- **Mesbah Z.**, Thureau S., Gouel P., Hapdey S., Rogé M., Jan O., Le Roux P., Langlais A., Lévêque É., Chaumet-Riffaud P., Amour É., Giroux-Leprieur É., Le Tinier F., Martin É., Pourel N., Boisselier P., Salem N., Le Pechoux C., Guillerm S., Audigier-Valette C., Zalcman G., Giraud P., Vera P., Modzelewski R. Strength of organ radiomics (organomics) for non small cell lung cancer (NSCLC) survival analysis based on mid-radiotherapy treatment PET/CT. A RTEP7 ancillary study. (Manuscript submitted to the International Journal of Radiation Oncology Biology Physics).

Communications

- **Oral Communications**
  - The 2024 International Conference on the use of Computers in Radiotherapy (ICCR), Oral Communication on *Fully Automatic Flap Segmentation through Deep Learning Methods to Refine Head and Neck Cancer Postoperative Radiotherapy, a GORTEC study* in Lyon
  - The 2025 MICCAI workshop, Oral Communication on *Promptable PET/CT segmentation* in Daejeon
  - The 2025 congress of the European Association of Nuclear Medicine (EANM), Oral Communication on *Decision support for early treatment adaptation. A case study for NSCLC: RTEP7-IFCT 14.02* in Barcelona

- **Posters**
  - The 2023 congress of Société Française de Radiothérapie Oncologique (SFRO), Poster on *Segmentation automatique des lambeaux dans la planification de la radiothérapie postopératoire du cancer de la tête et du cou, une étude GORTEC* in Montpellier
  - The 2025 Groupe de Recherche et d'Etudes de Traitement du Signal et des Images (GRETSI), Poster on *Dépasser les limites des traceurs en imagerie TEP/TDM : une segmentation universelle des lésions pour AutoPET III* in Strasbourg

Other works were conducted during this thesis, which are not presented here. To view these works, please refer to the author's Scholar profile `https://scholar.google.fr/citations?hl=fr&user=2HFzn_4AAAAJ`

Chapter 2

Background for Artificial Intelligence in Medical Image Analysis and Oncology

Chapter Overview

To understand how and why artificial intelligence (AI) is making its way in medical image analysis and oncology, it is necessary to precisely define AI and other connected concepts, such as machine learning and deep learning.

In this chapter, the names of tasks and paradigms commonly used in AI research will be presented, along with common tools used in this domain. Once the landscape of AI has been explored, existing applications to medical images and oncology will be presented and discussed.

Contents

2.1 Advances in Artificial Intelligence 9
2.1.1 Artificial Intelligence, Machine Learning and Deep Learning 9
2.1.2 The Origins of Deep Learning 9
2.1.3 Transformers and Large Language Models 10
2.2 Tasks & Paradigms for Artificial Intelligence 11
2.2.1 Tasks Nomenclature for Artificial Intelligence 11
2.2.2 Artificial Intelligence Training Paradigms 15
2.3 Tools for Medical Image Analysis 17
2.3.1 Medical Image Storage 17
2.3.2 Medical Image Processing 19
2.3.3 Tools for Machine learning with Medical Images 20
2.3.4 Tools for Deep Learning with Medical Images 21
2.4 Large Scale Oncology Imaging Datasets 22
2.4.1 Limiting factors for large medical imaging datasets in oncology 22
2.4.2 Examples of large datasets 24
2.5 Conclusion 26

2.1 Advances in Artificial Intelligence

2.1.1 Artificial Intelligence, Machine Learning and Deep Learning

A.I. is defined by the Oxford reference as *"the theory and development of computer systems able to perform tasks normally requiring human intelligence"*. In common language, the term AI has been used to mean various things, from algorithms controlling non-player characters in video games, to chatbots such as ChatGPT. In research, A.I. is a broad concept which gathers any algorithm which uses logic to perform a task.

It contains machine learning (ML), the collection of all AI methods which learn their logic from a given set of data. Multiple learning paradigms exist (see 2.2.2), some including human supervision, others not, but all of them are included under the ML denomination.

Deep Learning (DL) is a subset of ML, which includes only the methods using deep neural networks (Deep Neural Network (DNN)). This research field has been part of the research in AI almost since its beginning, mostly because of the focus on replicating the functioning of biological neurons. Thanks to the increasing availability of computing power, DL is now the main driver of both the research on A.I. and its commercial applications.

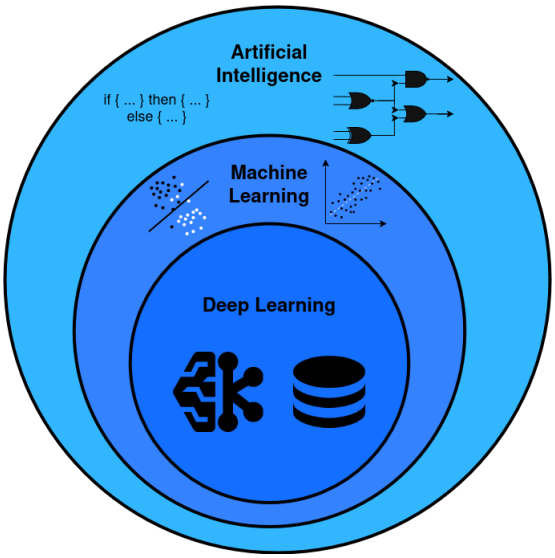

Figure 2.1: Artificial Intelligence covers methods relying on logic to process an input. Machine learning specifically designates methods which learn from sets of example data, trying to generalize to all possible cases. Machine learning with deep neural networks falls in the deep learning category which often requires large datasets for training.

Figure 2.1 emphasizes the fact that A.I., Machine Learning (ML) and DL are connected and that A.I. encompasses ML, which itself encompasses DL.

2.1.2 The Origins of Deep Learning

The foundations of Deep Learning were laid from the 1960's to the 1980's with the introduction of multiple keystones of the field.

First, the neural network: it consists of multiple artificial neurons, which are mathematical functions, applied to input data and gathered in layers. The mathematical groundwork for neural networks was laid over a long period, with some concepts dating back to Alan Turing's work. The most important breakthrough was the multi-layer perceptron (MLP), published in 1958[12]. MLPs were the first DNN, as they contained an input layer, hidden layer and output layer. Later the number of layers, as well as the number of neurons per layer, were both increased greatly.

Then, backpropagation: the concept of attributing parts of the error of an algorithm to each operation in this algorithm. Introduced by Seppo Linnainmaa in 1970 (republished in 1976)[13], this method allows to attribute parts of the error, which is computed between the output of a neural network and the ground truth, to each operation performed between the neurons of the network. Once a local error is obtained for each neuron, the value of its parameters can be modified to lower the error.

Convolution layers[14] were also a major improvement for neural networks. They enable the processing of inputs not as an independent set of values, but as order-dependent vectors (see figure 2.2). This is useful for time series processing (time-ordered data), as well as images and volumes processing (spatially-ordered data). These convolution layers rapidly became the most popular, as they enabled the processing of relationships between the data points, not only the values of the data points.

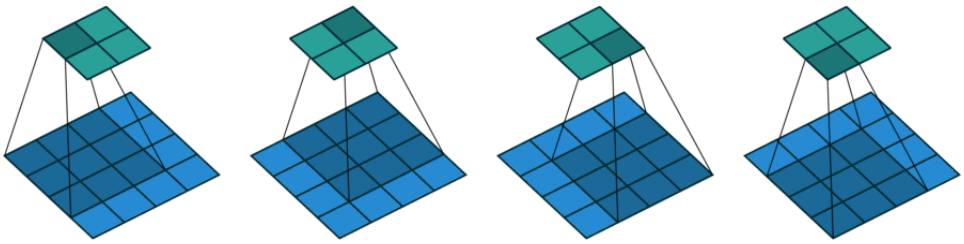

Figure 2.2: A $3 * 3$ convolution applied to a $4 * 4$ matrix, yielding a $2 * 2$ matrix. Taking the top left corner of that resulting matrix as an example: it is the result of the combination of values in the $3 * 3$ top left corner of the original matrix, by the weights of the 3x3 convolution kernel.

Finally, the combination of convolution-based neural networks (CNN) with back-propagation for error attribution gave birth to the modern neural networks. One of the earliest examples are LeNet[15], for digits classification in $32 * 32$ images (1998); AlexNet[16] for classification of 1000 objects in images (2012); UNet[17] for image segmentation (2015); ResNet[18], an improved classification network with residual connections (2016), still in use to this day.

2.1.3 Transformers and Large Language Models

The next breakthrough in DL occurred in 2017 when Attention is all you need[19], the paper introducing the Transformer architecture. This architecture, relying on a new mechanism, called attention, took the natural language processing language academic community by storm. This architecture first quickly demonstrated state-of-the-art performance for translation tasks, then was used for text generation, as it is at the core of the Generative Pre-trained Transformer (GPT) models[20].

These models powered the large language models which started an AI revolution, with commercial

products such as ChatGPT massively making their way in professional and personal uses. A specific category of models, named Vision-Language Model (Vision Language Model (VLM)), have included images in their input and/or output.

Although the models have demonstrated great text generation capabilities and some performance in logical tasks (text comprehension, mathematics, coding, reasoning...), they remain generative models, with no guarantees on their output. Incorrect generations, often referred to as hallucinations, have plagued these models and are often presented assertively by the model.

The depth of these models, emphasized by the Large in their names, is also a cause for problems: these models are almost un-explainable. Research on the explainability of Large Language Model (LLM)s is limited to either examining a group of a few hundred neurons[21] (out of billions!), or investigating model behavior at a macro level[22], with no certainty of the reproducibility of the observed behaviors in different contexts[23].

2.2 Tasks & Paradigms for Artificial Intelligence

2.2.1 Tasks Nomenclature for Artificial Intelligence

Artificial intelligence models are, still, to this day, specialized for a very specific task. In this section, the most common tasks AI models can solve will be presented, along with examples of what these tasks can look like in oncology and medical imaging.

2.2.1.1 Generation

As explained above, commercial chat-bots rely on generative models. This class of models execute a simple task: generating text from a given input. Generative models also exist for images[24], video[25] and audio[26], and have even been applied to medical images generation, in particular to create larger synthetic datasets when data-scarcity was an issue. For instance Kebaili et al.[27] trained a diffusion model to generate MRI scans of brain tumor patients, to improve the performance of downstream segmentation models. However, models trained on large amounts of synthetic data still tend to behave erratically and fail to generalize to unseen data[28]. This means that the output of generative A.I. models should not be considered as real data samples as it can introduce a bias which can lead to failure when presented with real-world data.

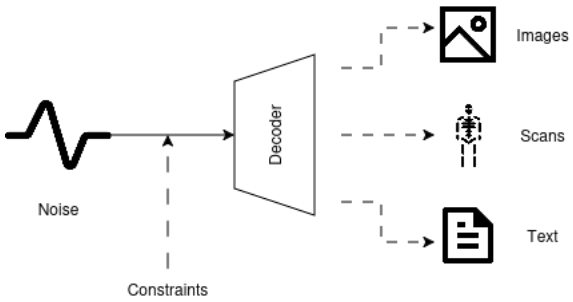

Figure 2.3: The main family of generation models: a noise is given to the decoder of a pre-trained auto-encoder, which generates a sample corresponding to its training distribution

In medical imaging and oncology, generative A.I. is still emerging, with specific challenges (complex images, need to account for rare cases...) which push the threshold for its adoption higher than for its current use in commercial chat bots and image and video generators.

2.2.1.2 Classification

Signals (text, audio, image, video) are often a representation of an object or a concept. A common use for A.I. is to automatically recognize an object, concept or situation when it is present in this signal. When a model is built for this type of behavior, it has often been trained to perform a task called classification.

Classification consists in analyzing a signal and assigning it a class. For an image it can be a "dog" or "cat" label, to classify animal photos, in a text a "positive" or "negative" label can be assigned for sentiment analysis and in a video, "movement" or "no movement" can be assigned to detect crowd movement, or an intrusion. Data required to train a model for classification needs to be at least partially labeled. The model then usually learns, from pairs of input signal and label, to assign the correct label to any given signal.

Classification of medical images is useful for disease classification, but can also be used at a different scale, by classifying nodules or tumors as benign or malignant. Using classification at patient-level is a weakly-supervised problem (see 2.2.2.2), with a lot of information which will not be used for the task. To avoid relying on weak supervision, one can break down the classification task into multiple, more supervised tasks (e.g.: nodule segmentation, then nodule classification, finally deciding if a patient has cancer or not). Some examples of classification using medical images in oncology include: breast cancer detection using DNNs and majority voting, from whole breasts imaging (ultrasound or mammography)[29], or lung nodule classification from CT images, classifying lung nodules malignant or benign, using a DNN for feature extraction and ML (binary decision tree) for the final decision[30].

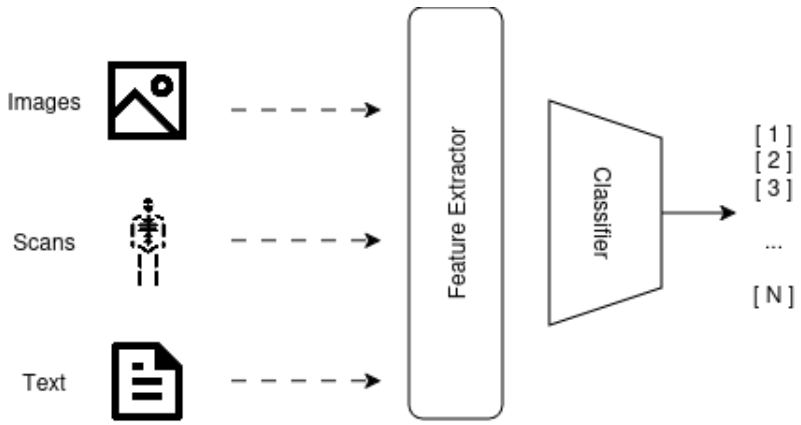

Figure 2.4: The objective goal of classification: assign a class to a signal: image, text, audio, scan... The classes are given semantic meaning by humans, models only handle integers.

2.2.1.3 Segmentation

Segmentation consists in performing a dense version of classification. This means that a class is assigned to each pixel (resp. voxel) in an image (resp. volume), resulting in a dense segmentation map of the same size as the input.

The segmentation task can be binary (2 classes) or multi-class, to handle the segmentation of multiple objects. The first class is often the background, and covers all the unannotated parts of an image. In situations where large parts of the images are not annotated, this background ensures that the model can at least learn to avoid generating false positive segmentations.

Segmentation is used in many domains:

- **Autonomous Driving**: it allows the detection, localization and identification of other vehicles in images from camera[31] or other modalities (LIDAR, radar).
- **Industrial Quality Control**: defects can be detected, localized and named using segmentation methods. For instance, segmentation has proven useful for the detection of defects on steel surfaces[32].
- **Medical Imaging**: segmentation is used to locate and delineate organs, tissues and lesions. The delineation of organs at risk for radiotherapy planning is increasingly supported by A.I. based automated segmentation methods.

Due to the density of supervision available for this task (voxel-wise supervision), it is often solved by training a DNN to reproduce human-made contours. However obtaining contours for a large amount of data can be complicated, because the annotation process is longer due to the need for pixel-level annotation. An additional hurdle exists for the application of segmentation to medical images, as an expert with knowledge of the domain (radiologist, nuclear physician, radiotherapy oncologist) is required to annotate the data.

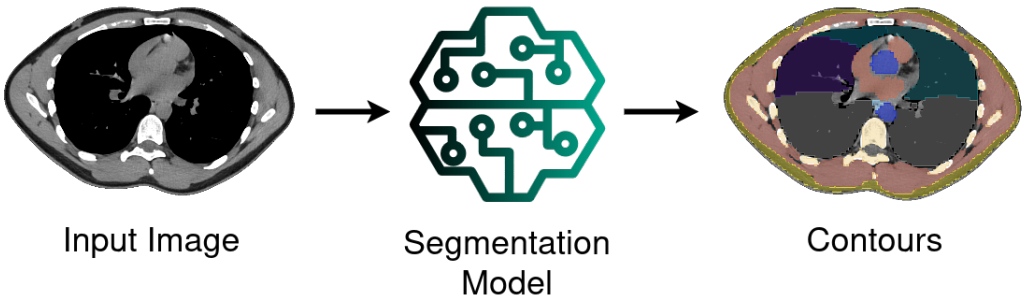

Figure 2.5: Segmentation consists in extracting the contours of objects from an image. Training a DNN for this task consists in comparing the predicted contours with the ground truth annotation.

2.2.1.4 Regression

Regression tasks are used to predict a continuous variable, from a given input. Its inner functioning is similar to that of classification: classification tasks consist in assigning a score to each class, then choosing the class with the highest scores.

The difference is that for classification, the scores (pseudo-probabilities, usually between 0 and 1) assigned to each class are not calibrated, which means they are not tied to any real value and therefore have no absolute meaning. They can only be used comparatively to the scores of the other classes. Regression, on the other hand, outputs a value that is continuous and the changes in that value should reflect the changes in the real target value.

To give an example, a classification task can consist of using a set of clinical features (such as age, life habits, diet..) to predict whether a patient has normal prostate specific antigen (PSA) levels (e.g.: below 5 ng/ml) or not; while the regression would predict the PSA level as a continuous variable. The difference between the score predicted for the PSA+ classification and the value outputted by the regression is calibration: PSA+ score may not be linearly related to the PSA values, but still relate to it, as a higher score corresponds to a PSA level <5 ng/ml.

While regression can be used in oncology to predict continuous values, such as PSA levels, a sub-type of regression is mostly used: survival prediction. Whether it is overall, progression-free, relapse-free, or any other conditional survival function, a regression can be learned using the survival duration as the target continuous value. However when the goal is only to rank patients for event-risk based stratification, an event-risk score can be predicted for each patient. This event score is inversely proportional to the survival time, and represents a proxy for survival, easier to predict than the actual number of survival days. Any ML model trained for this proxy can thus use a ranking objective rather than an error-based objective (MSE, MAE...).

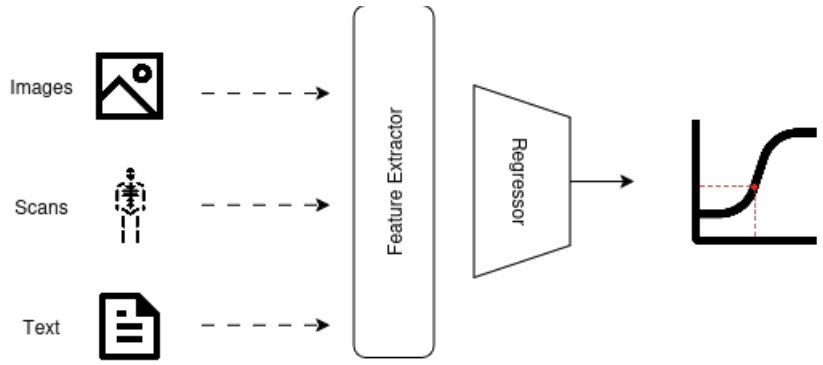

Figure 2.6: Regression methods are often similar to classification, with the difference that they predict continuous values. After the regression model has built a function to map the input to the output, any new data point can be assigned a value through this function.

2.2.2 Artificial Intelligence Training Paradigms

Regardless of the type of model used, the task to solve and the available data, training an AI model can be achieved through different learning paradigms. In the sections below, the requirements and common use cases for each training paradigm are described and discussed.

2.2.2.1 Supervised Learning

When the desired target prediction is available as a label for each sample in the training set, supervised learning can be used to train an AI model. It is the strongest learning paradigm, as it ensures the models are trained relying on true, often expert-made ground truth. Nevertheless, errors or noise could be introduced if the ground truth is not correct or missing information.

This training paradigm is the most basic and reliable one to train A.I. models. Practically, it consists of giving a signal to a model (ML, DNN), letting the model process it into an output, then comparing the obtained output to the expected ground truth using an error-computing function called a loss function. The loss is then used to modify the parameters of the model (for instance using backpropagation[33], in DNNs), in order to obtain a better output next time. To avoid teaching the model to only perform the task on the data it has been trained on (akin to "learning by heart"), the data should be split in a training and a testing set: training should only be performed on the training set, while the test set is held out to evaluate the model on unseen data. This is to ensure the model actually learns to generalize concepts from the training samples rather than overfitting on them. No overlap should exist between the training and test set, which means, for instance, that two CT scans from the same patient should always be in the same subset. In this case, splitting the data by patients and not by scan is better.

The learning / loss / parameters modification process is repeated multiple times, over the whole training dataset, one or multiple times (a full iteration over the training set is called an epoch), usually until no more improvement is obtained on the test set. In some configurations, the error on the test set starts going back up after some time, even if the error on the training keeps going lower. This is most often a sign of over-fitting, which means that, at this point, the model is not generalizing but learning to perfectly predict the samples from its training set. It is recommended to stop training before this behavior starts to happen (i.e.: at the test error's minimum), although there is emerging evidence that, in some specific cases, training past the overfitting point may lead to better solutions[34].

Using the test set to determine the best model could however lead to a sort of overfitting on the test data, called selection bias[35].

2.2.2.2 Weakly Supervised Learning

A subset of supervised learning is weakly supervised learning (WSL). This learning paradigm consists of situations where the supervision information is weak, with respect to the input information. Image classification is a good example of WSL: an image contains a large number of pixels, while the supervision consists only of the class assigned to the corresponding. This disparity in information between the input and the target means there is a large semantic gap between these two signals: the input is 2D, with possibly multiple channels (RGB), while the target is an integer drawn from a set of few discrete values.

When DNNs are used, this semantic gap is widened by the depth of the network, which makes it more difficult to learn a good representation of the input data.

To solve this problem, a popular approach is to use multi-task learning (MTL). MTL consists in accompanying the weakly supervised task (e.g.: classification) by other tasks with stronger supervision (e.g.: reconstruction, segmentation), to improve the performance of the former. In a paper[36], Aymyar et al. showed that COVID detection on CT scans (Normal / COVID+ / Other pathologies) was greatly improved when using MTL, specifically a reconstruction task (auto-encoding) and a segmentation task (lesion segmentation, when available). The logic behind MTL is similar to the one behind transfer learning (see 2.2.2.4): the additional, densely supervised tasks help shape the latent space with representations that extract important information from the input signal. These good representations of the data can be refined by, and used for the weakly supervised task, yielding better performance.

2.2.2.3 Semi-supervised Learning

In some cases, annotations for the task can be missing, or incomplete, preventing the use of traditional supervised learning methods. This learning configuration is called semi-supervised learning.

For instance, a dataset can have labels only for part of its samples, or the labels might be incomplete, covering only part of the input signal. An example of incomplete or partial labels in medical imaging is scribble supervision: in order to avoid the need for time-consuming 3D dense voxel annotations, researchers explored scribbles, which are small, rapidly drawn annotations, and tried to use them as ground-truth labels instead. This however breaks the premise from supervised learning, since a non-annotated voxel can no longer be considered as background and the learning methods need to account for unlabeled voxels.

When only partial annotations are available, solutions such as partial losses have been developed and showed great results without demanding too many modifications to the original DL methods[37].

In the case annotations are available only for part of the data, multiple strategies exist, mostly revolving around the creation of pseudo-labels for the unlabeled samples. This consists in training a model with the available labeled samples and using the predicted labels as pseudo ground-truth for unlabeled samples. It can be done either during training[38] or in multiple training cycles: first training only using the labeled samples to train a first model, then generating the pseudo labels for the unlabeled samples using this first model's predictions and, finally, training again on all the data.

If a human expert is available to correct the pseudo labels, active learning can be used to obtain a fully annotated dataset for a fraction of the time needed to annotate the dataset from scratch[39]. Using active learning, Qu et al.[40] obtained the segmentations for eight abdominal organs in 8 000 CT scans, in less than a month. This shows how semi-supervised learning can be used to rapidly complete the annotations in a dataset, thus improving both the quality of this dataset and the model trained on it.

2.2.2.4 Self Supervised Learning

Finally, data can be available without any annotation at all, often in the largest quantities. Using this data to train models (mostly DNNs) requires the use of self-supervised learning (SSL) methods.

A popular class of SSL methods is masked modeling. It is widely used to train the base generative models for LLMs on text: masked language modeling[41] consists in teaching a network to predict

one or multiple missing parts of a sentence. This is how the GPT models are trained[20].

Other variants of masked modeling exist for different data types, like masked image modeling[42], masked audio modeling[43], or masked video modeling[44]. This class of methods ensures that the DNN learns a representation of the input that is robust to the masking of a part of it. This also forces the network to learn non-trivial latent representations, because different, random parts of the input signal are masked each time, thus no part of it can be fully relied on.

Contrastive learning is a different approach to SSL, where multiple versions of the same image are presented to a DNN, with the objective to learn a similar representation for all of them. In the SimCLR[45] framework, each unlabeled data sample is processed using two different data augmentation transforms, yielding x1 and x2. The DNN then learns to encode x1 and x2 towards similar representations, by using a similarity loss between the two representations. This method can be extended by additionally maximizing the dissimilarity between the representations of different samples, although when no labels are available, this can be detrimental to the model's understanding (e.g.: maximizing the distance between two different pictures of a dog does not improve its representation of the concept of a dog).

SSL is used to pre-train networks, when large amounts of unlabeled data are available, before fine-tuning on smaller datasets containing labeled data, through supervised learning. The weights resulting from this pre-training contain knowledge which can then be transferred to part of a network (e.g.: an encoder), as a better-than-random initialization. This process of pre-training for a better initialization is called transfer learning.

2.3 Tools for Medical Image Analysis

Long before the advent of the latest AI methods for medical imaging, a community of researchers, medical professionals and engineers had been built around medical image analysis. Their goal was to create an ecosystem of tools, standards and libraries which would enable the storage, processing and analysis of medical images, using increasingly available and powerful computers.

In this section, some tools, which were used during this thesis, are presented. While this is not an exhaustive list, the tools presented below were selected for their usefulness and popularity in the medical imaging community.

2.3.1 Medical Image Storage

The first step was to define how medical images would be digitally stored. Medical images are not like natural images, they contain metadata, which contain information needed to describe the context of the image: the position in space (origin, spacing, direction), information about the patient, machine, medical exam or referring physician (see figure 2.7).

The most used standard format for medical image storage is the DIgitalCOmmunications in Medicines standard, DICOM. The American College of Radiology (ACR) and National Electrical Manufacturers Association (NEMA) initiated the discussion and work around the creation of a standard to store medical images in 1983. At that time, only imaging machine manufacturers could access the images and extract them from the machines. This prevented any analysis of these images by third-party researchers.

In 1985, the first version of their standard (ACR-NEMA 300) was released, but it was in 1993 that the standard was renamed DICOM and took off in adoption. The broader adoption came in part due to its integration with the standard networking protocol TCP/IP. Originally made for CT and MRI, the DICOM standard extended to include Ultrasound and Nuclear Imaging (1995), radiotherapy objects, such as structures, plan and dose (1997) or structured data (2000), extending the DICOM standard past images only[46].

The DICOM standard defines a file format to store and communicate medical information from point-to-point, usually from a machine to a storage or interpretation server for radiologists. While it is exhaustive, containing all the information included by the machine and/or by the technician; it is also very heavy, and bloated: not all the information contained is needed for some tasks. On top of that, DICOM stores the imaging volumes for most modalities as 2D slices.

Medical image analysis tasks require medical images to be represented as 3D voxel matrices in order to explore relations in the three dimensions. An intermediate object must then be created between DICOM files and the analysis task, to reconstruct a 3D volume from 2D slices. This is what the Neuroimaging Informatics Technology Initiative (NIfTI format) set out to do. Initially designed for MRI neuroimaging data, the NIfTI format has been widely adopted by the medical imaging community.

NIfTI stores any medical image, 2D, 3D, or 4D, in a single file, which can then be compressed to obtain .nii.gz files. Loading a volume from a NIfTI file is much faster and straightforward than reconstructing it from 2D DICOM slices. NIfTI retains the spatial information necessary for multi-modality registration and analysis, but discards the rest of the information contained in the fields of DICOM files. While this makes NIfTI a great file format for public sharing of medical images, it also means that any image transformation which requires values contained in the DICOM fields (e.g.: values rescaling, SUV computation) must be performed before changing the image format to NIfTI.

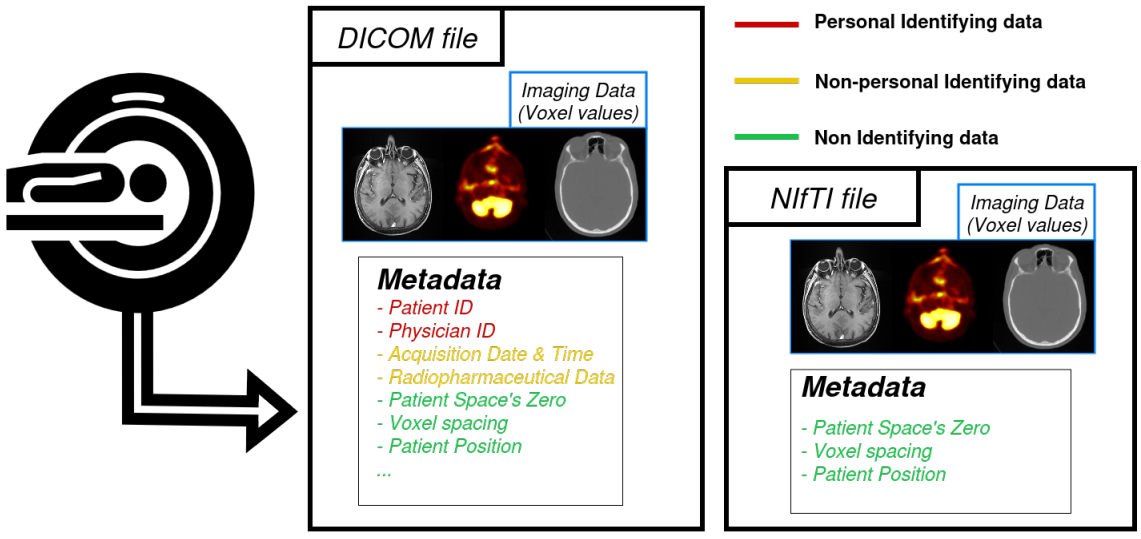

Figure 2.7: DICOM files store extensive metadata relating to the patient and image acquisition. NIfTI on the other hand only stores the information needed to reconstruct the patient's space in its metadata

2.3.2 Medical Image Processing

Along with standard formats for storage, tools were developed and shared to process medical images.

Some of these tools, such as PyDICOM and Nibabel, were designed to handle a specific data format. PyDICOM, for instance, offers a seamless solution to access DICOM tags[47]. However, to navigate from one format to another, it is recommended to use a library which can handle (open, read and write) different imaging formats.

2.3.2.1 SimpleITK: Interoperable Imaging Toolkit

To solve the inter-format communication problem, a library was built to interface the Insight ToolKit (ITK): SimpleITK[48]. SimpleITK can handle a large number of medical imaging formats (NifTI, DICOM, NRRD, MetaImage...) as well as regular image formats (PNG, JPEG 2000, BMP). It can read, write and modify images from all these formats, with an internal Image representation, which makes format conversion simple. SimpleITK is available both in C++ and Python, the latter being used a lot in research. SimpleITK also interfaces natively with NumPy[49], a mathematical library with support for fast array and matrix processing operations.

In addition to handling different formats, SimpleITK also offers numerous image processing tools, such as resampling (modifying the matrix of voxels, to shift the image's position in space), normalizing, cropping from a bounding box. It even offers image registration (displacing a source image to fit patient anatomy in the space of a target image), with the integration of the SimpleElastix[50] framework. Image analysis tools are also included in SimpleITK, such as obtaining maximum/ minimum values, morphological operations on binary masks (erosion, dilation, opening, closing), image thresholding (fixed threshold, Otsu[51]).

Designed with a C++ implementation and a python wrapper, the SimpleITK python library is both fast and easy to use. The core functionalities have been made available for multiple languages (R, C#, Java, Lua, TCL, Ruby, Python), allowing its integration in diverse projects. Alongside this, a community has been built around this library, encouraged by the availability of learning examples[52], a forum with active ITK members, and a regularly maintained and updated codebase[53]. Other tools, more targeted towards image manipulation have been developed.

2.3.2.2 ANTs: Advanced Normalization Tools

For instance, the Advanced Normalization Tools (ANTs[54]) is a C++ library which implements state-of-the-art registration and segmentation algorithms. This library is interfaced with R and Python (ANTsR and ANTsPy) to facilitate its use in research. Its main development goal was to implement registration algorithms such as rigid, affine, elastic deformation, with or without symmetric normalization. A command-line interface (CLI) is available in the terminal, along with functions (included in the python package) which can be used directly inside python code.

2.3.3 Tools for Machine learning with Medical Images

Machine learning (ML) is a field of computer science which consists in building statistical models of a system, using data samples consisting of examples of this system's behavior. These data samples (training samples) can be associated with ground-truth labels for supervised or weakly supervised learning. There is also a field of ML dedicated to learning from unlabeled data: unsupervised

learning. Multiple tasks can be solved using ML, most of them belonging to one of the tasks presented in 2.2.1.

To this end, ML libraries have been developed to make ML models and tools available to researchers, so they can easily apply them to their data. In medical imaging, once descriptive features have been obtained from images (e.g.: radiomic features), ML methods can be applied on these features to train a model for one of the tasks described above. Below are presented two frameworks which made ML methods available to researchers.

2.3.3.1 Scikit-Learn

Published in 2016, Scikit-Learn[55] implements standard ML algorithms, such as random forest, linear and logistic regression, support vector machine (SVM) for supervised learning, K-Means, K-Nearest Neighbors, principal components analysis, etc... Two types of ML tasks can usually be solved using Scikit-Learn: classification and regression. It also implements model selection algorithms (K-folds cross-validation) and model ensembling (soft and hard voting) to improve performance.

The philosophy behind the development of Scikit-Learn is to remain minimalist and implement only tools and methods which proved their importance and/or superior performance, over a large enough period of time. While this protects this library from becoming bloated, the latest algorithmic and modeling improvements take some time before being accepted. For instance, hierarchical density-based clustering[56] was presented in 2017, but its adoption in Scikit-learn's main codebase only occurred 6 years later, in 2023 (version 1.3.0).

To allow researchers to implement their own tools and methods and share them with other users, Scikit-Learn defined estimators, a basic interface which can be used to implement new ML algorithms, while ensuring compatibility with scikit-learn's other functionalities, like model selection and cross-validation. This interface has become accepted as the standard for the implementation of published ML estimators.

2.3.3.2 Scikit-Survival

Using the estimators interface, Sebastian Pölsterl released Scikit-survival[57], a library implementing a large number of survival prediction algorithms (survival SVMs[58],[59],[60], linear models[61],[62], ensemble models[63],[64]), along with metrics to evaluate their performance (Harrell's concordance index[65], cumulative-dynamic AUC[66]). This library offers an easy way to implement solutions for survival prediction tasks, by natively handling censored data and proposing regression models which can handle either risk scores, survival times, or a value proportional to the survival time.

2.3.4 Tools for Deep Learning with Medical Images

Deep Learning (DL) relies on DNNs, which are themselves composed of layers of multi-parameters operators and functions. To simplify the implementation of existing and new DNNs, tools have been developed, either by large AI companies, or by the AI and medical imaging communities. In this thesis, DL was used exclusively for segmentation purposes, as our approach is to tackle only densely supervised tasks with DL. Other tasks, such as classification have made use of the segmentation results, but were rather solved with ML methods.

2.3.4.1 PyTorch: The Foundation of Deep Learning for Research

PyTorch[67] is a deep learning python library, implemented in C++ for speed and with embedded Graphics Processing Unit (GPU) readiness[68]. Developed by Meta (ex Facebook), this deep learning framework has gained traction in the beginning of the 2020s, rapidly becoming the most used library for deep learning researchers to share the implementations of their work. Pytorch implements most of the training components (forward pass, backward pass, layers, optimizers, losses) of neural networks, leaving only the experiment itself to be developed by any researcher. Other libraries such as Tensorflow[69] offer similar tools, but although it is still widely used in industry (most importantly, used by Google), it has now been mostly abandoned by researchers, in favor of PyTorch[70]. Using the CUDA runtime, PyTorch defines GPU operations, which can be used to speed up the numerous matrix operations used in neural networks, compared to processing the same operations on a Central Processing Unit (CPU). The team behind PyTorch also developed multiple optimizations during the last 10 years (model compilation, 16-bit float computing, tensor cores, CUDA graphs) which reduce the duration of training and increase inference speed.

2.3.4.2 MONAI: NVIDIA's Deep Learning Library

MONAI is a deep learning framework, built on top of PyTorch by a team of mostly NVIDIA engineers. It offers a collection of data loading utilities for medical images, mostly similar to what is proposed by TorchIO. The main goal of MONAI is to provide seamless linking with PyTorch. MONAI also provides a large collection of neural network architectures for tasks such as reconstruction (Variational Auto-Encoder, ViT Auto-Encoder), classification (ResNet[18], EfficientNet[71]) and segmentation (UNet[17], UNETR[72], SwinUNETR[73]). This model collection enables faster and more reproducible research and rapidly available state-of-the-art models. Some models can be initialized using weights, obtained by pre-training the architectures on ImageNet[74] (for 2D models) or MedicalNet[75] (for 3D models). While the models and training tools proposed in MONAI can be used in any domain, they are prioritizing medical applications, with native support for 3D volumes, patch-based training and inference, as well as direct availability of medical images datasets.

2.3.4.3 TorchIO: Medical Image Processing for Deep Learning, by the Community

Building on top of SimpleITK, a group of researchers created TorchIO[76]. This Python library includes higher level medical images manipulation tools and is mainly directed at medical imaging and deep learning researchers.

Thanks to the use of SimpleITK, TorchIO handles most medical images formats and contains image pre-processing tools (resampling, normalization, cropping, padding, re-orientation). It also defines the concept of Label, a binary or multi-class mask, which can be used for ground-truths in segmentation tasks. Leaning heavily on the medical nomenclature, TorchIO uses the concept of Subject, representing a patient which can contain multiple images, labels and additional information and Dataset which can contain multiple Subjects and handle their reading from the disk. See 2.3.3.3 for the remaining capabilities of TorchIO.

TorchIO has been developed to interface seamlessly with PyTorch (hence the name). It stores data as torch.tensors, the base data type for PyTorch and the datasets it builds can be used by PyTorch's dataloader (Note: in the latest versions, TorchIO introduced the SubjectsLoader, a drop-

in replacement for PyTorch's Dataloader, however this new SubjectsLoader can still be included directly in any PyTorch-TorchIO training loop).

In addition, TorchIO implements multiple data augmentation transforms, from regular image transforms (flip, affine translation, noise, blurring) to augmentations specific to medical images in general (motion, anisotropy, elastic deformation) and MRI in particular (ghosting, field bias, motion, spiking). These transforms can be used to avoid over-fitting when training DNNs on small size datasets, or to improve model performance on unseen data in general. Unlike MONAI, TorchIO is not developed by a large company, but by researchers, engineers, students and other developers in the medical imaging and DL communities. This means that, while the development is led by a team of researchers at King's College in London, improvements, upgrades and bug fixes can be inserted in the codebase by anyone from the community.

2.4 Large Scale Oncology Imaging Datasets to Reach Expert Performance using AI

To increase performance, in a way similar to LLMs, it has been hypothesized that large scale datasets were needed. However, unlike for natural language, the collection of medical images datasets in oncology has been way more limited.

2.4.1 Limiting factors for large medical imaging datasets in oncology

The first limiting factor is the amount of data generated: for some diseases, only a few patients go through a given cancer diagnosis and treatment center per year. To obtain a large amount of imaging data in this configuration, collaboration between centers is necessary, but not always possible due to regulations on medical data collection and sharing.

In fact, due to the sensitive nature of medical data, regulations are often in place for the protection of patient privacy, making the collection of large amounts of medical imaging data more meticulous than web-scraping is (even considering only legal scraping). To collect medical imaging data without putting the patients' privacy at risk, it is necessary to remove identification data, such as names, birthdate, addresses and unique identifiers (UUID). Complete anonymization would also remove the acquisition date and time of exams, as these could be traced back to a patient, but this data can be needed during processing or analysis, SUV computation for instance. This is why in PET scans, a less strict alternative, pseudo-anonymization, is more often used: the data and metadata directly identifying the patient (names, birthdate) are removed, while fields related to the acquisition are kept.

If the file format is kept to DICOM, new UUID can also be generated and a correspondence table with original identifiers can be kept securely if it is required to be able to trace back to the patient. However, other spatially-aware medical imaging formats exist, which do not contain any identifying metadata (see 2.3.1). The most popular of these formats is the NIfTI format, which was, for instance, used for the public release of the TotalSegmentator, AutoPET and AbdomenAtlas datasets. NIfTI does not keep any of the DICOM fields, which ensures a complete anonymization, at least of the metadata.

Even once all personal information was removed from the metadata, anatomical imaging, such as CT scans and some MRI sequences still present a risk to leak the patient's identity. In recent studies,

researchers have applied facial recognition software to 3D volumetric renderings of CT scans and got matches for some of the patients[77]. To mitigate this issue, some public datasets have defaced[78] the scans to prevent a correct reconstruction of the face (see figure 2.8). This method could however be unusable for use cases which would need to analyze part of the face (e.g.: lesions in the eyes, nose, or skin of the face) or directly seeking to segment parts of the face (e.g.: organs-at-risk segmentation in head and neck cancer).

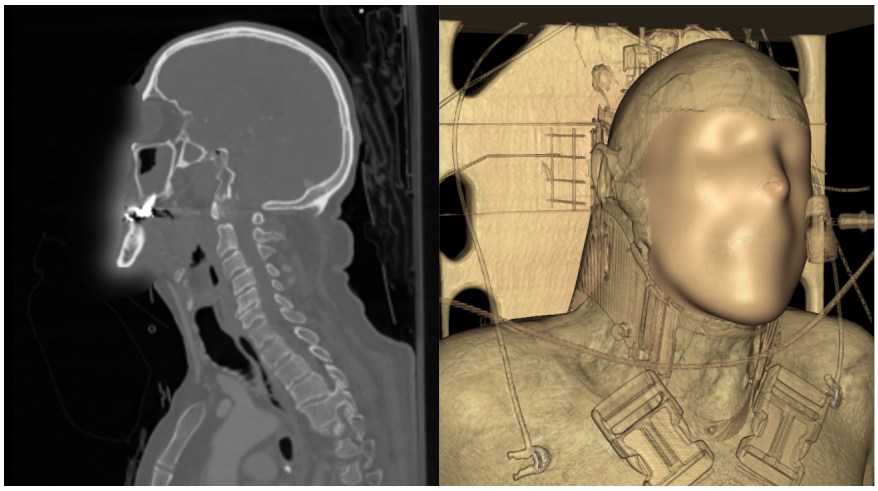

Figure 2.8: An example of a CT scan (s0591 from TotalSegmentator) which has been defaced for patient privacy. On the left: 3D volume rendering. On the right: 2D coronal slice, showing the blurring added to the face area.

2.4.2 Examples of large datasets

Despite these limitations, several large medical imaging datasets have already been gathered to train DNNs. Some of them contained only images (or volumes), limiting their usability to SSL pre-training. In this section, we will present a portion of these large datasets, summarized in table 2.1.

Among these datasets, the CT Foundation dataset (2022), gathered by Google in multiple US based hospitals, marked the scales as it gathered more than 500 000 CT volumes, to pre-train its Gemini model for medical image analysis[79],[80]. They demonstrated good performance in zero-shot report generation from this pre-training alone, with a human evaluation finding AI reports were as good or better at including important findings than human reports in more than 50% of the cases, on a 600 patients test set. While the report writing capabilities were obtained through Gemini's text pre-training, the CT analysis capabilities were brought by the SSL pre-training on CT Foundation.

More recently, in 2025, Pai et al.[81] released the data contained in CT-FM, an unlabeled CT dataset containing 148 000 CT volumes. Using label-agnostic contrastive learning, they trained a DNN which improved the performance of downstream tasks, compared to a random initialization of the weights. Unlike Google's CT Foundation dataset, CT-FM is composed of publicly available data and can be used by anyone, which means there is more reproducibility for the results obtained with this dataset.

Moving from SSL, Google also gathered more than 750 000 CT scans along with their corre-

sponding reports. Using these reports, Google refined the automated CT report generations of MedGemini-2D, Google's model with the most medical abilities. This dataset, CT-US1, was gathered from three major hospital regions (sic), with both contrast-enhanced and non-contrast-enhanced CTs, of different areas of the body (head, chest, abdomen, heart).

A similar, smaller dataset was made publicly available: AMOS-MM. This dataset contains more than 2000 CT scans, with structured reports reporting the human radiologist's findings and impressions across the chest, abdomen and pelvis. The research team behind AMOS-MM organized in 2024 a challenge to stimulate the medical imaging community around the task of automated report generation. Unfortunately, maybe because of the lack of maturity of 3D VLMs for medical images, very few teams have reported their participation in this challenge.

A third domain in which large scale datasets have begun appearing is medical image segmentation. Like report generation (and unlike classification), segmentation can be handled independently from any disease class: segmenting multiple organs, or lesions, can be done on virtually any patient. The first two large medical image segmentation datasets were AutoPET and TotalSegmentator. The first AutoPET dataset contained 1014 PET/CT images, including patients with cancerous lesions (lung cancer, melanoma, lymphoma) and healthy patients. This dataset and its subsequent versions are described in more detail in Chapter 4.

TotalSegmentator is a dataset containing 1228 CT volumes, with annotations for 117 anatomical structures, including bones, muscles, internal organs and parts of the cardio-vascular system. This dataset was used to train a DNN segmentation network using nnUNet[82], which was then released for public use. The TotalSegmentator segmentation tool made full body organs segmentation available to the medical imaging community, which led researchers to call it "A gift to the medical imaging community"[83].

Following in TotalSegmentator's footsteps, other large segmentation datasets were released. In particular, the AbdomenAtlas[40] dataset family, which gathered 9 262 CT scans with organ segmentation labels. The last version (AbdomenAtlas 3.0) also included both structured and narrative imaging analysis reports, which were generated with the help of RadGPT, a pipeline including LLMs and segmentation models, for faster annotation and report creation. Human experts were left to check the generated annotations and reports, and correct it when needed. The same team is behind the project BodyMaps, which is an internal dataset, not available to the public, but results from a collaboration of researchers from 145 hospitals and 19 countries. BodyMaps is declared to contain annotations for 377 anatomical structures and tumors, in more than 380 000 CT volumes.

To obtain datasets of this scale, it is necessary to automate most parts of the annotation process. Researchers involved in the BodyMaps project proposed ScaleMAI[39], an AI-integrated framework, capable of rapidly retrieving and annotating imaging volumes with a collection of segmentation models (model zoo). To limit the time human experts spend correcting the annotations, labels are refined through a label expert model, which picks the best segmentation out of the model zoo's propositions. Test on training is used to select the cases most likely to contain annotation errors for human revision. Using ScaleMAI, the team created PancreaVerse, a dataset of 25 362 CT scans, annotated with pancreatic cancerous and non cancerous lesions, as well as some anatomical structures. On this dataset, they trained an AI model to detect and classify the pancreatic lesions as cyst, pancreas neuro-endocrine tumor or pancreas duct adenocarcinoma. Their model outperformed both junior, senior and expert radiologists on classification, while matching their performance.

This work demonstrated how AI can be used to facilitate the creation of large scale imaging

datasets, particularly for oncology tasks, to then train models on par with human physicians.

	Volumes	Modalities	Annotations	Public
CT Foundation	500 000	CT	No	No
CT-FM	148 000	CT	No	No
CT-US1	500 000	CT	Reports	No
AMOS-MM	2 000	CT	Reports	Yes
AutoPET	1 611	PET/CT	Contours	Yes
TotalSegmentator	1 228	CT	Contours	Yes
AbdomenAtlas	9 262	CT	Contours/Reports	Yes
BodyMaps	380 000	CT	Contours	No
PancreaVerse	25 362	CT	Contours	No

Table 2.1: A summary of large datasets, their modalities, annotations and availability to the public

2.5 Conclusion

Since the 2010s, the landscape of tools for AI in medical imaging has greatly expanded, both for regular processing, ML and DL.

Different formats for the storage and sharing of medical images have been proposed and the community has mostly settled on DICOM and NIfTI, with an emphasis on the latter for public data sharing, as it removes identifying metadata. SimpleITK has acted as the go-to image handling and processing tool, thanks to its ability to seamlessly switch between numerous image storage formats, while conserving spatial localization information, as well as its native interface with NumPy, facilitating downstream ML and DL tasks.

Medical ML applications have relied on the Scikit-Learn library, a keystone for all types of research ML applications (5 million daily downloads), as well as on other libraries which used its interface to implement more recent and/or specific algorithms, such as Scikit-Survival.

Deep learning has also made the leap to medical imaging and oncology in the last two decades. While this progress relied on the evolution of DL, with the release and improvements of PyTorch making it the most used DL framework in research, it has mainly been enabled by the gathering of large scale datasets. These large datasets were gathered and annotated using innovative semi-automated methods to reduce the time experts needed to dedicate to annotation.

These datasets were mostly gathered for the same task: medical image segmentation. This densely supervised task is the only one which has seen performance reach expert level using DL and DNNs. In the next chapter, we will show how DL has been used to perform segmentation on medical images, then we will provide an example, with an automated segmentation algorithm for head and neck organs at risk on CT/MRI images.

Chapter 3

Medical Image Segmentation: State of the art and application to the segmentation of head and neck organs at risk

Chapter Overview

Medical image segmentation is a subdomain of image segmentation: it is an application of segmentation to images obtained by medical scans or other imaging modalities. It comes with unique challenges to be overcome, due to the nature of medical images. These are often acquired in three dimensions, using different imaging modalities, on machines provided by different vendors and have only recently (and rarely) become available in amounts large enough to use deep learning methods efficiently. In this chapter, the evolution of the state-of-the-art for medical image segmentation will be presented, starting from the theoretical concepts, towards existing tools and works in the domain. Finally, we present our application of this concept to the segmentation of organs at risk in the head and neck area for radiotherapy planning, using CT/MRI images.

Contents

3.1 Image Segmentation 29
3.1.1 Semantic Segmentation vs Instance Segmentation 29
3.1.2 Image segmentation using deep learning 29
3.1.3 Losses and Metrics 31
3.2 Medical Image Segmentation 32
3.2.1 Specific Challenges with medical images 32
3.2.2 Advances in automated medical images segmentation 32
3.2.3 A Call for Rigorous Validation 36
3.2.4 Advances in semi-automated medical images segmentation 36
3.3 Head and Neck OARs Segmentation 39
3.3.1 Context 39
3.3.2 The HaN-Seg dataset and challenge 42
3.3.3 Proposed Algorithm 44
3.3.4 Results 47
3.3.5 Discussion 48
3.3.6 Conclusion 49

3.1 Image Segmentation

3.1.1 Semantic Segmentation vs Instance Segmentation

Image segmentation is divided in two main objective categories: semantic segmentation and instance segmentation (see figure 3.1).

Semantic segmentation is a dense segmentation task which assigns to each pixel (or voxel) a class describing which type of object it is a part of. In natural image segmentation, this can mean assigning the class "*car*" to any pixel in the image which is part of a car, while in medical image segmentation, this can mean assigning the value "*left lung*" to each voxel belonging to a patient's left lung, in a CT scan for example.

This type of segmentation is optimal for uncountable objects, such as the sky, leaves, or in a medical context, emphysema, or fibrosis. Different instances of a countable object can also be segmented using semantic segmentation, however they will be grouped as a single class.

Instance Segmentation also assigns a value to each voxel which describes a given object, but it discriminates between different instances of the same object. Instance segmentation is used to segment countable objects in an image, for example multiple people in a crowd, or, in medicine, different cells in anatomical pathology slides[84] or multiple nodules in a patient's lung. It is limited by its conceptless segmentation: it is not possible to assign different object classes to segmented instances when instance segmentation is used on its own.

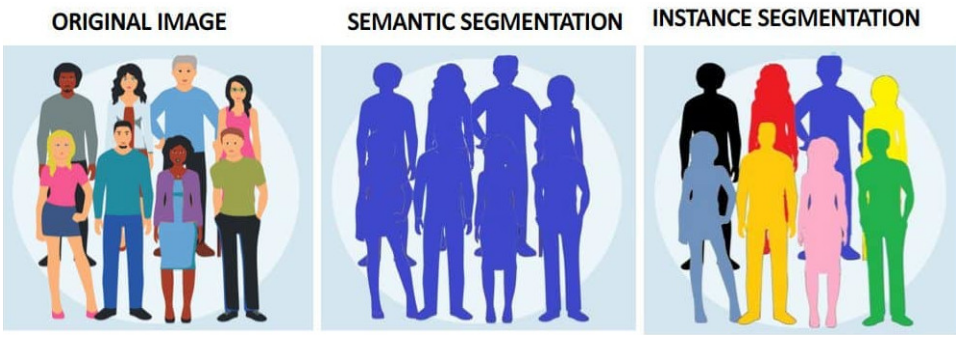

Figure 3.1: Starting from an image, segmentation consists in classifying the pixels of an image (or voxels of a volume) depending on the object they are a part of. Semantic segmentation groups all instances of this object in the same class (e.g.: *"person"*) whereas instance segmentation assigns different labels to each instance of the object (e.g.: each person). *Source: learnopencv.com*

3.1.2 Image segmentation using deep learning

In recent years, research on segmentation has largely moved from traditional computer vision methods, such as Otsu thresholding[51], towards segmentation methods relying on deep learning(DL) and deep neural networks (DNN). To this end, specific DNN architectures, training losses and evaluation metrics were designed.

3.1.2.1 Architectures

The most commonly used DNN architecture for segmentation tasks is the UNet[17]. Proposed in 2015, this Convolutional Neural Network (CNN) takes its name from its shape (see figure 3.2). It is composed of a downsampling convolutional encoder, which extracts high level features from the image, and a convolutional decoder, which reconstructs the segmentation map from these features. Between the encoder and the decoder are residual connections, called skip connections in the original paper and referred to as "*copy and crop*" in figure 3.2.

These residual connections transfer lower level, spatial features from the encoder to the decoder, thus allowing the network to better preserve spatial and low-level information, such as edges. Skip connections also help stabilizing gradients in deep architectures[85].

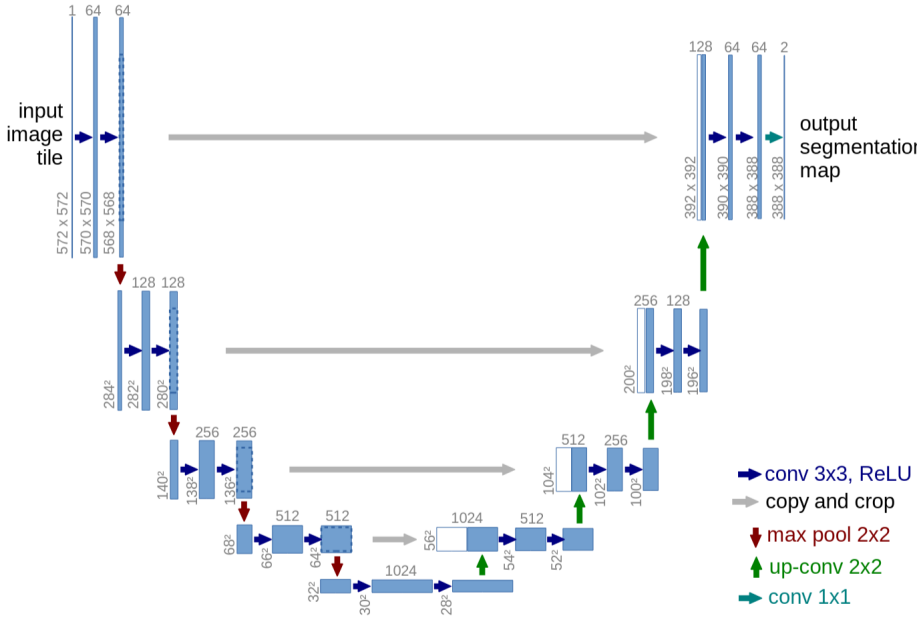

Figure 3.2: Original architecture of the UNet. Source: Ronneberger et al.[17]

After 2020 and the adaptation of the transformer architecture to images with the visual transformer (ViT)[86], transformer based segmentation networks were proposed. Most notably, the UNETR[72], a network using a ViT as encoder and a convolutional decoder, as well as Swin-UNetr[73], which used a sliding-window transformer to reduce the computational cost of the attention mechanism. Other research teams proposed to interleave the transformers and convolution layers, such as in the nnFormer architecture[87].

The latest fundamental shift in architecture arrived in 2024, with the release of state-space models, particularly the Mamba architecture[88], which was subsequently adapted for image segmentation with U-Mamba[89], a Mamba-based UNet. Although the building blocks and specific layer combinations have changed over the years, the U-shaped, encoder-decoder architecture, with skip connections across, has remained the core design for most proposed segmentation architectures.

The details, specificities and drawbacks for each of these DNN architectures in the context of medical image segmentation will be presented and discussed in 3.2.2.

3.1.3 Losses and Metrics

To train a DNN, an objective must be defined for this DNN to solve. This objective is defined as a mathematical error function, which quantifies the deviation between the network's output and the Ground Truth (GT). This function is called a loss function and must be chosen carefully, to give the network proper feedback on its output and steer it in the right direction.

The first type of losses used in image segmentation is information-based losses, which include the cross-entropy (CE)[90] based losses. Such loss functions compute the entropy between the different classes in the prediction and GT. Entropy relies on the pseudo-probabilities predicted by the network, after a non-linear function (such as softmax or sigmoid) is applied to the logits to limit their possible values between one and zero. Binary CE or regular CE can be used, depending on the number of output classes.

The second type of losses is geometric losses, which are used to compare the ground truth and network's output in space. Superposition losses, such as the Dice loss which relies on the Dice-Sorensen Similarity Coefficient (DSC)[91], are the most used. They compare the superposition between the output and the ground-truth to the total size of these objects.

When partial annotations are used to train the network, recall-based losses[92] can be used. These losses only include true positives and false negatives in its computation, leaving false positives out, as they can either actually be false positives or be missing because of the partial annotation.

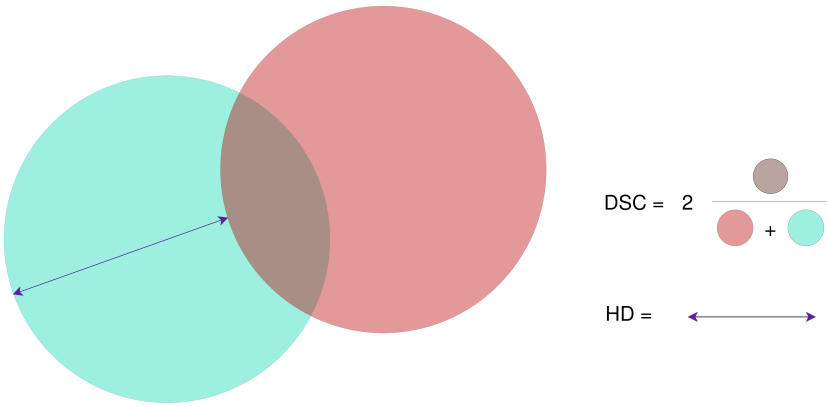

Figure 3.3: The Dice Similarity Coefficient (DSC) is a metric which compares the intersection of two geometric shapes with the summed area or volume of these two shapes. The Hausdorff Distance (HD) is the maximum of the set of minimum distances between any point in the predicted contour and a point in the reference contour.

To evaluate a DNN's performance, metrics need to be used either during its training, to oversee its progress and stop early if needed; or after it is done training, to evaluate its performance on hidden data.

The DSC can be used as a metric, along with Intersection over Union (IoU)[93], DSC variants and distance metrics. DSC and IoU measure the overlap between the ground truth and output segmentation. They can be computed for each class separately, or averaged to evaluate the overall performance. Moreover, the background can be included or excluded from computation, depending on its importance, which varies depending on the density of the annotations (low in lesion detection

tasks, high in road landscape segmentation tasks such as Cityscapes[94]). DSC variants have been engineered to fit specific use cases, such as normalized surface distance[95] which focuses on overlap of the boundaries, punishing less precise segmentation, or centerline DSC[96], which rewards network for conserving the topology of tubular structures.

These surface metrics however fail to capture distance between structures and should be complemented by a distance metric. Hausdorff's distance represents the maximum in the set of distances between a point in the ground truth segmentation and the closest point in the predicted segmentation. A variation of this metric, Hausdorff's distance 95th percentile, is more commonly used, because it is robust to very small outliers (<5% of the segmented volume).

A visual representation of these metrics is available in figure 3.3. Segmentation metrics, their uses and pitfalls are exhaustively discussed in the metrics reloaded paper and framework, along with metrics for other tasks[97].

3.2 Medical Image Segmentation

3.2.1 Specific Challenges with medical images

Training a DNN for an image segmentation task often requires dense (i.e.: pixel to pixel) and precise annotations for one or multiple objects in many images. Obtaining such an annotation is costly for medical images, as it requires the work of a trained expert, capable of recognizing anatomical structures (organs, prosthesis) or diseases (nodule, bleeding, cyst, malignant lesion) in medical images. On top of that, it requires a large amount of this expert's time, as annotating each voxel of a 3D image can be very time-consuming. To give an order of magnitude, manual segmentation annotation of the psoas in a 3D CT volume takes almost two hours for each scan[98].

In addition, precise and reproducible segmentation annotation is never guaranteed, as a study found significant variability, even when comparing the annotations from only two observers[99]. The authors of this study further warned that "*Two and even three observers capture only a fraction of the full range of inter-observer variability observed in common practice*". To add to this issue, gathering a large amount of medical imaging data is challenging in itself (as described in 2.4.2). Indeed, medical imaging data is mostly obtained through clinical trials, which often target specific problems (e.g.: pancreas tumor segmentation in CT scans), which means the number of patients included in these trials is limited (often below a few hundreds, rarely above a thousand). This, in turn, has resulted in a lack of large scale medical imaging segmentation datasets, until the last few years and the publication of the datasets presented in 2.4.2.

3.2.2 Advances in automated medical images segmentation

3.2.2.1 The UNet Architecture

In the last 10 years, both semantic segmentation and medical images segmentation took huge steps forward. First in 2015, when O. Ronneberger proposed a novel architecture, the U-Net [17], which was quickly adopted for segmentation tasks and adapted to 3D medical images segmentation, along with support for sparse annotations[100]. This architecture has been modified into a large number of variants[101] [102] [103][104] [105], which were mostly tailored to very specific use cases and transferred poorly to other tasks. The large number of variants also made it impractical to compare

a new method to all available methods in a single paper, which complicated the choice of a state-of-the-art method to compare new methods against.

3.2.2.2 The nnUNet framework

To fix this domain-wide issue, Isensee et al released the nnUNet framework[82],[106] in 2019. Their idea was to demonstrate that it is not so much the DNN's architecture which mattered the most, but the way it was trained. This framework lowered the access barrier to train state-of-the-art 3D and 2D segmentation DNNs, without the need to write a single line of code. It only required the training data to follow a specific format, as well as the inclusion of a JSON file describing the data. The nnUNet framework then handles the experiment planning, data preprocessing and training process, as well as the inference and evaluation of the trained DNNs.

On a technical level, nnUNet implements several training-related mechanisms to improve the learning process:

- A set of transformations are used for data augmentation. These transformations mostly modify either the voxel values of the images, or the geometry of both the images and segmentation masks simultaneously. They are designed to introduce variations in the data which helps the DNN to learn general solutions to its task rather than learning to solve only the training examples (see 5.3.1 for a presentation of the data augmentation principle).

- Mathematical improvements to the error backpropagation and optimization. First, gradients are clipped to a maximum amplitude of 12, to protect the training from failing because of exploding gradient values. Then, a weight decay of $3 * 10^{-5}$ is applied to all of the DNN's parameters, which has been shown to help prevent overfitting[107]. The learning rate, used to regulate how rapidly the optimizer modifies the DNN's parameters, is decreased as training progresses. This allows the training to make larger modifications at the beginning of the training process, while still being able to make finer adjustments towards the end of it.

- The supervision was also improved, using a compound loss and deep supervision. Compound loss (Dice loss + CE) enables both geometry and information based error-control during training. Deep supervision compares downsampled versions of the ground-truth to the feature maps obtained at each level of the decoder, using the same compound loss, although with a lower weight the deeper in the network the feature map is located. This additional supervision steers the DNN's internal representation of the data towards the desired segmentation output.

- Computational improvements such as automated mixed precision, which performs float operations on 16 bits instead of 32, or bias removal for the convolutions followed by a normalization layer.

All of this combines into a state-of-the-art DNN training pipeline, which only (optionally) delegates the choice of the architecture and hyper-parameters to the user. Originally, the default architecture in nnUNet was a UNet similar to the Ronneberger's UNet, but it was later replaced by a different UNet, with a residual encoder (i.e. an encoder traversed by residual connections). Any DNN architecture built for segmentation can be adapted to work with nnUNet, however the compatibility instructions, available on nnUNet's GitHub repository[106], must be followed.

While it can be used for any segmentation task, 2D or 3D, nnUNet was initially developed with medical image segmentation in mind[108] (precisely the Medical Segmentation Decathlon). This was initially visible in the "modalities" keyword, used in older versions of the dataset JSON file, and can still be perceived in the tasks its authors use to benchmark it[109]. Because of this, it contains specificities that are useful for medical imaging, such as its data augmentations (simulating low resolution, easily disabled mirroring transforms for anatomical left/right differentiation) and its preprocessing (handles spatial information from medical imaging formats such as NIfTI and metaimage, resamples image to a common, median spacing).

In 2020, nine out of ten MICCAI's 3D segmentation challenges were won by a team using nnUNet, confirming the power of this tool [110]. Among the challenges which were won this year using nnUNet, BraTs[111] and Hecktor[112] are the most important ones, since their datasets are still widely used to test the performance of segmentation algorithms to this day. In 2026, nnUNet is still recognized as the state-of-the-art framework for medical image segmentation and is widely used both for medical imaging challenges and medical studies involving automated segmentation[113].

3.2.2.3 Adaptation of Transformers for Segmentation

While CNNs and UNet have remained the best performing architecture, other architecture types were implemented for segmentation. In 2021, A. Dosovitskiy released *An image is worth 16x16 words*[86], a paper introducing the Visual Transformer (ViT), a novel neural network architecture for computer vision, based on the transformer architecture.

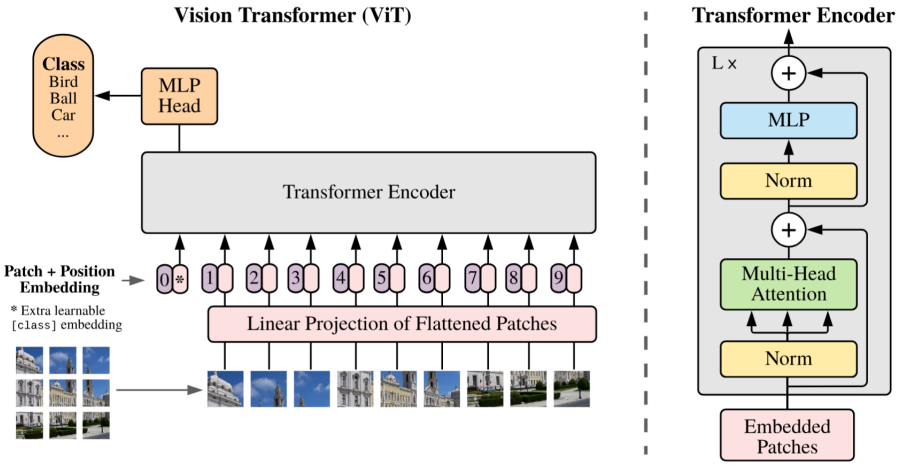

Figure 3.4: Visual Transformer (ViT) tokenizes images through the selection of patches, which are subsequently flattened and embedded. *Source: Dosovitskiy et al.[86]*

ViTs were used to process images, with the hope to achieve with images the same performance gain transformers brought to language processing. ViT behaves like an encoder, as it extracts a feature vector from images and can thus be used inside encoder-decoder DNN architectures, the most popular architecture class for segmentation. ViT brought forward the era of transformers in medical image segmentation methods, first with UNETR[72] and SwinUNETR[73]. To process medical imaging volumes, these networks divided them in 3D patches (usually cubes with 16 voxel sides), which are then flattened (put in one dimension), and forwarded to the transformer block.

While these DNNs showed promising results after being pre-trained using self-supervised learning on a large dataset[114], they failed to match the performance of CNNs and particularly nnUNet both on diverse datasets and when trained from scratch[109]. Hybrid architectures such as nnFormer[87] and CoTr[115] experimented with interleaving convolutions and transformers, to improve performance by combining the strengths of both these technologies. The performances demonstrated by these hybrid models were above that of transformer-only models, but remained below common fully-CNN baselines, such as nnUNet.

3.2.2.4 Mamba

In 2024, selective structured state space models (Mamba) were proposed [116][88][117] [118] as a more efficient replacement for the Transformer models. Mamba improves in efficiency over the transformer by replacing computations they need for regular attention, which scale quadratically with sequence length, by state space models, which scale linearly with respect to sequence length.

Mamba was adapted later that year to medical image segmentation with U-Mamba[89], a UNet-like architecture using Mamba either in the convolutional encoder (U-Mamba_enc) or in the bottleneck (U-Mamba_bot). The interest of Mamba-based models for medical image segmentation was the prospect of modelling long-range dependencies. In CNNs, the size of the spatial context depends on the convolution's kernel size and the depth of the network (number of convolutions in the encoder). While the transformer based models have the capacity to model long-range dependencies, they remain limited by the quadratic cost of their attention. U-Mamba introduced long range context with a linear computational cost, while keeping the capabilities for local information analysis of CNNs.

When benchmarked on medical image segmentation datasets, U-Mamba[89] performs better than transformer-based DNNs, and on par with CNN-based baselines[109].

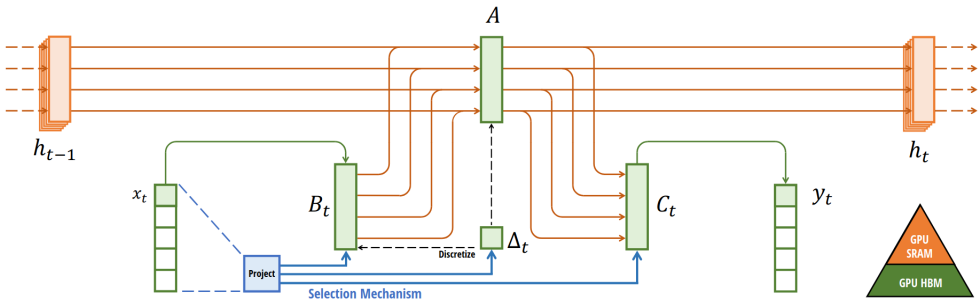

Figure 3.5: The selective structured state space model (Mamba) uses an architecture resembling a recurrent network with a hidden state h dynamically queried by the input x. *Source: Gu et al.[88]*

3.2.3 A Call for Rigorous Validation

At this point, nnUNet had appeared, in more and more papers, as a strong baseline, but was still not present in all of them. Some papers included only a comparison to the original UNet [119][120] or to a network trained with nnUNet, but on different data[121].

To address this situation, the team which built and maintains nnUNet published in 2024 *"nnU-Net revisited: A call for rigorous validation in 3d medical image segmentation"*. In this paper, they called for a systematic comparison of new medical image segmentation with nnUNet as a baseline[109]. They also show that the rapid adaptation of Mamba and Transformers to image segmentation led to a rapid adoption of these architectures in new research works, along with claims of superiority which lacked a thorough validation process. Moreover, they showed that, when submitted to a thorough comparative benchmark, these works did not meet the claims of superior performance over CNN architectures (such as the residual encoder U-Nets).

To solve this lack of proper validation, they proposed a benchmark composed of 6 popular 3D medical images segmentation datasets and used it to evaluate the performance of CNN, Transformer and Mamba based DNN architectures, by training all of them with nnUNet. They also included the performance of another DNN training framework, Auto3DSeg[122].

Results from this study showed that transformer based DNNs clearly under-perform when compared to nnUNet, on all benchmarking datasets, except for the ACDC dataset. Mamba-based models, while they rank closer to nnUNet, still fell short in most cases. Auto3DSeg also completely fails to reach similar performance to nnUNet, with the three tested architectures (DiNTS, SegResNet and SwinUNETR) [123] [124].

The CNNs evaluated in this paper were different sizes of the residual encoder UNet (ResEncUNet), STU-Net [125] and MedNeXt [126] architectures. STU-Net is a pre-trained, scalable modified U-Net, which showed good results with its version pre-trained on TotalSegmentator[6] but failed to converge when fine-tuned on BraTS. MedNeXt is a UNet modified to follow the recommendations proposed in ConvNeXt[127], and performed better than nnUNet and all other architectures (except on KiTs[128] where nnUNet takes first place), at the cost of a longer training time.

This article demonstrated that CNNs (UNet, MedNeXt) trained using a robust and validated framework (nnUNet) were the SoTA for medical images segmentation and should thus be used as baselines for performance comparison before any superiority claims. Additionally, producing novel architectures, a popular theme in the current publication landscape, should not be the focus of research teams. More time should rather be spent on ensuring the robustness, generalization capabilities and reproducibility of proposed methods through rigorous validation. This implies validating DNNs using large and diverse datasets comparing them with robust baselines and integrating them in available frameworks, such as nnUNet.

3.2.4 Advances in semi-automated medical images segmentation

Fully automated segmentation received lots of interest in recent years, which resulted in large improvements for both performance and accessibility of automatic segmentation tools. Despite this progress, segmentation methods which completely rely on algorithms were neither perfect, nor generalist enough to cover all medical use cases and rare occurrences. Additionally, for safety and accountability purposes, a human physician will always have to verify the contours returned by an algorithm.

Stemming from these observations, some researchers chose to shift their focus from fully automated to semi-automated segmentation, for more clinically realistic, human-in-the loop segmentation scenarios. The goal of semi-automated medical images segmentation is to reduce the time a doctor needs to spend contouring organs, lesions or any other volume of interest (VOI), while allowing them to remain in control of the process and have the possibility to correct any error.

3.2.4.1 Promptable Segmentation Models

The most popular method of semi-automated segmentation is promptable segmentation. Using a segmentation model and a prompt encoder, this method allows a user to input a prompt which will then guide the model in its segmentation task.

Prompts can take various forms: points in the image, bounding boxes, scribbles and more recently text (see Vision-Language Models for segmentation). The first breakthrough in this domain happened in 2023, with Meta's segment anything model (SAM)[129], trained on 11 million images and a billion segmentation masks to segment objects in natural, 2D images. The authors showed, via zero-shot transfer experiments (application to a new domain without re-training), that the SAM model could segment objects which were absent from its training set. SAM could also segment the same object with different levels of details (e.g.: a human, a face, an eye, the pupil). Meta later released SAM2[130], improving the performance of promptable segmentation on images and extending it to videos, then SAM3[131] which introduced object detection with concepts.

SAM was adapted for medical images, with MedSAM[132], a 2D medical image promptable segmentation model which used 1.5 million image-mask pairs to fine-tune SAM to the medical images domain. This work showed the great domain adaptability of SAM, precisely towards medical images. Other works quickly adapted SAM for 3D medical images: the same authors released MedSAM2[133], trained on both 3D medical volumes (CT/PET/MRI) and 2D videos (Ultrasound/Endoscopy). A key result published in the MedSAM2 paper is the reduced annotation time when using MedSAM2 in a human-in-the-loop annotation scheme. They showed 85.86% and 87.47% shorter annotation time at the third round of human-made contours refinement, for CT and MRI contours respectively. Other papers proposed similar methods, like SAM3D[134] and DSAM[135] as well as some redesigns for the same task of "Segment Anything" with different architectures[136], larger datasets [137] and new types of inputs [138].

Once again pushing for standard, accessible out-of-the-box models, the team behind nnUNet proposed nnInteractive[139], a promptable segmentation tool designed for all types of 3D medical images. Contrary to most of its predecessors, it does not handle 3D volumes as a sequence of 2D images, but directly with a 3D convolutional neural network (U-Net) and a patch based approach.

Additionally, instead of encoding prompts and images separately, the prompts are added as new channels to the input of the U-Net and directly fed to the encoder, thus bypassing the semantic barrier which previously existed between prompt encoding and image encoding. Training over 64 000 labeled volumes, including CT, PET, MRI scans, 3D ultrasound and 3D microscopy; they taught their U-Net to perform segmentation in a patch of the volume following a simulated user input, but also to correct an existing segmentation according to follow-up prompts. Points, scribbles, bounding boxes and lassos, all of them either positive or negative, can be used as prompts and are only needed in 2D, which makes this tool easy and intuitive to use.

Moreover, on its release, nnInteractive was already integrated in MITK[140] and napari[141]

viewers. Adoption of this tool kept rising later with its integration to 3D Slicer[142][143] and ITK-SNAP [144][145]. It has already been used to facilitate annotations of medical images in multiple studies[146],[147].

3.2.4.2 Vision Language Models for Segmentation

To improve the accessibility and intuitiveness of promptable segmentation algorithms, vision language models (VLM) were adapted to segment medical images, using text as a prompt. Originally built as a multi-modal improvement of LLMs, VLMs were trained to interface text and images by learning joint representations for these two modalities, often using image/captions pairs.

The first breakthrough in performance was obtained in 2023 when LLaVa[148] was trained on 158 000 image/text pairs, and showed the capacity to provide descriptions for new, unseen images. The following years, VLMs became common, with every major AI company releasing a VLM version of their flagship LLM.

The medical imaging community took interest in VLM, mostly to generate reports from medical images, but also to perform analysis of these images, with human guidance limited to a text prompt. The latter is an example of semi-automated segmentation taking advantage of existing text and image encoders to make the connection between text features and image features using projection heads[149][150] and train a VLM to perform segmentation guided by a text prompt. An integration in Napari is available for VoxTell[150], a text-guided segmentation model trained on 62 000 CT, MRI and PET scans along with corresponding segmentation masks and guidance text prompts.

3.3 Organs at risk segmentation using multiple anatomical modalities: The Head&Neck Segmentation challenge

In this section, we present a practical application of medical image segmentation to a real clinical task: organs-at-risk (OAR) segmentation for radiation therapy (RT) planning. Segmentation is performed using multi-modality imaging: unregistered CT/MRI.

3.3.1 Context

3.3.1.1 Computed Tomography

Computed tomography (CT) is a volumetric imaging method which uses a moving X-ray (XR) tube and detectors to acquire a 3-dimensional view of a patient's body. This imaging technique is used to obtain detailed anatomical information of a patient's internal organs.

A high energy electromagnetic ray (20 to 150 keV) is emitted by an XR tube and its attenuation by the matter in its path is measured by either an array of detectors moving at the same speed as the tube (third generation CT) or by a closed stationary ring array of detectors (fourth generation CT). XR is emitted by heating a tungsten wire at around 2400K to overcome electron-binding force and emit those electrons with a velocity which in turn determines the energy of the electromagnetic ray. This ray is then focused using electron optics devices to form a fan beam[151], passing through and attenuated by the patient's body, before being collected by sensors.

CT maps the attenuation value for each voxel using Hounsfield units (HU). These units are linearly assigned, with defined values for air (-1000 HU) and water in standard conditions (0 HU). These values are directly linked to the physical values, which allows quantitative analysis. This is not the case, for instance, with MRI, in which the values depend on acquisition parameters.

375 million CT scans were performed in 2021, according to a report by Frost & Sullivan[1], with a projected growth between 3 and 4% per year. This makes CT the third most used medical imaging modality behind X-ray and fluoroscopy (X-ray during surgery). It is used for internal imaging inside and outside oncology, for its ability to rapidly obtain 3D anatomical information. CT scans are however limited in their capacity to obtain physiological information and offer little contrast between structures with similar densities.

3.3.1.2 Magnetic Resonance Imaging

MRI does not rely on ionizing radiation, but on the properties of the hydrogen nuclei contained in organic matter. Indeed, when organic matter is placed in a powerful magnetic field (B_0), the nuclei of its hydrogen atoms align their spins with B_0. Then, when nuclei are perturbed, for instance by a radio-frequency (RF) pulse, a two step phenomenon occurs: first their spins are knocked out of their alignment with B_0, then they move back towards alignment with B_0, thus emitting a signal which is read by the MRI's antennas.

The variables measured are either the time it takes for the spin to reach 67% of its initial alignment with B_0 (T1) or the time it takes for the transverse component to decrease 67% (T2). Depending on the composition and physical properties of the body part in which the nuclei are situated, these times vary.

An MRI machine can be programmed to perform multiple measurements, mainly relying on T1 and T2 times. These measurements are called MRI sequences. Some of these sequences focus on enabling the visualization of anatomical structures, mostly fat and muscle, while others give more physiological information by measuring how easily water is moving inside the body (diffusion) or the density of protons in a tissue. Anatomical information is retrieved using T1 and T2 weighted MRI sequences (T1w and T2w respectively), both of which have similar contrast but inverted signal intensity: fat is in hyper-intensity in T1w but in hypo-intensity in T2w, while fluid shows in hypo-intensity in T1w and hyper-intensity in T2w. The most used physiological sequences are diffusion weighted imaging (DWI), Apparent Diffusion Coefficient (ADC) and Proton Density (PD). The first two are widely used in brain imaging since they offer a stark contrast for white, grey matter and cerebrospinal fluid. Proton density is nowadays mostly used for imaging of the joints, as it gives good contrast between fluid and cartilage[152].

Even though the images obtained with MRI give valuable insight to radiologists, quantifying them is a hard problem. The echo and response times measured depend on many factors, such as a patient's body weight, the MRI machine itself, the B_0 field (strength, irregularities) and the sequence's parameters. Because of this, MRI images are analyzed qualitatively by physicians and computer vision algorithms, meaning only relative differences in contrast can be established between tissue types.

To solve this problem, new MRI sequences were developed to obtain qMRI mappings, typically repeating the same sequence while varying only one parameter, then fitting a model on the multiple sequences to obtain quantitative values for the whole volume. This produces images called quantitative MRI mappings, or qMRI for short. T1_map, T2_map are qMRI mappings, which map the actual values of T1 and T2 in space, rather than relying on the relative difference in T1 between tissues (T1w).

qMRI mappings have been used to obtain functional information without needing ionizing radiation[153]. Unfortunately, the protocols for qMRI acquisition have remained mostly in the domain of research, or very specific applications such as cardiac MRI[154] or neuroimaging[155]. MRI thus remains mostly a qualitative, contrast-based imaging modality.

3.3.1.3 Multi-modal CT/MRI imaging

Both MRI and CT can give anatomical information, however CT can obtain more contrast between objects with different densities, while MRI can obtain more contrast between objects with similar density, but different compositions (i.e.: proton density). Combining these imaging modalities is therefore not redundant and gives a clearer picture of a patient's body.

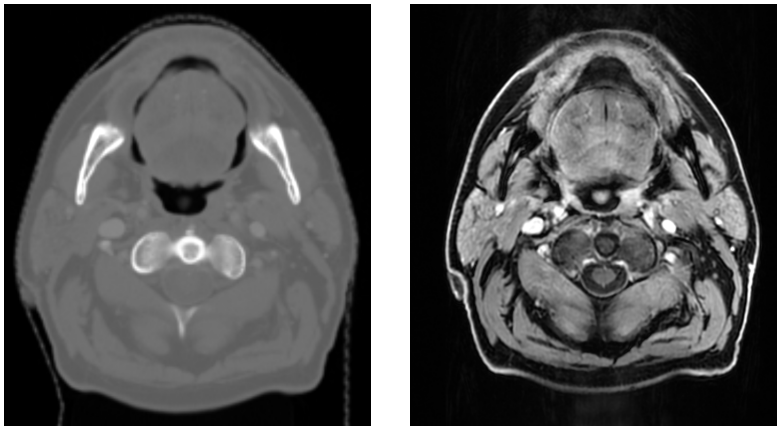

Figure 3.6: Side by side comparison of CT and MRI images for the same patient from the HaN-Seg dataset. While the CT has higher values in the high density areas, MRI gathers more signal in areas containing liquid, such as the brain and muscles.

The combination of CT and MRI acquisitions has shown good results, for instance, in the diagnosis of hepato-cellular carcinoma[156]. Nevertheless, clinical applications are still very rare and the research literature remains sparse. The main reason is the difficulty of combined acquisition of CT and MRI volumes, with no clinically available imaging device capable of such a combined acquisition, even though technological concepts are developed[157]. Indeed, CT scanners rely on a moving source of gamma rays, which makes it impossible to acquire a CT scan next to an MRI machine, due to the damage eddy currents would inflict to any hardware moving rapidly inside the MRI's intense magnetic field.

Using CT and MRI imaging together thus imply using a form of image registration, either automated (rigid, affine, elastic) or by hand, by a physician or technician. Registration is never perfect, but it can be improved by keeping the patient in the same position (e.g.: arms up) during both acquisitions when possible, or by using foam positioners if needed[158]. Such restraints are often already used in radiotherapy, to minimize patient movement between the planning CT treatment and RT sessions. This is done to ensure that the correct dose is delivered to the tumor and without exceeding the OARs' dose limits. This makes RT a good treatment target for the use of combined CT/MRI imaging.

3.3.1.4 Organs-at-risk in the head and neck area for radiation therapy

External radiation therapy (RT) consists of firing ionizing radiation to destroy or stop the spread of malignant cancer cells. It is a localized treatment, which focuses only on tumors and their surroundings, which consist of safety margins, named clinical target volume (CTV). Accounting for uncertainties in the treatment delivery, a larger volume is defined as the planning target volume (PTV). In this volume, radiation delivered should be as high as possible to destroy cancer cells.

However, ionizing rays pass through other organs, which contain healthy tissues and, while the RT technology for dose distribution has greatly improved over the years[159],[160], ionizing rays still deposit some of their energy in the organs on their path. When the target volume is surrounded by fragile and/or critical organs (i.e.: organs-at-risk), it creates constraints on the dose which can be safely delivered to the tumor, due to the dose these organs will be exposed to.

To make sure the maximal dose is delivered to the tumor, while minimizing exposure of the organs-at-risk (OAR), radiotherapy physicists go through a planning phase. RT planning consists in defining the number, intensity, shape and direction of the rays that will be shot during each RT session. This planning takes into account the dose limits deemed acceptable for each OAR to receive, by simulating the RT plan on a replica of the patient's body, obtained using the density values from the CT.

However, to measure the dose delivered to each OAR, contours for these organs have to be given to the planning software. The completely manual delineation of these organs was part of the work of RT oncologist, until commercial solutions started appearing[161]. The performance on these tasks was good enough that RT oncologists could use them as drafts and correct them. To further alleviate the burden of manual delineation or correction of automated delineation, RT oncologists look for better performing segmentation algorithms, covering all OARs, while running rapidly enough to be integrated in the RT workflow.

The head and neck area (referred to as HN) contains numerous, fragile organs (brainstem, eyes, teeth, jaws, optic nerves, buccal mucosa...). Some of these organs, such as optic nerves, are hard to delineate precisely on CT scans and require MRI imaging for delineation, whether automatic or manual[162].

For this reason, multi-modal (CT/MRI) imaging has been used to improve automatic segmentation of HN OARs[163]. However, MRIs and CTs are not acquired simultaneously, which means they cannot be directly super-imposed, before being processed by a DNN. Registration algorithms exist, but fail or give imperfect results when the patient's position changes between CT and MRI acquisition. Automated OAR segmentation remains a complex problem, with high applicability and demand in clinical practice.

3.3.2 The HaN-Seg dataset and challenge

In 2023, Podobnik, J. et al. proposed the Head and Neck Segmentation challenge (HaNSeg)[8]. A dataset containing T1 weighted (T1w) Magnetic Resonance (MR) and contrast enhanced Computed Tomography (CT) images for 42 HN cancer patients was made publicly available. These images are focused on the head and neck region (typically from the top of the skull to the bottom of the thyroid) and are taken on patients who are undergoing RadioTherapy (RT) treatment due to a cancer diagnosis.

30 Organs-At-Risk (OAR) have been delineated, (see figure 3.7) either by a RT technologist or by a radiation oncologist and the contours are then curated by a medical imaging researcher. The images and contours for 14 other patients are withheld to evaluate the segmentation algorithms trained on the publicly available set.

The HaN-Seg challenge defined the task to complete as the development of an automated segmentation algorithm, capable of contouring 30 OARs, using non-registered CT and MR scans.

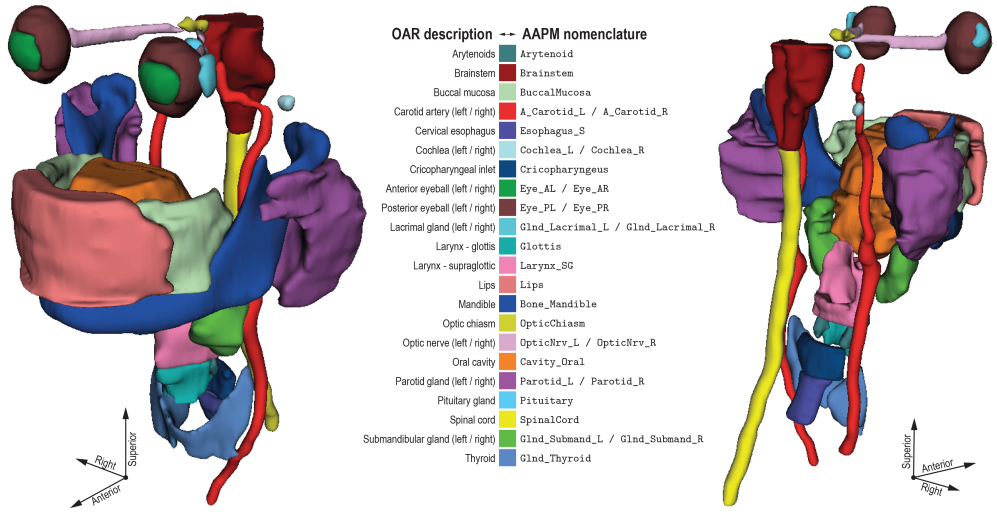

Figure 3.7: Organs at risk contoured for each patient in the HaN-Seg dataset. Left/right distinction is applied, bringing the number of segmentation labels to thirty. *Source: han-seg2023.grand-challenge.org*

The acquisition of CT images was performed using either Philips Brilliance Big Bore or Siemens Somatom Definition AS scanner, using standard clinical protocols (x-ray tube voltage: 100-140 kV; FOV: $500 * 500$ - $800 * 800 mm^2$). All MR images were acquired using GE Optima MR450w scanner with either spin echo or gradient echo T1w sequences, with FOV. 71% of the patients were male, in all the subsets and the average age was 59.5, 60 and 58.2 years in the training, validation and testing set respectively. Axial pixel size ranged between $0.51 * 0.51 mm^2$ and $1.56 * 1.56 mm^2$ for CT scans and between $0.47 * 0.47 mm^2$ and $0.82 * 0.82 mm^2$ for MRI scans, which was a very precise imaging protocol and will have consequences in the way we will have to handle the images in our algorithm (see 3.3.3.3).

The gender and age information for each patient were not included in the training, validation and testing subsets. No additional information was given on either image acquisition or patient clinical condition. The segmentation task had to be completed in under 15 minutes per patient, a time constraint put in place to exclude solutions which requires an unrealistically large amount of computation time. Six teams submitted their own method to solve this challenge, however only five of them submitted a complete method report. The paper going over the results of this challenge is published in Radiotherapy and Oncology[8].

The first difficulty to overcome in this challenge was to build a fast but correct registration between CT and MR volumes. These volumes are acquired on different machines and even though the use of 5-point thermoplastic RT immobilization masks has limited patient movement in the HN area, the patient's position was not the same in space, which means the CT and MRI volumes have spatially mismatched voxels.

To correct this, a registration algorithm had to be applied, with the goal to make anatomical structures in the MRI volume align with the ones in the CT volume. Since the sequence used is T1w, showing the difference in T1 relaxation time between tissues, anatomical structures are visible in the MRI volume and will be fairly easy to match to the corresponding anatomical structures in the CT volume, considering that the thermoplastic immobilization mask prevents large changes in

the patient's head and neck positions.

3.3.3 Proposed Algorithm

In this section, we will go over the different components of the solution we built and submitted for the HaN-Seg challenge. Each component is described separately in detail and figure 3.8 shows how all these components interact with each other. A summary of all the novelties introduced by our method is available in table 3.1, as well as the related improvements to the baseline.

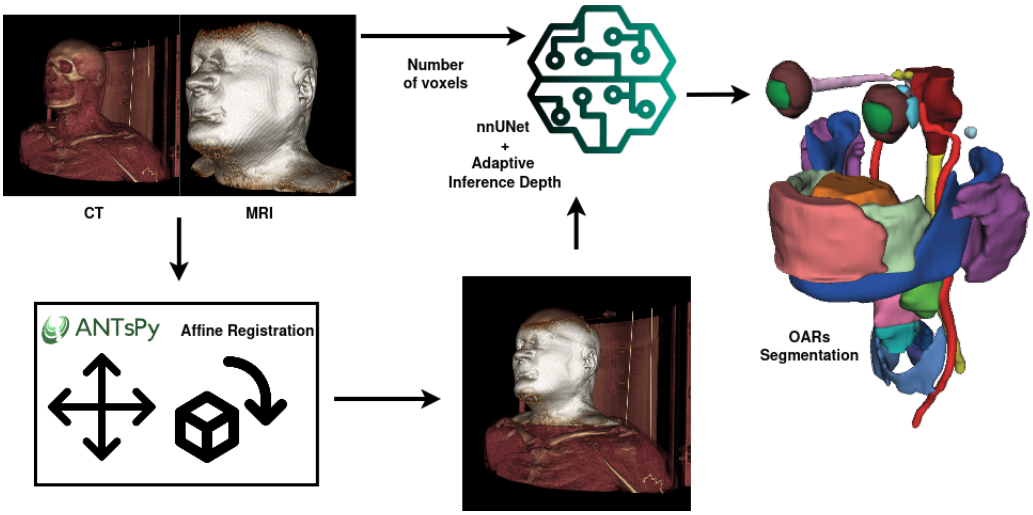

Figure 3.8: A summary of our proposed algorithm for fully automated HN OAR segmentation from unregistered CT/MRI. This figure represents the inference step of our algorithm once the UNet segmentation has been trained. Details on both the training and inference steps are provided below.

3.3.3.1 Registration

Like most teams, our solution relied on a simple affine registration, namely the one available in ANTsPy (v 0.4.2), a library relying on the Advanced Normalization Tools ecosystem[54]. The CT volume was used as reference, because it is the one on which radiologists delineated OARs, while only using the MRI volume as guidance.

Once the registration transform is computed, we apply it to the MR volume and obtain a registered MRI (regMRI) volume. This registration is computed and applied for each MRI volume in the training dataset. This yields a new dataset, containing CT, registered MRI and ground truth label volumes, for all the patients in the original dataset. As shown in figure 3.8, using this method, the registered MRI is very well aligned with the CT, even though no guarantee is available at a voxel level.

3.3.3.2 Model architecture and training

Using this dataset, we trained a U-Net[17] network with the nnUNet framework[82], a tool widely used to train neural networks to perform medical image segmentation and presented extensively in 3.2.2.2.

To start training the DNN with better initial weight values state than random, we used STU-Net, a modified U-Net pre-trained on 1204 CT volumes and 104 segmentation classes from the TotalSegmentator[6] dataset. By removing the first and last convolutions from this pre-trained network, it can be used for any medical image segmentation task, in our case with two input channels (CT and regMR) and 31 output classes (30 OARs + background). The STU-Net is trained following standard nnUNet training with modifications to the loss function and the choice of data augmentations.

A class-adaptive Dice loss[164] (ca-Dice) replaces the dice loss in the compound Dice/Cross Entropy loss. Class adaptive computes the Dice loss for each volume individually (i.e. it is not batched) and includes only the classes present in the current patch, in the ground truth label map. The ca-Dice is defined for a given volume patch treated by the DNN, as:

$$\frac{1}{N}\sum_{b,c}\frac{2\sum_v P_{bcv}G_{bcv}}{P_{bcv}+\sum_v G_{bcv}+\epsilon} \tag{3.1}$$

where the number of classes visible in the given patch N is defined by:

$$N=\sum_{b,c} min(\sum_v G_{bcv}, 1) \tag{3.2}$$

where b is the element index in the batch, c is the class label (excluding background), G is the ground truth and P is the predicted segmentation. The value of the Dice loss is 0 for all the classes not present in the processed patch.

To reduce variance in loss value, ca-Dice loss is computed across a training batch and not a single patch. Including only the classes visible in the local patch during training increases the weight of the classes which cover only a small number of voxels in the image, by setting to 0 the Dice loss for the classes not present in the patch, thus increasing their relative weight in the training loss.

Since some of the classes (Eyes, carotid arteries, glands) are differentiated between the left and the right side, all mirroring augmentations were removed from the training pipeline to allow the network to learn the concept of left and right in a medical image. This design choice is replicated in the inference pipeline.

3.3.3.3 Inference

Mirroring was also removed entirely from the test-time augmentations, for the same reason it was disabled during training. During inference, it was noted that our algorithm used a very large amount of Central Processing Unit (CPU) memory, which eventually led it to cause out-of-memory errors when tested on the challenge's infrastructure. The largest CPU memory allocation occurred during the last step of nnUNet's segmentation, when all patches had gone through the network and were gathered in a matrix with dimensions [c, h, w, d], with c the number of output channels (31 in our case) and h, w and d the spatial dimensions. This matrix then usually goes through the following processing steps:

1. A non-linearity (sigmoid function) is applied to it
2. The matrix is resampled to the original spacing and size of the CT scan
3. The *ArgMax* function is applied to keep only one class for each voxel

Due to the nature of the HaN-Seg dataset, spatial dimensions can be very large when entering the nnUNet inference step, i.e. once they have been resampled to the median spacing of the training dataset. Indeed, this spacing is around 0.5 by 0.5 millimeters in the axial dimensions, which means that when volumes with a thicker spacing would be resampled to this spacing, their dimensions resampled would become very large, particularly h and w.

Combining a large number of channels with these large dimensions, the application of a non-linearity function (which required a new tensor with the same dimensions to be allocated) used a lot of memory. The next step, resampling with trilinear interpolation, also used both lots of memory and runtime, since trilinear interpolation is applied on the float values resulting from the sigmoid function.

To alleviate this memory and runtime load, the processing steps were changed to:

- Applying the argmax function to choose the class for each voxel
- Resampling the volume using nearest neighbor interpolation

The output tensor created during the application of the argmax function consumed 62 times less memory than for the sigmoid function (only 1 channel instead of 31 channels and 8 bytes integers instead of 16 bytes floats). Removing the non-linearity function entirely and resampling using nearest-neighbour interpolation also sped our algorithm up a lot, giving us time to run more folds of the nnUNet ensembled models, with more overlap between patches.

Two models were built from these STU-Nets, specifically an ensemble of five STU-Nets (one for each fold of a five-fold cross-validation), the "5F" model and a single STU-Net trained on all the training data, the "all" model. By combining the predictions of the five STU-Nets, an error in segmentation caused by one of them can be corrected by the four others and overall segmentation performance improves, which makes the 5F model the most robust. However running multiple models increases the runtime and we are limited to 15 minutes for our whole algorithm.

To use the best model as often as possible, without crossing the time limit, the total number of voxels in the resampled image is computed and one of the following inference schemes is chosen accordingly:

- Model *all* and a step size of 0.6 if $nb_voxels > 1.7 * 10^8$
- Model *all* and step size of 0.4 if $1.3 * 10^8 < nb_voxels \leq 1.7 * 10^8$
- Ensemble model *5F* and step size of 0.6 if $1 * 10^7 < nb_voxels \leq 1.3 * 10^8$
- Ensemble model *5F* and step size of 0.4 if $nb_voxels \leq 1 * 10^7$

This adaptive inference depth, determined at runtime, allows our algorithm to optimize the amount of compute it uses for each case.

Method	Improvement
Registration	Spatial coherence is obtained between CT and MRI voxels
Custom Patch Agglomeration	Faster, less resource intensive process, remaining below the challenge's limits
Adaptive Inference Depth	Makes the most of the inference time

Table 3.1: Novelties implemented in our methods and the related improvements

3.3.4 Results

OAR	DSC: mean ± standard deviation (%)					
	eli1	cwlg102	CHB-QuantIF	Mamaa	m.m.gs	TurboMiki
Carotid artery (L)	82.8 ± 5.1	**85.2 ± 4.5**	79.6 ± 6.0	80.1 ± 5.8	71.2 ± 19.4	63.8 ± 9.5
Carotid artery (R)	85.2 ± 3.3	**86.8 ± 3.5**	82.4 ± 3.8	81.7 ± 5.2	69.8 ± 22.1	63.8 ± 7.3
Arytenoids	59.9 ± 14.3	52.3 ± 23.9	55.4 ± 13.1	**62.1 ± 12.5**	47.7 ± 26.5	39.6 ± 14.2
Mandible	94.3 ± 1.7	**95.0 ± 1.6**	94.2 ± 1.3	93.2 ± 2.0	94.6 ± 2.1	88.1 ± 2.7
Brainstem	**88.5 ± 4.7**	84.9 ± 3.1	86.9 ± 3.4	84.6 ± 3.6	85.4 ± 4.9	79.9 ± 3.4
Buccal mucosa	69.1 ± 8.9	**71.1 ± 7.3**	67.2 ± 9.5	68.0 ± 9.9	70.9 ± 8.4	59.7 ± 11.1
Oral cavity	89.4 ± 4.6	89.4 ± 4.5	**89.6 ± 4.5**	88.9 ± 4.0	87.6 ± 6.3	85.6 ± 4.4
Cochlea (L)	73.4 ± 10.1	**78.8 ± 7.8**	69.1 ± 8.7	72.5 ± 8.6	69.9 ± 14.7	61.6 ± 12.4
Cochlea (R)	74.1 ± 10.6	**78.2 ± 8.1**	68.6 ± 12.2	66.2 ± 12.5	67.4 ± 18.9	58.9 ± 19.1
Cricopharyngeal inlet	**64.5 ± 11.4**	62.8 ± 9.4	62.8 ± 11.2	62.4 ± 11.0	59.9 ± 18.0	55.5 ± 11.0
Cervical esophagus	**63.1 ± 13.2**	62.1 ± 12.3	61.8 ± 11.9	58.3 ± 14.6	54.8 ± 17.5	52.6 ± 15.2
Eyeball (A, L)	76.6 ± 7.5	**79.1 ± 7.0**	77.6 ± 4.8	76.3 ± 8.0	75.0 ± 14.0	62.9 ± 22.4
Eyeball (A, R)	78.5 ± 7.2	**80.6 ± 5.7**	78.6 ± 5.0	78.8 ± 5.1	72.3 ± 20.6	64.9 ± 13.9
Eyeball (P, L)	92.0 ± 2.0	**93.0 ± 1.4**	91.5 ± 1.9	91.5 ± 2.3	92.5 ± 2.2	86.0 ± 6.3
Eyeball (P, R)	91.2 ± 1.6	**92.6 ± 1.4**	91.1 ± 1.4	91.5 ± 1.9	91.7 ± 2.1	84.9 ± 6.3
Lacrimal gland (L)	60.8 ± 14.8	65.3 ± 10.4	61.4 ± 13.6	61.8 ± 9.6	**66.3 ± 10.3**	45.0 ± 16.2
Lacrimal gland (R)	61.9 ± 15.2	**64.2 ± 13.7**	61.1 ± 11.9	59.4 ± 13.2	61.2 ± 16.3	39.6 ± 24.3
Submandibular gland (L)	**85.7 ± 8.2**	82.2 ± 10.8	83.1 ± 10.8	82.7 ± 10.3	82.5 ± 10.5	74.7 ± 10.4
Submandibular gland (R)	84.5 ± 6.5	82.9 ± 7.0	83.1 ± 6.6	83.2 ± 9.0	**85.1 ± 5.3**	73.8 ± 6.6
Thyroid gland	**88.3 ± 4.7**	85.4 ± 15.3	86.9 ± 5.5	86.5 ± 7.0	86.9 ± 6.3	79.3 ± 8.9
Glottic larynx	**75.0 ± 6.9**	67.5 ± 11.2	73.3 ± 6.4	73.1 ± 6.8	70.6 ± 14.1	64.1 ± 11.0
Supraglottic larynx	**80.9 ± 5.1**	79.3 ± 7.0	79.7 ± 6.5	79.4 ± 4.9	79.0 ± 6.1	73.8 ± 8.6
Lips	**74.0 ± 8.0**	73.8 ± 10.2	73.5 ± 8.7	72.1 ± 8.8	71.0 ± 11.9	65.5 ± 8.5
Optic chiasm	44.6 ± 8.6	**45.6 ± 13.2**	41.1 ± 16.2	44.4 ± 10.1	35.3 ± 15.4	31.9 ± 16.1
Optic nerve (L)	67.7 ± 11.8	**71.7 ± 8.7**	64.3 ± 10.0	65.8 ± 11.2	65.8 ± 16.9	47.1 ± 11.0
Optic nerve (R)	72.5 ± 7.5	**75.5 ± 5.3**	67.2 ± 8.3	71.8 ± 5.6	68.5 ± 13.6	0.0 ± 0.0
Parotid gland (L)	**86.7 ± 2.2**	85.1 ± 3.8	86.1 ± 2.2	83.9 ± 3.1	83.0 ± 11.5	76.3 ± 7.8
Parotid gland (R)	**85.7 ± 3.7**	82.1 ± 6.6	84.0 ± 4.1	82.8 ± 4.6	82.1 ± 6.2	72.6 ± 8.3
Pituitary gland	**73.6 ± 9.2**	68.7 ± 12.4	68.0 ± 14.4	71.5 ± 12.7	63.5 ± 22.4	0.0 ± 0.0
Spinal cord	**83.3 ± 2.4**	82.0 ± 3.9	82.4 ± 2.9	80.2 ± 5.0	80.2 ± 7.5	74.6 ± 4.5
Mean performance	**76.9 ± 8.4**	76.8 ± 9.3	75.1 ± 8.6	75.2 ± 8.4	73.1 ± 14.0	60.9 ± 11.6

OAR	HD_{95}: mean ± standard deviation (mm)					
	eli1	cwlg102	CHB-QuantIF	Mamaa	m.m.gs	TurboMiki
Carotid artery (L)	3.5 ± 4.6	**3.0 ± 4.2**	4.7 ± 5.2	5.0 ± 5.3	44.7 ± 21.1	17.8 ± 3.6
Carotid artery (R)	**2.7 ± 4.6**	4.3 ± 9.5	3.2 ± 3.1	3.0 ± 4.1	55.7 ± 42.3	27.3 ± 16.4
Arytenoids	3.0 ± 1.4	4.7 ± 4.8	3.6 ± 1.4	**2.8 ± 1.3**	6.2 ± 5.4	5.0 ± 2.5
Mandible	1.3 ± 0.8	**1.2 ± 0.8**	1.4 ± 0.7	1.6 ± 0.8	1.7 ± 1.1	5.2 ± 2.3
Brainstem	**3.9 ± 2.1**	4.7 ± 1.6	4.6 ± 1.8	4.5 ± 1.6	4.8 ± 2.7	5.5 ± 1.3
Buccal mucosa	5.3 ± 2.4	**5.1 ± 2.5**	5.5 ± 2.6	5.8 ± 2.7	5.5 ± 2.4	6.5 ± 2.2
Oral cavity	5.3 ± 2.6	5.4 ± 2.4	**5.0 ± 2.1**	5.8 ± 2.7	15.8 ± 35.3	7.2 ± 2.2
Cochlea (L)	1.4 ± 0.7	**1.3 ± 0.8**	2.0 ± 0.6	1.7 ± 0.8	9.9 ± 30.3	9.6 ± 27.8
Cochlea (R)	1.9 ± 0.9	**1.6 ± 0.9**	2.3 ± 0.7	2.4 ± 1.3	1.9 ± 1.0	2.3 ± 0.9
Cricopharyngeal inlet	6.4 ± 3.5	6.5 ± 3.5	**6.2 ± 3.7**	7.2 ± 6.7	6.4 ± 4.0	7.5 ± 3.7
Cervical esophagus	8.0 ± 4.9	**7.5 ± 3.7**	7.6 ± 3.8	8.1 ± 3.3	7.9 ± 3.9	8.4 ± 3.9
Eyeball (A, L)	**2.3 ± 0.6**	2.4 ± 0.7	2.5 ± 0.5	**2.3 ± 0.7**	2.6 ± 0.9	3.2 ± 1.4
Eyeball (A, R)	2.3 ± 1.2	**2.1 ± 0.7**	2.3 ± 0.7	**2.1 ± 0.6**	2.5 ± 1.4	3.5 ± 1.5
Eyeball (P, L)	1.9 ± 0.5	**1.5 ± 0.3**	1.8 ± 0.6	1.8 ± 0.5	1.6 ± 0.7	2.9 ± 1.4
Eyeball (P, R)	1.9 ± 0.6	**1.6 ± 0.5**	2.1 ± 0.5	1.8 ± 0.5	1.9 ± 0.6	2.8 ± 1.0
Lacrimal gland (L)	3.9 ± 1.7	**3.1 ± 1.2**	**3.1 ± 1.6**	3.4 ± 1.7	3.3 ± 1.5	4.1 ± 1.7
Lacrimal gland (R)	3.7 ± 1.9	**3.6 ± 1.6**	3.8 ± 1.8	3.9 ± 1.7	4.1 ± 2.1	6.1 ± 2.9
Submandibular gland (L)	**3.1 ± 2.4**	6.5 ± 11.3	4.3 ± 4.4	5.1 ± 5.9	7.7 ± 12.4	5.0 ± 2.8
Submandibular gland (R)	3.9 ± 2.8	4.1 ± 2.8	4.1 ± 2.4	4.5 ± 3.2	**3.8 ± 2.1**	6.2 ± 2.3
Thyroid gland	**2.6 ± 2.0**	4.0 ± 7.0	2.9 ± 2.1	3.1 ± 3.2	3.3 ± 2.3	4.7 ± 2.6
Glottic larynx	**2.8 ± 1.2**	3.7 ± 1.9	3.0 ± 1.2	3.4 ± 1.3	14.4 ± 39.9	4.3 ± 2.1
Supraglottic larynx	3.5 ± 1.2	4.2 ± 1.7	**3.4 ± 1.3**	3.6 ± 1.3	4.4 ± 2.0	4.2 ± 1.9
Lips	6.5 ± 3.2	**6.3 ± 3.0**	6.5 ± 2.9	6.9 ± 3.0	8.1 ± 4.1	8.3 ± 4.3
Optic chiasm	4.2 ± 1.7	4.4 ± 2.1	4.4 ± 1.8	**4.1 ± 1.6**	5.4 ± 2.1	6.3 ± 2.5
Optic nerve (L)	2.6 ± 1.5	**2.2 ± 0.9**	2.4 ± 0.8	2.4 ± 0.9	6.4 ± 8.9	5.4 ± 3.3
Optic nerve (R)	2.7 ± 1.9	**2.5 ± 1.6**	2.8 ± 0.9	2.7 ± 0.9	5.5 ± 7.4	28.5 ± 0.0
Parotid gland (L)	**5.1 ± 2.6**	**5.1 ± 2.5**	5.3 ± 3.5	5.9 ± 3.0	8.3 ± 8.4	8.8 ± 3.5
Parotid gland (R)	**5.3 ± 3.1**	6.6 ± 5.1	6.2 ± 3.1	6.4 ± 3.5	6.0 ± 2.4	51.1 ± 36.1
Pituitary gland	**2.1 ± 0.7**	2.4 ± 1.0	2.6 ± 1.1	2.3 ± 1.0	18.6 ± 60.4	159.9 ± 73.2
Spinal cord	2.6 ± 2.4	2.1 ± 1.2	**2.0 ± 0.6**	2.3 ± 1.1	5.9 ± 6.5	3.7 ± 2.6
Mean performance	**3.5 ± 2.4**	3.8 ± 3.8	3.7 ± 2.3	3.9 ± 2.7	9.1 ± 18.4	14.0 ± 16.2

Figure 3.9: DSC and HD95 for all organs, for all algorithms submitted to HaN-Seg. Our solution is highlighted in red.

Our model was evaluated on the withheld test set (n=14) and scored third place with an average Dice score of 0.750 and Hausdorff distance of 3.714mm. It obtained the best precision on oral cavity (DSC=89.6(+/-4.5), HD^{95}=5.0(+/-2.1)) as well as the best localization (HD^{95}) for the larynx (3.4(+/-1.3)), cricopharynx (6.2(+/-3.7)) and spinal cord (2.0(+/-0.6)) respectively. Extensive results are presented in the challenge organizers publication of the challenge's results[8] and in figure 3.9.

3.3.5 Discussion

Leveraging a pre-trained U-Net-like model (STU-Net), we obtained top 3 performance in the Head and Neck segmentation challenge with an average Dice score of 0.750 and Hausdorff distance of 3.714mm.

A simple, low-compute affine registration achieved a satisfying registration of the MRI towards the CT scan, while the use of a pre-trained DNN, a common and effective strategy when dealing with a small dataset, proved useful for this challenge. The top 3 participants proposed solutions that reached a precision level similar to human inter-observer variability on a subset (n=27) of this data[165]. These models can be used to help curate large, high quality head and neck OARs segmentation datasets, to achieve performance better than that of human experts[39].

Registration methods were all similar, with a rigid or affine registration used mostly in 3D, but one team used 2D coronal cross-sections as reference. The time limit prevented the use of more precise but longer methods such as elastic deformation, although the results were already visually correct using rigid or affine registration. Registration frameworks SimpleElastix[50] and AntsPy[54] were used, with seemingly no major difference in registration quality or performance.

The architectures used and training processes were all variants of nnUNet (vanilla nnUNet, DynUNet, STU-Net, UID-Net), proving once more it is state-of-the-art in medical images segmentation. The only team which reported not using nnUNet performed way below the other teams (DSC was 60.9 vs 73.1-76.9). While using the 5-models ensemble proposed by default in nnUNet can prove time consuming, adaptive inference depth allowed our algorithm to respect inference speed requirements, even for large scans, while keeping the best performance on smaller scans.

Clinical implications for this challenge are meaningful, starting with having demonstrated the superiority of AI methods to inter-observer agreement on most organs[8], albeit on a small testing dataset. The size of the training dataset is also an important element: using only 42 training cases, participants demonstrated the possibility to build fully automated segmentation methods for 30 organs at risk in the head and neck area. Methods used in this challenge were all based on DL and DNNs, which were able to achieve convergence on such a small dataset.

The quality of the dataset, the density of supervision (30 organs per scan) and the specificity of the classes (left/right distinction, different parts of the respiratory system) gave the DNNs a lot of context and likely enabled their convergence, even with less than 100 samples, which is usually not considered enough data to obtain such results using DL[166].

The inclusion of multiple imaging modalities (MRI and CT) showed the possibility of the integration of MRI in segmentation models. In particular, the fields of view (FOVs) of the CT and MRI scan did not always entirely overlap, which meant the DNNs needed to learn to extract information from the MRI but also produce a reliable segmentation in the case some organs were not present in the FOV of the MRI. One team opted to discard the MRI scan altogether and focused on CT only,

after testing their method on both single and multi-modality data.

The registration of MRI and CT was however performed using simple registration algorithms (rigid or affine), with good results, likely due to the anatomical nature of both CT and T1w MRI. Landmarks such as the bones (skull, vertebrae, lower jaw) can constrain the earlier matching steps, preventing complete registration failure.

Since the scans were not acquired in the exact same position, perfect registration was not always possible, without using elastic deformation. But later work has shown that DNNs are capable of learning from data with small mismatches in registration[167].

3.3.6 Conclusion

Using a small, high-quality CT/MRI segmentation dataset, the HaN-Seg challenge has demonstrated the possibility of training DNNs to perform fully automated HN OAR segmentation, with performance on par with that of human radiologists. nnUNet has also shown that it remains the best tool to train DNNs for segmentation in such low data regimens.

Our algorithm used adaptive inference depth, which allowed it to always use the best of its capacities, while respecting computing time and memory limits. Pre-training on a larger dataset and training modifications during fine-tuning such as disabling mirroring and adapting the Dice loss have led to improved performance over the baseline.

Larger fine-tuning datasets with similar quality in imaging and annotations could lead to the creation of clinical practice-ready HN OAR segmentation algorithms, which would assist RT oncologists and physicists in the task of RT planning.

Chapter 4

Automated PET/CT Lesion Segmentation: four years of improvements with AutoPET

The works presented in this chapter have contributed to three papers:

- Dexl J., Gatidis S., Früh M., ..., Mesbah Z., Modzelewski R. et al. (2025) *AutoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging, Part 2: Domain Generalization* Published in Journal of Nuclear Medicine

  URL: `https://jnm.snmjournals.org/content/jnumed/early/2025/12/30/jnumed.125.270260`

- Mesbah Z., Mottay L. et al. (2025) *AutoPET III: The Tracer Frontier*

  URL: `https://arxiv.org/pdf/2410.02807`

- Mesbah Z., Perret S., Mottay L. (2025) *Semi-automatic PET/CT segmentation across tracers*

  URL: `https://autopet.org/docu/autopetiv/Task1/3_Zhack_AutoPETIV_GoodFormat.pdf`

Chapter Overview

PET/CT is a multi-modality imaging technique, which combines functional and anatomical images. The segmentation of lesions on PET/CT volumes is key to the computation of prognostic factors, but remains a time-consuming task for nuclear physicians. Medical imaging researchers set out to solve this issue, using automated segmentation algorithms.

This is the motivation behind the AutoPET challenge series, which has taken place for 4 consecutive years from 2022 to 2025. AutoPET aimed to improve the state-of-the-art for automated PET/CT lesion segmentation, by providing large datasets and involving the medical imaging and nuclear medicine community.

In this chapter, we will describe how these challenges took place, which novel methods were proposed in our contributions, how AutoPET improved the PET/CT segmentation SoTA and which were the key factors to foster this improvement.

Contents

4.1 Context 53
4.1.1 PET/CT imaging 53
4.1.2 PET Quantification: Standardized Uptake Value 55
4.1.3 PET Acquisition methods 55
4.1.4 Multimodal PET/CT acquisition 55
4.1.5 PET/CT Analysis 56
4.2 AutoPET: data and challenge history 57
4.2.1 Datasets 57
4.2.2 Challenges 57
4.3 Our methods 59
4.3.1 AutoPET II 59
4.3.2 AutoPET III 60
4.3.3 AutoPET IV 62
4.4 Discussion 66
4.4.1 The Winning Solutions from Each Year 66
4.4.2 Insights from AutoPET 67
4.5 Conclusion 68

4.1 Context

4.1.1 PET/CT imaging

PET/CT is a multi-modality imaging technique, which combines positron emission tomography (PET) and computed tomography (CT, see 3.3.1.3). The physics of PET imaging rely on radioactivity and will be presented alongside the underlying biological mechanism in this section.

4.1.1.1 The Physics and Biology of Positron Emission Tomography

PET is a volumetric, functional imaging technique, which relies on the injection of a radio-tracer inside the patient's body. A radio-tracer is a combination of a molecule called a ligand (e.g.: FluoroDeoxyGlucose (FDG), water, choline, PSMA) with a radioactive, positron-emitting isotope (e.g.: ^{18}Fluor, ^{68}Gallium, ^{15}Oxygen, ^{64}Copper, ^{11}Carbon).

The role of the ligand is to bind or be consumed by either a specific part of the body, specific cells, or tissues exhibiting a certain behavior. Ligands targeting the prostate-specific membrane antigen (PSMA), such as gozetotide[168], have for instance been designed to bind with an antigen which is mostly present on the membrane of benign and malignant prostate cells. FDG, the most used ligand in oncology, is a sugar analog, consumed by the cells of all the organs which use sugar as an energy source.

Other ligands, such as fluoro-misonidazole (FMISO) accumulate in cells which are deprived of oxygen (hypoxic cells). Some amount of the ligand accumulates in the body's cleaning organs (kidneys, bladder) but intense activity will be observed in the areas containing the target compound, where the radio-tracer has bound in large amounts.

Once the ligand has bound to its target, its location is given away by the emission of high energy rays. To trigger the emission of these rays, a positron-emitting isotope is attached to the ligand molecule. It either replaces a stable atom in the ligand (for instance, a ^{18}F isotope can replace a stable fluor atom in FDG, to obtain ^{18}F-FDG), or is attached to it by a chelator, which is the case in ^{68}Ga-PSMA (common name for the ^{68}Ga-gozetotide).

This isotope is an unstable atom, with a neutron count higher than its atomic number. Because of this, it will decay into an element with a lower atomic number by losing a proton, gaining a neutron and emitting a positron (which is a positively charged anti-electron particle) and a neutrino. This positron can travel around 2-3 millimeters in matter, before meeting an electron with a similar energy level, at which point these two particles annihilate[169]. The energy from these particles is mostly converted into two 511 keV gamma rays, emitted in opposite directions.

4.1.1.2 PET Image Acquisition

These gamma rays are then detected using scintillation crystals, which emit light when struck by ionizing radiation. This light is detected by photo-sensors and a hit is registered by the PET scanning device. When two hits are detected at a similar time by two of the machine's detectors, it is marked as a coincidence event. Because light travels in a straight line, this event means that a positron-electron annihilation likely took place in a straight line between the two detectors, called the line of response (LOR, see figure 4.1). Some recent PET scanners are equipped with sensors capable of precisely measuring the time a photon collided with the detector (typically under a nanosecond),

which allows them to localize the annihilation on a shorter segment of the LOR by computing the (very small) time difference between the two collisions events[170].

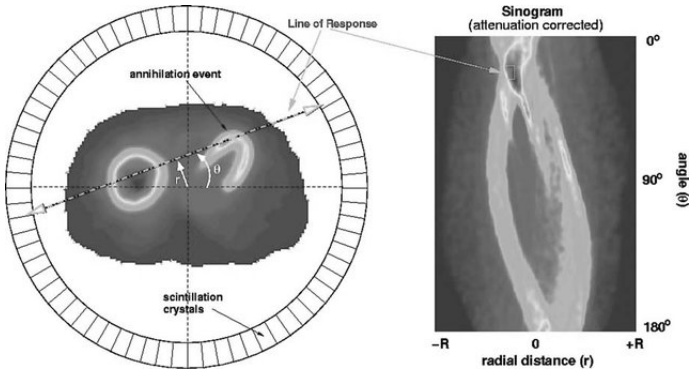

Figure 4.1: Line Of Response (LOR) and the sinogram obtained as a result of the addition of LORs. *Source: Volkow et al.[171]*

After a few minutes, using the list of detected coinciding collisions (gathered in a sinogram, see figure 4.1), a 3D map can be reconstructed, which represent the probable location of each detected positron-electron annihilation (see figure 4.2). From this probability map, the activity generated by the radio-tracer can be quantified spatially, which in turn shows how much of the tracer was accumulated in each part of the patient's body. This map is however often quite noisy, in part because of scatter, an interaction in which a photon created during the annihilation is deviated from its initial straight course.



Figure 4.2: A Maximum Intensity Projection (MIP) of a normal (i.e.: benign) 18F-FDG PET scan[172]. Radio-tracer uptake is visible in the brain and myocardium (heart muscle) which are constantly consuming the glucose analogue, as well as the kidneys and urinary bladder, which respectively filter and stores urine containing the radio-tracer. Thyroid uptake can also be seen in this example, while it is mostly not a cause for concern, it might be the sign of a non-cancerous disease[173]. *Source: radiopaedia.org*

4.1.2 PET Quantification: Standardized Uptake Value

The measured activity is influenced by multiple factors: the injected activity (usually measured in MBq), the time between the injection and the PET acquisition, the patient's body weight and body composition and the isotope's half-life.

To enable the comparison of PET scans at different points in time, or from different patients, PET scans are quantified using the standardized uptake value (SUV). This value is computed using the following formula:

$$SUV = \frac{Activity}{Activity_{Injected} * DecayFactor} * Weight \tag{4.1}$$

Activity is the measured $\beta+$ radioactivity, in Mbq/mL, weight is measured in grams, injected activity in Mbq and the decay factor is defined as:

$$DecayFactor = \exp \frac{-\Delta t * \log 2}{\lambda} \tag{4.2}$$

where λ is the half life of the isotope and Δt is the time between injection and acquisition.

The SUV PET volumes are quantified and can therefore be used to extract characteristics, such as the SUVmax, SUVmean or SUVpeak and lead studies on the impact of these characteristics on survival, disease progression or treatment response. Variants of the SUV such as SUL, a SUV normalized by the patient's lean body mass, have been developed to diminish inter-patient variability[174].

4.1.3 PET Acquisition methods

The most common method to acquire PET scans consists of a single 3D acquisition, performed usually one to two hours after the radio-tracer is injected (exceptions exist for PET tracers with short half-lives). However, new PET acquiqition modalities have emerged, such as parametric PET and 4D PET.

Parametric (often whole-body) PET is an imaging modality relying on successive, close PET acquisition. This type of PET scan lets the nuclear physician follow the evolution of tracer uptake over time in the body. Some applications of dynamic PET are simultaneous multi-tracer PET scans, using time-activity curves for tracer separation[175], better disease assessment through tracer perfusion analysis [176], or the dynamic analysis of a patient's metabolism [177]. While it is not yet widely used in clinical oncology, dynamic PET is already used in research and neurological PET and has been facilitated by the apparition of wide FOV PET machines.

4D PET relies on the cyclic acquisition of frames, which are synchronized with the patient's breathing. After reconstruction, a 4D PET scan suffers from less motion artifacts than regular PET scans, particularly less breathing induced motions around the lungs. 4D PET also enables precise metabolic tumor contours on PET, for each part of the breathing cycle. Radiotherapy can in turn be improved to better protect surrounding healthy organs, while delivering the same dose to the biological tumor volume.

4.1.4 Multimodal PET/CT acquisition

While PET scans have enabled functional imaging, they lack anatomical information. To complement them, a CT scan is acquired simultaneously. Indeed, PET/CT scans are the most frequent

multi-modal imaging medical exams where both modalities are acquired simultaneously. Unlike with MRI, there are no physical limitations to the acquisition of a CT scan inside a PET machine. This combined acquisition intrinsically ensures the correct registration of the two exams, which serves two main purposes:

- CT scans acquired alongside PET scans are used to offset a physical phenomenon called attenuation: photons emitted far away from the sensors will be more attenuated than the one emitted closest to the sensors (especially near the skin), which impacts the recorded activity and, downstream, the SUV values. Using the attenuation of the XR, collected during CT acquisition, the PET can be corrected for attenuation and a 3D map closer to the actual positron emission map is obtained.

- Using registered PET/CT scans, nuclear physicians can compare the functional imaging with anatomical landmarks. Compared to a single PET reading, this allows them to localize lesions and better assess a disease's extension and severity.

Overall, the PET/CT is the most common multi-modality imaging exam in oncology, as it lets physicians explore the metabolic function of a patient's body, while retaining anatomical information. There are limitations to this modality, for instance, PET spatial resolution is limited to around 3 millimeters, because of the distance traveled by the positron before it annihilates; while CT scans, being transmission imaging, can reach sub-millimetric spatial resolution. This often causes information loss when an automated analysis is performed and a common spatial resolution must be chosen.

4.1.5 PET/CT Analysis

To assess disease severity and progression, nuclear physicians rely on measurements which have been shown to influence prognosis such as total metabolic tumor volume (TMTV)[121]. To obtain such values, all lesions must be delineated in the PET/CT scan, which is a task nuclear physicians are trained for, but can still be very time-consuming. This is particularly the case with diseases such as lymphoma, melanoma and metastatic cancers, because of the high number of lesions visible on a PET/CT.

To alleviate this burden, automatic segmentation algorithms can be trained and deployed to assist nuclear physicians and radiologists in contouring, in one of two ways:

- **By performing a "first pass"** that a physician will then only have to correct

- **By interacting with the physician** and smartly propagating their inputs, using semi-automatic segmentation methods[139].

In the following sections, we will present the recent progress of automated and semi-automated PET/CT segmentation algorithms, through 4 years of AutoPET datasets and challenges as well as 3 contributions from our side.

4.2 AutoPET: data and challenge history

4.2.1 Datasets

For the first edition of this challenge, a dataset of 1014 whole-body PET/CT scans, with voxel level manual lesions segmentations, was made publicly available[178] in both DICOM and NifTi formats. 513 PET/CT scans are acquired on healthy patients, 188 on melanoma patients, 145 on lymphoma patients and 168 on lung cancer patients. All cancer subtypes have been histologically proven. Scans were collected from 900 patients, 43.8% of which were women and 56.2% were men with an average age of 59.4 years. Training and testing data was collected from University Hospital Tübingen and Ludwig Maximilians University (Münich) respectively.

The same public training dataset was used during the second edition, however the data used for testing is different, collected in multiple centers and include patients with varying diseases and profiles (see 4.2.2).

For the third edition, challenge organizers revealed a new dataset, this time including PSMA PET/CT imaging data, asking participants to create a multi-tracer segmentation algorithm.

This multi-tracer dataset was also used in the fourth edition of the AutoPET challenge, along with generated input clicks simulating a physician's interaction (see 4.2.2).

4.2.2 Challenges

AutoPET I The first challenge's goal was to create a fully automated algorithm which accurately delineates cancerous lesions in 18F-FDG PET/CT scans. In this edition, the test set was following a distribution similar to that of the training set. Algorithms proposed by the participants had to be submitted as a Docker container, with no access to external resources, thus they had to include the weights of any DNN used in the segmentation algorithm. The segmentation outputted by these algorithms on the data from the test set were scored using the following metrics:

- Dice Similarity Coefficient (DSC)
- False Positive Volume (FPV)
- False Negative Volume (FNV)

URL: `https://autopet.grand-challenge.org/`

AutoPET II While the training data and main objective were not changed for the second edition, the test set used to evaluate models was designed to test their generalization capabilities by including out-of-distribution data in the test set. While a quarter of the test data (n=50) followed the same distribution as the training data, the remaining data was obtained from different distributions.

Specifically:

- 25 cases represented similar diseases but were obtained in a different hospital (LMU)
- 25 cases were obtained in the same hospital, but from patients with different pathologies (oesophageal carcinoma, colon cancer, anal cancer, pancreatic cancer, seminoma, breast cancer, cervical cancer, endometrial cancer, and rectal cancer)

- 50 cases were pediatric oncology case, which makes for different, unseen anatomies
- 50 cases were cases of patients with prostate cancer, who had undergone PSMA PET/CT, which represents a major domain shift, due to the inclusion of a different tracer

The type of out-of-distribution data included in the test set was not communicated to the participants, to let them focus on building the most generalist algorithm, instead of building models to solve these specific cases. Two categories of algorithms were created for rewards, but the common, Docker-based submission system subsisted. The first category, *"best segmentation"*, scored methods using the regular segmentation metrics DSC, FPV and FNV, while the second category, *"robust segmentation"*, focused on robustness to the out-of-distribution data present in the test set, by dividing each metric (DSC, FPV, FNV) by their standard deviation over all patients in the test set. This category aimed to reward the algorithms showing the least variations in performance when out-of-distribution data was presented to them.

URL: `https://autopet-ii.grand-challenge.org/`

AutoPET III The third challenge tackled the multiplicity of tracers available for PET scans. The main objective remained the segmentation of cancerous lesions, only this time the PET images had been obtained through either FDG or PSMA PET.

The participants were free to choose between building one model capable of contouring lesion on both PET types, or to use pre-processing to determine the tracer type before performing segmentation.

Algorithm evaluation was performed for one half on data from the same distribution (FDG from Tübingen and PSMA from Munich) and, for the other half, on data from a different distribution (FDG from Munich and PSMA from Tübingen). Tracer and center information was not given to the algorithms during inference, they had to work without knowing this information *a priori*.

The practical uses for a multi-tracer segmentation model is the emergence of PSMA used in combination with ^{68}Gallium or ^{18}Fluor isotopes, mainly for prostate cancer screening, diagnosis and lesion localization. Additionally, a FDG PET scan can be acquired alongside PSMA PET, to determine if a patient only has PSMA-positive lesions and is therefore eligible for targeted therapies, such as radioligand therapy (RLT)[179].

Two categories were available, each with its own submission system:

- The "model-centric" category allowed participants to use any publicly available model and/or data to improve their algorithm's performance. In this category, participants were free to train a model capable of segmenting both tracers or two tracer-expert models.
- The "data-centric" category on the other hand disallowed the use of any data or pre-trained model additions, even from public sources, as well as modifications of the provided architecture. It instead encouraged participants to focus on preprocessing, data filtering and data augmentation strategies.

URL: `https://autopet-iii.grand-challenge.org/`

AutoPETIV In the fourth edition, organizers kept the same FDG/PSMA dataset, but shifted the focus from automated to semi automated segmentation through simulated human-in-the-loop

scenarios. Due to the nature of the challenge (asynchronous, lots of participants), the interaction for the evaluation was offline, with a constant set of 20 points representing expert clicks in each PET/CT volume. A list of automatically generated clicks were proposed to the participants and clicks on the test set were inputs from actual experts.

Participants needed to design a network capable of performing segmentation in a fully automated manner, while integrating a number between 1 and 10 of positive (lesion) and negative (background) clicks as a human guidance. Since clicks were determined beforehand, the task is not an iterative correction (which would be closer to what happens in a clinical setting). It does however still challenge participants to find the best way to include the expert input and steer their model's output.

The two categories presented were more different from each other this time, with the first one focused on semi-automated PET/CT segmentation while the second one tackled longitudinal CT segmentation. *"Semi-automated PET/CT segmentation"* was a regular segmentation task allowing the use of publicly available data and pre-trained models. Algorithms were tested in a fully automated setup (0 clicks) then in semi-automated setups, adding 1 pair of clicks at a time, until all available clicks are used.

"Longitudinal CT segmentation" was a completely different task, relying only on CT scans, with an added longitudinal component. A baseline CT and its corresponding lesions segmentation mask is given, alongside a follow-up CT for the same patient. The goal for this task is to produce the segmentation of lesions in the follow-up CT scan, guided by lesion and background clicks.

URL: `https://autopet-iv.grand-challenge.org/`

4.3 Our methods

During my PhD, I participated in the second, third and fourth editions of AutoPET. Below, I describe the methods used in my submissions for each year, focusing on the novelties our methods and methods from the other winning teams brought to the research community. The researchers who participated in the development of each year's method will be named at the beginning of each subsection.

4.3.1 AutoPET II

I participated in AutoPET II on my own in 2023.

Default DNN architecture For this edition, a new model was developed, based on a ResidualEncoderUNet architecture, one of the two architectures natively implemented at the time in the nnUNet framework (equivalent to the ResidualEncoderUNetM architecture in the current nnUNetv2 framework). More complex than the vanilla alternative (PlainConvUNet), because of the residual connections spanning its encoder, it was hypothesized that this DNN architecture would be capable of retaining more information and learning more complex relations from the large amount of data available for this challenge. The choice of the nnUNet framework stemmed from the growing body of literature which backed it as the state-of-the-art for segmentation tasks. While no modifications were made to the architecture of the neural network itself, I adapted the training loop from nnUNet to the AutoPET dataset, specifically to its duration.

Custom training scheme The total number of epochs for each fold was increased from 1000 (default value) to 1500 and the number of iterations per epoch from 250 and 50 to 1250 and 250, for one training and one validation epoch respectively. Considering a batch size of 2 (which was the highest power of 2 my 24 GB GPU could handle), the default value of 250 iterations per epoch meant only 500 patches were shown to the neural network at each epoch. This modification increased the number of sampled patches the network could learn from at each step to 2500, beyond the number of cases in the dataset.

The same logic was applied for the modification of the number of validation iterations, this time to obtain better validation metrics and be more robust when evaluating the best model to select (early stopping). Increasing the number of epoch, I sought to slightly improve the segmentation performance by giving the U-Net more time to improve at each step of the training, from big weight modifications at the beginning (high learning rate) to finer adjustments towards the end (lower learning rate).

4.3.2 AutoPET III

I participated in AutoPET III alongside Léo Mottay in 2024.

The approach we chose for this edition was first to automatically find out which ligand was used for the PET scan (FDG or PSMA), then forward the PET/CT volumes to one of two tracer-specific neural networks.

Tracer Discriminator After researching signal analysis methods based on a 1D projection of the PET scan and instead of running the whole 3D volumes through a DNN, a 2D CNN was used for this task. Indeed, we found that using 2D maximum intensity projections (MIP) of the PET scan on the coronal plane (i.e.: maximum along the front-back axis) performed almost perfectly and was way faster and required less memory and compute than 3D processing. MIPs were obtained by first resampling the TEP volumes to a common spacing ($3 * 3 * 3mm^3$), then computing the MIP on the coronal plane and selecting a 224 by 224 zero padded window centered around the center of the MIP (cropped or padded as needed). This $224 * 224$ image was then fed to our 2D CNN.

The CNN was composed of 6 2D convolutions (kernel size = [8, 3, 3, 3, 3, 6], stride = [2, 2, 2, 2, 2, 0], padding = [3, 0, 0, 0, 0, 0], channels=[8, 16, 32, 64, 128, 256]) followed by 5 fully connected layers. It reached 99.64% accuracy in the 5-folds cross-validation we ran over the entire AutoPET III dataset. Additionally, we measured the whole process (MIP + window selection + CNN) to last on average 2.18 seconds per scan on an Intel i7-11850H CPU, which was considered to be an acceptable amount of time, compared to the actual segmentation time (several minutes).

nnU-Net based Segmentation A similar architecture to the one used the previous year is used again in this edition: a U-Net CNN with a residual encoder (ResidualEncoderUNet M). The reason the name is different is simply a change in the nnUNet version (v1 to v2). The hyper-parameters were kept as defined by default using nnUNet's experiment planner.

The main architectural modification was in the input passed to the CNN: in addition to the PET and CT, provided in two channels, windowed versions of the PET and CT volumes were also provided, as two additional channels. The rationale behind this decision was that radiologists and

nuclear physicians have often used windowed images to enjoy a better contrast and consequently perform a better delineation of volumes of interest (organs, lesions).

The values in the PET scan volumes were clipped between 0 and 20 SUV, which excluded abnormally high SUV values (from injection tubing, filled bladder, spills). Similarly, the values in CT scans were clipped between -300 (higher HU values in the lungs) and 400 HU (lower HU values in the bones), which was optimal to analyze soft tissues (20 to 40) and fat (-70 to -30) [180]. A 5 CNNs ensemble was trained using nnUNet's default training pipeline (one model per fold, 5 folds cross-validation). However, instead of using only the lesions as segmentation targets in the ground truth, some organs were added as additional segmentation classes, to improve training supervision.

Additional Organs Supervision This method was introduced in AutoPET II by team BAMF Health [181] and took first place in this edition (see below). Using TotalSegmentator, we extracted the contours for some organs from the CT scans (see figure 4.3). These organs were selected because they were deemed important in the context of PET/CT segmentation. This includes organs which exhibit physiological FDG uptake because of glycolysis (brain and heart), organs excreting waste, thus filling with FDG (kidneys and urinary bladder), organs filtering or conveying the blood pool (aorta, spleen and liver), organs expressing PSMA and thus exhibiting PSMA-targeting tracer uptake (digestive system and prostate), as well as some additional landmark organs (skeleton, lungs and pancreas).

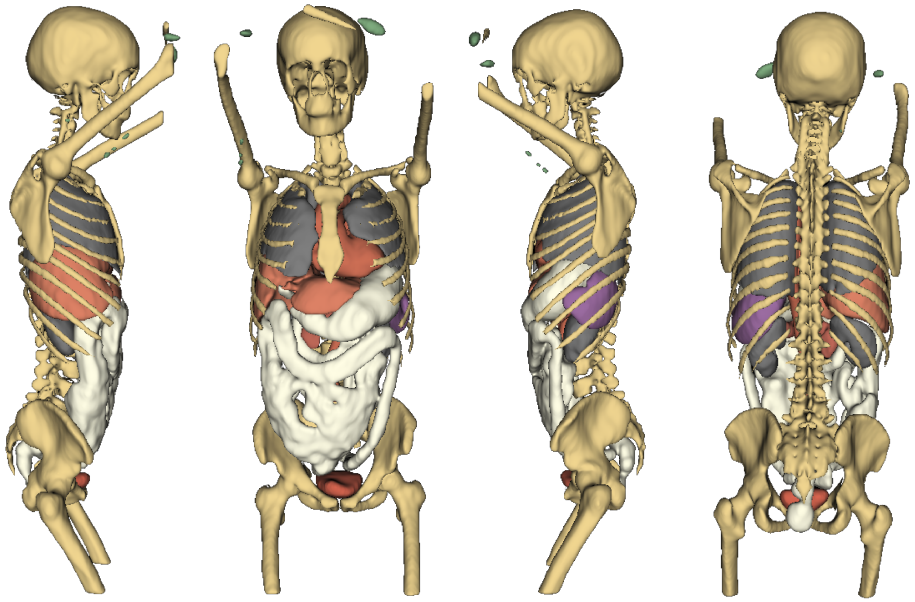

Figure 4.3: 3D rendering of the organ contours used as additional supervision. These contours were segmented on the CT of each patient from the training set, using TotalSegmentator

Adding these organs to our ground truth map, the goal was to teach the CNN to avoid classifying the voxels inside these organs as malignant lesions, which would reduce the amount of false positives in the segmentation, without needing any post-processing. Additionally, it was meant to teach the

CNN an informative representation of the human body, both in CT (anatomical knowledge) and PET scans (functional knowledge). Two CNNs, a PSMA-expert and a FDG-expert, were trained, each only on the data of their specific tracer. While this reduced the data available to train each network, our initial thought was that focusing on one PET tracer at a time would simplify the learning task.

Fully Automatic Segmentation Algorithm The pipeline participants had to build was required to take one PET and one CT volume and output the segmentation of the lesions, with the same spacing, origin, direction and size. Our pipeline would first load the images, then compute both the windowed versions, (PET_w and CT_w) as well as the PET MIP using SimpleITK[53]. The 2D MIP-CNN was then used to determine the PET tracer. Finally, the inference of the selected CNN was performed on the 4 channels [CT, PET, CT_w, PET_w] using nnUNet's nnUNetPredictor. Once the multi-class segmentation output was obtained, only the two first channels (0: background, 1: lesions) were kept since organ segmentations were not needed any further.

4.3.3 AutoPET IV

I participated in AutoPET IV alongside Léo Mottay and Solène Perret in 2025.

At the beginning of the fourth edition, we started building from the state-of-the-art, which was the winning solution of the AutoPETIII challenge[167]. They used a pre-trained ResidualEncoderUNet L, a larger variant of the UNet with a residual encoder, which was pre-trained on a diverse medical imaging dataset[182] using the Multi-Talent framework[183]. The weights of the pre-trained network were made publicly available, to be used as initial weights for further training and fine-tuning.

Using these weights for initialization, our solution integrated two fine-tuned expert models (PSMA and FDG). Building on top of our algorithm from the last edition, one of these tracer-expert models was chosen by a tracer discriminator (MIP-CNN), tasked with classifying the PET volumes as FDG or PSMA using a 2D MIP and routing the input to the correct expert U-Net.

Input Heatmaps The main addition to our algorithm was the inclusion of user input, in the form of positive and negative clicks. To process these clicks, organizers propose using gaussian "heatmaps", one for the lesion clicks and one for the background clicks, as additional 3D channels concatenated to the CNN's input. Although this boilerplate design is a good way of providing the information from the clicks to the network, we modified it to better fit the specific use case of PET/CT segmentation.

First, instead of two separated input maps, only one likelihood heatmap ($H_{likelihood}$) per case was generated, as shown in figure 4.4. This heatmap was filled with both negative values (for voxels more likely to be background) and positive values (for voxels more likely to be part of lesions). Having a joint representation allows us to better process cases which contain a positive and negative click close to one another. While only one heatmap was built, it was the result of the combination of two heatmaps: the lesion likelihood heatmap (H_{lesion}) and the background likelihood heatmap ($H_{background}$).

The lesion likelihood heatmap was built using the lesion *clicks* and the PET SUV values. It was initialized as a volume the same size as the PET, filled with zeroes, named H^0. At the coordinate of a lesion *click*, the voxel's value was set to one, in a new zero-filled volume H^0_{local} and

a gaussian filter (sigma=3, order=0) is applied to H^0_{local}. The PET volume is then used to restrain positive values in the heatmap to the voxels with a significant tracer uptake. This was defined as any voxel below an adaptive threshold defined as $min(4, SUV_{click})$. All voxels falling below this threshold in PET SUV values are set to 0 in H^0_{local}. Finally, H^0_{local} was added to H^0 and these steps are repeated for all lesion *clicks*.

This yielded a H_{lesion} heatmap where voxels close to one lesion *click* have a value greater than zero, but lower than the value of a voxel close to two lesion *clicks*, etc...) and voxels with SUV values too low to be considered as being tracer avid are not given importance (i.e.: they are set to zero).

The background likelihood map, was also initialized to a zero-filled volume, the same size as the PET. In this volume, all background *clicks* are set to 1 and a gaussian filter (sigma=6, order=0) was applied to the whole volume at once. This is because PET values were not used to adjust the background heatmap. The combination of H_{lesion} and $H_{background}$ occurs in two steps: first, all voxels which are non-zero in H_{lesion} are set to 0 in $H_{background}$. This was done to prevent the area of effect of background *clicks* to spill over the voxels labeled as likely being part of a lesion: lesion *clicks* take priority over background *clicks*. H_{lesion} and $H_{background}$ are then combined as $H_{likelihood} = H_{lesion} - H_{background}$.

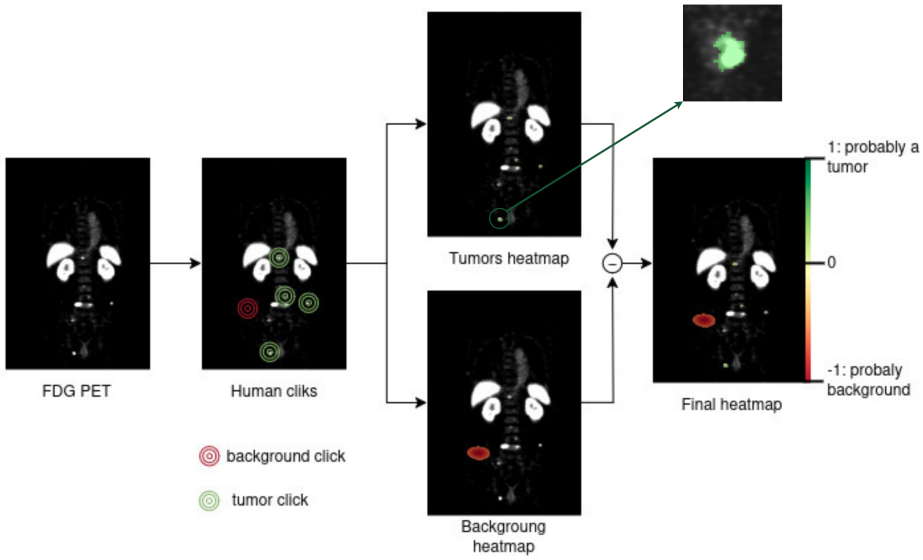

Figure 4.4: The heatmap creation algorithm. While the final heatmap is a single volume, the positive and negative clicks were processed separately. *Credit: Solène Perret.*

Training Hyperparameters Both models were trained as a 5 folds/models ensemble for 1500 epoch per fold with 350 iterations per epoch for the FDG model and 250 iterations per epoch for the PSMA model. Patch size was set to a cube of 192 voxels sides, for both models, as was done in the previous year's winning solution[167], consequently the batch size was set to 3 to fit in our GPUs memory (2*NVIDIA H100 (96 GB), used to train folds in parallel). Default data augmentations

were used during training, with the addition of an augmentation simulating a misalignment between PET and CT volumes[167].

4.3.3.1 Additional Organs Supervison

From the algorithm we built the previous year (see 4.3.2.3), we picked up the additional labels strategy, for additional landmarks and better representation learning. This time we included new organs as labels: the parotid and sub-mandibular glands (see figure 4.5), two organs exhibiting physiological PSMA tracer uptake. Parotids and sub-mandibular glands were grouped as one class each (no right/left separation, since nnUNet mirroring data augmentations were kept during training and at inference time, see 4.3.3.4).

Segmentation of organs was performed in a separate segmentation head (i.e.: a different last convolution) from the one performing lesion segmentation[167]. This allowed our model to predict an organ and a lesion for the same voxel, which is helpful for cases including organ metastasis (bone, liver, lung metastasis). In addition, inference is simpler this way: the convolution for organ segmentation is simply discarded and the output only consists of the binary lesion segmentation map.

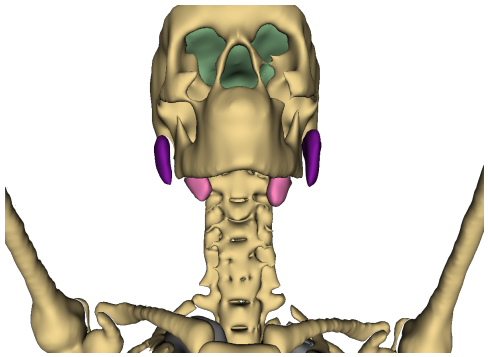

Figure 4.5: 3D rendering of the parotid glands (purple) and sub-mandibular glands (pink). These organs were added to the list of supervision organs, as they often exhibit an uptake of PSMA tracer.

Post-Processing To improve our algorithm's performance, a post-processing step was added to remove small, low uptake components from the lesions segmentation map. This was inspired by the method presented in AutoPET II by team Alloula [184], in which the authors showed that the nnUNet-based model they train had a tendency to output much more small components than are initially present in the dataset.

To counteract this effect, they determined a threshold of 10 voxels as the best solution to keep most of the non false positive single connected components. However, this threshold was determined on the FDG data only and the prostate cancer patients undergoing PSMA PET are more likely to have metastasis, which can be small and should not be missed. Considering this, a second condition was added before a component was removed: its SUVmax must be lower than 4, meaning it is a small component with low tracer uptake.

Dynamic Inference Depth When using patch-based inference, such as with nnUNet's predictor, the number of patches passed through the model to segment one case depends on the spatial size

of the PET/CT volumes. This size can vary importantly from a patient to another, which means different cases may have very different numbers of patches to process. To make use of as much computation time as possible for each case, without crossing the time limit for inference, mirroring axes were added or removed dynamically, depending on the total number of voxels in the input volume ($H * W * D$).

Experimenting with different sizes of volumes and taking into account the difference in performance between the GPU used for this test (NVIDIA Quadro P600, 24 GB) and the one used by the organizers to run submissions (Tesla T4, 16GB), the optimal cutoffs to add one, two and three mirroring axes were estimated to be respectively less than 90, 39 and 18 million voxels. The inference algorithm takes these thresholds into account when the nnUNetPredictor object is parameterized for mirroring axes: the number of voxels in the volume is computed and compared to these thresholds to set the number of mirroring axes dynamically for each case. This allows us to make the most of the given time while respecting the time limit.

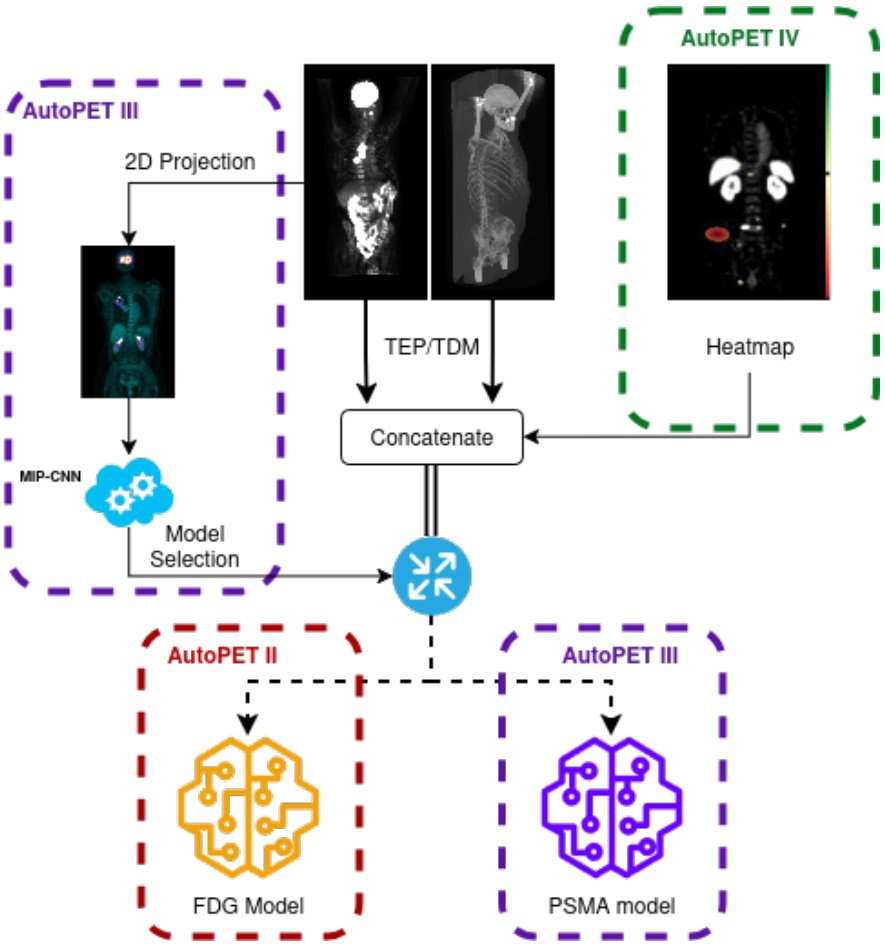

Figure 4.6: A summary of the novelties proposed and included in our methods. While each edition led to its own improvements, these improvements were iteratively integrated to build a solutions which benefited from all of them.

4.4 Discussion

4.4.1 The Winning Solutions from Each Year

2023 AutoPET II's "*best segmentation*" category was won by team BAMF Health[181], with their most notable contribution being the inclusion of additional organs (contoured using TotalSegmentator) as additional supervision during training. This had only a small effect on Dice score (3rd place) but caused a major decrease in the false positive volume (FPV) given them 1st place in this category, with half the FPV of the team in 2nd place overall, without increasing the false negative volume (FNV), in which they scored better than all the other teams in the top five. This solution was broadly adopted in the following AutoPET editions, with consistent success.

Their solution also took the palm in the "robust segmentation" category, where the leaderboard was roughly the same as the first category's, with a notable exception: team cv hci jumped from 5th to 3rd place. Their method consisted in using a large $(224 * 224 * 224)$ patch size during training then reducing it (to $128 * 128 * 128$) and using a sliding window during inference, similar to nnUNet but with an overlap of 75% between patches (eq. to $step_size = 0.25$). CT volumes were discarded entirely from their method.

Insights from other proposed solutions and exploratory analysis of the results can be read in the challenge report paper, published in 2025 by the challenge authors, in collaboration with the challenge participants[9].

2024 AutoPET III rewarded the LesionTracer[167] team for the "model-centric segmentation" category, in which additional public data and models could be used. Their team made full use of this authorization and gathered a large imaging dataset (containing MR, CT, PET volumes) to pre-train a ResidualEncoderUNet L (larger than the default architecture in nnUNet) using MultiTalent[183]. They then used the weights from this pre-trained CNN as initialization to fine-tune a five-models ensemble, capable of segmenting malignant lesions in both FDG and PSMA PET/CT volumes. A misalignment data augmentation method was also introduced in the same work, which brought a significant improvement and was used again in AutoPET. Like many teams, they also used organ contours from TotalSegmentator to improve supervision during training.

For the "data-centric" category, the best submission[185] (by team LennonLYChan) adapted the DiffTumor[186] framework to generate additional synthetic samples using diffusion. These synthetic data samples, along with regular data augmentation transforms, have greatly improved the performance of the CNN, suggesting the suitability of synthetic data to increase dataset size, as was already explored for the segmentation of other modalities, with dataset of similar sizes to AutoPET[27].

2025 AutoPET IV's semi-automated lesion PET/CT segmentation task was won by team BIRTH[187], who presented a comprehensive analysis of different strategies and selected the best performing combination. They trained two tracer-expert models, with different strategies (see below) and used a 2D classifier, extracting features from coronal and sagittal PET MIPs, forwarding both through a ResNet backbone and concatenating them before using a MLP to classify the PET as FDG or PSMA. Their PSMA-expert model was trained from[167] only on PSMA data and used a balanced stochastic click sampling strategy to enable the model to work properly both with few and lots of clicks. Their FDG expert was trained on all the AutoPET data, also using a balanced stochastic click sampling strategy for one model (V3) and full guidance (all clicks) for another model (V4).

The best model to use was then determined during inference, depending on the number of clicks available: V3 when 0-4 clicks are available and V4 when 5-10 clicks are available.

4.4.2 Insights from AutoPET

4.4.2.1 CNNs trained with nnUNet are the SoTA

As for the medical imaging segmentation landscape in general, nnUNet remains the best performing framework for neural network training. Residual encoder U-Nets are used by almost all the winning submissions (DynUNet was used by team cv:hci in AutPET II and team LennonLYChan in AutoPET III, which is directly inspired from nnUNet). No transformer-based, mamba-based or other CNN networks reached performance comparable to nnUNet in the AutoPET competitions. This correlates with the findings in a recent study[109] that nnUNet is still necessary as a baseline comparison and still consistently outperforms all other CNNs, especially in limited runtime settings.

4.4.2.2 Large and Relevant Pre-Training is Important

The AutoPET I challenge was won by team blackbean, who proposed STU-Net, an architectural variation of the U-Net, trained with nnUNet and pre-trained using the anatomical data from the TotalSegmentator dataset.

In AutoPET II, STU-Net earned the second position, behind the solution using organ supervision; while the algorithm which placed third[184] included 200 additional segmented PET/CT scans from the HECKTOR22 challenge dataset, which are taken from head and neck cancer patients. Although this additional data was not exactly pre-training, it improved the breadth of their model's domain knowledge and led to better results on the out-of-distribution test set.

AutoPET III's winning solution pre-trained a large residual encoder U-Net using MultiTalent[182] on a large, publicly available medical imaging dataset and made the trained weights for this model also publicly available. The difference between the scores obtained by this proposition and all the others in the same edition demonstrated the power of a large, well conducted model pre-training on relevant data to improve the segmentation performance after fine-tuning.

The winning solutions in AutoPET IV's PET/CT semi-automated lesion segmentation task all used these publicly-available weights as an initialization of their networks' weights, for the automated part of their method. This extended the demonstration of how important proper initialization is for neural networks[188], in particular for medical image segmentation.

4.4.2.3 Semi-Automated Segmentation is Taking Off

AutoPET IV showed how human supervision could be integrated in existing state-of-the-art segmentation methods to improve performance, using lesion likelihood heatmaps. Our team was the only one to build heatmaps using the PET values to set a prior, rather than only using a gaussian. Fusing the background clicks and lesion clicks influence maps also prevented the influence of background clicks from spilling over voxels likely belonging to tumors. We did not however use a different number of clicks (sampling) to account for the cases in which less clicks are presented, which we hypothesize caused the lower performance of our model. Combining a custom heatmap-building algorithm which integrates prior knowledge, such as ours, with a clicks sampling strategy and the

improvements presented in the previous editions should yield a state of the art semi-automated PET/CT segmentation tool.

4.4.2.4 Global improvements for PET/CT lesion segmentation

One of the rules for the AutoPET challenges was that participants had to include a short paper describing their method, as well as a link to their code and models, which must be made publicly available. When combined with the rule stating that only publicly available data and/or models can be used in the design of the methods, this makes methods presented in the challenge easily reproducible. It is therefore common for the methods presented in a challenge to be built on top, or using parts of the methods presented in the previous editions. This is how iterative research works: learning how the best methods work, reproducing the state of the art and improving from it. Organizers introduced a new challenge each year, tackling a specific aspect of PET/CT segmentation (robustness to out of domain data, multiple PET tracers, semi-automated segmentation using human inputs). New context is added and pushes innovation from participants, to adjust existing methods to new segmentation setups.

Real world data is still more diverse, with more tracers (68Ga-DOTA-peptides[189], 68Ga-FAPI[190], 18F-Misonidazole[191]), more out-of-distribution cases (physiological uptake of tracer in rare areas[192]) and more technical specificities (other constructors, 4D Gated PET/CT[193], dynamic PET[194]). But the will to progressively include specific aspects of PET/CT is interesting not only to the deep learning community, but to nuclear medicine physicians, both in research in clinical practice, looking to automate part of their PET/CT segmentation workloads.

4.5 Conclusion

This chapter went over the improvements of the state-of-the-art automated PET/CT segmentation algorithms in the last four years.

The AutoPET challenges have demonstrated the efficiency of publicly releasing datasets to motivate the medical imaging community to improve the state-of-the-art. The challenges proposed ranged from simple semantic segmentation, to use cases more specific to PET/CT segmentation, such as multi-tracer segmentation and semi-automated segmentation with user input. Year after year, organizers included new data and tasks, not only for the sake of obtaining better DL models, but to solve real-world adjacent tasks, which reflect the problems from clinical practice.

Beyond the large, more complete datasets produced and published by the organizers, trained and pre-trained model weights were released to the public, particularly, the pre-trained weights for the ResidualEncoderUNetL. The publication of these objects both greatly advanced and made accessible the SoTA for PET/CT segmentation to other researchers. This is thanks to the rules for participation in AutoPET which included a requirement to publish the algorithm, model weights and any additional data used to train the model, along with a short paper describing the method. Sharing model weights has been a driver of innovation and improvements since it allowed participants to shift their focus from DNN training towards innovative inference algorithms. This has also allowed the improvements in the DL process (organs supervision, large pre-training) to be carried on from one edition to another and compound over the years.

Through three years of participation in these challenges, we have not only witnessed, but also

participated in the improvement to PET/CT segmentation in these challenges. The main technical bricks we provided were the automatic tracer-expert model selection with a two-dimensional MIP-CNN, for fully automated segmentation and a PET-aware heatmap construction algorithm to better process human input in semi-automated segmentation.

In 2025 the number of participants in AutoPET greatly decreased. According to the organizers, this could have been caused by the presence of other PET/CT segmentation challenges (DEEP-PSMA, HECKTOR 2025) this year, along with the increasing number and complexity of the proposed tasks in these challenges. Discussions are currently under way to group similar challenges and tasks together, which would centralize the efforts of the research community and, as a side effect, could also bring datasets together, to obtain larger, more diverse datasets.

Open challenges such as AutoPET have lowered the accessibility barrier in medical image segmentation, by ensuring the data, DNN weights and algorithms are openly available. While this progress has occurred in research, the performance improvement with limited inference time is a great step towards clinical adoption.

There are two main ways this technology could be included in oncology workflows:

- Fully automated segmentation of PET/CT: once the images are obtained, automated segmentation of the lesions and subsequent computations of prognostic values (TMTV, TLG) could be performed and communicated to the referring physician.
- Semi-automated segmentation in radiotherapy treatment planning: Human operators could be helped to contour cancerous lesions more rapidly, by providing input which would be propagated by the algorithm.

While extensive deployment and validation studies are still needed before PET/CT segmentation tools could be deployed, the technology to make them a reality is becoming more mature every year.

Chapter 5

Radiomics for survival prediction using PET/CT: state of the art and improved reliability using data augmentations

Chapter Overview

Outside of radiotherapy planning, lesion contours obtained from PET/CT imaging are often used to compute lesion descriptors. These descriptors are then used for tasks such as patient classification and survival prediction.

In this chapter, we discuss radiomic features, a high dimensional class of descriptors, how they are obtained from medical images and how they can be used in machine learning to predict survival and patient response to treatment. Additionally, we present how we applied data augmentation to radiomics-based models to improve the performance of ML models on the task of survival prediction.

Contents

5.1 Radiomics: high-dimensional medical imaging descriptors 73
5.1.1 Medical imaging descriptors 73
5.1.2 Radiomics: high dimensional descriptors 73
5.1.3 Extracting Radiomics 74
5.2 Survival Prediction using Radiomics 75
5.2.1 Feature Selection 76
5.2.2 Machine learning for survival prediction 76
5.2.3 Metrics for Survival Prediction Evaluation 77
5.3 Radiomics Augmentation against overfitting 80
5.3.1 Context 80
5.3.2 Material and methods 81
5.3.3 Results 87
5.3.4 Discussion 90
5.3.5 Conclusion 91

5.1 Radiomics: high-dimensional medical imaging descriptors

5.1.1 Medical imaging descriptors

Once medical images are processed and the lesion contours are obtained, descriptors can be computed to extract information on the image, from inside these contours. In nuclear medicine, PET descriptors, sometimes as simple as the metabolic tumor volume (MTV), SUV_{max} value or SUV_{mean} value, have been shown to be clinically relevant to patient prognosis[195],[196]. Combined descriptors such as total metabolic tumor volume (TMTV), the product of MTV and SUV_{mean}, also showed interesting performance in identifying high risk patients, for instance in high tumoral burden diseases such as diffuse B-cell lymphoma[197].

However, even such simple descriptors can be sensitive to variations in both the image texture and in the contour shapes. These variations originate from multiple sources and are hard to completely control or eliminate.

Variations in the image can be caused by the acquisition being performed on different machines, in different imaging sites[198], with different reconstruction parameters[199], even because of the variability which exists in a patient's metabolism[200]. This variability can cause some descriptors, such as texture descriptors to be considered not robust enough to be reliably used for patient prognosis.

Variations in contours are caused by inter-observer variability, a well-documented phenomenon where manual contours obtained from two or more observers, even experts, rarely completely overlap. This problem, which was discussed earlier for organ contouring on CT scans, also exists for lesion contouring on PET scans[201], although some methods relying on the specificities of PET scans can be used to limit this variability. The most popular of these methods is semi automated contouring, using a manually defined VOI and SUV thresholding[202].

5.1.2 Radiomics: high dimensional descriptors

To take advantage of recent, more complex models, researchers started producing more and more descriptors, extracting shape and texture information from the VOI in the image. These descriptors were named radiomics[203], or radiomic features.

There are different feature types, which can mostly be sorted in the following categories:

- **Simple descriptors**, such as mean value, quartiles, or features from the first order histogram of the image. They are similar to the original, older descriptors described earlier.

- **Shape descriptors**, such as the volume, surface or cross-sectional diameter of the contoured structures. These features describe the geometry of a contour and, although some shape descriptors were also used before the term radiomics was coined, most of them were volume-related features.

- **Texture descriptors**, which describe texture as the relationships between the values of neighboring pixels. These features can describe more complex patterns present in the data, however their complexity makes them less intuitively understandable.

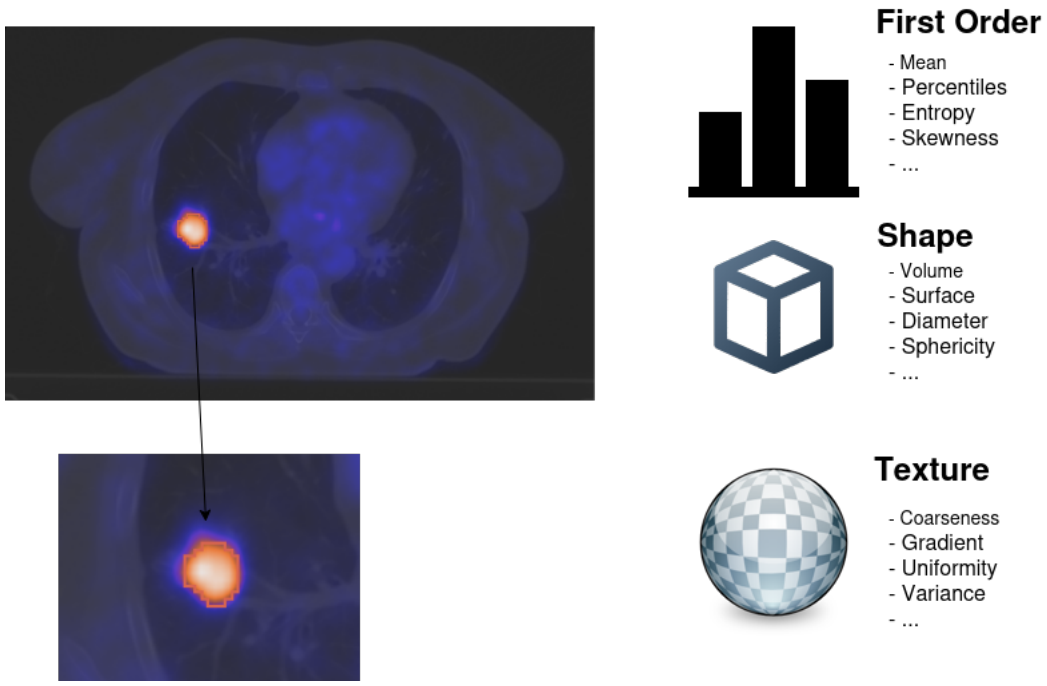

Figure 5.1: Radiomic features are computed using the values of the medical images inside contours. Different classes of radiomics exist, each measuring a specific aspect of the contour and/or the image

Some radiomics have been proposed, which do not fit in any of these categories, such as distances between lesions[204] and the number of lesions[205]. While they have shown promising result, they have not been as widely used and evaluated as the radiomics from the three categories presented above.

Even though lots of radiomic features can be computed for a given image/VOI pair, radiomics remain a dimensionality reduction technique, as it extracts only a limited amount of features from images containing a large amount of voxels. Descriptors also increase the robustness and semantics of the information, as it elevates it from the voxel.

This shift is helpful in any downstream tasks, as performing the analysis of all the voxels from a 3D image takes a lot of time and requires deep, heavy models, which would in turn require large datasets for training. On top of that, some voxels have more importance than other, while other voxels have zero interest (voxels outside of the patient's body, padding voxels). Using the contours, attention is focused on the information available inside the parts of the image which have been deemed important: in many cases, the lesions.

Radiomic features have been used in disease classification[206], patient survival[207][208] and treatment response prediction[209].

5.1.3 Extracting Radiomics

In an effort to make radiomic features accessible to the medical imaging community, then to build a standard for reproducible radiomics definitions and computation, PyRadiomics[210] was published in 2017 as a python package. It has offered a way to automatically extract a large amount of radiomic

features, without the need to re-define these features every time.

To this day, PyRadiomics has been the most used radiomics extraction framework, with 6700 citations of its paper, 1300 stars[211] on its GitHub repository and hundreds of daily downloads[212] through PyPi, python's package manager. PyRadiomics computes 107 descriptors in its default configuration, which can be modified to exclude some features, or some feature types altogether. Additional descriptors, which are excluded by default, can still be obtained by forcing PyRadiomics to include them in its computations. Reasons for a feature to be excluded from the default configuration can be: it is equal or proportional to an existing included feature, it is always equal to the same value or it is part of the group of shape features computed in 2D. Along with the descriptors, PyRadiomics includes pre-processing tools, some of which were developed to help with feature standardization (IBSI, see below). Other pre-processing tools can be used to apply filters on the input images, such as Laplacian of Gaussian, wavelet transform, arithmetic modifications, gradient or local binary pattern. These filters can be enabled by the user, in order to make some specific parts of the image stand out.

In 2020, researchers from multiple labs worked together to propose IBSI[213],[214]: Imaging Biomarker Standardization Initiative. This paper lays out definitions and rules to harmonize practices in radiomics computation, across different studies, teams and software (LifeX[215], MITK[140], PyRadiomics[210]). A manual for reproducible radiomics studies was issued, with recommendations such as resampling the images to an isotropic voxel spacing, using trilinear interpolation. PyRadiomics participated in this standardization initiative and some of the recommendations were included in its code and documentation.

As mentioned above, other radiomics extraction software exist, although less popular than PyRadiomics. LifeX, MITK, CERR[216], MIRP[217] or RaCat[218] are examples of frameworks which include a radiomics extraction pipeline. Some of these frameworks were developed and are maintained by universities or research labs, which makes initiatives such as IBSI necessary to keep radiomics values coherent from one framework to another. However, as shown in the IBSI report paper, no framework has implemented exactly the same set of radiomics, which leaves a difference in the information contained in the radiomics vector, depending on the framework which computed it.

5.2 Survival Prediction using Radiomics

Medical imaging is used by physicians to choose the best course of action to treat a patient, whether it is further surveillance or a treatment. In the latter case, the treatment delivered to a patient can also be adapted based on imaging. To support physicians in their decision, image analysis through a ML model is a promising research domain. Learning from large cohorts of patients and looking at subtle variations in imaging data, these models offer a complementary, quantitative insight to the physician's qualitative assessment of the imaging. Radiomics have been used in numerous research studies and have demonstrated their ability to at least partially differentiate between patients, based on their future response to treatment, or the risk of death or relapse. In this section, a basic framework to use radiomics ML methods for survival prediction will be presented.

5.2.1 Feature Selection

Once radiomic features were extracted as described above, the resulting vector often remained highly dimensional. The first step in any radiomics study should be feature selection (FS), i.e.: selecting the variables carrying the most information, while discarding the remaining ones.

Radiomic features often exhibit inter-feature correlation, which means that redundancy exists between features. A first pass reducing redundancy in features is recommended[219], which can consist of either deleting of certain features which are equal or highly correlated (e.g.: spearman correlation coefficient >0.90) to another feature , which will be kept in the set, or creating meta-features, which are a combination of multiple features[220] (e.g. through principal component analysis) as one. Though the latter does reduce dimensionality while keeping part of the information from the features, it makes the downstream ML model harder to interpret, as the exact meaning of the meta-feature can be impossible to obtain.

Once redundant features are dealt with, further dimensionality reduction can be achieved by selecting the most important features. These can either be the ones with the highest inter-patient variability (with information allowing to differentiate patients), or the ones correlating the most with the target to predict[219]. When this target is available entirely (e.g.: no censoring), correlation coefficients such as Spearman's[221] or Pearson's[222] can be used.

If part of the target is missing (e.g.: right censored data, very common in survival analysis), univariate models can be trained, with one feature at a time, to assign an importance score to each feature, based on how well they solve the problem on their own. This method has been used in multiple research papers, often as a baseline to compare with new methods[223],[224]. Although these new methods improved the downstream model's performance slightly, univariate FS methods showed stable performance while it used lower computation time[223].

It is interesting to note that the implementations of FS algorithms for survival data presented research papers were sometimes found in R[225] packages, but were absent from Python. This shows that survival prediction is often only considered, in litterature from a statistical standpoint, rather than from a machine learning one.

5.2.2 Machine learning for survival prediction

Most ML models, such as support vector machines[226] (SVM), logistic regressions[227], decision trees[228] or random forests[229] are built to solve classification tasks. Survival tasks are therefore often reduced to classification tasks[230],[231] (e.g.: overall survival becomes 6-months survival, with a *survival/death* binary classification task). Using this shift in target definition, regular ML models can be used directly for survival tasks. However, this causes precision to be reduced, because of the shift from a continuous to a discrete, broader target. Patients can thus only be stratified in classes and, once the time used as discretization threshold is passed, the model has no additional information to learn from.

To keep the continuous nature of the survival information, the next logical choice is to use regression models. These models use a continuous value as target and output, which allows to model variables such as time-to-event (death, relapse, disease progression) or risk of an event happening.

Ths issue with regression models is that they do not natively handle an important aspect of survival data: **censoring**. Indeed, medical datasets are often gathered for a limited amount of time, after which the database is frozen and the patients who had not experienced the monitored

event are right-censored. The value obtained for time-to-event for these patients is then not the real value, but a lower bound: for example, a study with monitoring extended to three years will set the survival times of all patients alive after three years to three years, but these patients will be marked as censored. This is called type I censoring (see figure 5.2).

Sometimes, contact might be lost with a patient, without them being dropped from the study. In this case, the patient is marked as censored and time-to-event is set to the time this patient was last seen alive. This more loose type of censoring lets the investigators keep more patients, but makes the censoring non-informative for these patients, as the censoring time is independent from both time-to-event and the study duration. This in turn modifies the assumptions underlying models trained on such data[232].

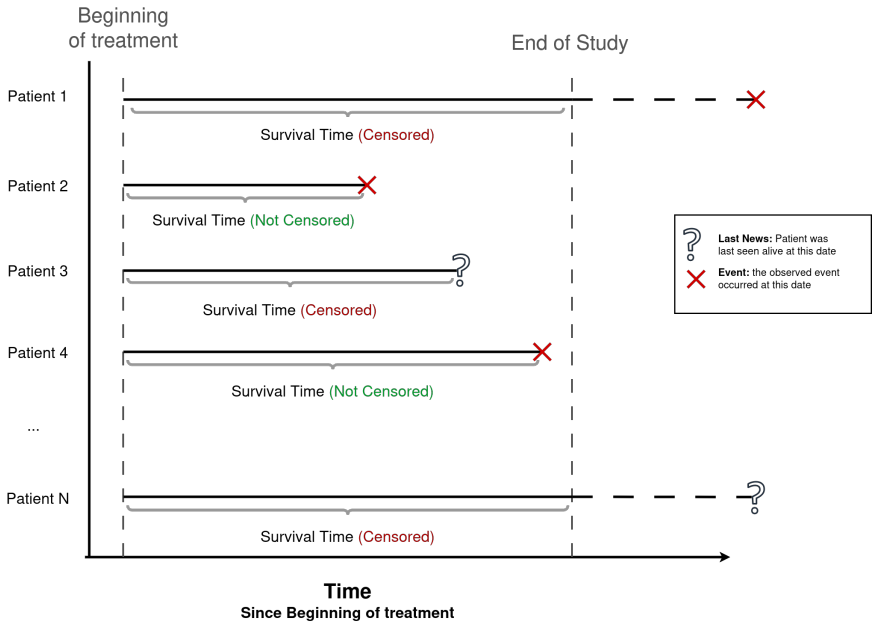

Figure 5.2: Survival is defined as the time (often in days) between the beginning of a treatment and the occurrence of an event (death, relapse, progression). A patient survival can be censored if they did not undergo an event before the end of the study, or if they dropped out of the study. In these cases, the patient is marked as *censored* and the survival time is registered as the time to end of study or the time to last news, respectively.

To handle data censoring, likelihood functions are adapted to include only the cases with an event happening through the probability density function, while the censored cases are included directly through the survival function, at the time of their censoring.

5.2.3 Metrics for Survival Prediction Evaluation

To evaluate the performance of survival prediction models, a suitable metric should be used. Survival prediction handles time-to-event values, thus categorical metrics cannot be directly used to evaluate models and metrics should be able to handle continuous values.

Different types of performance can be evaluated, depending on the model and its output. Ranking

metrics evaluate the capacity of the model to rank a set of cases based in their risk or survival times; they are crucial for models without a regression, which assign ranks to patients based on an output with arbitrary unit. Accuracy metrics on the other hand, should be used on models which provide an actual time-to-event prediction (e.g.: models including a regression), since they compare the outcome with the predicted time-to-event value. Below, one metric is presented for each of these categories.

5.2.3.1 The Concordance Indices

Concordance indices are metrics which evaluate how well patients are ranked by their time of survival. Harrell's C-Index[65] has been shown to be an extension of the receiver operating curve (ROC) for time-dependent data[233]. This metric compares, for each pair of patients , the relative rank of their survival time and of their predicted survival as obtained from the model. To take into account data censoring, the data pairs for patient i and j are considered concordant if $y_i < y_j$ and $\hat{y}_i < \hat{y}_j$ or if $y_i > y_j$ and $\hat{y}_i > \hat{y}_j$ and discordant only if $y_i < y_j$ and $\hat{y}_i > \hat{y}_j$, if patient i's survival is not censored or if $y_i > y_j$ and $\hat{y}_i < \hat{y}_j$, if patient j's survival not censored. Harrell's C-Index is then computed, for a set of cases processed by the model, with the formula:
$CI = \frac{nb_{concordant}}{nb_{concordant}+nb_{discordant}}$.

As it is a rank-based metric, the C-Index can only be computed for multiple patients, with a growing robustness as the number of cases increase. Different versions of the C-Index have been proposed, to account for censoring and ties in patient survival times. Using inverse proportional censoring weights (IPCW), such as in Uno's C-Index[234], gives more weight to non-censored samples for which samples with a similar recorded time-to-event are censored. Ties can also be treated in different ways as described in [235], either considering samples with tied outcomes or tied predictions as non-comparable, completely comparable (for tied outcomes, with one of the samples uncensored) or partially comparable, with an assigned weight lower than one for example.

This metric is the most used to evaluate survival prediction models[236], although it is limited since it is not measuring the accuracy of the time-to-event predictions, only their order and does not assess the model's calibration at all. Additionally, non-IPCW variants, like Harrell's C-Index, are often used, because they remain the most popular. This can lead to overly optimistic C-Index values[237], as samples with time-to-event closer to many censored samples will be drowned in the more numerous uncensored samples at more frequent times, which are more easily learnt by the network.

5.2.3.2 The Brier Score

Brier score[238] (or Brier Index) consists of the mean squared difference between the distribution of outcomes and the distribution of the predicted survival times. This squared difference is computed, in the original version of the Brier score, at a given set of time points, which makes it dependent on these chosen evaluation times. To solve this issue, the integrated Brier score (IBS) theoretically considers the integral of the Brier score, which means the Brier score at all time points in the study. In real-world experiments, IBS consists of computing the Brier score for all available points.

Both metrics measure the accuracy of a model, by computing this non-IPCW by defaulte differences between observed and predicted survival times. Because of this, they can only be used to evaluate models which output a survival time and not an arbitrary unit (such as a risk score). If the

number of patients in the test set is large enough to only consider the uncensored patients, mean squared error (MSE) and mean absolute error (MAE) can be used to assess the accuracy of survival models.

The accuracies measured by Brier, IBS, MSE or MAE represent how well a model estimates the time-to-event for a patient, but is tied to the distribution of events in a given dataset, thus both Brier and IBS are best suited to compare the performance of multiple models on the same dataset.

5.3 Radiomics Augmentation: Using data augmentations to combat overfitting in radiomics studies

5.3.1 Context

5.3.1.1 Overfitting and data augmentations

Medical imaging datasets have often been limited in size, which has in turn limited the performance of ML methods when applied to specific medical tasks[239]. In particular, models trained with a limited amount of data have been shown to be subject to overfitting[240], a phenomenon in which a model performs well on its training data, but fails to generalize on new, unseen testing data.

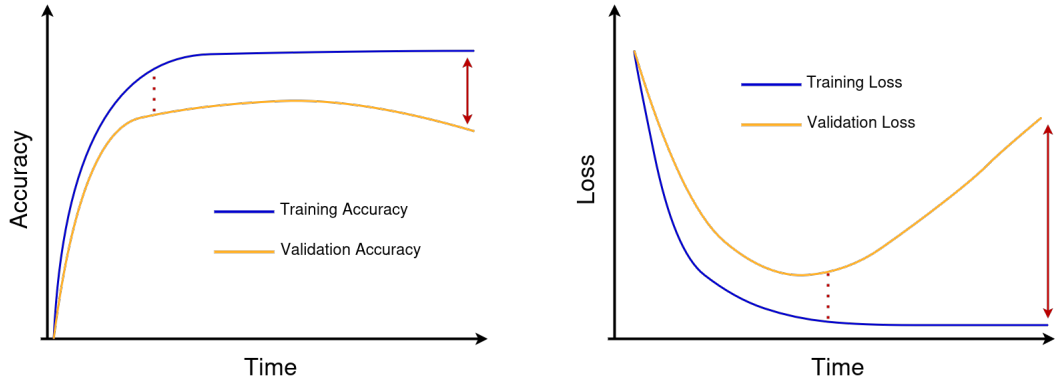

Figure 5.3: A toy demonstration of what happens when a model over-fits. Over the course of its training, it will first start improving on both its training and validation set (loss decreases, accuracy increases). After some time, performance on the training data will keep on improving, while the progress on the validation set will start to slow down, stop and eventually revert. This is a clear sign of overfitting, quantified by the difference between training and validation accuracies (red arrow). In the figure above, the moment the model starts overfitting is marked by a red dotted line.

Additionally, the number of input features submitted to the ML model can sometimes be larger than the number of samples in the dataset, which creates a situation with more unknowns (the importance of each radiomic feature) than equations (the set of features and the endpoint for each sample/patient case in the dataset). Deep learning models were always sensitive to overfitting when trained on small datasets, due to the large number of trainable parameters in their architecture. This problem was mitigated in part thanks to the use of data augmentations (DA), which are transforms applied to the data to create slightly different samples, which act as new samples. These DA should be realistic and represent the variabilities found in the real distribution of the data.

Using data augmentations thus enables the model to learn from data better fitting the distribution found in the real world, while decreasing the importance of specificities found only in the training data, which helps the model to generalize, increasing its performance when presented with data it has never seen before. This method is as old as deep learning, as the inclusion of distorted images were shown to lower the test error for digit reconstruction using LeNet[15]. Different DA are available for different types of data[241] and new data augmentations are created to fit specific use cases, such as medical imaging[242], autonomous driving[243] and speech recognition[244]. Generic DA include the

addition of noise (usually gaussian noise, but can be other types of noise for specific data types, such as poisson noise for PET sinograms[245]), geometric transformations, such as translations rotations and scaling, on one or multiple axes, values modifications (contrast modifications, exponential, clamping). DA based on geometric transformations can be applied to all of the input modalities, or only to some of them, to simulate mis-alignement[167].

5.3.1.2 Data augmentations and radiomics

In the field of radiomic features, DA were mostly applied directly on radiomics, after they were computed and stored as tabular data. For instance, the synthetic minority over-sampling technique (SMOTE)[246] and its high dimensional variant[247] select samples from minority classes with a higher probability, to deal with datasets containing class imbalance.

Other methods were presented to generate synthetic radiomic data, using auto-encoders[248], generative adversarial networks[249] (GAN) and conditional tabular GAN[250] (CTGAN). DA have also been applied to images, either to measure a feature's sensitivity to perturbation for feature selection[251] or to improve robustness in data-scarce regimes, for instance when training a model for Gleason score classification in prostate cancer[252].

We can also cite the winning submission[205] in the HECKTOR22[253] challenge, where the authors used image and mask transformations, although not as DA, but to create new features, instead of new samples. While this looks similar to DA, this is not DA and instead increases the amount of features, while the number of samples stays the same.

In this study, we present the Radiomics Augmentation framework, in which DA are used on images and segmentation masks before radiomic features are computed, to measure whether they lead to increased performance and reduced overfitting. To this end, a ML benchmark was built, to evaluate eight different models on the task of RFS prediction, with and without radiomics augmentations.

5.3.2 Material and methods

5.3.2.1 The HECKTOR 2022 Dataset

The dataset provided for the second task of the HECKTOR22 challenge[253] was used in this study. In this dataset, PET/CT scans for 489 patients suffering from head and neck cancer and treated using radiotherapy were collected. For each of these patients, contours for the gross target volume (GTV) of the primary tumor (GTV_p) and affected lymph nodes (GTV_n) were delineated. This delineation was performed by a first annotator, then cross-checked by a second one, while both had contouring guidelines at their disposal.

Using this data, the first task proposed in the HECKTOR challenge was to design a fully-automated segmentation algorithm for both GTVs. For the purpose of our study, only the manual delineation were used, so as not to introduce biases through the implementation of a specific automated segmentation method.

The second task consisted of predicting the recurrence-free survival (RFS), which is the time (in days) between the end of the radiotherapy treatment and the patient's death or cancer relapse. In the original challenge task, a different set of data was kept hidden from participants, for which the tumor and lymph nodes contours were not available. This pushed participants to use an automated

segmentation model (trained for the first task) to obtain tumor and GTV_n contours.

To use only manual contours GTV_p and GTV_n when extracting radiomic features, only the training set was used in this study and the generalization performance is evaluated thanks to a cross validation. This choice is also motivated by experimental constraints: because the test set was private and allowed only three successful submissions per month, it was not possible to use it to evaluate multiple models and perform multiple runs for each model (for results stability purposes).

For each patient, the PET and CT scans are provided as 3D volumes, along with the ROI contours, all stored in compressed NIfTI format. While registration was not performed after the acquisition, the PET and CT volumes were assumed to be registered to one another, since they were acquired simultaneously. The contours were provided as multi-class integer masks (0: background, 1: GTV_p, 2: GTV_n). The dimensions of the mask volumes were the same as the dimensions of the PET volume.

In addition to the imaging data, clinical information was provided. This clinical info always included gender, age and whether the patient underwent chemotherapy. It sometimes also included the weight, performance status (i.e.: WHO status), HPV status of the patient, their alcohol and/or tobacco consumption (*yes/no*) and whether they had surgery for their cancer or not. Some of these elements were missing from some centers while others were missing for specific patients. In this study, clinical data was not included, as we focused on radiomic features and the effect of data augmentations.

RFS, the endpoint for our study, was given for each patient, as the number of days between radiotherapy treatment and the occurrence of an event. An event was described as *"locoregional failures and distant metastases"* or patient death. When no event was recorded for a patient, this patient was noted as **censored** and the RFS value was set to the time between radiotherapy and the last time that patient was seen (*time of last news*).

Models were benchmarked on the second task, which consisted in assigning a risk score to each patient, based on the imaging and clinical data available for this patient. This risk score did not have to be similar in magnitude, or inversely proportional to the time-to-event, but had to be able to correctly rank patients, from the most to the least likely to experience an event after N days.

5.3.2.2 Data augmentations

Generating new samples from slightly modified versions of the existing ones is called data augmentation. In our study, multiple data augmentations (DA) were chosen to be applied to the images (PET and CT) and to the ROI masks (GTV_p and GTV_n). A visualization of each of these data augmentation transforms is presented in figure 5.4.

DA applied to images were the following:

- **Blur:** a gaussian filter is applied on the image, with a kernel size (in millimeters) chosen from a uniform probability distribution between 0 and 4.
- **Noise:** adding a gaussian noise to the image, with mean of 0 and a standard deviation randomly sampled from a uniform probability distribution between -0.1 and 0.1.
- **Gamma:** changing the contrast of an image by raising its values to the power γ. The γ value is chosen for each image from a uniform probability distribution between $-\exp 0.9$ and $\exp 0.9$.

- **ContrastShift:** modifying the contrast of an image by normalizing the image using Z-Score normalization and multiplying the normalized image by a multiplier before multiplying the image back by its original standard deviation and adding its original mean. Two multiplier values are used: 0.9 and 1.1
- **MultiplicativeBrightness:** modifying the image values by multiplying them by a given number. Multipliers used in this study are 0.8 and 1.2.
- **SimulateLowResolution:** downsampling the image using nearest neighbor interpolation then upsampling it back to its original dimensions. This simulates a low quality image and can even introduce artefacts due to using BSpline interpolation during upsampling. A downsampling factor of 0.8 is used for all dimensions.

DA were also applied on the GTV_p and GTV_n contours, each treated separately as a binary mask. The DA used on masks were:

- **Dilation:** applying a dilation operation on the mask, with a kernel size of 4mm.
- **SUVRelativeThreshold:** re-segmenting the contour using the PET scan's SUV values. Contours are defined using a threshold set to 40% of the SUVmax value (maximum SUV value inside the contour).
- **SUVAbsoluteThreshold:** re-segmenting the contours using the PET SUV values but with an absolute threshold. Two thresholds are chosen: 2.5 and 4.

Some of these DA were inspired from the ones implemented in the default trainer class of nnUNet[82], while others were imported from the TorchIO library[76] (see *2.3.4.3*). In total, 8 image transforms and 4 mask transforms were used in this work, which meant that, including the non-modified versions of the images and masks, there were $9 * 5 = 45$ times more samples in the augmented dataset than in the original one. For the RFS prediction task in HECKTOR,**22 005 samples** (including 21516 augmented samples) were obtained from the original 489 patients.

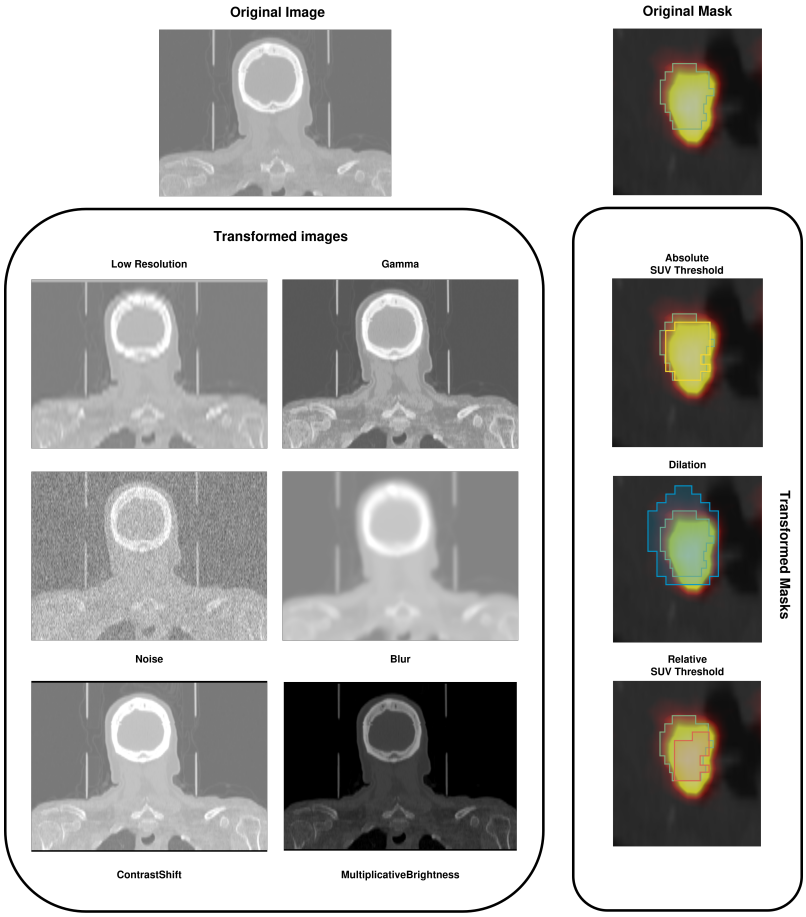

Figure 5.4: Visual representation of the transforms applied on the images and masks. The green contour is the original manual contour, while the yellow, blue and red are the contours modified by DA methods. The parameters of all the image transformations were set to extreme values for visualization purposes.

5.3.2.3 Visualizing the impact of augmentations on data distribution

Our first experiment with this data was to visualize how the distribution of augmented samples compares to the original distribution.

The goal was to qualitatively evaluate how DA affected the distribution, most notably if a distribution shift was visible, or whether the magnitude of the modifications was large enough to bury the information with noise.

To obtain such a visualization, the extracted radiomics vectors were projected in a two-dimensional space using uniform manifold approximation projection[254] (UMAP). UMAP is a recent, state-of-the-art algorithm for dimensionality reduction, which is faster than its similar counterparts (t-SNE, PCA) thanks to a non-quadratic complexity of $n*\log(n)$[255]. UMAP has also shown better structure conservation than its conterpart, thanks to a repulsion parameter which keeps points corresponding to distant n-dimensional vectors far from each other in the low-dimensional projection.

The projected 2-dimensional vectors were then displayed using the Matplotlib[256] Python library, with a color code representing the RFS value in days. Original samples were represented using

points with a larger diameter than the points representing augmented samples. This representation gave us an intuitive look into the effects of the DAs on the distribution of the data and figure out if the augmentation process filled the gaps in the data distribution, thus bringing it closer to real-world data, or if the addition of the augmented samples to the dataset was irrelevant, or caused too much shift from the original distribution.

5.3.2.4 Radiomics

107 radiomics descriptors were used in this study and were extracted using the PyRadiomics[210] Python library. These radiomics included 18 first-order features, 14 3-dimensional shape descriptors and 75 texture-based features. Radiomics were obtained separately from GTV_p and GTV_n masks, as well as from PET and CT images, for a total of 4 image/mask pairs, thus 428 radiomic features per sample.

To reduce dimensionality and computation time, radiomics pairs which values were equal for all samples were un-duplicated. This included for instance some shape descriptors which were extracted multiple time for the same mask, but different images. Once these features were removed, 221 unique radiomic features were left for each sample. These features were then used to train the machine learning models included in our benchmark.

5.3.2.5 Machine Learning

Models To ensure that our method works independently from the model used, a collection of models and different parameters settings were integrated to the benchmark experiment and trained on the radiomic features described above.

- Fast Survival SVM[60] (FS-SVM): a SVM built for survival prediction, fast to train and demonstrating good performance. This model was trained with default parameters ($\alpha = 1$, $max_iter = 20$).
- CoxNet Survival Analysis[61] (CoxSA): A combination of a popular survival model, Cox proportional hazards, with a very basic regression model, ElasticNet (L_1 and L_2) regularization.
- Bagged Icare[205]: an ensemble model which bootstraps binary weighted estimators. This model has been used in the winning solution of task 2 during thr HECKTOR22 challenge. The authors claimed that using binary weighted estimators reduces overfitting. A notable difference with their method was that no parameters sets were defined *a priori*. This model was benchmarked with 10 and 100 estimators.
- Random Survival Forest (RSF): an adaptation of random forests[257] for survival data. Two configurations of this model were benchmarked: RSF with 10 and 100 estimators respectively.
- Gradient Boosting Survival Analysis[64] (GB-SA): an ensemble model relying on an iterative process to learn multiple, weak classifiers and adapting the weight of the training samples for the nth classifier depending on the errors of the previous $n-1^{th}$ classifier. Again, two versions of this model, with 10 and 100 estimators, were benchmarked.

For all of these models, no regression task was learned, which meant the models output a risk score with an arbitrary unit. Each model was only trained to rank patients by their risk score,

which is inversely related to the survival time. Consequently, only rank-based metrics were suited to evaluate these models.

Feature Selection For each run, the best model was selected using an exhaustive univariate C-Index (UCI) feature selection. This selection consisted in **1)** ranking each variable by importance, through the fitting and evaluation a CoxNet model using a single variable as input **2)** Running the subsequent training and evaluation of the multivariate models, with each number of variables k (from one to all variables) **3)** Selecting the k_{best} most important variables for each model, (i.e.: the model using k_{best} variables had obtained the highest value in the CI_{train} metric).

Using this method for the model selection, while keeping the same metric for evaluation, we voluntarily steered our benchmark towards the selection of the models which tend to overfit the most on their training data.

Evaluation Harrell's Concordance Index was also used to evaluate each model's performance. This metric is a rank based metric, described as the number of pairs of samples ranked correctly. This means: $C_{Index} = \frac{\sum_{i,j}^{N} R}{C(N,2)}$ with $R = 1$ if $y^i_{true} > y^j_{true}$ and $y^i_{pred} > y^j_{pred}$, for any two patients i and j, else $R = 0$; N being the number of patients and $C(N, 2)$ the number of pairs without repetition ($i \neq j$). When a sample had its RFS censored, it was only compared with samples having a shorter RFS than itself, as it was only in this scenario that one can be certain it should be assigned a smaller risk factor than these samples.

In order to obtain results with greater numerical stability, a 10-repeat, 5-fold cross-validation (CV) was designed to evaluate the performance of each model and compare the performance of basic and augmented radiomics. Each fold was used once for evaluation, and four times, in conjunction with three other folds, for training. The C_{Index} was averaged over the 10 repeats of the CV, to obtain more robust CI_{train} and CI_{test} values.

During the cross-validation, regardless of whether basic or augmented radiomics were used during training, the fold used to test the model only contained the original, unmodified samples. This is so the models were evaluated on unseen, real data, not augmented data. Using Augmented samples during evaluation would have given an unfair advantage to the models also trained on augmented data.

5.3.2.6 Overfitting Visualization

Additionally to the impact on model performance for the best selected model (selection described in *Feature Selection*), our goal was also to study the impact of radiomics augmentation on overfitting. To this end, performance (CI_{train} and CI_{test}) was collected for all variable number thresholds. This data was then used to visualize the difference between CI_{train} and CI_{test} at each point, to determine whether the gap between training and test performance widens, tightens or stays the same.

A model would be considered as suffering from major overfitting when feature selection has resulted in the selection of a model with a CI_{test} significantly lower than the highest CI_{test} observed over all of the possible versions of the model.

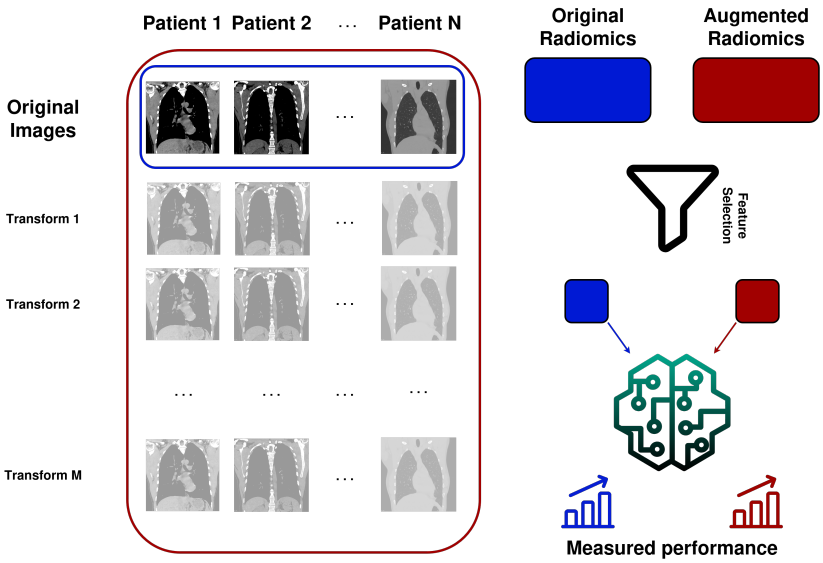

Figure 5.5: The *Radiomics Augmentation* Framework. Transformed are applied to masks and images to create augmented samples, which are added to the original samples. Augmented radiomics and original radiomics are then benchmarked, on the task of RFS prediction using ML models. *Masks are not represented here.*

5.3.3 Results

5.3.3.1 Data Distribution

The distribution of the augmented samples resembles the original distribution of data, with emphasis being added on some existing structures (red circles). In some places, augmented vectors filled space where no original vector was, suggesting these points either bridged a missing part of the distribution, or wandered out of distribution (green circles). Finally, structures which were empty in the original distribution were also mostly empty in the augmented distribution of augmented samples as well (black circles).

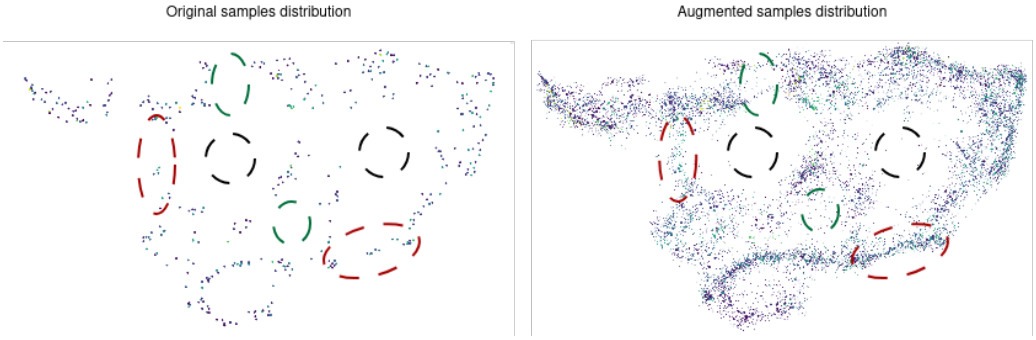

Figure 5.6: Results of the UMAP projection applied on the original and augmented radiomic features from the HECKTOR22 dataset. In the picture on the left, the 2D projections of radiomics vectors for all the original cases are represented. On the right, the projection for the augmented cases are also represented. Color coding of points with RFS is mainly used to enhance readability.

These observations suggest that DA had not shifted the data significantly out-of-distribution and had instead generated samples which completed the original distribution.

5.3.3.2 Impact on Model Performance

For 7 out of 8 models, a better prediction of RFS was obtained when augmented radiomics were included in the training set. The only exception was the GB-SA model with 10 estimators. Performance improvements ranged from **+0.2%** for the smallest Random Forest, to **+8.6%** for the largest Bagged Icare. The non-ensemble models, FS-SVM and CoxSA both had their performance increased using augmented samples, by 0.8% and 2.5% respectively.

Model	Basic Radiomics	Augmented Radiomics	Performance Change
FS-SVM	0.571	**0.576**	+0.8%
CoxSA	0.580	**0.595**	+2.5%
Bagged Icare (10 estimators)	0.569	**0.618**	+8.6%
Bagged Icare (100 estimators)	0.593	**0.614**	+3.5%
RSF (10 estimators)	0.606	**0.607**	+0.2%
RSF (100 estimators)	0.647	**0.652**	+0.8%
GB-SA (10 estimators)	**0.621**	0.614	-1.1%
GB-SA (100 estimators)	0.632	**0.636**	+0.6%

Table 5.1: C-Index for each model, using basic and augmented radiomics. In the last column the percent change in performance is added.

Interestingly, the performance of both bagged Icare versions, with 10 and 100 classifiers respectively, largely benefited from the inclusion of augmented samples in their training set, with an improvement of 8.6% and 3.5% respectively. Performance improvements were also measured for the RSF models, even though they were marginal (both less than 1%). An RSF with 100 estimators trained on the augmented data obtained the best overall performance, with a **C-Index of 0.652**, an 0.8% improvement over the best result with non-augmented data. The GB-SA model was the only one to exhibit performance loss, -1.1% for its smaller version (10 estimators), although the trend reversed to +0.6% with its larger version (100 estimators).

5.3.3.3 Impact on Feature Selection and Overfitting

As described earlier, UCI is a feature selection method which is very sensitive to overfitting. In figure 5.7, CI_{train} and CI_{test} were plotted against the number of features included by UCI, for basic (dotted lines) and augmented (solid lines) radiomics.

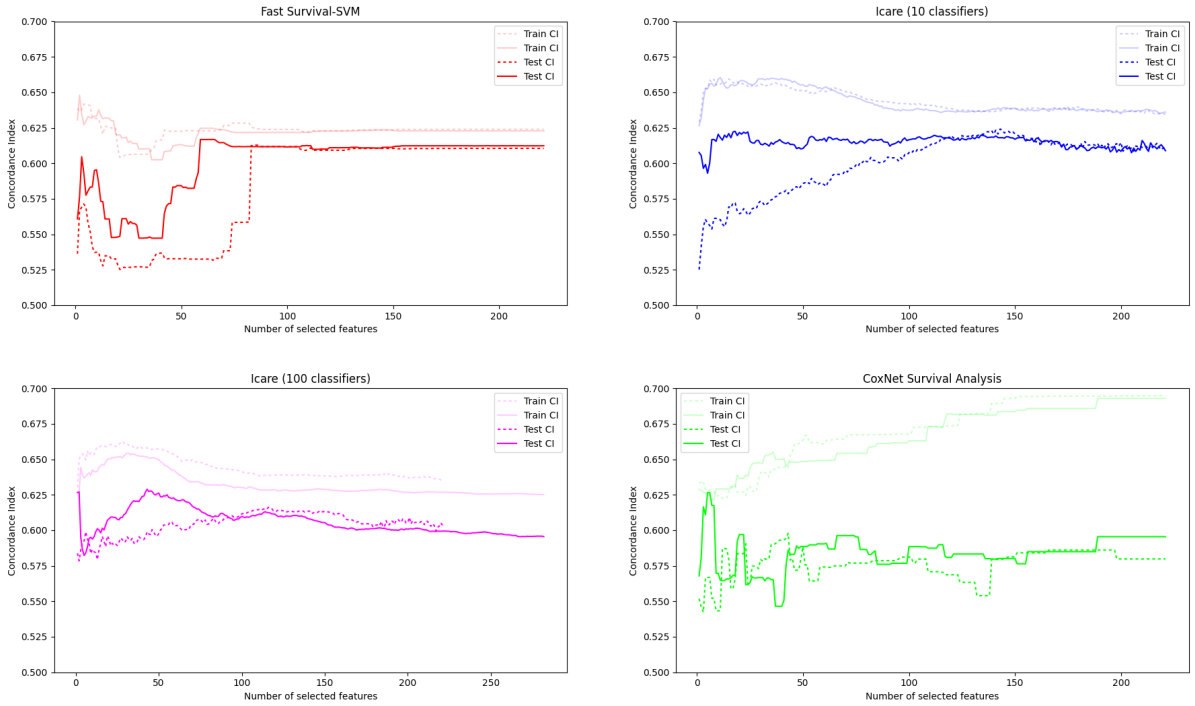

Figure 5.7: Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for some of the benchmarked models. Dashed and solid lines represent the performance using basic radiomics lines and augmented radiomics respectively. CI_{train} values are displayed with more transparency. See Appendix 1 for the remaining models.

FS-SVM, CoxSA and bagged Icare showed the same trend: when augmented radiomics were included in the training set, the maximum values of CI_{train} and CI_{test} tend to correspond more, while in their absence, the higher CI_{train} values do not correspond to an optimal number of variables and the models perform poorly on the test set.

This is particularly visible for FS-SVM (in red) and bagged Icare with 10 classifiers (in blue). When augmented samples were included, CI_{train} and CI_{test} values were also closer than without those samples, at many points other than the optimum of CI_{train}. This demonstrates improved robustness: even if the number of selected variables k_{best} was different, the decrease in CI_{test} performance would likely be more limited with augmented radiomics than without.

Other models did not show large difference in overfitting between basic and augmented training setups. Similar graphs are available for all of the evaluated models in Appendix 1.

5.3.4 Discussion

5.3.4.1 Insights

Radiomics Augmentation has shown a great potential to improve the performance of ML models trained on radiomic features, when using DA on images and masks before features are computed.

Specifically, features extracted from manually delineated lesion and lymph nodes contours in PET/CT imaging for head and neck cancer patients were used to test our radiomics augmentation pipeline. Evaluating different models on the task of RFS prediction, we measured an improvement for 7 out of 8 models, when using augmented samples to train them.

The only model which did not improve with the use of DA is a gradient boosting survival analysis with 10 estimators. The reason behind this might be the inability of the GB-SA model to use the additional information from the augmented samples with only 10 estimators, especially considering that GB algorithms are initialized using a naive solution, such as an uniform probability distribution[258]. Applying the image and mask transforms commonly used in deep learning before the extraction of radiomic features, we obtained up to 8.6% improvement in C-Index, as well as the best overall performance at 0.652, although the impact on performance did vary depending on the model used.

Our initial hypothesis that this performance gain would be driven by a reduction in overfitting seems to be strengthened by the results obtained. Specifically, a smaller gap between CI_{train} and CI_{test} at the point of feature selection and at neighboring values has been noticed when augmented samples were used. This last point is interesting, as it shows that robustness to variations in feature selection can be achieved using DA.

Therefore, it was concluded that improved generalization from data augmentation, using images and masks transforms, was not limited to deep learning but could be extended to most radiomics-based ML methods.

5.3.4.2 Limitations

In complement to the study described here,a more comprehensive study could include other transforms such as affine translations, elastic deformations, erosions, or different versions of the transforms described above (different parameterization). Combinations of multiple image and mask transforms could also be combined to create even more modified samples.

In our study, the parameters of the transforms were chosen randomly from an arbitrary range; exploring the range of parameters, with fixed values, could lead to more accurate measurements of each DA's impact on performance, along with optimal parameters, which could be adapted to different diseases and imaging modalities.

An other possible complementary study could include a selection of augmented samples. Indeed, in our study, all the augmented samples were included in the training set, without a selection process. In *5.5.1*, we demonstrated that some augmented could end up outside of the original distribution. In order to limit the inclusion of unrealistic, out-of-distribution samples, a possible solution is to remove samples too far away from either their original data point or any point in the original distribution. For this purpose, distances metrics (e.g.: euclidean distance), nearest neighbors (e.g.: K-NN) or similarity measures (e.g.: cosine similarity) could be used to determine which samples are too far away from the original distribution.

While our objective was to study the performance increase from overfitting reduction specifically, other feature selection methods, less sensitive to overfitting, could be explored. Minimal depth[259], variable hunting[259], or mutual information[260] are examples of FS methods for survival data.

Additionally, while we used multiple already existing and implemented survival, other models, such as survival DNNs[261], or neural ODEs[262] have been developed for the same task. Some of the model architectures we tested were limited in their configuration (bagged Icare was used with 1000 estimators in[205]), because the computation time needed by larger configurations was too large to include them in our benchmark.

Finally, while the interest of radiomics augmentations was demonstrated on a rather large medical imaging dataset, this method should be applied to other datasets, with other types of imaging data (MRI, 3D ultrasound, CT) and for other ML tasks, such as other survival tasks (overall survival, progression-free survival) and classification (patient response, optimal treatment, N-months relapse). Information other than lesion radiomics could also be added to the prediction task. Clinical information, describing the patient's general state or more specific clinical values (breath quality, genetic profile) can inform the prediction of ML and DL models, including in survival analysis. Organomics, a method extracting radiographic information using the contours of organs instead of lesions, have also shown promising results[7]. Organomics could also be enhanced using DA, which could create both a lot of features and a lot of samples for a given problem. Nevertheless, this amount of data would require an important amount of computation for processing, analysis and ML.

5.3.5 Conclusion

This study demonstrated that data augmentations could be applied to images before radiomic features were extracted, to create augmented samples. These augmented samples followed a distribution similar to the one of the original samples.

Seven out of the 8 ML models we benchmarked showed an improved performance in ranking the patients by RFS time when using the samples obtained using DA.

Additionally, we witnessed a reduced risk of overfitting in four models, characterized by a reduction of the gap between the values of training and evaluation metrics.

These results serve as a first methodological validation of radiomics augmentation. Validation on more diverse, larger datasets as well as other tasks and models would strengthen the importance of our findings.

Different types of medical imaging features, such as organomics, could also benefit from this augmentation method.

The code used for this work is available at `https://www.github.com/Zhack47/radiomics-augmentation`.

Chapter 6

Organomics: Extracting radiographic information from organs imaging

Chapter Overview

Radiomics derived from manually delineated lesion contours are widely used in medical imaging studies, but this approach has many limitations. First, the human contouring is not always reproducible[165] from one observer to another. Even though the availability of precise guidelines improves contours reproducibility, manual delineation remains a bottleneck in many studies, due to its time-intensive nature. Additionally, only a limited amount of information is obtained when only the lesions, such as the primary tumor and malignant lymph nodes, are analyzed.

In 2024, a team of researchers explored the possibility to use automated organ segmentation tools in conjunction with radiomics, to extract the information contained in the organs instead of the lesions. They introduced the term organomics (organ-radiomics) to describe their method[7]. However, to the best of our knowledge, no publicly available end-to-end pipeline has automated the extraction of organ-based radiomic features from medical imaging datasets.

To fill this gap, we have developed and made publicly available **Organ-omics**, an automated, configurable Python pipeline. This pipeline (1) generates organ contours (2) groups organs according to a user-defined configuration and (3) extracts reproducible organ-level radiomic features.

To validate the usability and the performances of our pipeline, Organ-omics was applied to the task of survival prediction, using PET/CT imaging from a lung cancer trial: RTEP7 IFCT-1402. Specifically, we compared the performance of organomic features, lesion radiomics and clinical data for this task.

Contents

6.1 The Organ-omics pipeline: Overview 95
6.2 Detailed description 95
6.2.1 Structure and Configuration 95
6.2.2 Organs Segmentation 97
6.2.3 Radiomics Extraction 97
6.3 Applications and Limitations 98
6.3.1 Applications 98
6.3.2 Limitations and Future Improvements 99
6.4 Application of organomics to survival prediction of NSCLC patients: The RTEP7-IFCT 1402 protocol 100
6.4.1 Introduction 100
6.4.2 Material and Methods 101
6.4.3 Results 106
6.4.4 Discussion 109
6.4.5 Conclusion 111

6.1 The Organ-omics pipeline: Overview

Organ-omics is a command-line ready, Python package which extracts and stores organ-level radiomic features in a CSV file. It follows a user-provided JSON configuration file (see 6.2.1), which specifies the segmentation tasks from TS that should be used and how TS labels should be grouped to form the final labels: **organ groups**.

Once that configuration file is parsed, TS is ran compute-efficiently, by limiting the organs to the necessary minimum (through the *roi_subset* parameter, when available) and assembles labels according to the user-defined configuration. All freely available tasks from TS can be used, while the tasks requiring a license[263] can only be used once the license has been set up on the machine running Organ-omics.

The multi-label segmentation masks are then stored in NIfTI format in a folder chosen by the user. Organ-omics always uses the NIfTi format to store and read medical images, which makes it easy to integrate into larger preprocessing workflows, even though it can also be used as a standalone tool for organ radiomics studies.

The second part of Organ-omics uses the contours previously obtained, alongside the images, to compute the organ-level radiomic features for each organ group and each imaging modality, using PyRadiomics, then stores these features in a single CSV file.

6.2 Detailed description

In this section, each component of Organ-omics is described in detail. These components are: the initial data structure and configuration file, the organs segmentation algorithm and the organomics computation. In the rest of this chapter, *Organ-omics* refers to our pipeline, while *organomics* refer to the organ-radiomic features.

6.2.1 Structure and Configuration

The pipeline was intentionally developed to be split in two parts: segmentation *then* organomics computation. This design was chosen to make contouring independent, so that different contours could be used: either manual contours, contours obtained with a different segmentation algorithm or the contours from TS, reviewed by an expert radiologist.

For the two parts of the pipeline to function properly, only two requirements must be met: a specific organization of the data using the NIfTI file format and a JSON configuration file.

Structure The data structure users need to follow is a folder structure similar to the one used in the nnUNet[82] framework:

```
Root Folder
├─ imagesTr
│  ├─ P01_0000.nii.gz
│  ├─ P01_0001.nii.gz
│  ├─ P02_0000.nii.gz
│  └─ P02_0001.nii.gz
└─ labelsTr
   └─ <empty>
```

Configuration To allow users to adapt Organ-omics, without the need for additional code, our pipeline ingests and parses a user-defined JSON configuration file. This file supports multiple fields, some of which are required:

- **channel_names**: A dictionary of the imaging modalities, with their corresponding index (0, 1, 2 ..., *nb_modalities*). One of these modalities should be named *"CT"* and will be used for the organ contouring *via* TS. The remaining modality names are stored, until the radiomics extraction is performed, since a set of organomic features is extracted per modality/mask pair.
- **labels**: A dictionary containing the names of all the organ groups with their user-assigned label number (+ background which must correspond to label number 0). These labels are the final segmentation labels, which can correspond to one or multiple TS VOIs grouped together.
- **tasks:** A list of all tasks from TS which need to be ran in order to obtain the organs included in the VOI defined above.

  Each element of this list must contain:

  - **name**: The name of the task, as defined in the TS documentation. The most popular task is named *total* and contours 117 individual organs.
  - **correspondence**: A dictionary, using the **labels** defined earlier as keys, and lists of the organs from that task which must be combined to obtain the final label, as values.

Other fields can be added to the JSON file, but are not required:

- **name** and **description**: Name and description for the dataset and/or the experiment. These fields are not used by the pipeline and is only added for identification purposes
- **spacing**: The spacing towards which all imaging volumes and segmentation masks will be resampled before organomic features are extracted. To respect IBSI[213] recommendations, this spacing should be isotropic. Interpolation towards this spacing was performed using third order BSpline interpolation on images and nearest neighbour interpolation on masks.

This configuration file is read and parsed using Python, to parameterize the subsequent organ segmentation and organomic feature extraction.

6.2.2 Organs Segmentation

Once Organ-omics has been installed, from the TestPyPi Python package manager, and when the data structure and configuration file presented above are properly defined, two commands are available to the user. The first one, **Organomics_contour_dataset**, performs the segmentation of the organs defined in the configuration file.

It needs to be provided with the following arguments:

- **-d <dataset_path>**: The path to the root folder of the dataset (see below)
- **-o <contours_output_path>**: Path to the folder which should be used to store for the NIfTI segmentation files
- **–json-file-path <json_file_path>**: Path to the JSON configuration file

Organs segmentation is performed using TotalSegmentator[6] on the CT volume. Parsing the configuration file, the channel number identifying the CT volume is obtained and used to select the CT file for each patient. This volume is then processed using TotalSegmentator's Python API:

1. **The tasks are initialized**: All the tasks defined in the JSON configuration file are initialized, using the *roi_subset* parameter to limit the amount of labels returned, which in turn reduces the number of sub-tasks TS needs to perform.
2. **Segmentation inference is ran**: Each patient is processed by each of the TS tasks, through the TS Python API. This step takes the most time, as the UNets are processing the image to obtain the organ segmentations. The labels returned for each of the tasks are stored separately at this step (1 segmentation mask per task).
3. **Labels are grouped**: Final labels are obtained, through the combination of the task-specific TS labels, in the way defined in the JSON configuration file. An important point to keep in mind is that, since the contours can be obtained from different tasks and grouped afterwards, overlap may occur between the labels. Assuming that the labels are mutually exclusive, this was resolved by prioritizing the labels with the smallest label number value. Users should thus include the most important organs towards the beginning of the list.
4. **Organs segmentation masks are saved**: The segmentation masks, grouped and resolved for overlaps, are saved in one single NIfTI file for each patient.

From this point on, organ segmentation masks are available to the user, in a folder they defined in the command line options. Depending on the user's needs, these contours can be modified, or directly used to compute organomic features.

6.2.3 Radiomics Extraction

The second command available to the user is **Organomics_extract_radiomics**. This command uses images and organ contours to compute organomic features for each organ/image pair. These organomic features are then stored in a CSV file.

The command needs the following arguments:

- **-d <dataset_path>**: The path to the root folder of the dataset, this time including the segmentation files in the labelsTr folder

- **-o <csv_file_organomics_output_path>**: Path where the CSV file containing the computed organ-radiomics should be stored

- **–json-file-path <json_file_path>**: Path to the JSON configuration file

Organ radiomics extraction occurs in three steps, which are repeated for each patient:

1. **Images and masks are loaded**: Images from different modalities can be included alongside the CT image, as long they are all co-registered. These images are loaded using SimpleITK[53]. The organ segmentation masks are also loaded using SimpleITK, then split in one binary mask per organ

2. **Resampling is applied (optional)**: If a spacing was defined in the JSON configuration file, all images and masks are resampled towards this spacing. This resampling is performed using SimpleITK's *ResampleImageFilter*. The remaining spatial information (origin, direction, size) is set to the values of the CT, as this is the space in which organ masks are already placed.

3. **Organ-radiomics are computed**: Using PyRadiomics, our pipeline goes over each modality / mask pair, computing all 107 radiomic features offered by the library.

Once this processed has been repeated for all patients, the resulting features are stored in a CSV file, which can then be used for downstream tasks.

Organ-omics source code and installation guidelines are available at: `https://www.github.com/Zhack47/Organ-omics`.

6.3 Applications and Limitations

6.3.1 Applications

Organ-omics allows clinicians, AI researchers and other medical imaging professionals to investigate the links between information contained in the imaging values of organs and multiple endpoints, such as patient survival and treatment response. Enabling the discovery of new, universal (i.e.: applicable to patients without lesions) imaging biomarkers, it aims to facilitate personalized medicine and treatment, powered by a completely automated analysis of patient imaging.

Oncology imaging was the main application field for which Organ-omics was designed. This field is ideal for radiomic features in general, as it uses multiple quantitative imaging modalities already (PET, CT, SPECT, qMRI). Oncology has also started to shift from a one-size-fits-all to a personalized treatment workflow, due in part to the heterogeneity found in cancers[264].

Outside of oncology, Organ-omics could be used on any CT imaging, or any body imaging which can be registered with a CT of the same patient. For instance, non-oncologic musculo-skeletal diseases could be evaluated using organomic features computed on bone and muscle contours[265][266].

6.3.2 Limitations and Future Improvements

While it was designed to be able to extend organomic features to non-oncologic imaging, Organ-omics still suffers from domain limitations. First, a CT scan is required, and any additional imaging modality should be registered to this CT, previous to the computation of organomic features. In the future, this restriction could be partly lifted, thanks to organ segmentation models which use MRI volumes instead of CT[267]. Although the performance of these models is lower than the performance of CT-based organ segmentation models, they are capable of contouring bones and tissues, as MRI is more often used to analyze such tissues.

Another limitation lies in the organ contours: the quality of these contours depend on both the imaging quality and acquisition field-of-view (FOV). Indeed, if an organ is partly or completely missing from the image volume, the associated organomic feature value will be respectively missing or wrong. This means that having images acquired with a small FOV, leading to missing or partly cropped out organs after segmentation, will negatively impact the performance of downstream tasks. Therefore, it is recommended to use Organ-omics on data acquired with the same FOV, ideally a broad, whole body FOV, which includes all the target organs for each patient.

Additionally, human corrections and/or validation of the generated contours are needed if adherence to segmentation guidelines is required.

As described above, Organ-omics does not currently support overlapping structures. Multiple, separated segmentation masks could be saved for each structure, in order to support simultaneous organomic features extraction from overlapping structures, such as lung *and* lung lobes for instance.

On the radiomics extraction side, other features of the PyRadiomics library, such as re-segmentation and image filters were neither used nor re-implemented in Organ-omics. PyRadiomics' re-segmentation is a threshold based segmentation, which parameters can be specified in a configuration file. Such a re-segmentation does not make much sense in the case of organ contours, especially on CT, where threshold based segmentation is not commonly used (outside of air and bones).

Image filters on the other hand could be used to create different features. The additional information produced using such filters would multiply the number of features per sample, which, in the case of Organ-omics, is already multiplied by both the number of modalities and the number of organs. This would create an imbalance between the number of features and the number of samples, short of having a very large number of samples as well. A data augmentation framework such as radiomics augmentation (presented in chapter 5) could reduce the imbalance. It would however further increase the amount of computing power required to process the data. The combination of new, numerous organ features with data augmentation methods remains an interesting experiment, which should be led to study performance improvement for downstream prediction tasks.

6.4 Application of organomics to survival prediction of NSCLC patients: The RTEP7-IFCT 1402 protocol

To demonstrate the usefulness of Organ-omics, it was applied to a PET/CT imaging dataset, with two time-points, to predict overall survival (OS). This dataset contained PET/CT imaging before and during radiotherapy, for 138 patients with non-small celle lung cancer (NSCLC), treated using radio-chemotherapy. The predictive performance of Organ-omics was evaluated and compared with the performance of lesion radiomic features and clinical data.

6.4.1 Introduction

Non-Small Cell Lung Cancer (NSCLC) the third most prevalent and the deadliest cancer worldwide[268]. Effective treatment of NSCLC is critical to improve cancer patients' outcomes and overall survival (OS). Radiotherapy (RT) plays an important role in the treatment for NSCLC, and optimizing its efficacy for each patient is one of the keys for the future of personalized cancer treatment. ^{18}F-FDG PET-CT is used in lung cancer diagnosis, to measure glucose metabolism and assess tumor extent and aggressiveness. In recent years, studies have shown that PET/CT can optimize radiation therapy doses and volumes [269][196][270][191].

In the last decade, radiomic features were introduced as a powerful tool to predict treatment outcomes and patient survival. Radiomics involves the extraction of quantitative features from medical images, usually to characterize tumor or lesion phenotypes. This quantitative approach has recently been extended to healthy organs. Using these organs as VOIs to extract organomic features further quantifies the patient's imaging data. Organomic features have shown strong predictive power in survival[271][7].

The RTEP7 IFCT 14.02 trial is a phase II/III clinical trial designed to study the safety and efficiency of dose escalation (DE) in the RT, using mid-treatment PET/CT (42Gy). Imaging data from this protocol is the foundation for our study. Data from this multi-site protocol is gathered and used in the present study to predict the Overall Survival (OS) endpoint.

The goal of our work is to compare OS predictive strength of organomic features to radiomic features and clinical features, using the data from the RTEP7 trial. This paper will go over the methodology employed for feature extraction and for prediction, present the results of our analysis, and discuss the implications of our findings for clinical practice and future research.

6.4.2 Material and Methods

6.4.2.1 Dataset (from RTEP7 IFCT-1402 trial)

The RTEP7 trial (NCT02473133) was a multi-center, randomized, controlled phase 2 trial RTEP7 IFCT-1402 aiming to prove the efficiency and safety of dose-escalated radiotherapy in 19 French hospitals.

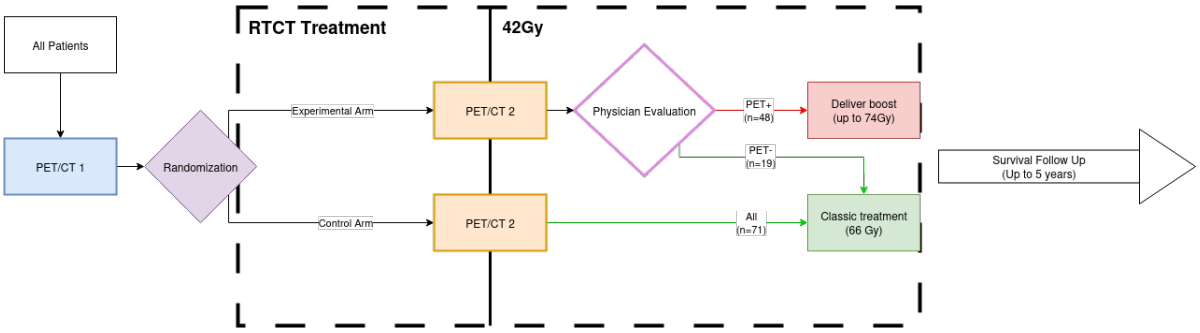

Figure 6.1: Flowchart describing the RTEP7 IFCT-1402 protocol: all patients included undergo a first PET/CT scan (PET/CT1) before being randomly assigned into either the experimental or the control arm. In the control arm the classical 66Gy RT treatment was delivered to all patients but a second PET/CT was acquired (PET/CT2). In the experimental arm, the second PET/CT was used to decide a boost of RT to 74Gy.

Two PET/CT scans were conducted for each patient: one before treatment (PET/CT1) and one during the treatment, after 42 Gy (PET/CT2). Patients were randomized between two treatment arms, before the start of radiotherapy. In the control arm, patients received 66Gy in 33 fractions regardless of the contents of PET/CT2. For patients in the experimental arm, radiation oncologists decided whether dose-escalated radiotherapy was necessary and/or feasible, based on the analysis of PET/CT2. If it was deemed necessary, these patients received up to 74Gy of radiation in total during radiotherapy. The flowchart for the RTEP7 trial is presented in figure 6.1.

The primary endpoint was local control rate, 15 months after randomization. Secondary endpoints included progression free survival (PFS) and Overall Survival (OS). The RTEP7-IFCT14.02 trial involved 19 hospitals in France, each providing 1 to 29 patients (See figure 3 in Appendix). Further details on the method and outcomes are available in the main publication[272] and its ancillary study[273].

6.4.2.2 Details on the PET/CT procedure

PET/CT scans were acquired following the EANM/EARL guidelines[274]. By design, an harmonization strategy of PET reconstruction was done: PET reconstruction was optimized in each center for the protocol. CT reconstruction was not standardized and was left to the practice of the local Nuclear Medicine (NM) department.

A patient always had all their scans acquired on the same PET/CT scan model. Field Of View (FOV) varies between patients: a whole body FOV was mandatory for PET/CT1, but the FOV may be centered on lungs for PET/CT2. All patients had at least their lungs area visible in the scan regardless.

6.4.2.3 Tumor and organs segmentation

For this ancillary analysis, three nuclear physicians manually segmented cancerous lesions on the PET/CT1 and PET/CT2 exams for each patient. A Volume Of Interest (VOI) was applied on the tumor burden and malignant nodes, then a manually adapted threshold was used to create a united Biological Tumor Volume (BTV) segmentation (PETVCARE V3.2 GEMS, Buc, France). Detailed description of the procedure can be found in a previous ancillary study[273].

TotalSegmentator (TS)[6] (version 2.8.0) was used to extract 117 organ contours from the CT scans acquired during PET/CT1 and PET/CT2. To avoid missing data due to FOV variations, only 7 thoracic organ groups were kept (trachea, esophagus, heart, aorta, spinal cord, thoracic vertebrae and lungs). Going forward and for readability purposes, we will refer to these organ groups as *"organs"*. The "thoracic vertebrae" structure was created by gathering T1 to T12 contours and the "lungs" structure gathered all 5 lobes of the lungs.

Some of these organs, namely aorta, spinal cord, trachea and esophagus, can sometimes be cropped shorter than their full length, because of a limited scan FOV. This was the main source of variability we identified in automated organ contours.

6.4.2.4 Automated Features Extraction

The features provided by the PyRadiomics library were used, both for radiomic and organomic features. All of the default categories were included, namely shape features, first order features and texture features (GLCM, GLSZM, GLRLM, NGTDM and GLDM).

In this work, radiomics will reference the lesion radiomics, meaning the radiomic features extracted from the PET and CT images using the manual BTV (tumor + lymph nodes) contours. Our study follows the recommendations from the CLEAR checklist.

Organomics were extracted automatically using the Organ-omics pipeline presented above. It was configured to use a common, isotropic spacing of $3 * 3 * 3mm$. All volumes (PET, CT and contours) were resampled to this spacing. Resampling to an isotropic spacing was necessary to preserve the meaning of texture-based radiomic features, as per the IBSI recommendations. This spacing was chosen to stay close to the PET spacing, therefore minimizing information loss on this scan, although it impacted the CT scans, which have a smaller spatial resolution and were thus overly smoothed.

All 3D features classes were used (FirstOrder, Shape, GLCM, GLSZM, GLRLM, NGTDM, GLDM), the same as for radiomics. No filters (LoG, Wavelet etc. . .) were applied to create secondary images.

Organ-radiomic (organomic) features were also computed on the PET and CT volumes, using the TS contours for each of the organs defined above.

The ComBat feature harmonization method[275] was not used, although it has widely been used in radiomics studies. Indeed, PET reconstruction was optimized at each site for the RTEP7 trial and no impact has been found on the tumor volume when different reconstructions are used[276]. Moreover, our dataset contains data from numerous sites which each provided few patients. In a recent article[277] it was recommended to use ComBat harmonization when at least 20 to 30 patients were available per site, which is not our case (See figure 3 in Appendix).

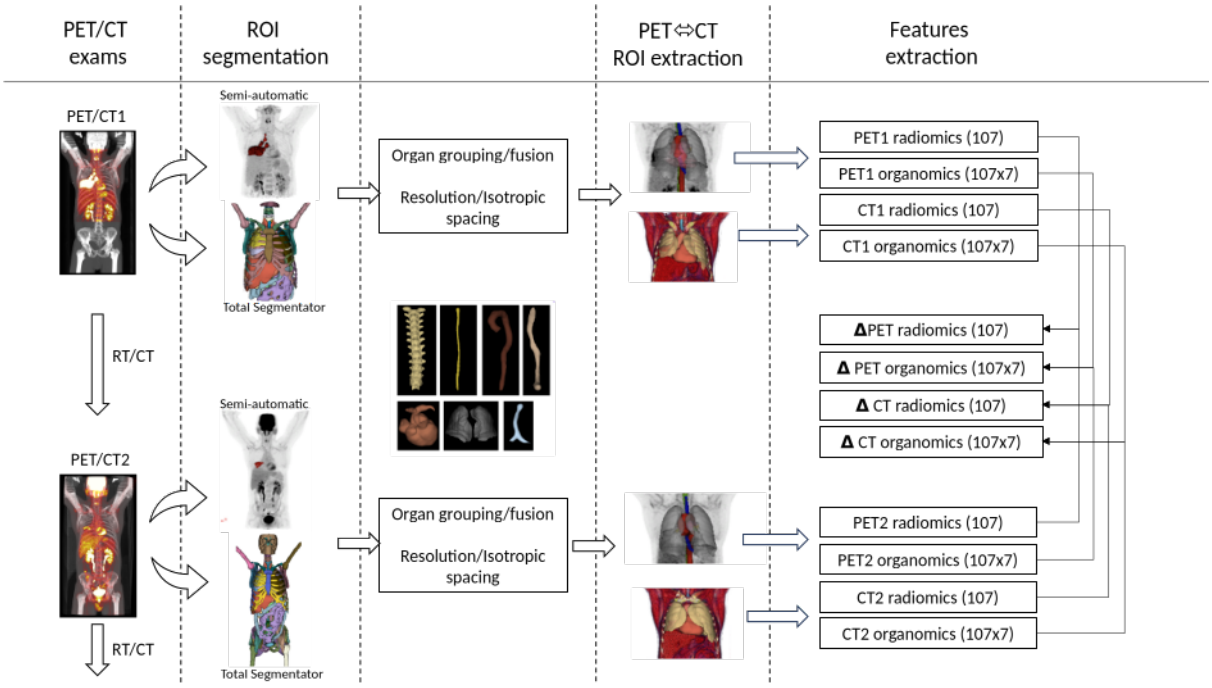

Figure 6.2: Overview of the feature extraction pipeline. Organ contours were extracted from the CT scans using TS. The BTVs were manually annotated by nuclear physicians. Using the PET and CT volumes as 3D matrices of voxels and the lesion and organ contours, we extracted the radiomics and organomics respectively. We obtained two sets of each, one for the pre-treatment PET/CT and one for the per-treatment PET/CT. A third set (Δ F), was computed as the difference between the features at the two time points.

To analyze the temporal variations of the features, we extracted a set of features for PET1 (F1) and PET2 (F2) separately. Additionally, a set of feature quantifying the variation in features was subsequently computed: (ΔF).

6.4.2.5 Machine Learning benchmark

To compare the predictive power of organomics and radiomics, we built a fixed Machine Learning (ML) training and scoring benchmark. This benchmark first used Univariate C-Index (UCI) to select the most relevant features: each feature was used to train a univariate CoxNet model[61] on the training data, which was then scored on this same data using the C-Index (CI). This approach has been shown to be robustly selecting the most important variables, regardless of the model used afterwards[278], albeit with a risk of overfitting.

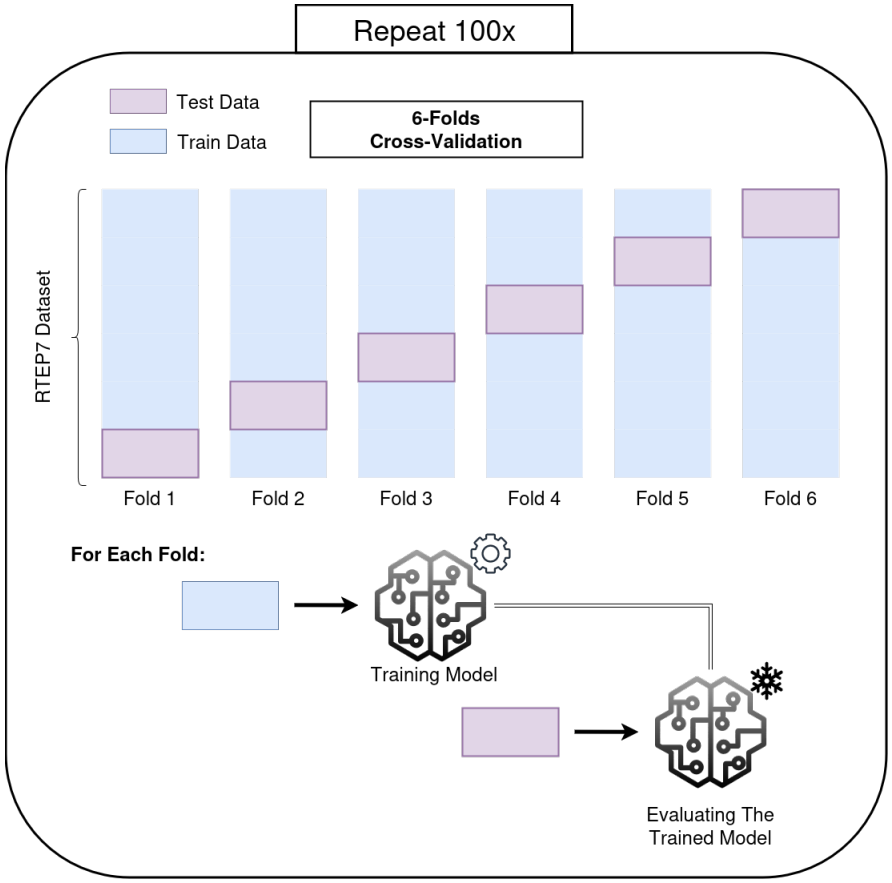

Figure 6.3: We built a 6-folds cross-validation to evaluate the performance of our approach (organomic features) against radiomic features and clinical data. The data was split in 6 folds, each used once for testing and five times as part of the training set. Once the model was trained on 5 folds, it was then evaluated on the remaining fold. This CV was repeated 100 times, to average the results for more stability.

The second step was a 6-fold cross-validation (CV). Each of the 6 folds was used once as a test set (23 patients), to evaluate a model which was trained on the 5 remaining folds (115 patients). A feature selection threshold was found by training and scoring the ML model on the training data, using all features, then dropping the least predictive features, one by one. The feature selection threshold yielding the highest C-Index on the training set was deemed optimal and selected for training and evaluation.

To obtain robust results, the CV is then repeated 100 times and the average Harrel C-Index (C-Index) and cumulative dynamic Area Under the Curve (cdAUC) are computed, along with confidence intervals (CI) for both these metrics. The ML model we used was a Fast Survival SVM ($\alpha = 1e-4$, max iteration=1000)[60] from the scikit-survival(v0.22.2)[57] library. Before inputting the selected features in the model they are scaled to a mean of 0 and standard deviation of 1 using a Standard Scaler. Scaling normalizes features ranges, which will later be useful to analyze the relative importance of the selected variables in the model.

6.4.2.6 Model categories

Four different categories of models are benchmarked:

1. **Clinical model:** a model using only clinical data (Age, Gender, Cancer stage, histological subtype, Performance Status (PS), arm assignation (Experimental / Control), RT boost (yes / no), type of induction Chemotherapy (indCT), type of Radio Chemotherapy (RTCT), number of induction cycles, number of RTCT cycles and RT technique (IMRT or not).
2. **Radiomics models:** models using radiomic features extracted from the lesion VOI (+clinical data)
3. **Organomics models:** models using features extracted from the organ VOIs (+ clinical data).
4. **Combined models:** models using both radiomic and organomic features.

Models from categories 2 to 4 were trained on three different imaging features subsets:

A CT features only

B PET features only

C PET + CT features

6.4.2.7 Feature Importance

Once all models were scored using this benchmark, the coefficients attributed to each selected feature by the SVM was retrieved along with the name of that feature. This was done to show how each selected features influenced the model's prediction of a patient's survival. On top of that, the coefficients reflect the relative importance of each feature, explaining in part how the model works.

6.4.3 Results

6.4.3.1 Demographic Data

Gender		
	Male	100 (72.5%)
	Female	38 (27.5%)
Age		62.5 years (38-76.7)
Histology		
	Adenocarcinoma	71 (51.4%)
	Squamous Cell Carcinoma	53 (38.4%)
	Other	14 (10.2%)
Stage		
	IIIa	74 (53.6%)
	IIIb	63 (45.6%)
	Other	1 (0.8%)
Treatment Arm		
	Control	71 (51.4%)
	Experimental	67 (48.6%)
Induction Chemotherapy		
	1 cycle	1 (0.8%)
	2 cycles	127 (99.2%)
RadioChemotherapy		
	1 cycle	8 (5.8%)
	2 cycles	53 (38.4%)
	3 cycles	73 (52.9%)
Radiotherapy		
	3D	27 (19.6%)
	IMRT	111 (80.4%)
Censored		
	Yes	83 (60.1%)
	No	55 (39.9%)

Table 6.1: Demographic data from the patients included in the RTEP7 protocol.

A total of 138 patients with stage III non-small cell lung cancer (NSCLC) were included in the analysis. The average age was 62.5 years (range: 38-76.7 years). Out of these, 100 patients were male (72.5%) and 38 were female (27.5%). Histological cancer type confirmation was available for all patients: 71 had adenocarcinoma (51.4%) and 53 had squamous cell carcinoma (38.4%). Disease staging at inclusion was either stage IIIA (n = 74, 53.6%) or stage IIIB (n = 63, 45.6%). Performance status (PS) was 0 in 52 patients (37.7%) and 1 in 86 patients (62.3%) (table 1).

Patients were randomized into two treatment arms. The control arm included 71 patients (51.4%)

who received standard radiotherapy (66 Gy). The experimental arm included 67 patients (48.6%), of whom 48 (71.6%) received dose escalation to 74 Gy based on PET/CT 2 evaluation. Regarding chemotherapy, 137 patients (99.3%) received two cycles of induction chemotherapy (indCT), while one patient received a single cycle. Concurrent chemoradiotherapy (RTCT) was administered in 1 to 3 cycles: 8 patients (5.8%) received one cycle, 53 patients (38.4%) received two cycles, and 73 patients (52.9%) received three cycles.

The most frequently used chemotherapies were Vinorelbine-Cisplatin (n=70, 50.7%) and Paclitaxel-Carboplatin (n=57, 41.3%). Radiotherapy technique was 3D conformal RT for 27 patients (19.6%) and intensity-modulated radiotherapy (IMRT) in 111 patients (80.4%). Out of the 138 patients who underwent the full treatment and had PET/CT1(before RT) and PET/CT2 (42Gy from RT) imaging available, 83 (60.1%) had their OS data right-censored, and 55 (39.9%) had uncensored OS data. In the original study, it is stated Median OS was not reached, which explained in part the large number of right-censored patients.

6.4.3.2 Available Features

Radiomic and organomic features were extracted from both PET and CT volumes at two time points: pre-treatment (PET/CT1) and mid-treatment (PET/CT2). For each volume of interest (VOI) and each imaging modality, 107 quantitative features were computed using the PyRadiomics library, following IBSI-compliant standards. These features included 3D shape descriptors, statistics from the first order histogram and texture-based metrics (GLCM, GLRLM, GLSZM, GLDM, NGTDM).

Seven thoracic VOIs were analyzed: lungs, heart, aorta, trachea, oesophagus, spinal cord, and thoracic vertebrae. Each VOI was quantified on both PET and CT scans, resulting in 14 VOI-modality combinations and features subsets. For each patient, three distinct feature sets were generated:

- F1: Features from PET/CT1
- F2: Features from PET/CT2
- ΔF: Temporal change in features, computed as $\Delta F = F_2 - F_1$

This yielded a total of 4,494 organomic features per patient (1,498 features in each of F1, F2, and ΔF).

Radiomic features were similarly extracted from a manually segmented biological tumor volume (BTV) on both PET and CT scans, for a total of 642 features.

Table 2 summarizes the number of features available across different input combinations and their corresponding model performance metrics obtained on the benchmark.

6.4.3.3 Quantitative Results

The results from the benchmark described above were collected and presented in table 6.2.

Using clinical features alone (specifically age, treatment arm and RT delivered dose), the model achieved a CI of 0.573 ± 0.008 and a cdAUC of 0.592 ± 0.012, establishing a baseline for comparison.

			Radiomics		Organomics		Results	
Model	Nb Feat.	Clin.	PET	CT	PET	CT	CI (±*Cint*) ↑	cdAUC (±*Cint*) ↑
Clinical	6	✓(3)					0.573 (±0.008)	0.592 (±0.012)
Radiomics	327	✓(3)	✓(20)				0.517 (±0.012)	0.515 (±0.017)
	327	✓		✓(29)			0.591 (±0.012)	0.612 (±0.017)
	648	✓	✓(8)	✓(29)			0.584 (±0.012)	0.602 (±0.016)
Organomics	2253	✓			✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2253	✓				✓**(2)**	**0.663** (±0.008)	0.688 (±0.012)
	4500	✓			✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
Combined (3)	2574	✓	✓		✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2574	✓	✓			✓**(2)**	**0.663** (±0.008)	0.688 (±0.012)
	2574	✓		✓	✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2574	✓		✓		✓**(2)**	**0.663** (±0.008)	**0.688** (±0.012)
Combined (4)	2895	✓	✓	✓	✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2895	✓	✓	✓		✓**(2)**	**0.663** (±0.008)	0.688 (±0.012)
	4821	✓	✓		✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
	4821	✓		✓	✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
All	5142	✓	✓	✓	✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)

Table 6.2: All data combinations tested as input to our benchmarking algorithm are presented in this table. A check mark means this type of data was used as input to the model, red if this type was not selected at all, green if some features were selected (with the number of selected features in parenthesis). The total number of features included in each model is indicated in the second column.

When PET radiomics features were added (20 additional features selected) to the clinical data, performance declined, with C-Index and cdAUC of 0.513 ± 0.011 and 0.512 ± 0.014, respectively. In contrast, including both PET (8 features selected) and CT radiomic features (29 features selected) alongside clinical data (no feature selected) improved performance to a CI of 0.584 ± 0.012 and

cdAUC of 0.602 ± 0.016. Notably, clinical features were not selected when radiomic features were included. This was the best performance obtained using radiomics, as well as our imaging baseline.

Models using only organomics features showed the strongest predictive performance. The best-performing model achieved CI and cdAUC reaching values of 0.663 ± 0.008 and 0.688 ± 0.012 respectively using only CT organomics features (2 features selected). When both PET (35 selected features) and CT (15 selected features) organomic features were included, the model also performed well with a CI of 0.656 ± 0.011 and cdAUC of 0.689 ± 0.011.

The performance of the model using CT organomics was significantly better than the clinical data model ($p<<0.001$), all three radiomics models ($p<<0.001$) and the PET organomics model ($p<<0.001$). The performance difference between the CT and PET/CT organomics models was not significative ($p=0.78$).

In all organomic-inclusive models, neither clinical nor radiomic features were selected during feature selection, underscoring the dominant prognostic value of organomic descriptors. All possible combinations between the feature subsets were benchmarked and their results were reported in table 6.1.

All selected features along with their coefficients were listed, sorted by models (see table 1 in Appendix).

6.4.4 Discussion

In this study, we gathered imaging and clinical data for 138 stage III NSCLC patients from the RTEP7 trial who underwent radio-chemotherapy. This includes pre- and mid-treatment PET/CT imaging, from both of which radiomic features were extracted using lesions (BTV) and organomic contours, the latter derived from thoracic organs using automated segmentation software (TotalSegmentator).

Using a standardized machine learning benchmark, we compared the prognostic value of clinical variables, lesion-based radiomics, and healthy-organ radiomics (organomics) for overall survival (OS) prediction. Our results suggest that radiomic features derived from healthy thoracic organs (organomics), particularly CT-based descriptors, may provide stronger prognostic information for OS than conventional lesion-based radiomics ($p<<0.001$) in patients with stage III NSCLC undergoing radio-chemotherapy.

The strongest predictive model relied solely on CT organomics, achieving a C-Index of 0.663 ± 0.008 and a cdAUC of 0.688 ± 0.012. Notably, whenever organomic features were included, neither radiomics nor clinical features were included by the feature selection algorithm, suggesting a dominant contribution of organ-level imaging biomarkers within this specific modeling framework. This finding aligns with recent literature suggesting that the physiological response of non-tumoral tissues to treatment may reflect systemic resilience and treatment tolerance, thereby influencing survival outcomes[271][7].

Several hypotheses may explain the observed prognostic value of organomics. Healthy organ textures and shape descriptors may indirectly capture inter-patient variability related to baseline physiological reserve, inflammatory status, or early radiation-induced changes. Such information is not directly accessible through tumor-centric radiomics alone, especially in a mid-treatment setting where lesion response may be heterogeneous or confounded by treatment-induced inflammation.

From a clinical perspective, these results suggest that organomics could be integrated into routine

imaging workflows to enhance prognostic stratification. Importantly, CT-based organomics showed the best performance, which is of practical relevance since CT imaging is widely available and less resource-intensive than PET imaging. Automated organ segmentation using open-source tools such as TotalSegmentator (or CE/FDA-approved commercial solutions) makes this approach scalable and potentially transferable across institutions. This also raises the possibility of longitudinal organomics monitoring during treatment or follow-up, even in the absence of functional imaging.

Our findings contrast with earlier lesion-centric studies such as RTEP2[196], where lesion SUVmax was shown to predict short-term outcomes (death or progression at 1 year). In the present analysis, lesion radiomics did not contribute meaningfully to censored 5-year OS prediction, and inclusion of organomics substantially improved model performance. SUV_{max} features, whether assessed at baseline, mid-treatment, or as temporal changes, were consistently not selected and yielded poor predictive performance when evaluated separately (see table 2 in Appendix). This discrepancy likely reflects differences in endpoints (short-term vs. long-term survival), modeling strategies, and the added complexity introduced by censored survival data.

Among the selected organomic features, esophagus shape descriptors, particularly the change in esophagus volume between scans, emerged as the most important feature in the best-performing models. This observation is difficult to understand given that no severe toxicity or high grade adverse events were recorded during the RTEP7 trial, unlike in PET-Boost. In the PET-Boost trial[279], esophageal toxicity played a significant role in patient outcomes[280]. In that trial, severe dysphagia and esophago-tracheal fistula were associated with treatment-related mortality, highlighting the potential of organ-level imaging biomarkers to capture clinically meaningful toxicity signals.

Several limitations of this study should be acknowledged. First, our analysis was restricted to patients enrolled in the RTEP7 protocol, all of whom had stage III NSCLC treated with radiochemotherapy, with optional dose escalation. Therefore, the generalizability of our findings to other disease stages, treatment strategies, or cancer sites remains to be established.

Second, while PET reconstruction was harmonized across centers, CT acquisition parameters were not standardized, which may have introduced additional variability into CT-based features. Regarding lesion delineation, we relied on the Biological Tumor Volume derived from PET imaging. A more comprehensive lesion-based radiomics analysis would ideally include CT-based GTV or CTV contours, which were unavailable for patients who did not undergo replanning after PET/CT2. This limitation may partly explain the relatively modest performance of lesion PET radiomics in our study.

Organ contours were automatically generated using TotalSegmentator. Alternative organ segmentation tools such as MOOSE[281] have been proposed, and recent studies have shown good agreement between these methods for major thoracic organs[282][283]. While segmentation reproducibility appears acceptable, further work is warranted to assess the robustness of organomic features across segmentation algorithms and imaging protocols.

Compared with previous organomics studies[7][271], our approach differs in that we grouped thoracic organs into anatomically and functionally relevant regions rather than analyzing all available organs individually. This strategy was motivated by dimensionality reduction considerations and by the variability in tumor location across lung lobes, and may facilitate clinical interpretability.

6.4.5 Conclusion

In conclusion, this study supports the concept that imaging features extracted from healthy organs provide complementary and potentially superior prognostic information compared with traditional lesion-centered radiomics in stage III NSCLC. By shifting the focus from the tumor alone to the surrounding organs, organomics may capture treatment tolerance and systemic response, opening new avenues for personalized risk stratification and treatment adaptation. Prospective validation in independent cohorts and integration with clinical and biological biomarkers will be essential before clinical translation. The code used for organomic features extraction is available at:

`https://www.github.com/Zhack47/Organ-omics`

Chapter 7

Conclusion and perspectives

7.1 Conclusion

This thesis has presented the current state of AI technologies for multi-modality imaging in oncology. Fully automated and semi-automated segmentation are at the forefront of AI integration in the clinical practice, while imaging descriptors are still an open research domain. The combination of automated segmentation tools with these descriptors offers a new avenue for survival prediction and the evaluation of a treatment's efficiency.

While DNN and DL techniques are available and show the best performances for segmentation, survival prediction is still using ML methods. We have demonstrated how an end-to-end pipeline could still be created, using SoTA segmentation tools to contours VOIs, extracting radiomic descriptors from these contours, then using ML models to predict survival from these descriptors and even obtaining meaningful features.

Below is a summary of the most important works contributed during this thesis:

- **Automated PET/CT segmentation algorithms**: In chapter 4, the contibutions to the SoTA in PET/CT segmentation were presented. Starting with a simple ^{18}F-FDG PET/CT segmentation model, trained with a modified version of the nnU-Net algorithm, which outperformed the baseline model and ranked 4th in AutoPET II **(2023)**.

  The training recipe was then updated to include additional organs supervision, a residual U-Net architecture as well as multi-tracer segmentation (FDG and PSMA). A PET tracer classification CNN, which proved almost infallible (99.64% accuracy), was introduced to route segmentation to one of two tracer-expert models, for AutoPET III **2024**.

  Finally, human input was forwarded to the U-Net, through the construction of PET-aware heatmaps as a domain informed prior to segmentation. The combination of all these improvements culminated in a team contribution to AutoPET IV **2025**.

  All of this work is publicly available, both in papers[9][10][11] and repositories containing the code and weights necessary to train and run these models on PET/CT images[284][285][286].

- **The Radiomics Augmentation framework**: A framework to apply data augmentations on images and masks before the extraction of radiomic features. This framework aims to transpose the concept of data augmentation in medical data from DL to ML. A comparative

study measuring the performance of 8 ML models, on the task of RFS prediction using the HECKTOR 2022 dataset, found that Radiomics Augmentation improved the performance for 7 out of these 8 models and reduced overfitting at the feature selection step in 4 of them.

The code used to apply Radiomics Augmentation on the Hecktor2022 dataset has been made publicly available[287].

- **The Organ-omics pipeline**: A pipeline extracting organ-based descriptors from medical images. This end-to-end pipeline embodies the philosophy of this thesis: first DL-based segmentation is used to obtain the contours of relevant VOIs , then high level descriptors are extracted from these VOIs and used to train ML models for less densely supervised tasks. Organ-omics was built to make the concept of organomics[7] accessible to researchers, to foster new research using these more holistic descriptors. The code and instructions to run Organ-omics on any dataset are publicly available online[288].

  This pipeline has been validated on a private dataset, collected retrospectively from a clinical trial, which contained PET/CT imaging data from two time points. The predictive power of organomics, radiomics and clinical data was compared using this data and the collected endpoint: right-censored OS. This experiment has demonstrated the clear superiority of organomics over clinical data (0.663 vs 0.573) and radiomics (0.663 vs 0.591).

  Additionally, it was noted that organomics computed on the CT scan offered the best performance, which has interesting clinical implications: OS seems to be related to imaging features visible with a simple CT scan, rather than a PET scan. Organomic features with the highest importance in the model (retrieved from its weights) showed that the lungs and esophagus were the organs containing the most relevant information.

The works produced during this thesis have been made publicly available to the research community, both in paper, code and tools. They provide researchers with a starting platform to lead new studies, with SoTA concepts, without the burden of implementation.

7.2 Perspectives

The research presented in this thesis has also served to define future perspectives for the use of A.I. to automatically analyze medical imaging for oncology. In this section, we will present the main trends for progress in this domain.

7.2.1 Building foundational medical image segmentation models

From the numerous small medical segmentation datasets, the trend seems to move towards large scale datasets, with annotations for a large number of classes. This trend is correlated with the increasing performance of DNNs in the domain of medical image segmentation:

- The performance of DNNs increases with more training data, medical image segmentation is no exception to this rule[289]. While the quality of the images, annotations and imaging protocol may influence the performance, real world data is messy and requires segmentation algorithms which great generalization capabilities. It is impossible to enclose the entirety of

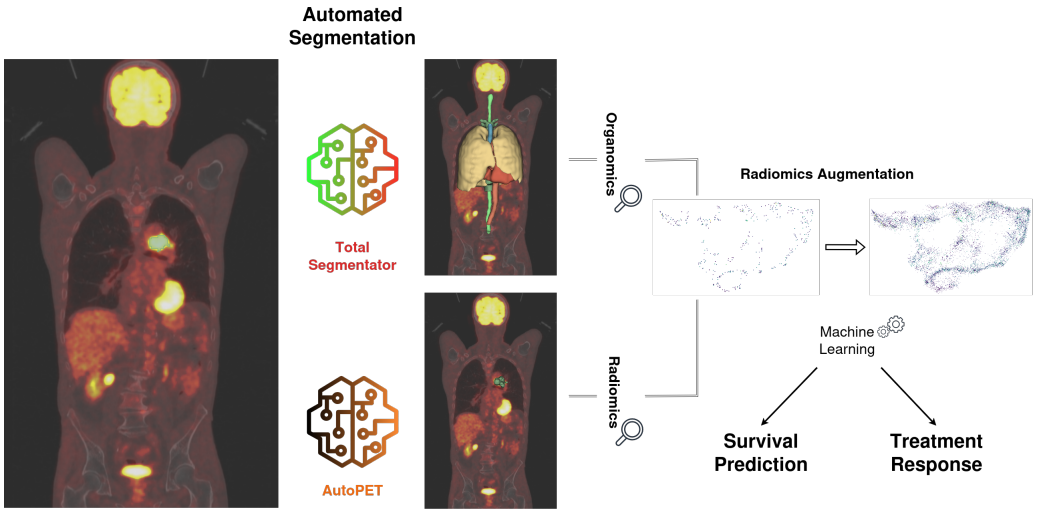

Figure 7.1: Using the tools we made publicly available, the automated processing of multi-modality medical images is made easier. Deep learning models can contour organs using anatomical imaging and lesions using PET/CT imaging. These contours can then be used to extract descriptive features, lesion radiomics and organomics, which can in turn be used to train machine learning models to predict outcomes such as survival and treatment response.

possible images in a training set, therefore obtaining a segmentation model fit for clinical practice relies on the generalization capabilities of DNNs.

- DNNs trained for segmentation on small datasets can be used in iterative learning or active learning, to annotate large datasets at a lower cost. A DNN trained on a small, high-quality dataset, can infer segmentation masks on a large number of medical images. These masks can then be used as a pseudo-annotation, to train a large DNN on all the previously unannotated images. Pseudo-labels can also be corrected by human experts, thus allowing an efficient annotation of large databases.

This two-way beneficial relationship between large scale dataset and DNN segmentation models is increasingly exploited, with the BodyMaps[290] project being at the forefront of this effort. However, large datasets are still only emerging for the task of segmentation, leaving other tasks scattered across multiple, small, often private datasets. We are still missing large datasets, built from real-world cohorts, including temporal clinical endpoints, such as survival, relapse and progression, as well as other clinical data (e.g.: anatomical pathology data, considered the ground truth in many cases).

7.2.2 Improving the state of survival prediction

Once such datasets will be made publicly available, the next step to improve the clinical usability of AI will be to use these datasets for survival prediction tasks. Indeed, these datasets would be necessary to enable the development and validation of reliable, robust and open survival prediction models. We have seen this process unfold for medical image segmentation: from the first breakthrough (UNet, 2015), to the first few large public datasets (AutoPET (2022/2025), TotalSegmen-

tator (2022)), SoTA models and training framework(nnUNet and MedNeXt, 2019/2023), a rigorous validation framework[109] and, finally, very large public datasets (AbdomenAtlas 3.0[291], 2026) and segmentation models performing better than humans (pancreas tumor segmentation, 2025[39]).

When it comes to survival prediction, models have been made available (Scikit-Survival[57], Icare[205]), but a clear consensus on the SoTA and the best training and evaluation practices has not been established yet. This lack of a stringent training and evaluation framework, along with the absence of a baseline, leads to the proliferation of methods, which further complicates the elaboration of a rigorous validation, as each new method should be evaluated.

Additionally, these methods are mostly developed and evaluated on small and private datasets, which does not ensure they can generalize to new data and prevents the research from being reproduced by other researchers. This absence of open, large datasets hinders the emergence of a SoTA method for survival prediction. Nevertheless, building large survival datasets is not out of reach, especially not for research teams working within a clinical environment.

Indeed, some countries, like France, maintain an open service which registers the dates of birth and death of deceased individuals[292]. Using these databases, clinicians could gather imaging exams from deceased patients and associate to it the survival time in days, starting from this exam. Once the imaging metadata is removed (and defacing was applied, if needed), these exam/time-to-death pairs could even be made publicly available, in accordance with privacy laws.

Considering the results presented in chapter 6, this method could be applied to CT scans, which are performed in large and growing amounts, in both oncology medical centers and imaging clinics. The inclusion of additional data types could strengthen the performance of developed methods: for instance, one of the perspectives for the RTEP7 study was the inclusion of dosiomics, radiomic features describing the dose delivered to different regions of a patient's body. Once large scale public datasets are available, research efforts should focus on the evaluation of the available survival prediction methods, to establish a robust SoTA. Some attempts have been made with clinical data from electronic health records (Electronic Health Records (EHR)), but to the best of our knowledge, not with imaging data.

7.2.3 Quantifying the whole human body

Our motivation to include organ-level imaging data in predictions stems from an observation: imaging, especially multi-modal imaging, gives a complete view and understanding of the human body to a physician. While focusing on lesions has helped understanding the relations between their features (shape, texture..) and the state of the underlying disease (stage, progression, aggressiveness), measuring the function and anatomy through the whole body gives a more holistic view of the patient's health.

To foster research towards the quantification of the whole human body, we published Organomics, a tool to extract organ-radiomic features from any imaging which included a CT scan. Possible improvements were discussed in chapter 6, to extend this tool to other imaging modalities. Beyond that, structured research could define one common set of organs, to contour and analyze, across multiple databases from clinical trial or early detection cohorts.

Such research could uncover systemic patterns, linked to increased risk of adverse events, or better treatment response. This information could enhance the information physicians get from imaging exams, further refining their treatment decisions.

Instead of considering each organ on its own, the relationships between organs (distance, differences in values, surface connection) could also be analyzed. Lesions could be included in this analysis, as it has been shown that the location of tumors, relative to surrounding organs, has an impact on the severity of the disease[293].

Tumors and organs could also be used to build graphs, a spatial representation of the function of a patient's organs as well as the distribution of the lesions. Graphs have already been used to encompass the spatial relationships between multiple lesions, in the analysis of PET imaging from lymphoma patients[294].

7.3 Closing Remarks

The works presented, discussed and referenced in this thesis show the impact that A.I. technologies have when applied to medical images, in particular in oncology. A.I. has made the analysis (segmentation, survival prediction) of large 3D medical images possible, especially through the use of quantified imaging, such as CT and PET. The integration of multiple registered imaging modalities (PET/CT, CT/MRI) has also shown promising results, by considering different functional and anatomical measures of the body.

Still, most of these methods never make it to clinical practice. Taking PET/CT segmentation as an example, threshold based methods are still used more often than DNNs.

This can be in part explained by the following roadblocks to A.I. adoption in clinical routine:

- **Limited performance**: In spite of their tirelessness, DNNs are still often falling short from, or only replicating human performance. This downgrade in performance is the first hangup to the adoption of these A.I. methods. Through large scale, efficiently curated datasets, research teams have already started to break this barrier. While perfect accuracy is rarely achievable in medicine, even by humans, robust performance have been thoroughly demonstrated for some use cases[39].

- **Processing speed and computational load**: Neural networks often require performant hardware (CPU/GPU) to be used and are slower than simple methods such as thresholding. However, optimizations are coming to DNNs, which let them run in near real-time[143].

- **Result explainability**: DNNs have often been described as a black box, which inner workings are undecipherable to the user. This property of DNNs has often hindered their adoption in critical domains. Nevertheless, if human supervision and validation is built into the applications (human-in-the-loop), this concern can be mitigated, as a human will always make the final decision.

The inclusion of A.I. in oncology requires models with super-human performance, fast enough to integrate seamlessly in the clinical workflow and encompassed in applications requiring human validation.

In conclusion, we have demonstrated that multi-modality imaging can be used for survival analysis and treatment response prediction. Methods have been developed to help with the next step of progress: the collection of large scale, high quality datasets.

Utilisation de l'intelligence artificielle en Imagerie Multi-modalité pour la Prédiction de la Réponse au Traitement et la Survie en Cancérologie - Application au Protocole RTEP7-IFCT 14.02

Intelligence Artificielle, Oncologie, Cancérologie, Imagerie Médicale, Traitement d'Image

Abstract

Multi-modality imaging has helped physicians to understand the human body and how it is affected by oncologic diseases. A growing number of imaging exams is performed each year and the practice of using these images for treatment has evolved towards personalized, precise care protocols.

To handle this large amount of data, research on artificial intelligence has been extended to medical images, to solve simplified versions of tasks from clinical practice. The contouring of structures, such as organs and lesions, in single or multi-modality imaging, has seen the most progress, through the collection of large annotated datasets and the publication of state-of-the-art segmentation tools. Automated and semi-automated contouring tools were the first successful applications of A.I. on medical images and were made possible thanks to community efforts.

This thesis presents how A.I. can be applied to accelerate the oncology workflow, using deep neural networks on multi-modality imaging for organs and lesions segmentation; and machine learning on imaging features to predict survival and support physician decision.

We developed an improved segmentation algorithm to **delineate organs-at-risk** in the head and neck area, on CT/MRI imaging. Our algorithm reached the third place in the HaN-Seg challenge, with a performance on par with human experts. This convincing result was obtained, in spite of the small size of the training dataset (42 patients), thanks to a specifically designed algorithm which relies on large scale pre-training and modified inference. Constraints on the runtime and inference hardware demonstrated the adaptability of SoTA segmentation methods based on deep neural networks to the conditions of clinical practice.

Focusing on nuclear imaging, we then tackled the task of **automated lesion segmentation on PET/CT** scans. For 3 years, we participated in the AutoPET challenge series and developed novel algorithms to improve over the SoTA and solve new problems inspired from clinical reality.

Our first algorithm achieved a higher performance than the baseline model trained with nnUNet, for automated lesion segmentation on FDG PET/CT scans.

The next year, we developed an algorithm capable of performing lesion segmentation on both PSMA and FDG PET/CT scans, using a tracer discriminator.

Finally, we adapted our algorithm to handle user input and perform semi-automated lesion segmentation. To feed user inputs to the DNN, we designed PET-aware lesion probability heatmaps which guided the segmentation.

Radiomic features extracted from the PET/CT contours can be used to forecast events, such as death or relapse. To improve the performance of machine learning models which use radiomic features as input, we developed **Radiomics Augmentation**.

This framework applies data augmentation transforms on the images and contours to create new synthetic samples. The goal of this framework is to reduce the over-fitting phenomenon, by filling holes in the data distribution, through the addition of modified but realistic samples.

Out of the 8 models we benchmarked, 7 showed an improved performance when using Radiomics Augmentation (from 0.2% to 8.6%) and 4 models exhibited a globally reduced over-fitting.

Going beyond lesion radiomics, we explored organ-level radiomics: **organomics**.

To this end, we developed and published Organ-omics, an automated organomics extraction pipeline, which relies on TotalSegmentator and PyRadiomics.

We applied Organ-omics to an internal dataset of NSCLC patients, which gathered PET/CT scans at 2 time-points for each patient. Using this dataset, we compared the predictive power of organomics, clinical data and lesion radiomics for overall survival.

We found that organomics largely outperformed clinical data ($p \ll 0.001$) and lesion radiomics ($p \ll 0.001$). On top of that, we showed that organomics extracted from CT images performed better than PET organomics ($p \ll 0.001$).

Through these contributions, we showed how A.I. could be applied to medical images in oncology. Deep Learning methods can generate accurate contours of organs and lesions from medical images. These contours can then be used to extract descriptive radiographic features such as radiomics or organomics. Radiomics and organomics can serve as inputs to machine learning models, to predict survival and treatment response.

The novel methods we have developed were made easily accessible, through the release of public tools.

Résumé

L'imagerie multi-modalité a ouvert la voie à une meilleure compréhension du corps humain, ainsi que de l'impact que les pathologies oncologiques ont sur celui-ci. Le nombre de ces examens a augmenté rapidement ces dernières années et les médecins les ont utilisés pour construire des protocoles de traitement adaptés à chaque patient.

Pour traiter cette quantité grandissante d'images, la recherche en intelligence artificielle a été étendue à l'imagerie médicale. Le contourage de structures, telles que les organes et les lésions, dans des images simples ou multi-modalités est le domaine qui a connu le plus de progrès. Ce progrès a été rendu possible notamment par la récolte d'importants jeux de données annotés, ainsi que par l'émergence d'outils puissants, servant de *baseline* robuste. Des outils de segmentation automatique et semi-automatique, la première application réussie de l'I.A. sur des images médicales, ont été construits grâce aux efforts collectifs de la communauté.

Cette thèse présente comment l'I.A. peut être appliquée pour accélérer le travail des médecins en oncologie, en utilisant des réseaux de neurones profonds pour contourer les organes et les lésions; ainsi que de l'apprentissage machine sur les caractéristiques extraites de ces contours, pour assister la décision médicale.

Nous avons développé un algorithme capable de contourer les organes à risque de la tête et du cou sur des images TDM/IRM. Cet algorithme a obtenu la troisième place dans le challenge HaN-Seg, avec une performance au niveau de celle d'experts humains. Ce résultat a été obtenu en dépit de la taille réduite du jeu de données d'entraînement (42 patients), grâce à l'utilisation de pré-entraînement à grande échelle et une inférence adaptative. Des contraintes placées sur le temps d'exécution et le matériel d'inférence ont permis de démontrer que les algorithmes à l'état de l'art peuvent être adaptés pour se conformer aux conditions de la routine clinique.

Nous nous sommes par la suite concentrés sur l'imagerie nucléaire, avec en premier lieu le contourage automatique des lésions sur les scanners TEP/TDM. Durant 3 années consécutives, nous avons contribué aux challenges AutoPET en développant des algorithmes innovants, afin d'améliorer l'état de l'art et de l'adapter aux nouveaux problèmes proposés par les organisateurs, inspirés de la réalité de la pratique clinique. Notre premier algorithme a obtenu une performance meilleure que la *baseline*, un modèle entraîné avec nnUNet, pour le contourage de lésions sur TEP/TDM au FDG. Pour le second algorithme, nous avons étendu les capacités de segmentation aux TEP/TDM au PSMA, avec une détermination automatique du type de traceur utilisé. Enfin, nous avons ajouté la capacité pour un utilisateur de guider l'algorithme, rendant le contourage semi-automatique, grâce à des cartes de probabilité innovantes, informées par la TEP.

Une fois les lésions contourées, des caractéristiques radiomiques peuvent être extraites de l'image pour prédire des évènements, tels que le décès et la rechute. Pour améliorer les performances des modèles d'apprentissage automatique qui utilisent les radiomiques comme entrées, nous avons développé *Radiomics Augmentation*. Ce framework crée de nouvelles données d'entraînement en appliquant des augmentations de données sur les images et les contours, avant d'extraire les radiomiques. L'objectif de ce framework est de réduire le sur-apprentissage afin d'améliorer les per-

formances des modèles, en remplissant les parts de la distribution initialement absentes du jeu d'entraînement. Sur 8 modèles que nous avons évalués, 7 ont montré des performances améliorées (de +0.2% à +8.6%) et 4 ont montré une réduction du sur-apprentissage.

Pour aller au-delà des lésions, nous avons exploré les radiomiques des organes: les organomiques. Pour cela, nous avons développé et publié Organ-omics, un pipeline d'extraction des organomiques, qui repose sur TotalSegmentator et PyRadiomics. Nous avons appliqué Organ-omics à un jeu de données issu d'une étude clinique incluant des patients souffrant de CBNPC. Deux imageries TEP/TDM ont été récoltées pour chaque patient, avant et pendant le traitement par radiothérapie. En comparant les performances prédictives des données cliniques, radiomiques de lésions et des organomiques, nous avons trouvé que les organomiques étaient beaucoup plus performantes que les données cliniques ($p \ll 0.001$) et les radiomiques ($p \ll 0.001$) pour prédire la survie globale. De plus, les organomiques extraites du CT ont obtenu de meilleurs résultats que celles extraites du PET ($p \ll 0.001$).

Grâce à ces travaux, nous avons démontré comment l'I.A. pouvait être appliquée à l'imagerie médicale en oncologie. Les méthodes d'apprentissage profond permettent de générer les contours des organes et des lésions à partir de ces images. Ces contours peuvent ensuite être utilisés pour extraire des caractéristiques radiographiques afin d'entraîner des modèles d'apprentissage automatique, pour prédire la survie et la réponse au traitement. Ainsi l'I.A. peut assister les décisions médicales.

Les nouvelles méthodes que nous avons développées sont facilement accessibles grâce à la mise à disposition d'outils publics.

List of Figures

1.1 A structural improvement can be achieved using A.I. to help doctors with tedious, repetitive and time-consuming tasks. Some tasks, such as lesion detection, organ segmentation or even imaging report generation can be automated, saving the physician's time for the more complex medical decisions. 3

2.1 Artificial Intelligence covers methods relying on logic to process an input. Machine learning specifically designates methods which learn from sets of example data, trying to generalize to all possible cases. Machine learning with deep neural networks falls in the deep learning category which often requires large datasets for training. 9

2.2 A $3 * 3$ convolution applied to a $4 * 4$ matrix, yielding a $2 * 2$ matrix. Taking the top left corner of that resulting matrix as an example: it is the result of the combination of values in the $3 * 3$ top left corner of the original matrix, by the weights of the 3x3 convolution kernel. 10

2.3 The main family of generation models: a noise is given to the decoder of a pre-trained auto-encoder, which generates a sample corresponding to its training distribution . . 11

2.4 The objective goal of classification: assign a class to a signal: image, text, audio, scan... The classes are given semantic meaning by humans, models only handle integers. 12

2.5 Segmentation consists in extracting the contours of objects from an image. Training a DNN for this task consists in comparing the predicted contours with the ground truth annotation. 13

2.6 Regression methods are often similar to classification, with the difference that they predict continuous values. After the regression model has built a function to map the input to the output, any new data point can be assigned a value through this function. 14

2.7 DICOM files store extensive metadata relating to the patient and image acquisition. NIfTI on the other hand only stores the information needed to reconstruct the patient's space in its metadata 19

2.8 An example of a CT scan (s0591 from TotalSegmentator) which has been defaced for patient privacy. On the left: 3D volume rendering. On the right: 2D coronal slice, showing the blurring added to the face area. 23

3.1 Starting from an image, segmentation consists in classifying the pixels of an image (or voxels of a volume) depending on the object they are a part of. Semantic segmentation groups all instances of this object in the same class (e.g.: *"person"*) whereas instance segmentation assigns different labels to each instance of the object (e.g.: each person). *Source: learnopencv.com* 29
3.2 Original architecture of the UNet. Source: Ronneberger et al.[17] 30
3.3 The Dice Similarity Coefficient (DSC) is a metric which compares the intersection of two geometric shapes with the summed area or volume of these two shapes. The Hausdorff Distance (HD) is the maximum of the set of minimum distances between any point in the predicted contour and a point in the reference contour. 31
3.4 Visual Transformer (ViT) tokenizes images through the selection of patches, which are subsequently flattened and embedded. *Source: Dosovitskiy et al.[86]* 34
3.5 The selective structured state space model (Mamba) uses an architecture resembling a recurrent network with a hidden state h dynamically queried by the input x. *Source: Gu et al.[88]* 35
3.6 Side by side comparison of CT and MRI images for the same patient from the HaN-Seg dataset. While the CT has higher values in the high density areas, MRI gathers more signal in areas containing liquid, such as the brain and muscles. 41
3.7 Organs at risk contoured for each patient in the HaN-Seg dataset. Left/right distinction is applied, bringing the number of segmentation labels to thirty. *Source: han-seg2023.grand-challenge.org* 43
3.8 A summary of our proposed algorithm for fully automated HN OAR segmentation from unregistered CT/MRI. This figure represents the inference step of our algorithm once the UNet segmentation has been trained. Details on both the training and inference steps are provided below. 44
3.9 DSC and HD95 for all organs, for all algorithms submitted to HaN-Seg. Our solution is highlighted in red. 47

4.1 Line Of Response (LOR) and the sinogram obtained as a result of the addition of LORs. *Source: Volkow et al.[171]* 54
4.2 A Maximum Intensity Projection (MIP) of a normal (i.e.: benign) 18F-FDG PET scan[172]. Radio-tracer uptake is visible in the brain and myocardium (heart muscle) which are constantly consuming the glucose analogue, as well as the kidneys and urinary bladder, which respectively filter and stores urine containing the radio-tracer. Thyroid uptake can also be seen in this example, while it is mostly not a cause for concern, it might be the sign of a non-cancerous disease[173]. *Source: radiopaedia.org* 54
4.3 3D rendering of the organ contours used as additional supervision. These contours were segmented on the CT of each patient from the training set, using TotalSegmentator 61
4.4 The heatmap creation algorithm. While the final heatmap is a single volume, the positive and negative clicks were processed separately. *Credit: Solène Perret.* 63
4.5 3D rendering of the parotid glands (purple) and sub-mandibular glands (pink). These organs were added to the list of supervision organs, as they often exhibit an uptake of PSMA tracer. 64

4.6 A summary of the novelties proposed and included in our methods. While each edition led to its own improvements, these improvements were iteratively integrated to build a solutions which benefited from all of them. 65

5.1 Radiomic features are computed using the values of the medical images inside contours. Different classes of radiomics exist, each measuring a specific aspect of the contour and/or the image 74

5.2 Survival is defined as the time (often in days) between the beginning of a treatment and the occurrence of an event (death, relapse, progression). A patient survival can be censored if they did not undergo an event before the end of the study, or if they dropped out of the study. In these cases, the patient is marked as *censored* and the survival time is registered as the time to end of study or the time to last news, respectively. 77

5.3 A toy demonstration of what happens when a model over-fits. Over the course of its training, it will first start improving on both its training and validation set (loss decreases, accuracy increases). After some time, performance on the training data will keep on improving, while the progress on the validation set will start to slow down, stop and eventually revert. This is a clear sign of overfitting, quantified by the difference between training and validation accuracies (red arrow). In the figure above, the moment the model starts overfitting is marked by a red dotted line. . . . 80

5.4 Visual representation of the transforms applied on the images and masks. The green contour is the original manual contour, while the yellow, blue and red are the contours modified by DA methods. The parameters of all the image transformations were set to extreme values for visualization purposes. 84

5.5 The *Radiomics Augmentation* Framework. Transformed are applied to masks and images to create augmented samples, which are added to the original samples. Augmented radiomics and original radiomics are then benchmarked, on the task of RFS prediction using ML models. *Masks are not represented here.* 87

5.6 Results of the UMAP projection applied on the original and augmented radiomic features from the HECKTOR22 dataset. In the picture on the left, the 2D projections of radiomics vectors for all the original cases are represented. On the right, the projection for the augmented cases are also represented. Color coding of points with RFS is mainly used to enhance readability. 87

5.7 Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for some of the benchmarked models. Dashed and solid lines represent the performance using basic radiomics lines and augmented radiomics respectively. CI_{train} values are displayed with more transparency. See Appendix 1 for the remaining models. 89

6.1 Flowchart describing the RTEP7 IFCT-1402 protocol: all patients included undergo a first PET/CT scan (PET/CT1) before being randomly assigned into either the experimental or the control arm. In the control arm the classical 66Gy RT treatment was delivered to all patients but a second PET/CT was acquired (PET/CT2). In the experimental arm, the second PET/CT was used to decide a boost of RT to 74Gy. . 101

6.2 Overview of the feature extraction pipeline. Organ contours were extracted from the CT scans using TS. The BTVs were manually annotated by nuclear physicians. Using the PET and CT volumes as 3D matrices of voxels and the lesion and organ contours, we extracted the radiomics and organomics respectively. We obtained two sets of each, one for the pre-treatment PET/CT and one for the per-treatment PET/CT. A third set (Δ F), was computed as the difference between the features at the two time points. 103
6.3 We built a 6-folds cross-validation to evaluate the performance of our approach (organomic features) against radiomic features and clinical data. The data was split in 6 folds, each used once for testing and five times as part of the training set. Once the model was trained on 5 folds, it was then evaluated on the remaining fold. This CV was repeated 100 times, to average the results for more stability. 104

7.1 Using the tools we made publicly available, the automated processing of multi-modality medical images is made easier. Deep learning models can contour organs using anatomical imaging and lesions using PET/CT imaging. These contours can then be used to extract descriptive features, lesion radiomics and organomics, which can in turn be used to train machine learning models to predict outcomes such as survival and treatment response. 115

2 Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for all of the benchmarked models. Dashed lines represent the performance using basic radiomics and solid lines represent the performance using augmented radiomics. CI_{train} values are displayed with more transparency. 129
3 Number of patients per site involved in the RTEP7 IFCT-1402 clinical trial. In our study, 138 patients with PET/CT 1 and 2 are used. These patients are distributed as seen in this figure, across 19 sites. 130

List of Tables

2.1 A summary of large datasets, their modalities, annotations and availability to the public 25

3.1 Novelties implemented in our methods and the related improvements 47

5.1 C-Index for each model, using basic and augmented radiomics. In the last column the percent change in performance is added. 88

6.1 Demographic data from the patients included in the RTEP7 protocol. 106

6.2 All data combinations tested as input to our benchmarking algorithm are presented in this table. A check mark means this type of data was used as input to the model, red if this type was not selected at all, green if some features were selected (with the number of selected features in parenthesis). The total number of features included in each model is indicated in the second column. 108

1 Weight of the features in all tested configurations. 136

2 Results from the biomarkers reported in Vera et al.[273] 136

Appendix 1

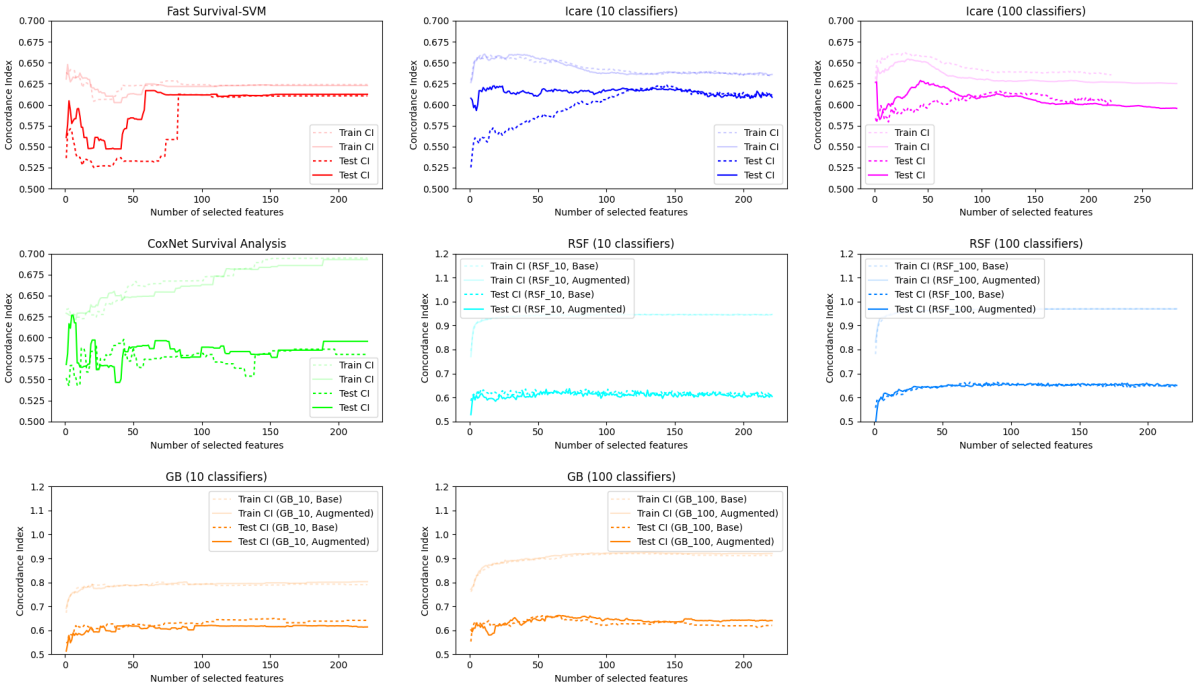

Figure 2: Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for all of the benchmarked models. Dashed lines represent the performance using basic radiomics and solid lines represent the performance using augmented radiomics. CI_{train} values are displayed with more transparency.

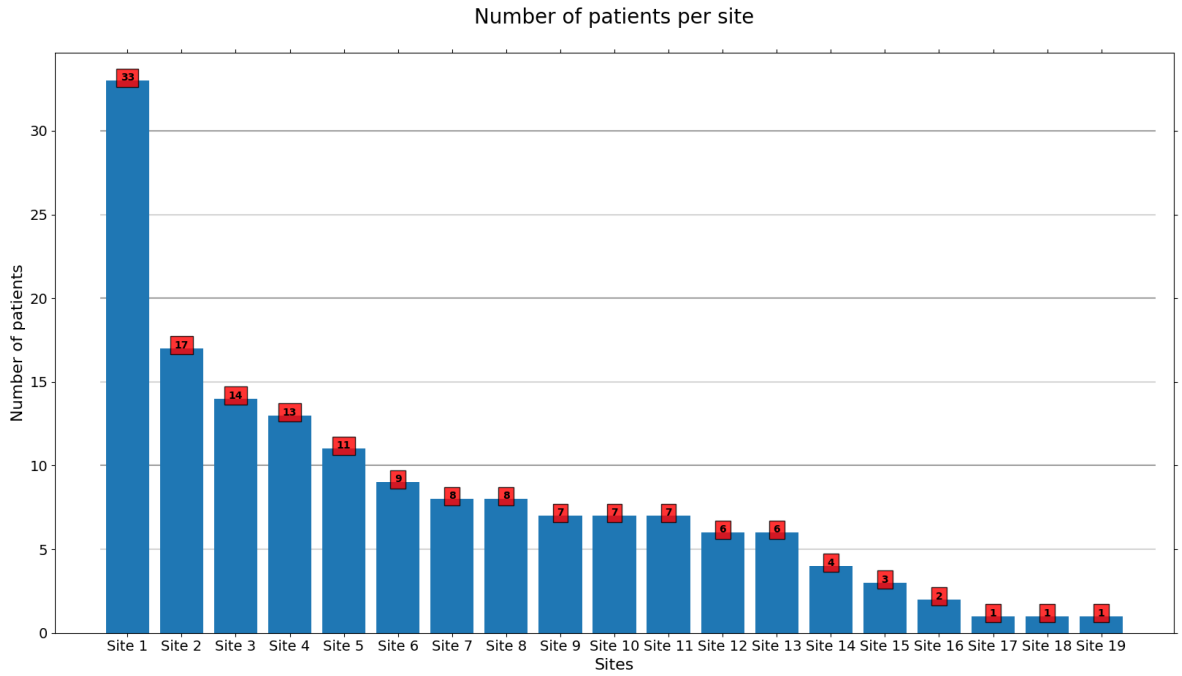

Figure 3: Number of patients per site involved in the RTEP7 IFCT-1402 clinical trial. In our study, 138 patients with PET/CT 1 and 2 are used. These patients are distributed as seen in this figure, across 19 sites.

Data type	Selected features: Weight in the model
Clinical	Boosted(Yes/No): 0.171288273505359 Experimental: 0.1266781247255013 Age: -0.009231904454067517
Radiomics_CT	Delta_CT_lesion_FirstOrder_Kurtosis: -0.0032568771536971375 Delta_CT_lesion_Shape_LeastAxisLength: -0.00039481734740006943 Delta_CT_lesion_Shape_Maximum2DDiameterColumn: -0.0001350578662420905 Delta_CT_lesion_Shape_Maximum2DDiameterSlice: -0.0016998602619281659 Delta_CT_lesion_GLCM_Imc1: -1.0303926436434996e-05 Delta_CT_lesion_GLCM_Imc2: 1.1481533776979401e-05 Delta_CT_lesion_GLSZM_ZonePercentage: -3.3800031393478217e-06 Delta_CT_lesion_GLRLM_LongRunEmphasis: -3.498460551430704e-05 Delta_CT_lesion_GLRLM_RunPercentage: -1.2208911867992713e-06 Delta_CT_lesion_GLRLM_RunVariance: -2.2156686137317874e-05 Delta_CT_lesion_NGTDM_Coarseness: 2.5011040909550846e-06 Delta_CT_lesion_GLDM_LargeDependenceEmphasis: -5.2166898207662085e-05 CT1_CT_lesion_FirstOrder_10Percentile: -0.0003840627338612997 CT1_CT_lesion_FirstOrder_90Percentile: 0.0022559887149257106 CT1_CT_lesion_FirstOrder_InterquartileRange: -0.0016095446293638155 CT1_CT_lesion_FirstOrder_MeanAbsoluteDeviation: 0.0005209625420466566 CT1_CT_lesion_FirstOrder_Mean: 0.0015778784981024855 CT1_CT_lesion_FirstOrder_Median: 0.001690273082394749 CT1_CT_lesion_FirstOrder_Minimum: 0.0007713914699874162 CT1_CT_lesion_FirstOrder_RobustMeanAbsoluteDeviation: -0.0002854281930327365 CT1_CT_lesion_Shape_Sphericity: 1.5287338506441425e-05 CT1_CT_lesion_GLCM_ClusterTendency: 0.000853885756135136 CT1_CT_lesion_GLRLM_LongRunLowGrayLevelEmphasis: 1.1534879446499722e-06 CT1_CT_lesion_GLSZM_SizeZoneNonUniformityNormalized: -1.100897457037568e-06 CT1_CT_lesion_GLSZM_SmallAreaEmphasis: -1.10107794242329e-06 CT1_CT_lesion_GLDM_GrayLevelVariance: 0.0002617452227458239 CT2_CT_lesion_Shape_Sphericity: 1.3960118206453879e-05 CT2_CT_lesion_GLCM_Imc1: -1.3244254635373948e-05 CT2_CT_lesion_NGTDM_Coarseness: 2.702887148588082e-06
Radiomics_PET + Clinical	Boosted(Yes/No): 0.0006096753140544884 Experimental: 0.0005238675087472743 Age: -0.018636101978460818 CT1_PET_lesion_FirstOrder_Median: 0.0069287535866368564 CT1_PET_lesion_FirstOrder_Skewness: -0.0012409306531805113

	CT1_PET_lesion_Shape_Sphericity: 0.0003724910241868513 CT2_PET_lesion_FirstOrder_Skewness: 0.0008195294870407045 CT2_PET_lesion_FirstOrder_Minimum: 0.0006901713748275661 CT2_PET_lesion_Shape_Sphericity: -6.323508745199023e-05 CT2_PET_lesion_Shape_Elongation: 2.3327838594512688e-05 CT2_PET_lesion_GLSZM_SizeZoneNonUniformityNormalized: -0.00020021863003267122 CT2_PET_lesion_GLSZM_ZonePercentage: 5.4885642431640434e-06 CT2_PET_lesion_GLDM_DependenceNonUniformityNormalized: 2.7571468714682573e-05 Delta_PET_lesion_FirstOrder_Range: -0.004368820910178274 Delta_PET_lesion_Shape_LeastAxisLength: -0.0036352682152495483 Delta_PET_lesion_Shape_Maximum2DDiameterColumn: -0.0028169884755085517 Delta_PET_lesion_Shape_Maximum2DDiameterRow: 0.00787731805122173 Delta_PET_lesion_Shape_Maximum2DDiameterSlice: -0.0026880430842161113 Delta_PET_lesion_Shape_MinorAxisLength: -0.004744492414281289 Delta_PET_lesion_Shape_Sphericity: -3.737277297193055e-05 Delta_PET_lesion_GLSZM_ZonePercentage: 8.462817151719552e-06 Delta_PET_lesion_GLRLM_RunEntropy: -0.00038969404068608037 Delta_PET_lesion_GLRLM_RunLengthNonUniformity: -0.00027745705189633364
Radiomics_PET + Radiomics_CT	CT1_PET_lesion_Shape_Sphericity: 2.1493136208493814e-05 CT1_CT_lesion_FirstOrder_10Percentile: -0.0006578560868302722 CT1_CT_lesion_FirstOrder_90Percentile: 0.0038673858780106938 CT1_CT_lesion_FirstOrder_InterquartileRange: -0.0025961901346676242 CT1_CT_lesion_FirstOrder_MeanAbsoluteDeviation: 0.0015359408743502426 CT1_CT_lesion_FirstOrder_Mean: 0.0014227842513486874 CT1_CT_lesion_FirstOrder_Median: 0.0009134501840995246 CT1_CT_lesion_FirstOrder_Minimum: 0.0009640326261549107 CT1_CT_lesion_FirstOrder_RobustMeanAbsoluteDeviation: -0.00045094667497230665 CT1_CT_lesion_Shape_Sphericity: 2.1493136208493814e-05 CT1_CT_lesion_GLCM_ClusterTendency: 0.0009776229782893396 CT1_CT_lesion_GLRLM_LongRunLowGrayLevelEmphasis: 1.9368369022262248e-06 CT1_CT_lesion_GLSZM_SizeZoneNonUniformityNormalized: 8.909323014434347e-06 CT1_CT_lesion_GLSZM_SmallAreaEmphasis: 5.2624599527825166e-06 CT2_PET_lesion_Shape_Sphericity: 2.629865773106539e-05 CT2_PET_lesion_FirstOrder_Minimum: 0.0001618207996398443 8 CT2_PET_lesion_FirstOrder_Skewness: 8.673240079539086e-05

	CT2_PET_lesion_GLDM_DependenceNonUniformityNormalized: 1.1832890000454763e-05 CT2_CT_lesion_Shape_Sphericity: 2.629865773106539e-05 CT2_CT_lesion_GLCM_Imc1: -2.6681546052600872e-05 CT2_CT_lesion_GLSZM_SmallAreaLowGrayLevelEmphasis: 2.9904106418649724e-06 CT2_CT_lesion_NGTDM_Coarseness: 7.318850987978018e-06 Delta_PET_lesion_Shape_LeastAxisLength: -0.00013144561062011946 Delta_PET_lesion_Shape_Maximum2DDiameterSlice: -0.0020071546934004236 Delta_PET_lesion_Shape_Maximum2DDiameterColumn: 0.0002976925575858717 Delta_CT_lesion_FirstOrder_Kurtosis: -0.0041461855705722745 Delta_CT_lesion_Shape_LeastAxisLength: -0.00013144561062011946 Delta_CT_lesion_Shape_Maximum2DDiameterColumn: 0.0002976925575858717 Delta_CT_lesion_Shape_Maximum2DDiameterSlice: -0.0020071546934004236 Delta_CT_lesion_GLCM_Imc1: -2.0533514059037472e-05 Delta_CT_lesion_GLCM_Imc2: 1.8431305988523626e-05 Delta_CT_lesion_GLSZM_ZonePercentage: 8.755021508268755e-06 Delta_CT_lesion_GLRLM_LongRunEmphasis: -5.705377590129127e-06 Delta_CT_lesion_GLRLM_RunPercentage: 2.997145851995086e-06 Delta_CT_lesion_GLRLM_RunVariance: 6.213222342450179e-06 Delta_CT_lesion_NGTDM_Coarseness: 6.900844894628752e-06 Delta_CT_lesion_GLDM_LargeDependenceEmphasis: -0.0007196936529842788
Organomics_PET	CT1_PET_heart_FirstOrder_10Percentile: 9.791396502471734e-21 CT1_PET_esophagus_FirstOrder_10Percentile: 6.5435899426036505e-21 CT1_PET_heart_GLSZM_ZonePercentage: 8.901856193645206e-24 CT1_PET_heart_GLSZM_ZoneVariance: 7.761995176254995e-14 Delta_PET_lungs_Shape_Elongation: 3.6199691917419645e-23 Delta_PET_lungs_Shape_Maximum2DDiameterColumn: 1.0984488104675428e-20 Delta_PET_esophagus_Shape_MeshVolume: 6.070310208136325e-18 Delta_PET_esophagus_Shape_SurfaceVolumeRatio: -3.1840553452068144e-23 Delta_PET_esophagus_Shape_VoxelVolume: 6.130553607547314e-18 Delta_PET_esophagus_GLSZM_LargeAreaEmphasis: 7.786413303672785e-16 Delta_PET_esophagus_GLSZM_LargeAreaHighGrayLevelEmphasis: 7.785744872512045e-16 Delta_PET_esophagus_GLSZM_LargeAreaLowGrayLevelEmphasis: 7.786419820452987e-16 Delta_PET_esophagus_GLRLM_LongRunEmphasis: 4.108533126358524e-21 Delta_PET_esophagus_GLRLM_LongRunHighGrayLevelEmphasis: -1.5426700486802076e-21 Delta_PET_esophagus_GLRLM_LongRunLowGrayLevelEmphasis: 4.157216609114309e-21 Delta_PET_esophagus_GLRLM_RunPercentage: -5.087818592371667e-23

	Delta_PET_esophagus_GLRLM_ShortRunEmphasis: -6.893172316631042e-23 Delta_PET_esophagus_GLRLM_ShortRunHighGrayLevelEmphasis: -4.458711241246022e-21 Delta_PET_esophagus_GLDM_DependenceEntropy: -1.472581442018825e-22 Delta_PET_esophagus_GLDM_GrayLevelNonUniformity: 2.70903692558928e-19 Delta_PET_esophagus_GLDM_LargeDependenceEmphasis: 4.118265642224025e-20 Delta_PET_esophagus_GLDM_LargeDependenceHighGrayLevelEmphasis: -1.3022372005008692e-21 Delta_PET_esophagus_GLDM_LargeDependenceLowGrayLevelEmphasis: 4.15961221178467e-20 Delta_PET_esophagus_GLDM_SmallDependenceEmphasis: -1.1708038523289427e-23 Delta_PET_esophagus_GLDM_SmallDependenceHighGrayLevelEmphasis: -1.5491785480720967e-21 Delta_PET_esophagus_GLDM_SmallDependenceLowGrayLevelEmphasis: -5.349820480857711e-25 Delta_PET_heart_FirstOrder_Energy: -2.3314528521837266e-13 Delta_PET_heart_FirstOrder_Maximum: -6.093350592874768e-20 Delta_PET_heart_FirstOrder_MeanAbsoluteDeviation: -6.5165810455090574e-21 Delta_PET_heart_FirstOrder_Mean: -1.7807459558243763e-20 Delta_PET_heart_FirstOrder_Median: -1.5308594169118138e-20 Delta_PET_heart_FirstOrder_Range: -5.994172717986347e-20 Delta_PET_heart_FirstOrder_RootMeanSquared: -1.9774179018247235e-20 Delta_PET_heart_FirstOrder_TotalEnergy: -6.29492270089606e-12 Delta_PET_heart_GLSZM_ZoneVariance: -6.740865552856904e-14
Organomics_CT	Delta_CT_lungs_GLSZM_GrayLevelNonUniformity: 0.001480608200934843 Delta_CT_esophagus_Shape_MeshVolume: 3.476076284381402e-05
Organomics_PET + Organomics_CT	CT1_PET_esophagus_FirstOrder_10Percentile: 4.426930510587905e-18 CT1_PET_heart_FirstOrder_10Percentile: 8.115473675218833e-18 CT1_PET_heart_GLSZM_ZonePercentage: 2.602516865485936e-22 CT1_PET_heart_GLSZM_ZoneVariance: 1.7710091256193482e-09 Delta_CT_thoracic_vertebrae_GLCM_Imc2: -1.0233652652751714e-18 Delta_CT_thoracic_vertebrae_GLCM_MCC: -2.3217375678424126e-18 Delta_CT_thoracic_vertebrae_GLDM_DependenceEntropy: -2.045771695473691e-18 Delta_CT_lungs_Shape_Elongation: 8.246324537294979e-19 Delta_CT_lungs_Shape_Maximum2DDiameterColumn: 3.2579888399839104e-16 Delta_CT_lungs_Shape_MinorAxisLength: 2.726689624706136e-16 Delta_CT_lungs_GLSZM_GrayLevelNonUniformity: 4.729342444658447e-15 Delta_CT_lungs_GLSZM_SizeZoneNonUniformity: 9.307406765785618e-14 Delta_CT_lungs_GLRLM_RunLengthNonUniformity: 2.5232535722702586e-13

	Delta_CT_esophagus_Shape_MeshVolume: 1.514549205993131e-13 Delta_CT_esophagus_Shape_SurfaceVolumeRatio: -6.351702334844627e-19 Delta_CT_esophagus_Shape_VoxelVolume: 1.5294292901317486e-13 Delta_CT_esophagus_GLSZM_LargeAreaEmphasis: 1.5449199005026775e-14 Delta_CT_esophagus_GLSZM_ZoneVariance: 1.484892560395159e-14 Delta_CT_heart_GLCM_Imc2: -3.12815712988472e-18 Delta_PET_lungs_Shape_Elongation: 5.255912464084415e-19 Delta_PET_lungs_Shape_Maximum2DDiameterColumn: 2.468202218870801e-16 Delta_PET_esophagus_Shape_MeshVolume: 1.1958729797121086e-13 Delta_PET_esophagus_Shape_SurfaceVolumeRatio: -6.192831052948948e-19 Delta_PET_esophagus_Shape_VoxelVolume: 1.2005020971765432e-13 Delta_PET_esophagus_GLSZM_LargeAreaEmphasis: 1.8100450042516475e-11 Delta_PET_esophagus_GLSZM_LargeAreaHighGrayLevelEmphasis: 1.8100450042516475e-11 Delta_PET_esophagus_GLSZM_LargeAreaLowGrayLevelEmphasis: 1.8100450042516475e-11 Delta_PET_esophagus_GLRLM_LongRunEmphasis: 8.920671489091207e-17 Delta_PET_esophagus_GLRLM_LongRunHighGrayLevelEmphasis: 8.920671489091207e-17 Delta_PET_esophagus_GLRLM_LongRunLowGrayLevelEmphasis: 8.920671489091207e-17 Delta_PET_esophagus_GLRLM_RunPercentage: -5.904783389872639e-19 Delta_PET_esophagus_GLRLM_ShortRunEmphasis: -1.0958243240792942e-18 Delta_PET_esophagus_GLRLM_ShortRunHighGrayLevelEmphasis: -1.0958243240792942e-18 Delta_PET_esophagus_GLDM_DependenceEntropy: -1.8573182655752213e-18 Delta_PET_esophagus_GLDM_GrayLevelNonUniformity: 4.446304063616826e-15 Delta_PET_esophagus_GLDM_LargeDependenceEmphasis: 5.549072467113804e-16 Delta_PET_esophagus_GLDM_LargeDependenceHighGrayLevelEmphasis: 5.549072467113804e-16 Delta_PET_esophagus_GLDM_LargeDependenceLowGrayLevelEmphasis: 5.549072467113804e-16 Delta_PET_esophagus_GLDM_SmallDependenceEmphasis: -1.1166101354353392e-20 Delta_PET_esophagus_GLDM_SmallDependenceHighGrayLevelEmphasis: -1.1166101354353392e-20 Delta_PET_esophagus_GLDM_SmallDependenceLowGrayLevelEmphasis: -1.1166101354353392e-20 Delta_PET_heart_FirstOrder_Energy: -3.902136265766758e-12 Delta_PET_heart_FirstOrder_Maximum: -1.5127075682949365e-16

	Delta_PET_heart_FirstOrder_MeanAbsoluteDeviation: -1.1055737068081332e-17 Delta_PET_heart_FirstOrder_Mean: -1.7198263056317378e-17 Delta_PET_heart_FirstOrder_Median: -9.551843437991132e-18 Delta_PET_heart_FirstOrder_Range: -1.4733492970559412e-16 Delta_PET_heart_FirstOrder_RootMeanSquared: -2.1527788620860757e-17 Delta_PET_heart_FirstOrder_TotalEnergy: -1.0535767917570244e-10 Delta_PET_heart_GLSZM_ZoneVariance: -1.9640574584296235e-09

Table 1: Weight of the features in all tested configurations.

Included Features	C-Index	cdAUC
PET 2 SUV_{Max}	0.413 ($\pm$0.015)	0.398($\pm$0.019)
$(PET1PET2\Delta)PETSUV_{Max}$	0.505($\pm$0.012)	0.504($\pm$0.016)

Table 2: Results from the biomarkers reported in Vera et al.[273]

Bibliography

[1] Mariana Fernandez. High-end Global Computed Tomography Purchases to Propel the High-end CT Segment Revenue, September 2021.

[2] The National Lung Screening Trial: Overview and Study DesignRadiology. https://pubs.rsna.org/doi/abs/10.1148/radiol.10091808.

[3] Jens Vogel-Claussen, Florian Lasch, Benjamin-Alexander Bollmann, Katharina May, Alexander Kuhlmann, Gerald Schmid-Bindert, Rudolf Kaaks, Jörg Barkhausen, Sabine Bohnet, and Martin Reck. Design and Rationale of the HANSE Study: A Holistic German Lung Cancer Screening Trial Using Low-Dose Computed Tomography. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 194(12):1333–1345, December 2022. ISSN 1438-9029, 1438-9010. doi: 10.1055/a-1853-8291.

[4] Alain Bernard and Hani Shennib. Resection of pulmonary nodules using video-assisted thoracic surgery. *The Annals of Thoracic Surgery*, 61(1):202–205, January 1996. ISSN 0003-4975. doi: 10.1016/0003-4975(95)01014-9.

[5] Cancer Tomorrow. https://gco.iarc.who.int/today/, .

[6] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5):e230024, September 2023. doi: 10.1148/ryai.230024.

[7] Yazdan Salimi, Ghasem Hajianfar, Zahra Mansouri, Amirhosein Sanaat, Mehdi Amini, Isaac Shiri, and Habib Zaidi. Organomics: A Concept Reflecting the Importance of PET/CT Healthy Organ Radiomics in Non–Small Cell Lung Cancer Prognosis Prediction Using Machine Learning. *Clinical Nuclear Medicine*, 49(10):899, October 2024. ISSN 0363-9762. doi: 10.1097/RLU.0000000000005400.

[8] Gašper Podobnik, Bulat Ibragimov, Elias Tappeiner, Chanwoong Lee, Jin Sung Kim, Zacharia Mesbah, Romain Modzelewski, Yihao Ma, Fan Yang, Mikołaj Rudecki, Marek Wodziński, Primož Peterlin, Primož Strojan, and Tomaž Vrtovec. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation challenge. *Radiotherapy and Oncology*, 198:110410, September 2024. ISSN 0167-8140. doi: 10.1016/j.radonc.2024.110410.

[9] Jakob Dexl, Sergios Gatidis, Marcel Früh, Katharina Jeblick, Andreas Mittermeier, Anna Theresa Stüber, Balthasar Schachtner, Johanna Topalis, Matthias P. Fabritius, Sijing Gu, Gowtham Krishnan Murugesan, Jeff VanOss, Jin Ye, Junjun He, Anissa Alloula, Bartłomiej W. Papież, Zacharia Mesbah, Romain Modzelewski, Matthias Hadlich, Zdravko Marinov, Rainer Stiefelhagen, Fabian Isensee, Klaus H. Maier-Hein, Adrian Galdran, Konstantin Nikolaou, Christian la Fougère, Moon Kim, Nico Kallenberg, Jens Kleesiek, Ken Herrmann, Rudolf Werner, Michael Ingrisch, Clemens C. Cyran, and Thomas Küstner. AutoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging, Part 2: Domain Generalization. *Journal of Nuclear Medicine*, December 2025. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.125.270260.

[10] Zacharia Mesbah, Léo Mottay, Romain Modzelewski, Pierre Decazes, Sébastien Hapdey, Su Ruan, and Sébastien Thureau. AutoPETIII: The Tracer Frontier. What Frontier?, September 2025.

[11] Zacharia Mesbah and Solène Perret. QuantIF: Semi-automatic PET/CT segmentation across tracers.

[12] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 1939-1471. doi: 10.1037/h0042519.

[13] Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160, June 1976. ISSN 1572-9125. doi: 10.1007/BF01931367.

[14] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. ISSN 1432-0770. doi: 10.1007/BF00344251.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

[21] Decomposing Language Models Into Understandable Components. https://www.anthropic.com/news/decomposing-language-models-into-understandable-components.

[22] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don't Always Say What They Think.

[23] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, December 2024.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[25] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic Video Generation with Diffusion Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 393–411, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72986-7. doi: 10.1007/978-3-031-72986-7_23.

[26] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation, March 2023.

[27] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. 3D MRI Synthesis with Slice-Based Latent Diffusion Models: Improving Tumor Segmentation Tasks in Data-Scarce Regimes. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. doi: 10.1109/ISBI56570.2024.10635533.

[28] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The Curse of Recursion: Training on Generated Data Makes Models Forget, April 2024.

[29] Adyasha Sahu, Pradeep Kumar Das, and Sukadev Meher. An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. *Biomedical Signal*

Processing and Control, 87:105377, January 2024. ISSN 1746-8094. doi: 10.1016/j.bspc.2023.105377.

[30] Devinder Kumar, Alexander Wong, and David A. Clausi. Lung Nodule Classification Using Deep Features in CT Images. In *2015 12th Conference on Computer and Robot Vision*, pages 133–138, June 2015. doi: 10.1109/CRV.2015.25.

[31] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, March 2021. ISSN 1558-0016. doi: 10.1109/TITS.2020.2972974.

[32] Dejene M. Sime, Guotai Wang, Zhi Zeng, and Bei Peng. Deep learning-based automated steel surface defect segmentation: A comparative experimental study. *Multimedia Tools and Applications*, 83(1):2995–3018, January 2024. ISSN 1573-7721. doi: 10.1007/s11042-023-15307-y.

[33] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Internal Representations by Error Propagation. (ICS8506), September 1985.

[34] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, January 2022.

[35] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of Significance: Model selection and overfitting. *Nature Methods*, 13(9):703–705, September 2016. ISSN 15487091.

[36] Amine Amyar, Romain Modzelewski, Hua Li, and Su Ruan. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. *Computers in Biology and Medicine*, 126:104037, November 2020. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.104037.

[37] Karol Gotkowski, Klaus H. Maier-Hein, and Fabian Isensee. Revisiting 3D Medical Scribble Supervision: Benchmarking Beyond Cardiac Segmentation, August 2025.

[38] Dong-Hyun Lee. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

[39] Wenxuan Li, Pedro R. A. S. Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, Xiaoxi Chen, Lizhou Wu, Qilong Wu, Yannick Kirchhoff, Maximilian Rokuss, Saikat Roy, Yuxuan Zhao, Dexin Yu, Kai Ding, Constantin Ulrich, Klaus Maier-Hein, Yang Yang, Alan L. Yuille, and Zongwei Zhou. ScaleMAI: Accelerating the Development of Trusted Datasets and AI Models, January 2025.

[40] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan L. Yuille, and Zongwei Zhou. AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks. *Advances in Neural Information Processing Systems*, 36:36620–36636, December 2023.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[42] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked Image Modeling: A Survey. *International Journal of Computer Vision*, 133(10):7154–7200, October 2025. ISSN 1573-1405. doi: 10.1007/s11263-025-02524-1.

[43] Yifei Xin, Xiulian Peng, and Yan Lu. Masked Audio Modeling with CLAP and Multi-Objective Learning, January 2024.

[44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems*, 35:10078–10093, December 2022.

[45] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020.

[46] History. https://www.dicomstandard.org/history.

[47] DICOM Library - Anonymize, Share, View DICOM files ONLINE. https://www.dicomlibrary.com/dicom/dicom-tags/.

[48] SimpleITK/SimpleITK: SimpleITK: A layer built on top of the Insight Toolkit (ITK), intended to simplify and facilitate ITK's use in rapid prototyping, education and interpreted languages. https://github.com/SimpleITK/SimpleITK.

[49] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2649-2.

[50] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A User-Friendly, Multi-Lingual Library for Medical Image Registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 134–142, 2016.

[51] Nobuyuki Otsu. A Tlreshold Selection Method from Gray-Level Histograms.

[52] Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimpleITK Image-Analysis Notebooks: A Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3):290–303, June 2018. ISSN 1618-727X. doi: 10.1007/s10278-017-0037-8.

[53] SimpleITK/SimpleITK. SimpleITK, December 2025.

[54] Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11(1):9068, April 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-87564-6.

[55] Fabian Pedregosa, Fabian Pedregosa, Gael Varoquaux, Gael Varoquaux, Normalesup Org, Alexandre Gramfort, Alexandre Gramfort, Vincent Michel, Vincent Michel, Logilab Fr, Bertrand Thirion, Bertrand Thirion, Olivier Grisel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, Alexandre Tp, and David Cournapeau. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

[56] Leland McInnes, John Healy, and Steve Astels. Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, March 2017. ISSN 2475-9066. doi: 10.21105/joss.00205.

[57] Sebastian Pölsterl. Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. ISSN 1533-7928.

[58] V Van Belle, K Pelckmans, J A K Suykens, and S Van Huffel. Survival SVM: A Practical Scalable Algorithm.

[59] Ludger Evers and Claudia-Martina Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, July 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn253.

[60] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. An Efficient Training Algorithm for Kernel Survival Support Vector Machines, November 2016.

[61] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39 (5), 2011. ISSN 1548-7660. doi: 10.18637/jss.v039.i05.

[62] Bradley Efron. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 72(359):557–565, September 1977. ISSN 0162-1459. doi: 10.1080/01621459.1977.10480613.

[63] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS169.

[64] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, July 2006. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj011.

[65] Frank E. HARRELL Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996. ISSN 1097-0258. doi: 10.1002/(SICI)1097-0258(19960229)15:4⟨361::AID-SIM168⟩3.0.CO;2-4.

[66] Jérôme Lambert and Sylvie Chevret. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research*, 25(5):2088–2102, October 2016. ISSN 0962-2802. doi: 10.1177/0962280213515571.

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[68] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. ACM, April 2024.

[69] TensorFlow. https://www.tensorflow.org/.

[70] Zakariya Ba Alawi. A Comparative Survey of PyTorch vs TensorFlow for Deep Learning: Usability, Performance, and Deployment Trade-offs, August 2025.

[71] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, May 2019.

[72] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[73] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08999-2. doi: 10.1007/978-3-031-08999-2_22.

[74] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

[75] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis, July 2019.

[76] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, September 2021. ISSN 0169-2607. doi: 10.1016/j.cmpb.2021.106236.

[77] Connie L. Parks and Keith L. Monson. Automated Facial Recognition of Computed Tomography-Derived Facial Images: Patient Privacy Implications. *Journal of Digital Imaging*, 30(2):204–214, April 2017. ISSN 1618-727X. doi: 10.1007/s10278-016-9932-7.

[78] M. Lindholz, R. Ruppel, S. Schulze-Weddige, G. L. Baumgärtner, I. Schobert, A. Panten, R. Schmidt, T. A. Auer, J. Nawabi, A. M. Haack, L. Stepansky, L. Poggi, R. Hosch, C. A. Hamm, and T. Penzkofer. Analyzing the TotalSegmentator for facial feature removal in head CT scans. *Radiography*, 31(1):372–378, January 2025. ISSN 1078-8174. doi: 10.1016/j.radi.2024.12.018.

[79] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing Multimodal Medical Capabilities of Gemini, May 2024.

[80] Imaging-research/ct-foundation at master · Google-Health/imaging-research. https://github.com/Google-Health/imaging-research/tree/master/ct-foundation, .

[81] Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision Foundation Models for Computed Tomography, February 2025.

[82] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z.

[83] Ronnie Sebro and John Mongan. TotalSegmentator: A Gift to the Biomedical Imaging Community. *Radiology: Artificial Intelligence*, 5(5):e230235, August 2023. ISSN 2638-6100. doi: 10.1148/ryai.230235.

[84] Antonella Santone, Francesco Mercaldo, Luca Brunese, Antonella Santone, Francesco Mercaldo, and Luca Brunese. A Method for Real-Time Lung Nodule Instance Segmentation Using Deep Learning. *Life*, 14(9), September 2024. ISSN 2075-1729. doi: 10.3390/life14091192.

[85] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem: If resnets are the answer, then what is the question?

[86] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

[87] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Interleaved Transformer for Volumetric Segmentation, February 2022.

[88] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling*, August 2024.

[89] Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation, January 2024.

[90] George Cybenko, Dianne P. O'Leary, Jorma Rissanen, Avner Friedman, and Robert Gulliver, editors. *The Mathematics of Information Coding, Extraction and Distribution*, volume 107 of *The IMA Volumes in Mathematics and Its Applications*. Springer, New York, NY, 1999. ISBN 978-1-4612-7178-9 978-1-4612-1524-0. doi: 10.1007/978-1-4612-1524-0.

[91] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26 (3):297–302, 1945. ISSN 1939-9170. doi: 10.2307/1932409.

[92] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1): 168–192, July 2020. ISSN 2634-1964. doi: 10.1016/j.aci.2018.08.003.

[93] Paul Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, 1912. ISSN 1469-8137. doi: 10.1111/j.1469-8137.1912.tb05611.x.

[94] Cityscapes Dataset – Semantic Understanding of Urban Street Scenes, October 2020.

[95] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R. Ledsam, and Olaf Ronneberger. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research*, 23(7):e26151, July 2021. doi: 10.2196/26151.

[96] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16555–16564, June 2021. doi: 10.1109/CVPR46437.2021.01629.

[97] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02151-z.

[98] Woorim Choi, Chul-Ho Kim, Hyein Yoo, Hee Rim Yun, Da-Wit Kim, and Ji Wan Kim. Development and validation of a reliable method for automated measurements of psoas muscle volume in CT scans using deep learning-based segmentation: A cross-sectional study. *BMJ Open*, 14(5):e079417, May 2024. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2023-079417.

[99] Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, March 2019. ISSN 1432-1084. doi: 10.1007/s00330-018-5695-5.

[100] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46723-8. doi: 10.1007/978-3-319-46723-8_49.

[101] Yue Cao, Shigang Liu, Yali Peng, and Jun Li. DenseUNet: Densely connected UNet for electron microscopy image segmentation. *IET Image Processing*, 14(12):2682–2689, 2020. ISSN 1751-9667. doi: 10.1049/iet-ipr.2019.1527.

[102] Haigen Hu, Yixing Zheng, Qianwei Zhou, Jie Xiao, Shengyong Chen, and Qiu Guan. MC-Unet: Multi-scale Convolution Unet for Bladder Cancer Cell Segmentation in Phase-Contrast Microscopy Images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1197–1199, November 2019. doi: 10.1109/BIBM47256.2019.8983121.

[103] Sakib Reza, Ohida Binte Amin, and M.M.A. Hashem. TransResUNet: Improving U-Net Architecture for Robust Lungs Segmentation in Chest X-rays. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 1592–1595, Dhaka, Bangladesh, 2020. IEEE. ISBN 978-1-7281-7366-5. doi: 10.1109/TENSYMP50017.2020.9230835.

[104] Yongjin Zhou, Weijian Huang, Pei Dong, Yong Xia, and Shanshan Wang. D-UNet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3):940–950, May 2021. ISSN 1557-9964. doi: 10.1109/TCBB.2019.2939522.

[105] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, June 2020. ISSN 1558-254X. doi: 10.1109/TMI.2019.2959609.

[106] MIC-DKFZ/nnUNet. MIC-DKFZ, December 2025.

[107] Anders Krogh and John A Hertz. A Simple Weight Decay Can Improve Generalization.

[108] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation, September 2018.

[109] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 488–498, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72114-4. doi: 10.1007/978-3-031-72114-4_47.

[110] Jun Ma. Cutting-edge 3D Medical Image Segmentation Methods in 2020: Are Happy Families All Alike?, January 2021.

[111] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation

Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015. ISSN 1558-254X. doi: 10.1109/TMI.2014.2377694.

[112] Vincent Andrearczyk, Valentin Oreiller, Mario Jreige, Martin Vallières, Joel Castelli, Hesham Elhalawani, Sarah Boughdad, John O. Prior, and Adrien Depeursinge. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. In Vincent Andrearczyk, Valentin Oreiller, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation*, pages 1–21, Cham, 2021. Springer International Publishing. ISBN 978-3-030-67194-5. doi: 10.1007/978-3-030-67194-5_1.

[113] Juliette Thariat, Zacharia Mesbah, Youssef Chahir, Arnaud Beddok, Alice Blache, Jean Bourhis, Abir Fatallah, Mathieu Hatt, and Romain Modzelewski. Auto-Segmentation via deep-learning approaches for the assessment of flap volume after reconstructive surgery or radiotherapy in head and neck cancer. *Scientific Reports*, 15(1):21136, July 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-08073-4.

[114] Medical image segmentation on Medical segmentation decathlon - ORKG. https://orkg.org/benchmarks/PWC_MEDICAL_SEGMENTATION_DECATHLON_BENCHMARK_DATASET/pro

[115] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation, March 2021.

[116] Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces, August 2022.

[117] Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks.

[118] Chapter 50 State-space models. In *Handbook of Econometrics*, volume 4, pages 3039–3080. Elsevier, January 1994. doi: 10.1016/S1573-4412(05)80019-4.

[119] Yun Liu, Shuanglong Yao, Xing Wang, Ji Chen, and Xiaole Li. MD-UNet: A medical image segmentation network based on mixed depthwise convolution. *Medical & Biological Engineering & Computing*, 62(4):1201–1212, April 2024. ISSN 1741-0444. doi: 10.1007/s11517-023-03005-8.

[120] Xin Zhong, Gehao Lu, and Hao Li. Vision Mamba and xLSTM-UNet for medical image segmentation. *Scientific Reports*, 15(1):8163, March 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-88967-5.

[121] Fereshteh Yousefirizi, Ivan S. Klyuzhin, Joo Hyun O, Sara Harsini, Xin Tie, Isaac Shiri, Muheon Shin, Changhee Lee, Steve Y. Cho, Tyler J. Bradshaw, Habib Zaidi, François Bénard, Laurie H. Sehn, Kerry J. Savage, Christian Steidl, Carlos F. Uribe, and Arman Rahmim. TMTV-Net: Fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images — a multi-center generalizability analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(7):1937–1954, June 2024. ISSN 1619-7089. doi: 10.1007/s00259-024-06616-x.

[122] Tutorials/auto3dseg at ed8854fa19faa49083f48abf25a2c30ab9ac1c6b · Project-MONAI/tutorials. https://github.com/Project-MONAI/tutorials/tree/ed8854fa19faa49083f48abf25a2c30ab9ac1c6b/auto3dseg.

[123] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021.

[124] Monai.networks.nets.segresnet_ds — MONAI 0 Documentation. https://monai-dev.readthedocs.io/en/latest/_modules/monai/networks/nets/segresnet_ds.html.

[125] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, and Yu Qiao. STU-Net: Scalable and Transferable Medical Image Segmentation Models Empowered by Large-Scale Supervised Pre-training, April 2023.

[126] Saikat Roy, Yannick Kirchhoff, Constantin Ulrich, Maximillian Rokuss, Tassilo Wald, Fabian Isensee, and Klaus Maier-Hein. MedNeXt-v2: Scaling 3D ConvNeXts for Large-Scale Supervised Representation Learning in Medical Image Segmentation, December 2025.

[127] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[128] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjan Sathianathen, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis*, 67:101821, January 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101821.

[129] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[130] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, October 2024.

[131] SAM 3: Segment Anything with Concepts | Research - AI at Meta. https://ai.meta.com/research/publications/sam-3-segment-anything-with-concepts/.

[132] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z.

[133] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment Anything in 3D Medical Images and Videos. https://arxiv.org/abs/2504.03600v1, April 2025.

[134] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. SAM3D: Segment Anything Model in Volumetric Medical Images, March 2024.

[135] Zhaoyang Tan and Qing Cai. DSAM: A Faster SAM for 3D Medical Image Segmentation. In *2024 International Annual Conference on Complex Systems and Intelligent Science (CSIS-IAC)*, pages 891–895, September 2024. doi: 10.1109/CSIS-IAC63491.2024.10919382.

[136] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images, September 2024.

[137] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. SegVol: Universal and Interactive Volumetric Medical Image Segmentation. *Advances in Neural Information Processing Systems*, 37:110746–110783, December 2024. doi: 10.52202/079017-3516.

[138] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 207–229, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73661-2. doi: 10.1007/978-3-031-73661-2_12.

[139] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, Jonathan Deissler, Ralf Floca, and Klaus Maier-Hein. nnInteractive: Redefining 3D Promptable Segmentation, March 2025.

[140] Ivo Wolf, Marcus Vetter, Ingmar Wegner, Thomas Böttger, Marco Nolden, Max Schöbinger, Mark Hastenteufel, Tobias Kunert, and Hans-Peter Meinzer. The Medical Imaging Interaction Toolkit. *Medical Image Analysis*, 9(6):594–604, December 2005. ISSN 1361-8415. doi: 10.1016/j.media.2005.04.005.

[141] Napari: A fast, interactive viewer for multi-dimensional images in Python. https://napari.org/index.html.

[142] 3D Slicer. https://ieeexplore.ieee.org/abstract/document/1398617.

[143] Coen de Vente, Kiran Vaidhya Venkadesh, Bram van Ginneken, and Clara I. Sánchez. SlicerNNInteractive: A 3D Slicer extension for nnInteractive, April 2025.

[144] ITK-SNAP DLS Quick Start — ITK-SNAP Deep Learning Segmentation Server documentation. https://itksnap-dls.readthedocs.io/en/latest/quick_start.html.

[145] Paul A. Yushkevich, Yang Gao, and Guido Gerig. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3342–3345, August 2016. doi: 10.1109/EMBC.2016.7591443.

[146] Maximilian Rokuss, Benjamin Hamm, Yannick Kirchhoff, and Klaus Maier-Hein. Divide and Conquer: A Large-Scale Dataset and Model for Left-Right Breast MRI Segmentation, July 2025.

[147] Tom Julius Blöcker, Nikolaos Delopoulos, Miguel A Palacios, Sebastian Klüter, Juliane Hörner-Rieber, Carolin Rippke, Lorenzo Placidi, Luca Boldrini, Vincenzo Frascino, Nicolaus Andratschke, Michael Baumgartl, Riccardo Dal Bello, Sebastian N Marschner, Claus Belka, Stefanie Corradini, Denis Dudas, Marco Riboldi, Christopher Kurz, and Guillaume Landry. GTV segmentation in MRI guided radiotherapy with promptable foundation models. *Physics in Medicine & Biology*, 71(1):015006, December 2025. ISSN 0031-9155. doi: 10.1088/1361-6560/ae2db9.

[148] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023.

[149] Yang Xing, Jiong Wu, Savas Ozdemir, Ying Zhang, Yang Yang, Wei Shao, and Kuang Gong. MedVL-SAM2: A unified 3D medical vision-language model for multimodal reasoning and prompt-driven segmentation, January 2026.

[150] Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. VoxTell: Free-Text Promptable Universal 3D Medical Image Segmentation, November 2025.

[151] Thorsten M. Buzug. Computed Tomography. In Rüdiger Kramme, Klaus-Peter Hoffmann, and Robert S. Pozos, editors, *Springer Handbook of Medical Technology*, pages 311–342. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-540-74657-7 978-3-540-74658-4. doi: 10.1007/978-3-540-74658-4_16.

[152] Frank Gaillard. MRI sequences (overview) | Radiology Reference Article | Radiopaedia.org. https://radiopaedia.org/articles/mri-sequences-overview.

[153] Pierrick Gouel, Françoise Callonnec, Franchel-Raïs Obongo-Anga, Pierre Bohn, Emilie Lévêque, David Gensanne, Sébastien Hapdey, Romain Modzelewski, Pierre Vera, Sébastien Thureau, Pierrick Gouel, Françoise Callonnec, Franchel-Raïs Obongo-Anga, Pierre Bohn, Emilie Lévêque, David Gensanne, Sébastien Hapdey, Romain Modzelewski, Pierre Vera, and Sébastien Thureau. Quantitative MRI to Characterize Hypoxic Tumors in Comparison to FMISO PET/CT for Radiotherapy in Oropharynx Cancers. *Cancers*, 15(6), March 2023. ISSN 2072-6694. doi: 10.3390/cancers15061918.

[154] William Warnica, Amna Al-Arnawoot, Aleksandra Stanimirovic, Paaladinesh Thavendiranathan, Rachel M. Wald, Mini Pakkal, Gauri Rani Karur, Bernd J. Wintersperger, Valeria Rac, and Kate Hanneman. Clinical Impact of Cardiac MRI T1 and T2 Parametric Mapping in

Patients with Suspected Cardiomyopathy. *Radiology*, 305(2):319–326, November 2022. ISSN 0033-8419. doi: 10.1148/radiol.220067.

[155] Gaspare Saltarelli, Giovanni Di Cerbo, Antonio Innocenzi, Claudia De Felici, Alessandra Splendiani, and Ernesto Di Cesare. Quantitative MRI in Neuroimaging: A Review of Techniques, Biomarkers, and Emerging Clinical Applications. *Brain Sciences*, 15(10), October 2025. ISSN 2076-3425. doi: 10.3390/brainsci15101088.

[156] Chul-min Lee, Sang Hyun Choi, Jae Ho Byun, So Jung Lee, So Yeon Kim, Hyung Jin Won, Yong Moon Shin, and Pyo-Nyun Kim. Combined computed tomography and magnetic resonance imaging improves diagnosis of hepatocellular carcinoma ≤ 3.0 cm. *Hepatology International*, 15(3):676–684, June 2021. ISSN 1936-0541. doi: 10.1007/s12072-021-10190-x.

[157] Ge Wang, Mannudeep Kalra, Venkatesh Murugan, Yan Xi, Lars Gjesteby, Matthew Getzin, Qingsong Yang, Wenxiang Cong, and Michael Vannier. Vision 20/20: Simultaneous CT-MRI — Next chapter of multimodality imaging. *Medical Physics*, 42(10):5879–5889, 2015. ISSN 2473-4209. doi: 10.1118/1.4929559.

[158] Daniel M. Lindberg, Nicholas V. Stence, Joseph A. Grubenhoff, Terri Lewis, David M. Mirsky, Angie L. Miller, Brent R. O'Neill, Kathleen Grice, Peter M. Mourani, and Desmond K. Runyan. Feasibility and Accuracy of Fast MRI Versus CT for Traumatic Brain Injury in Young Children. *Pediatrics*, 144(4):e20190419, October 2019. ISSN 0031-4005. doi: 10.1542/peds.2019-0419.

[159] Anders Brahme. Optimization of stationary and moving beam radiation therapy techniques. *Radiotherapy and Oncology*, 12(2):129–140, June 1988. ISSN 0167-8140. doi: 10.1016/0167-8140(88)90167-3.

[160] Karl Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35(1):310–317, 2008. ISSN 2473-4209. doi: 10.1118/1.2818738.

[161] Paul J. Doolan, Stefanie Charalambous, Yiannis Roussakis, Agnes Leczynski, Mary Peratikou, Melka Benjamin, Konstantinos Ferentinos, Iosif Strouthos, Constantinos Zamboglou, and Efstratios Karagiannis. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Frontiers in Oncology*, 13:1213068, August 2023. ISSN 2234-943X. doi: 10.3389/fonc.2023.1213068.

[162] Carla Xena-Bosch, Srikirti Kodali, Nitin Sahi, Declan Chard, Sara Llufriu, Ahmed T Toosy, Eloy Martinez-Heras, and Ferran Prados. Advances in MRI optic nerve segmentation. *Multiple Sclerosis and Related Disorders*, 98:106437, June 2025. ISSN 2211-0348. doi: 10.1016/j.msard.2025.106437.

[163] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Multimodal CT and MR Segmentation of Head and Neck Organs-at-Risk. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 745–755, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8. doi: 10.1007/978-3-031-43901-8_71.

[164] Elias Tappeiner, Martin Welk, and Rainer Schubert. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 17(11):2103–2111, November 2022. ISSN 1861-6429. doi: 10.1007/s11548-022-02649-5.

[165] Gašper Podobnik, Bulat Ibragimov, Primož Peterlin, Primož Strojan, and Tomaž Vrtovec. vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. *Medical Physics*, 51(3):2175–2186, 2024. ISSN 2473-4209. doi: 10.1002/mp.16924.

[166] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?, January 2016.

[167] Maximilian Rokuss, Balint Kovacs, Yannick Kirchhoff, Shuhan Xiao, Constantin Ulrich, Klaus H. Maier-Hein, and Fabian Isensee. From FDG to PSMA: A Hitchhiker's Guide to Multitracer, Multicenter Lesion Segmentation in PET/CT Imaging. https://arxiv.org/abs/2409.09478v2, September 2024.

[168] Gozétotide : substance active à effet thérapeutique. https://www.vidal.fr/medicaments/substances/gozetotide
100484.html.

[169] Electron–positron annihilation. *Wikipedia*, January 2026.

[170] Artem Zatcepin and Sibylle I. Ziegler. Detectors in positron emission tomography. *Zeitschrift für Medizinische Physik*, 33(1):4–12, February 2023. ISSN 0939-3889. doi: 10.1016/j.zemedi.2022.08.004.

[171] Nora Volkow, Nizar Mullani, and Bernard Bendriem. Positron emission tomography instrumentation: An overview. *American journal of physiologic imaging*, 3:142–53, February 1988.

[172] David Little. Normal F-18 FDG PET-CT | Radiology Case | Radiopaedia.org. https://radiopaedia.org/cases/normal-f-18-fdg-pet-ct.

[173] K. Agrawal, J. Weaver, R. Ngu, and H. Krishnamurthy Mohan. Clinical significance of patterns of incidental thyroid uptake at 18F-FDG PET/CT. *Clinical Radiology*, 70(5):536–543, May 2015. ISSN 0009-9260. doi: 10.1016/j.crad.2014.12.020.

[174] Andrei Gafita, Jeremie Calais, Charlott Franz, Isabel Rauscher, Hui Wang, Andrew Roberstson, Johannes Czernin, Wolfgang A. Weber, and Matthias Eiber. Evaluation of SUV normalized by lean body mass (SUL) in 68Ga-PSMA11 PET/CT: A bi-centric analysis. *EJNMMI Research*, 9:103, December 2019. ISSN 2191-219X. doi: 10.1186/s13550-019-0572-z.

[175] Léo Mottay, Pierre Decazes, Sébastien Hapdey, and Su Ruan. In Silico Dynamic Dual-Tracer PET Image Separation for Prostate Cancer Diagnosis using Deep Learning. In *2024 IEEE 17th International Conference on Signal Processing (ICSP)*, pages 678–681, October 2024. doi: 10.1109/ICSP62129.2024.10846781.

[176] Mei Xin, Yihan Wang, Xinlan Yang, Lianghua Li, Cheng Wang, Yue Gu, Chenpeng Zhang, Gang Huang, Yun Zhou, and Jianjun Liu. Exploring the nigrostriatal and digestive interplays in Parkinson's disease using dynamic total-body [11C]CFT PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(8):2271–2282, July 2024. ISSN 1619-7089. doi: 10.1007/s00259-024-06638-5.

[177] Konrad Klimek, Xinyu Chen, Takanori Sasaki, Daniel Groener, Rudolf A. Werner, and Takahiro Higuchi. PET imaging of sodium-glucose cotransporters (SGLTs): Unveiling metabolic dynamics in diabetes and oncology. *Molecular Metabolism*, 90:102055, December 2024. ISSN 2212-8778. doi: 10.1016/j.molmet.2024.102055.

[178] Sergios Gatidis, Marcel Früh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Rainer Stiefelhagen, Jan Egger, Jens Kleesiek, Ludovic Sibille, Lei Xiang, Simone Bendazolli, Mehdi Astaraki, Bernhard Schölkopf, Michael Ingrisch, Clemens Cyran, and Thomas Küstner. The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging, June 2023. ISSN 2693-5015.

[179] Kerstin Michalski, Juri Ruf, Christian Goetz, Anna Katharina Seitz, Andreas K. Buck, Constantin Lapa, and Philipp E. Hartrampf. Prognostic implications of dual tracer PET/CT: PSMA ligand and [18F]FDG PET/CT in patients undergoing [177Lu]PSMA radioligand therapy. *European Journal of Nuclear Medicine and Molecular Imaging*, 48(6):2024–2030, June 2021. ISSN 1619-7089. doi: 10.1007/s00259-020-05160-8.

[180] Hounsfield Scale - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/hounsfield-scale.

[181] Gowtham Krishnan Murugesan, Diana McCrumb, Eric Brunner, Jithendra Kumar, Rahul Soni, Vasily Grigorash, Stephen Moore, and Jeff Van Oss. Improving Lesion Segmentation in FDG-18 Whole-Body PET/CT scans using Multilabel approach: AutoPET II challenge, November 2023.

[182] Constantin Ulrich. MultiTalentV2 Challenge Edition. September 2024. doi: 10.5281/zenodo.13753413.

[183] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H. Maier-Hein. MultiTalent: A Multi-dataset Approach to Medical Image Segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 648–658, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43898-1. doi: 10.1007/978-3-031-43898-1_62.

[184] Anissa Alloula, Daniel R. McGowan, and Bartłomiej W. Papież. Autopet Challenge 2023: nnUNet-based whole-body 3D PET-CT Tumour Segmentation. https://arxiv.org/abs/2309.13675v2, September 2023.

[185] Lap Yan Lennon Chan, Chenxin Li, and Yixuan Yuan. One-Shot Tumour Synthesis for PET-CT Imaging. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, April 2025. doi: 10.1109/ISBI60581.2025.10981197.

[186] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards Generalizable Tumor Synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11147–11158, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01060.

[187] Junwei Huang, Yingqi Hao, Yitong Luo, Ziyu Wang, Mingxuan Liu, Yifei Chen, Yuanhan Wang, Lei Xiang, and Qiyuan Tian. autoPET IV challenge: Incorporating organ supervision and human guidance for lesion segmentation in PET/CT, September 2025.

[188] Lisa Torrey and Jude Shavlik. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global Scientific Publishing, 2010. ISBN 978-1-60566-766-9. doi: 10.4018/978-1-60566-766-9.ch011.

[189] Valentina Ambrosini, Davide Campana, Paola Tomassetti, Gaia Grassetto, Domenico Rubello, and Stefano Fanti. PET/CT with 68Gallium-DOTA-peptides in NET: An overview. *European Journal of Radiology*, 80(2):e116–e119, November 2011. ISSN 0720048X. doi: 10.1016/j.ejrad.2010.07.022.

[190] Martina Sollini, Margarita Kirienko, Fabrizia Gelardi, Francesco Fiz, Noemi Gozzi, and Arturo Chiti. State-of-the-art of FAPI-PET imaging: A systematic review and meta-analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, 48(13):4396–4414, December 2021. ISSN 1619-7089. doi: 10.1007/s00259-021-05475-0.

[191] Sébastien Thureau, Bernard Dubray, Romain Modzelewski, Pierre Bohn, Sébastien Hapdey, Sabine Vincent, Elodie Anger, David Gensanne, Nicolas Pirault, Gouel Pierrick, and Pierre Vera. FDG and FMISO PET-guided dose escalation with intensity-modulated radiotherapy in lung cancer. *Radiation Oncology*, 13(1):208, October 2018. ISSN 1748-717X. doi: 10.1186/s13014-018-1147-2.

[192] S Ahmad Sarji. Physiological uptake in FDG PET simulating disease. *Biomedical Imaging and Intervention Journal*, 2(4):e59, October 2006. ISSN 1823-5530. doi: 10.2349/biij.2.4.e59.

[193] S. A. Nehmeh, Y. E. Erdi, T. Pan, A. Pevsner, K. E. Rosenzweig, E. Yorke, G. S. Mageras, H. Schoder, Phil Vernon, O. Squire, H. Mostafavi, S. M. Larson, and J. L. Humm. Four-dimensional (4D) PET/CT imaging of the thorax. *Medical Physics*, 31(12):3179–3186, 2004. ISSN 2473-4209. doi: 10.1118/1.1809778.

[194] Arman Rahmim, Martin A. Lodge, Nicolas A. Karakatsanis, Vladimir Y. Panin, Yun Zhou, Alan McMillan, Steve Cho, Habib Zaidi, Michael E. Casey, and Richard L. Wahl. Dynamic whole-body PET imaging: Principles, potentials and applications. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(2):501–518, February 2019. ISSN 1619-7089. doi: 10.1007/s00259-018-4153-6.

[195] Ercan Kurtipek, Mustafa Çayc, Nuri Düzgün, Hdr Esme, Yüksel Terzi, Süleyman Bakdk, Murat Serhat Aygün, Yasar Unlü, Cengiz Burnik, and Taha Tahir Bekci. 18F-FDG PET/CT Mean SUV and Metabolic Tumor Volume for Mean Survival Time in Non–Small Cell Lung Cancer. *Clinical Nuclear Medicine*, 40(6):459, June 2015. ISSN 0363-9762. doi: 10.1097/RLU.0000000000000740.

[196] Pierre Vera, Sandrine Mezzani-Saillard, Agathe Edet-Sanson, Jean-François Ménard, Romain Modzelewski, Sebastien Thureau, Marc-Etienne Meyer, Khadija Jalali, Stéphane Bardet, Delphine Lerouge, Claire Houzard, Françoise Mornex, Pierre Olivier, Guillaume Faure, Caroline Rousseau, Marc-André Mahé, Philippe Gomez, Isabelle Brenot-Rossi, Naji Salem, and Bernard Dubray. FDG PET during radiochemotherapy is predictive of outcome at 1 year in non-small-cell lung cancer patients: A prospective multicentre study (RTEP2). *European Journal of Nuclear Medicine and Molecular Imaging*, 41(6):1057–1065, June 2014. ISSN 1619-7089. doi: 10.1007/s00259-014-2687-9.

[197] Anne-Ségolène Cottereau, Hélène Lanic, Sylvain Mareschal, Michel Meignan, Pierre Vera, Hervé Tilly, Fabrice Jardin, and Stéphanie Becker. Molecular Profile and FDG-PET/CT Total Metabolic Tumor Volume Improve Risk Classification at Diagnosis for Patients with Diffuse Large B-Cell Lymphoma. *Clinical Cancer Research*, 22(15):3801–3809, July 2016. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-15-2825.

[198] Frederic H. Fahey, Paul E. Kinahan, Robert K. Doot, Mehmet Kocak, Harold Thurston, and Tina Young Poussaint. Variability in PET quantitation within a multicenter consortium. *Medical Physics*, 37(7Part1):3660–3666, 2010. ISSN 2473-4209. doi: 10.1118/1.3455705.

[199] Paulina E. Galavis, Christian Hollensen, Ngoneh Jallow, Bhudatt Paliwal, and Robert Jeraj. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncologica*, 49(7):1012–1016, October 2010. ISSN 0284-186X. doi: 10.3109/0284186X.2010.498437.

[200] Nancy Paquet, Adelin Albert, Jacqueline Foidart, and Roland Hustinx. Within-Patient Variability of 18F-FDG: Standardized Uptake Values in Normal Tissues. *Journal of Nuclear Medicine*, 45(5):784–788, May 2004. ISSN 0161-5505, 2159-662X.

[201] F. Yang, G. Simpson, L. Young, J. Ford, N. Dogan, and L. Wang. Impact of contouring variability on oncological PET radiomics features in the lung. *Scientific Reports*, 10(1):369, January 2020. ISSN 2045-2322. doi: 10.1038/s41598-019-57171-7.

[202] Angela van Baardwijk, Geert Bosmans, Liesbeth Boersma, Jeroen Buijsen, Stofferinus Wanders, Monique Hochstenbag, Robert-Jan van Suylen, André Dekker, Cary Dehing-Oberije, Ruud Houben, Søren M. Bentzen, Marinus van Kroonenburgh, Philippe Lambin, and Dirk De Ruysscher. PET-CT–Based Auto-Contouring in Non–Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes. *International Journal of Radiation Oncology*Biology*Physics*, 68(3):771–778, July 2007. ISSN 0360-3016. doi: 10.1016/j.ijrobp.2006.12.067.

[203] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G. P. M. van Stiphout, Patrick Granton, Catharina M. L. Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J. W. L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, March 2012. ISSN 0959-8049. doi: 10.1016/j.ejca.2011.11.036.

[204] Pierre Decazes, Vincent Camus, Elodie Bohers, Pierre-Julien Viailly, Hervé Tilly, Philippe Ruminy, Mathieu Viennot, Sébastien Hapdey, Isabelle Gardin, Stéphanie Becker, Pierre Vera,

and Fabrice Jardin. Correlations between baseline 18F-FDG PET tumour parameters and circulating DNA in diffuse large B cell lymphoma and Hodgkin lymphoma. *EJNMMI Research*, 10(1):120, October 2020. ISSN 2191-219X. doi: 10.1186/s13550-020-00717-y.

[205] Louis Rebaud, Thibault Escobar, Fahad Khalid, Kibrom Girum, and Irène Buvat. Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT. In Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation and Outcome Prediction*, pages 121–134, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-27420-6. doi: 10.1007/978-3-031-27420-6_13.

[206] Fangyang Jiao, Min Wang, Xiaoming Sun, Zizhao Ju, Jiaying Lu, Luyao Wang, Jiehui Jiang, Chuantao Zuo, Fangyang Jiao, Min Wang, Xiaoming Sun, Zizhao Ju, Jiaying Lu, Luyao Wang, Jiehui Jiang, and Chuantao Zuo. Based on Tau PET Radiomics Analysis for the Classification of Alzheimer's Disease and Mild Cognitive Impairment. *Brain Sciences*, 13(2), February 2023. ISSN 2076-3425. doi: 10.3390/brainsci13020367.

[207] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 287–297, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75238-9. doi: 10.1007/978-3-319-75238-9_25.

[208] Giulia Polverari, Francesco Ceci, Valentina Bertaglia, Maria Lucia Reale, Osvaldo Rampado, Elena Gallio, Roberto Passera, Virginia Liberini, Paola Scapoli, Vincenzo Arena, Manuela Racca, Andrea Veltri, Silvia Novello, Désirée Deandreis, Giulia Polverari, Francesco Ceci, Valentina Bertaglia, Maria Lucia Reale, Osvaldo Rampado, Elena Gallio, Roberto Passera, Virginia Liberini, Paola Scapoli, Vincenzo Arena, Manuela Racca, Andrea Veltri, Silvia Novello, and Désirée Deandreis. 18F-FDG Pet Parameters and Radiomics Features Analysis in Advanced Nsclc Treated with Immunotherapy as Predictors of Therapy Response and Survival. *Cancers*, 12(5), May 2020. ISSN 2072-6694. doi: 10.3390/cancers12051163.

[209] Madhurima R. Chetan and Fergus V. Gleeson. Radiomics in predicting treatment response in non-small-cell lung cancer: Current status, challenges and future perspectives. *European Radiology*, 31(2):1049–1058, February 2021. ISSN 1432-1084. doi: 10.1007/s00330-020-07141-9.

[210] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, October 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0339.

[211] AIM-Harvard/pyradiomics. Artificial Intelligence in Medicine (AIM) Program, December 2025.

[212] PyPI Download Stats. https://pypistats.org/packages/pyradiomics.

[213] The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based PhenotypingRadiology. https://pubs.rsna.org/doi/abs/10.1148/radiol.2020191145, .

[214] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. *Radiology*, 295(2):328–338, May 2020. ISSN 0033-8419, 1527-1315. doi: 10.1148/radiol.2020191145.

[215] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Frédérique Frouin, and Irène Buvat. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Research*, 78(16):4786–4789, August 2018. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-18-0125.

[216] Joseph O. Deasy, Angel I. Blanco, and Vanessa H. Clark. CERR: A computational environment for radiotherapy research. *Medical Physics*, 30(5):979–985, 2003. ISSN 2473-4209. doi: 10.1118/1.1568978.

[217] Alex Zwanenburg and Steffen Löck. MIRP: A Python package for standardised radiomics. *Journal of Open Source Software*, 9(99):6413, July 2024. ISSN 2475-9066. doi: 10.21105/joss.06413.

[218] Elisabeth Pfaehler, Alex Zwanenburg, Johan R. de Jong, and Ronald Boellaard. RaCaT: An open source and easy to use radiomics calculator tool. *PLOS ONE*, 14(2):e0212223, February 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0212223.

[219] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to Radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, April 2020. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.118.222893.

[220] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Mechthild Krause, Michael Baumann, Esther G. C. Troost, Steffen Löck, and Christian Richter. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):13206, October 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13448-3.

[221] C. Spearman. *The Proof and Measurement of Association Between Two Things.* Studies in Individual Differences: The Search for Intelligence. Appleton-Century-Crofts, East Norwalk, CT, US, 1961. doi: 10.1037/11491-005.

[222] Karl Pearson. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, December 1895. ISSN 0370-1662. doi: 10.1098/rspl.1895.0041.

[223] Michael Laimighofer, Jan Krumsiek, Florian Buettner, and Fabian J. Theis. Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. *Journal of Computational Biology*, 23(4):279–290, April 2016. doi: 10.1089/cmb.2015.0192.

[224] Shahrbanoo Goli, Hossein Mahjub, Javad Faradmal, Hoda Mashayekhi, and Ali-Reza Soltanian. Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression. *Computational and Mathematical Methods in Medicine*, 2016(1):2157984, 2016. ISSN 1748-6718. doi: 10.1155/2016/2157984.

[225] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, September 1996. ISSN 1061-8600. doi: 10.1080/10618600.1996.10474713.

[226] John Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft, April 1998.

[227] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, July 1958. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1958.tb00292.x.

[228] William A. Belson. Matching and Prediction on the Principle of Biological Classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75, 1959. ISSN 1467-9876. doi: 10.2307/2985543.

[229] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, August 1995. doi: 10.1109/ICDAR.1995.598994.

[230] Jin-zhou Feng, Yu Wang, Jin Peng, Ming-wei Sun, Jun Zeng, and Hua Jiang. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of Critical Care*, 54:110–116, December 2019. ISSN 0883-9441. doi: 10.1016/j.jcrc.2019.08.010.

[231] Margarita Kirienko, Luca Cozzi, Lidija Antunovic, Lisa Lozza, Antonella Fogliata, Emanuele Voulaz, Alexia Rossi, Arturo Chiti, and Martina Sollini. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. *European Journal of Nuclear Medicine and Molecular Imaging*, 45(2):207–217, February 2018. ISSN 1619-7089. doi: 10.1007/s00259-017-3837-7.

[232] Shankar Prinja, Nidhi Gupta, and Ramesh Verma. Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine*, 35(2):217–221, April 2010. ISSN 0970-0218. doi: 10.4103/0970-0218.66859.

[233] Patrick J. Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2005.030814.x.

[234] Alfensi Faruk, Jan Palczewski, and Georgios Aivaliotis. A new discrimination measure for assessing predictive performance of non-linear survival models, April 2025.

[235] Begoña B. Sierra, Colin McLean, Peter S. Hall, and Catalina A. Vallejos. The C-index Multiverse, August 2025.

[236] Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Stop Chasing the C-index: This Is How We Should Evaluate Our Survival Models, June 2025.

[237] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in medicine*, 30(10):1105–1117, May 2011. ISSN 0277-6715. doi: 10.1002/sim.4154.

[238] Wayback Machine. https://web.archive.org/web/20171023012737/https://docs.lib.noaa.gov/rescue/mwr/078/mwr-078-01-0001.pdf, October 2017.

[239] Alessandro Stefano. Challenges and limitations in applying radiomics to PET imaging: Possible opportunities and avenues for research. *Computers in Biology and Medicine*, 179:108827, September 2024. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.108827.

[240] Xue Ying. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2):022022, February 2019. ISSN 1742-6596. doi: 10.1088/1742-6596/1168/2/022022.

[241] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A Comprehensive Survey on Data Augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 38(1):47–66, January 2026. ISSN 1558-2191. doi: 10.1109/TKDE.2025.3622600.

[242] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, January 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.106391.

[243] Lisa Jöckel, Michael Kläs, and Silverio Martínez-Fernández. Safe Traffic Sign Recognition through Data Augmentation for Autonomous Vehicles Software. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 540–541, July 2019. doi: 10.1109/QRS-C.2019.00114.

[244] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617, September 2019. doi: 10.21437/Interspeech.2019-2680.

[245] A. Teymurazyan, T. Riauka, H.-S. Jans, and D. Robinson. Properties of Noise in Positron Emission Tomography Images Reconstructed with Filtered-Backprojection and Row-Action Maximum Likelihood Algorithm. *Journal of Digital Imaging*, 26(3):447–456, June 2013. ISSN 0897-1889. doi: 10.1007/s10278-012-9511-5.

[246] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

[247] Rok Blagus and Lara Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14(1):106, March 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-106.

[248] Muhammad Asif Munir, Reehan Ali Shah, Urooj Waheed, Muhammad Aqeel Aslam, Zeeshan Rashid, Mohammed Aman, Muhammad I. Masud, Zeeshan Ahmad Arfeen, Muhammad Asif Munir, Reehan Ali Shah, Urooj Waheed, Muhammad Aqeel Aslam, Zeeshan Rashid, Mohammed Aman, Muhammad I. Masud, and Zeeshan Ahmad Arfeen. Sparse Regularized Autoencoders-Based Radiomics Data Augmentation for Improved EGFR Mutation Prediction in NSCLC. *Future Internet*, 17(11), October 2025. ISSN 1999-5903. doi: 10.3390/fi17110495.

[249] Ye Zhang, Zhixiang Wang, Zhen Zhang, Junzhuo Liu, Ying Feng, Leonard Wee, Andre Dekker, Qiaosong Chen, and Alberto Traverso. GAN-based one dimensional medical data augmentation. *Soft Computing - A Fusion of Foundations, Methodologies & Applications*, 27(15):10481, August 2023. ISSN 1432-7643. doi: 10.1007/s00500-023-08345-z.

[250] Haiyang WANG and Luca Mainardi. CTGAN in Augmentation of Radiomics Features Classification from Narrow Band Imaging for Laryngeal Cancer. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5, June 2024. doi: 10.1109/MeMeA60663.2024.10596771.

[251] F. Lo Iacono, R. Maragna, G. Pontone, and V. D. A. Corino. A Novel Data Augmentation Method for Radiomics Analysis Using Image Perturbations. *Journal of Imaging Informatics in Medicine*, 37(5):2401–2414, October 2024. ISSN 2948-2933. doi: 10.1007/s10278-024-01013-0.

[252] Marcus R. Makowski, Keno K. Bressem, Luise Franz, Avan Kader, Stefan M. Niehues, Sarah Keller, Daniel Rueckert, and Lisa C. Adams. De Novo Radiomics Approach Using Image Augmentation and Features From T1 Mapping to Predict Gleason Scores in Prostate Cancer. *Investigative Radiology*, 56(10):661, October 2021. doi: 10.1097/RLI.0000000000000788.

[253] Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Cheze Le Rest, Pierre Decazes, Ricardo Correia, Dina El-Habashy, Hesham Elhalawani, Clifton D. Fuller, Mario Jreige, Yomna Khamis, Agustina La Greca, Abdallah Mohamed, Mohamed Naser, John O. Prior, Su Ruan, Stephanie Tanadini-Lang, Olena Tankyevych, Yazdan Salimi, Martin Vallières, Pierre Vera, Dimitris Visvikis, Kareem Wahid, Habib Zaidi, Mathieu Hatt, and Adrien Depeursinge. Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. In Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation and Outcome Prediction*, pages 1–30, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-27420-6. doi: 10.1007/978-3-031-27420-6_1.

[254] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020.

[255] A Survey: Potential Dimensionality Reduction Methods. https://arxiv.org/html/2502.11036v1.

[256] Matplotlib: A 2D Graphics Environment. https://www.computer.org/csdl/magazine/cs/2007/03/c3090/13rR .

[257] Steven J. Rigatti. Random Forest. *Journal of Insurance Medicine*, 47(1):31–39, NaN/NaN/NaN. ISSN 0743-6661. doi: 10.17849/insm-47-01-31-39.1.

[258] Cheng Li. A Gentle Introduction to Gradient Boosting.

[259] Hemant Ishwaran, Udaya B. Kogalur, Eiran Z. Gorodeski, Andy J. Minn, and Michael S. Lauer. High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, 105(489):205–217, March 2010. ISSN 0162-1459. doi: 10.1198/jasa.2009.tm08622.

[260] Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni, and Marcello Restelli. Feature Selection via Mutual Information: New Theoretical Insights, July 2019.

[261] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: A review. *Artificial Intelligence Review*, 57(3):65, February 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10681-3.

[262] Stefan Groha, Sebastian M Schmon, and Alexander Gusev. Neural odes for multi-state survival analysis, 2020.

[263] TotalSegmentator License. https://backend.totalsegmentator.com/license-academic/.

[264] Richard L. Schilsky. Personalized medicine in oncology: The future is now. *Nature Reviews Drug Discovery*, 9(5):363–366, May 2010. ISSN 1474-1784. doi: 10.1038/nrd3181.

[265] J. Oh, S. H. Hong, J.-Y. Choi, H. J. Yoo, and H.-D. Chae. Radiomics in non-oncologic musculoskeletal diseases: From pixels to practice. *Clinical Radiology*, 92, January 2026. ISSN 0009-9260, 1365-229X. doi: 10.1016/j.crad.2025.107106.

[266] Run Tian, Xudong Duan, Fangze Xing, Yiwei Zhao, ChengYan Liu, Heng Li, Ning Kong, Ruomu Cao, Huanshuai Guan, Yiyang Li, Xinghua Li, Jiewen Zhang, Kunzheng Wang, Pei Yang, and Chunsheng Wang. Computed tomography radiomics in predicting patient satisfaction after robotic-assisted total knee arthroplasty. *International Journal of Computer Assisted Radiology and Surgery*, 20(2):237–248, February 2025. ISSN 1861-6429. doi: 10.1007/s11548-024-03192-1.

[267] Tugba Akinci D'Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Daniel Boll, Hanns-Christian Breit, Andrew Phillip Nicoli, Martin Segeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI. *Radiology*, 314(2):e241613, February 2025. ISSN 0033-8419, 1527-1315. doi: 10.1148/radiol.241613.

[268] Cancer Today. https://gco.iarc.who.int/today/, .

[269] Agathe Edet-Sanson, Bernard Dubray, Kaya Doyeux, Adeline Back, Sebastien Hapdey, Romain Modzelewski, Pierre Bohn, Isabelle Gardin, and Pierre Vera. Serial assessment of FDG-PET FDG uptake and functional volume during radiotherapy (RT) in patients with non-small cell lung cancer (NSCLC). *Radiotherapy and Oncology*, 102(2):251–257, February 2012. ISSN 0167-8140. doi: 10.1016/j.radonc.2011.07.023.

[270] J. Ganem, S. Thureau, I. Gardin, R. Modzelewski, S. Hapdey, and P. Vera. Delineation of lung cancer with FDG PET/CT during radiation therapy. *Radiation Oncology*, 13(1):219, November 2018. ISSN 1748-717X. doi: 10.1186/s13014-018-1163-2.

[271] Mehdi Amini, Ghasem Hajianfar, Yazdan Salimi, Zahra Mansouri, and Habib Zaidi. MetaPredictomics: A Comprehensive Approach to Predict Post-Surgical NSCLC Recurrence Using Clinicopathological, Radiomics, and Organomics Data. *Journal of Nuclear Medicine*, 66 (supplement 1):251261–251261, June 2025. ISSN 0161-5505, 2159-662X.

[272] Pierre Vera, Sébastien Thureau, Florence Le Tinier, Philippe Chaumet-Riffaud, Sébastien Hapdey, Hélène Kolesnikov-Gauthier, Etienne Martin, Alina Berriolo-Riedinger, Nicolas Pourel, Jean Marc Broglia, Pierre Boisselier, Sophie Guillemard, Naji Salem, Isabelle Brenot-Rossi, Cécile Le Péchoux, Céline Berthold, Etienne Giroux-Leprieur, Damien Moreau, Sophie Guillerm, Khadija Benali, Laurent Tessonnier, Clarisse Audigier-Valette, Delphine Lerouge, Elske Quak, Carole Massabeau, Frédéric Courbon, Patricia Moisson, Anne Larrouy, Romain Modzelewski, Pierrick Gouel, Nadia Ghazzar, Alexandra Langlais, Elodie Amour, Gérard Zalcman, and Philippe Giraud. Adaptive radiotherapy (up to 74 Gy) or standard radiotherapy (66 Gy) for patients with stage III non-small-cell lung cancer, according to [18F]FDG-PET tumour residual uptake at 42 Gy (RTEP7–IFCT-1402): A multicentre, randomised, controlled phase 2 trial. *The Lancet Oncology*, 25(9):1176–1187, September 2024. ISSN 1470-2045, 1474-5488. doi: 10.1016/S1470-2045(24)00320-6.

[273] Pierre Vera, Philippe Giraud, Sébastien Hapdey, Pierrick Gouel, Orianne Jan, Paul Le Roux, Alexandra Langlais, Emilie Lévêque, Florence Le Tinier, Anaïs Olivier, Etienne Martin, Alina Berriolo-Riedinger, Nicolas Pourel, Jean Marc Broglia, Pierre Boisselier, Sophie Guillemard, Naji Salem, Isabelle Brenot-Rossi, Camilo Garcia, Céline Berthold, Etienne Giroux-Leprieur, Damien Moreau, Sophie Guillerm, Khadija Benali, Laurent Tessonnier, Clarisse Audigier-Valette, Delphine Lerouge, Elske Quak, Carole Massabeau, Frédéric Courbon, Maxime Loo, Anne Larrouy, Nadia Ghazzar, Philippe Chaumet-Riffaud, Elodie Amour, Gérard Zalcman, Romain Modzelewski, and Sébastien Thureau. Prognostic Value of FDG PET Metabolic Parameters Before and After 42 Gy of Radiochemotherapy in Patients with Inoperable Stage III Nonsmall Cell Lung Cancer. *Journal of Nuclear Medicine*, 66(4):516–524, April 2025. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.124.268499.

[274] Ronald Boellaard, Roberto Delgado-Bolton, Wim J. G. Oyen, Francesco Giammarile, Klaus Tatsch, Wolfgang Eschner, Fred J. Verzijlbergen, Sally F. Barrington, Lucy C. Pike, Wolfgang A. Weber, Sigrid Stroobants, Dominique Delbeke, Kevin J. Donohoe, Scott Holbrook, Michael M. Graham, Giorgio Testanera, Otto S. Hoekstra, Josee Zijlstra, Eric Visser, Corneline J. Hoekstra, Jan Pruim, Antoon Willemsen, Bertjan Arends, Jörg Kotzerke, Andreas Bockisch, Thomas Beyer, Arturo Chiti, and Bernd J. Krause. FDG PET/CT: EANM procedure guidelines for tumour imaging: Version 2.0. *European Journal of Nuclear Medicine and Molecular Imaging*, 42(2):328–354, February 2015. ISSN 1619-7089. doi: 10.1007/s00259-014-2961-x.

[275] Fanny Orlhac, Sarah Boughdad, Cathy Philippe, Hugo Stalla-Bourdillon, Christophe Nioche, Laurence Champion, Michaël Soussan, Frédérique Frouin, Vincent Frouin, and Irène Buvat. A

Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *Journal of Nuclear Medicine*, 59(8):1321–1328, August 2018. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.117.199935.

[276] I. Buvat, C. Nioche, A. Dupont, S. Thureau, R. Modzelewski, P. Gouel, C. Massabeau, P. Giraud, P. Vera, and S. Hapdey. Variabilité multicentrique de l'estimée du volume tumoral métabolique total en TEP au FDG dans le cadre du PHRC multicentrique RTEP7-IFCT-1402. *Médecine Nucléaire*, 41(3):172, May 2017. ISSN 0928-1258. doi: 10.1016/j.mednuc.2017.02.089.

[277] Fanny Orlhac, Jakoba J. Eertink, Anne-Ségolène Cottereau, Josée M. Zijlstra, Catherine Thieblemont, Michel Meignan, Ronald Boellaard, and Irène Buvat. A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine*, 63(2):172–179, February 2022. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.121.262464.

[278] Róbert Bata, Amr Sayed Ghanem, and Attila Csaba Nagy. Feature Selection and Model Optimization for Survival Prediction in Patients with Angina Pectoris. *Journal of Clinical Medicine*, 14(22), November 2025. ISSN 2077-0383. doi: 10.3390/jcm14228111.

[279] Saskia A. Cooke, Dirk de Ruysscher, Bart Reymen, Maarten Lambrecht, Gitte Fredberg Persson, Corinne Faivre-Finn, Edith M. T. Dieleman, Rolf Lewensohn, Judi N. A. van Diessen, Karolina Sikorska, Ferry Lalezari, Wouter Vogel, Wouter van Elmpt, Eugène M. F. Damen, Jan-Jakob Sonke, and José S. A. Belderbos. 18F-FDG-PET guided vs whole tumour radiotherapy dose escalation in patients with locally advanced non-small cell lung cancer (PET-Boost): Results from a randomised clinical trial. *Radiotherapy and Oncology*, 181:109492, April 2023. ISSN 0167-8140. doi: 10.1016/j.radonc.2023.109492.

[280] Judi van Diessen, Dirk De Ruysscher, Jan-Jakob Sonke, Eugène Damen, Karolina Sikorska, Bart Reymen, Wouter van Elmpt, Gunnar Westman, Gitte Fredberg Persson, Edith Dieleman, Hedvig Bjorkestrand, Corinne Faivre-Finn, and José Belderbos. The acute and late toxicity results of a randomized phase II dose-escalation trial in non-small cell lung cancer (PET-boost trial). *Radiotherapy and Oncology*, 131:166–173, February 2019. ISSN 0167-8140. doi: 10.1016/j.radonc.2018.09.019.

[281] Lalith Kumar Shiyam Sundar, Josef Yu, Otto Muzik, Oana Kulterer, Barbara J. Fueger, Daria Kifjak, Thomas Nakuz, Hiyung Min Shin, Annika Katharina Sima, Daniela Kitzmantl, Ramsey Derek Badawi, Lorenzo Nardo, Simon R. Cherry, Benjamin A. Spencer, Marcus Hacker, and Thomas Beyer. Fully-automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. *Journal of Nuclear Medicine*, June 2022. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.122.264063.

[282] Julie Auriac, Christophe Nioche, Narinée Hovhannisyan-Baghdasarian, Charlotte Loisel, Romain-David Seban, Nina Jehanno, Lalith Kumar Shiyam Sundar, Thomas Beyer, Irène Buvat, and Fanny Orlhac. Comparison of organ volumes and standardized uptake values in [18F]FDG-PET/CT images using MOOSE and TotalSegmentator to segment CT images. *Medical Physics*, 52(10):e70025, 2025. ISSN 2473-4209. doi: 10.1002/mp.70025.

[283] Cameron Wheeler, Phyo H. Khaing, Eleonora D'Arnese, and Adriana A.S. Tavares. Comparison of multi-organ CT image segmentation tools for whole-body [18F]FDG-PET/CT clinical imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, November 2025. ISSN 1619-7089. doi: 10.1007/s00259-025-07666-5.

[284] Zhack47. Zhack47/autopetii, December 2023.

[285] Zhack47. Zhack47/AutopetIII_CHB, December 2025.

[286] Zhack47. Zhack47/autopetiv, February 2026.

[287] Zhack47/radiomics-augmentation: Radiomics augmentation code repository. https://github.com/Zhack47/radiomics-augmentation.

[288] Zacharia Mesbah. Organ-omics, November 2025.

[289] Harrison C. Gottlich, Adriana V. Gregory, Vidit Sharma, Abhinav Khanna, Amr U. Moustafa, Christine M. Lohse, Theodora A. Potretzke, Panagiotis Korfiatis, Aaron M. Potretzke, Aleksandar Denic, Andrew D. Rule, Naoki Takahashi, Bradley J. Erickson, Bradley C. Leibovich, and Timothy L. Kline. Effect of Dataset Size and Medical Image Modality on Convolutional Neural Network Model Performance for Automated Segmentation: A CT and MR Renal Tumor Imaging Study. *Journal of Digital Imaging*, 36(4):1770–1781, August 2023. ISSN 1618-727X. doi: 10.1007/s10278-023-00804-1.

[290] Dataset. https://www.zongweiz.com/dataset.

[291] Zongwei Zhou. MrGiovanni/RadGPT, February 2026.

[292] matchID - Moteur de recherche des personnes décédées. https://deces.matchid.io/search, .

[293] Junichi Shindoh, Xabier de Aretxabala, Thomas A. Aloia, Juan Carlos Roa, Ivan Roa, Giuseppe Zimmitti, Milind Javle, Claudius Conrad, Dipen M. Maru, Taku Aoki, Luca Vigano, Dario Ribero, Norihiro Kokudo, Lorenzo Capussotti, and Jean-Nicolas Vauthey. Tumor Location Is a Strong Predictor of Tumor Progression and Survival in T2 Gallbladder Cancer: An International Multicenter Study. *Annals of Surgery*, 261(4):733, April 2015. ISSN 0003-4932. doi: 10.1097/SLA.0000000000000728.

[294] Oriane Thiery, Mira Rizkallah, Clément Bailly, Caroline Bodet-Milin, Emmanuel Itti, René-Olivier Casasnovas, Steven Le Gouill, Thomas Carlier, and Diana Mateus. Graph-Based Multimodal Multi-lesion DLBCL Treatment Response Prediction from PET Images. In Jonghye Woo, Alessa Hering, Wilson Silva, Xiang Li, Huazhu Fu, Xiaofeng Liu, Fangxu Xing, Sanjay Purushotham, Tejas S. Mathai, Pritam Mukherjee, Max De Grauw, Regina Beets Tan, Valentina Corbetta, Elmar Kotter, Mauricio Reyes, Christian F. Baumgartner, Quanzheng Li, Richard Leahy, Bin Dong, Hao Chen, Yuankai Huo, Jinglei Lv, Xinxing Xu, Xiaomeng Li, Dwarikanath Mahapatra, Li Cheng, Caroline Petitjean, and Benoît Presles, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops*, pages 103–112, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47425-5. doi: 10.1007/978-3-031-47425-5_10.

Thèse en Français

Liste d'acronymes

I.A. Intelligence Artificielle
CNN Convolutional Neural Network
CT *Computed Tomography*
CPU *Central Processing Unit*
DL *Deep Learning*
DNN *Deep Neural Network*
EHR *Electronic Health Records*
FOV *Field Of View*
FS *Feature Selection*
GPT *Generative Pre-trained Transformer*
IP *Internet Protocol*
IRM Imagerie par Résonance Magnétique
LLM *Large Language Model*
ML *Machine Learning*
MRI *Magnetic Resonance Imaging*
MTL *Multi-Task Learning*
NLP *Natural Language Processing*
OAR *Organes À Risque*
OS *Overall Survival*
PET *Positron Emission Tomography*
RGB *Red, Green, Blue*
SoTA *State of The Art*
SSL *Self-Supervised Learning*
TOF *Time-Of-Flight*
TCP *Transport Control Protocol*
TDM TomoDensitoMétrie
UUID *Universally Unique Identifiers*
ViT *Visual Transformer*
VLM *Vision Language Model*
XR *X-rays*

Table des matières

1 Introduction **1**

1.1 Utilisation de l'intelligence artificielle dans l'analyse d'images médicales 2
1.2 Énoncé du problème 3
1.3 Structure de la thèse 4
1.4 Contributions et Publications 4

2 I.A. pour l'Imagerie Médicale **7**

2.1 Avancées en Intelligence Artificielle 9
2.1.1 Intelligence Artificielle, *Machine Learning* et *Deep Learning* 9
2.1.2 Les Origines du *Deep Learning* 9
2.1.3 *Transformers* et *Large Language Modèles* 10
2.2 Paradigmes & Tâches de l'Intelligence Artificielle 11
2.2.1 Nomenclature des Tâches pour l'Intelligence Artificielle 11
2.2.2 Paradigmes d'Entraînement pour l'Intelligence Artificielle 13
2.3 Outils pour l'Analyse d'Images Médicales 14
2.3.1 Stockage des Images Médicales 14
2.3.2 Traitement des Images Médicales 14
2.3.3 Outils pour l'Intelligence Artificielle 15
2.4 Grands Jeux de Données en Oncologie 15
2.4.1 Facteurs Limitants 15
2.4.2 Exemples de Grands Jeux de Données 16
2.5 Conclusion 16

3 Segmentation des Images Médicales **17**

3.1 Segmentation d'images 19
3.1.1 Segmentation sémantique vs segmentation d'instance 19
3.1.2 Segmentation d'images par apprentissage profond 19
3.1.3 Fonctions de perte et métriques 21
3.2 Segmentation d'images médicales 21
3.2.1 Défis spécifiques aux images médicales 21
3.2.2 Avancées en segmentation médicale automatique 21
3.2.3 Adaptation des Transformers et de Mamba 22
3.2.4 Segmentation semi-automatisée 22
3.3 Segmentation des organes à risque de la tête et du cou 22

3.3.1 Contexte technique . . . 22
3.3.2 Le challenge HaN-Seg . . . 23
3.3.3 Algorithme proposé . . . 23
3.3.4 Résultats et Conclusion . . . 23

4 PET/CT Lesion Segmentation 25
4.1 Contexte . . . 27
4.1.1 L'imagerie TEP/TDM . . . 27
4.1.2 Quantification en TEP : le Standardized Uptake Value (SUV) . . . 29
4.1.3 Méthodes d'acquisition TEP . . . 30
4.1.4 Acquisition multimodale TEP/TDM . . . 30
4.1.5 Analyse TEP/TDM . . . 30
4.2 AutoPET : données et historique des challenges . . . 31
4.2.1 Jeux de données . . . 31
4.2.2 Challenges . . . 31
4.3 Nos méthodes . . . 31
4.3.1 AutoPET II . . . 31
4.3.2 AutoPET III . . . 31
4.3.3 AutoPET IV . . . 32
4.4 Discussion et Conclusion . . . 32
4.4.1 Enseignements d'AutoPET . . . 32
4.4.2 Conclusion . . . 33

5 Data Augmentation for Radiomics 35
5.1 Radiomiques : descripteurs d'imagerie médicale de haute dimension . . . 37
5.1.1 Descripteurs d'imagerie médicale . . . 37
5.1.2 Radiomiques : descripteurs de haute dimension . . . 37
5.1.3 Extraction des caractéristiques radiomiques . . . 38
5.2 Prédiction de la survie à l'aide de la radiomique . . . 39
5.2.1 Sélection de caractéristiques . . . 39
5.2.2 Apprentissage automatique pour la prédiction de survie . . . 40
5.2.3 Métriques pour l'évaluation de la prédiction de survie . . . 41
5.3 *Radiomics Augmentation* : augmentations de données pour combattre le surapprentissage dans les études radiomiques . . . 42
5.3.1 Surapprentissage et augmentations de données . . . 42
5.3.2 Augmentations de données et radiomique . . . 43
5.4 Matériel et méthodes . . . 43
5.4.1 Le jeu de données HECKTOR 2022 . . . 43
5.4.2 Augmentations de données . . . 44
5.4.3 Radiomique . . . 47
5.4.4 Apprentissage automatique . . . 47
5.5 Résultats . . . 47
5.5.1 Distribution des données . . . 47
5.5.2 Impact sur la performance des modèles . . . 48

5.5.3 Impact sur la sélection de caractéristiques et le surapprentissage 48
5.6 Discussion 49
5.6.1 Enseignements 49
5.6.2 Limites 49
5.7 Conclusion 50

6 Extracting radiographic information from organs 51
6.1 Vue d'ensemble 53
6.2 Description détaillée 53
6.2.1 Structure et Configuration 53
6.2.2 Segmentation des organes 54
6.2.3 Extraction de la radiomique 54
6.3 Applications et Limites 55
6.3.1 Applications 55
6.3.2 Limites et Améliorations futures 55
6.4 Application des organomiques à la prédiction de la survie des patients atteints de CBNPC 55
6.4.1 Introduction 55
6.4.2 Matériel et Méthodes 55
6.4.3 Résultats 57
6.4.4 Discussion 57
6.4.5 Conclusion 58

7 Conclusion and perspectives 59
7.1 Conclusion 59
7.2 Perspectives 61
7.2.1 Vers des modèles fondationnels pour la segmentation d'images médicales . . . 61
7.2.2 Améliorer l'état de la prédiction de la survie 61
7.2.3 Quantifier le corps humain dans sa globalité 62

List of Figures 63

List of Tables 65

Appendix 1 67

Bibliography 69

Chapitre 1

Introduction

Contents

1.1 Utilisation de l'intelligence artificielle dans l'analyse d'images médicales 2
1.2 Énoncé du problème 3
1.3 Structure de la thèse 4
1.4 Contributions et Publications 4

1.1 Utilisation de l'intelligence artificielle dans l'analyse d'images médicales

Au fil des années, le nombre d'examens d'imagerie réalisés n'a cessé de croître. Pour les scanners (*Computed Tomography* (CT)), un rapport prévoit une croissance annuelle de 4 % du nombre d'examens effectués à l'échelle mondiale [1]. Cette augmentation du volume d'imagerie est portée par le développement de programmes de dépistage [2], [3], visant à détecter des pathologies comme le cancer du poumon suffisamment tôt pour maximiser les chances de survie des patients, tout en réduisant le recours à des traitements lourds. Par exemple, une détection précoce des nodules pulmonaires cancéreux permet d'envisager une chirurgie mini-invasive, telle que la thoracoscopie vidéo-assistée [4], alors qu'un cancer du poumon de stade III est souvent inopérable et nécessite une chimiothérapie.

Parallèlement à cette augmentation, les examens ont aussi évolué : si l'on prend l'exemple de la TEP (*Positron Emission Tomography* (PET)), la technologie TOF (Time-of-Flight) a permis d'affiner la résolution d'image, augmentant par conséquent la quantité d'informations visibles.

Une meilleure résolution pour chaque examen et une multiplication du nombre d'actes se traduisent, en informatique, par un signal plus riche et une quantité de données accrue. Ces caractéristiques constituent un terreau idéal pour l'application des méthodes d'intelligence artificielle, l'apprentissage automatique (Machine Learning) et l'apprentissage profond (Deep Learning) pour entraîner des modèles de prédiction relatifs au devenir des patients ou pour extraire des informations cliniques pertinentes à partir des images.

Les cas d'usage médicaux de ces modèles prédictifs sont variés : classification des patients (sains ou malades), détermination de la malignité d'un nodule ou d'une tumeur, contourage d'organes avant la planification d'un traitement de radiothérapie, ou encore estimation du risque de survenue d'un événement indésirable (décès, effet secondaire, récidive). Pour la plupart de ces applications, l'entraînement d'un modèle nécessite l'accumulation de paires de données composées :

— L'imagerie (volumes TEP, CT, IRM)

— Du résultat clinique (statut TNM, temps de survie) ou de la décision médicale (choix du traitement, contours des organes)

À partir de ces paires de données, les modèles apprennent des relations statistiquessur lesquels ils s'appuient pour extrapoler sur de nouvelles données, devenant ainsi des outils d'aide à la décision pour les médecins, radiologues et oncologues.

À long terme, ces modèles prédictifs pourraient accroître l'efficacité des spécialistes en automatisant des tâches chronophages, telles que le contourage des organes sains et des volumes cibles sur les scanners de planification en radiothérapie, ou encore la détection, la localisation et l'évaluation de la malignité des nodules pulmonaires lors de programmes de dépistage par scanner. Le temps ainsi économisé par les praticiens pourrait être réinvesti dans la recherche clinique et le développement de traitements innovants.

Libérer du temps médical pourrait également, si cela s'accompagne de politiques de financement adéquates, permettre aux médecins de se consacrer davantage au soin :

— Augmentation du temps passé avec chaque patient

— Amélioration de la qualité de l'écoute et de la prise en charge

— Possibilité de traiter un plus grand nombre de patients.

Ce dernier point est crucial compte tenu de l'augmentation prévue du nombre de cas en oncologie : l'OMS prévoit que le nombre de nouveaux cas de cancer dépassera les 35 millions d'ici 2050 [5].

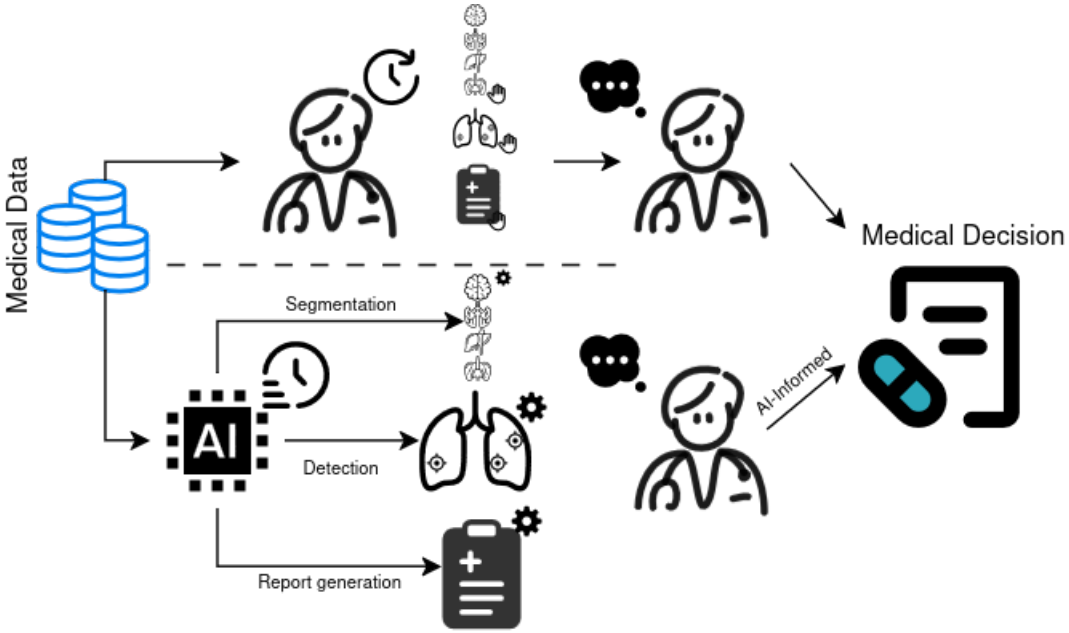

Figure 1.1 – Une amélioration structurelle peut être obtenue en utilisant l'I.A. pour assister les médecins dans les tâches chronophages. Des tâches telles que la détection de lésions, la segmentation d'organes ou même la génération de comptes rendus d'imagerie, pourront ainsi être automatisées, réservant le temps du médecin aux décisions médicales les plus complexes.

En complément, la capacité de ces modèles à extraire et traiter l'information issue de larges volumes d'imagerie pourrait être utilisée pour identifier des caractéristiques radiomiques pertinentes pour la santé du patient et l'évolution de sa maladie. Ces caractéristiques cliniques et d'imagerie pourraient contribuer à un score de risque individuel, pour aider les oncologues à choisir le traitement le mieux adapté à chaque patient.

1.2 Énoncé du problème

Les réseaux de neurones profonds (Deep Neural Networks, DNN) ont atteint un niveau de performance tel qu'ils peuvent accomplir de nombreuses tâches, soit de manière autonome, soit en collaboration avec un utilisateur humain. En oncologie, les médecins et radiologues pourraient grandement bénéficier du traitement automatisé des images médicales pour se concentrer davantage sur le traitement. Cependant, bien que les applications des DNN semblent illimitées, ces modèles commettent encore des erreurs, ce qui est inacceptable pour des tâches critiques. La vie des patients étant en jeu, ce traitement automatisé doit être fiable pour garantir une adoption large et sécurisée par les experts médicaux.

Le Machine Learning (ML) a également réalisé des progrès majeurs au cours de cette dernière

décennie, notamment dans sa capacité à agréger des informations provenant de sources multiples, pour assister la décision humaine. Contrairement aux DNN, les modèles de ML classique ne sont pas capables de gérer de grandes quantités de données brutes, mais ils nécessitent en revanche moins de données pour leur entraînement et tendent à être plus explicables. En oncologie, leur utilité réside souvent dans la sélection des patients pour un traitement, sur la base d'un score de risque.

Nos travaux de recherche visent à répondre à la question suivante :

Comment appliquer les réseaux de neurones profonds pour accélérer le flux de travail en oncologie, en particulier pour la segmentation des organes et des lésions ; et comment utiliser le ML sur ces segmentations pour soutenir la décision médicale en prévoyant la survie des patients et leur réponse à un traitement ?

1.3 Structure de la thèse

Les chapitres de cette thèse sont structurés comme suit :

Le **Chapitre 2** présente les concepts structurels de l'intelligence artificielle ainsi que de ses applications en imagerie médicale et en oncologie. Nous y définissons la nomenclature des différentes tâches de l'I.A. et les outils utiles pour analyser et apprendre à partir des images médicales.

Le **Chapitre 3** présente l'état de l'art de la segmentation d'images médicales. Les différentes architectures et frameworks utilisés par la communauté scientifique pour cette tâche y sont présentés, comparés et discutés. Nos travaux sur la segmentation des organes à risque (OAR) de la tête et du cou sont présentés, comme exemple de segmentation automatisée atteignant des performances comparables à l'annotation humaine.

Le **Chapitre 4** détaille les travaux menés pour faire progresser la segmentation automatisée et semi-automatisée en PET/CT. Lors de nos participations aux défis AutoPET, nous avons exploré les algorithmes de pointe dans ce domaine. Nous avons ensuite développé des nouvelles méthodes pour améliorer la segmentation automatisée des lésions en PET/CT au PSMA et au FDG, ainsi qu'un nouvel algorithme de segmentation semi-automatique.

Le **Chapitre 5** décrit un framework appliquant le concept d'augmentation de données à l'extraction de caractéristiques radiomiques. L'augmentation de données est largement utilisée en apprentissage profond pour lutter contre le surapprentissage. Nous démontrons que l'application de ces augmentations aux images et aux contours avant l'extraction de radiomiques améliore les performances de prédiction de survie sans récidive (RFS), sur le jeu de données HECKTOR.

Le **Chapitre 6** introduit notre contribution majeure : **Organ-omics**. En nous appuyant sur des travaux très récents [6][7], nous avons implémenté et rendu public un outil permettant l'extraction de caractéristiques radiomiques décrivant sur les organes. Nous avons appliqué Organ-omics aux données de l'essai clinique RTEP7, démontrant sa supériorité par rapport à l'utilisation exclusive de la radiomique des lésions pour la prédiction de la survie globale (OS).

1.4 Contributions et Publications

Les principales contributions présentées dans cette thèse sont :

— **Un algorithhme de segmentation automatisée des organes à risque de la tête et du cou en CT/*Magnetic Resonance Imaging* (MRI) :** Un algorithme de segmentation

automatisée, reposant sur un DNN entraîné sur le jeu de données HaN-Seg. L'algorithme, incluant une inférence dynamique pour maîtriser le temps d'exécution, est disponible via un conteneur Docker. Code source : `https://github.com/Zhack47/HaNSeg-QuantIF`

— **Une collection de modèles de segmentation PET/CT :** Développés lors des challenges AutoPET II-III-IV. Trois algorithmes ont été publiés, accompagnés de scripts Docker :
  — **AutoPET II :** segmentation automatisée de PET/CT au 18F-FDG. `https://github.com/Zhack47/autopetii`
  — **AutoPET III :** segmentation automatisée de deux traceurs (PSMA et FDG) sans information préalable sur le traceur utilisé. `https://github.com/Zhack47/AutopetIII_CHB`
  — **AutoPET IV :** algorithme de segmentation PET/CT semi-automatique utilisant les clics d'un expert humain pour guider le DNN. Une méthode innovante d'encodage des clics basée sur des cartes de probabilité informées par la TEP est proposée. `https://github.com/Zhack47/autopetiv`

— **Radiomics-augmentation :** Un framework permettant d'extraire des caractéristiques radiomiques augmentées, visant une meilleure généralisation des modèles prédictifs en complétant la distribution des données.

— **Organ-omics :** Un outil d'extraction de caractéristiques organomiques, utilisant TotalSegmentator pour les contours d'organes et PyRadiomics pour le calcul des caractéristiques. Nous avons montré que les organomiques améliorent la prédiction de la survie. `https://github.com/Zhack47/Organ-omics/`.

Publications

— Podobnik, G., Ibragimov, B., Tappeiner, E., Lee, C., Kim, J. S., **Mesbah, Z.**, Modzelewski, R., Ma, Y., Yang, F., Rudecki, M., Wodzinski, M., Peterlin, P., Strojan, P., & Vrtovec, T. (2024). HaN-Seg : The head and neck organ-at-risk CT and MR segmentation challenge. Radiotherapy and Oncology, 198, 110410. https ://doi.org/10.1016/j.radonc.2024.110410[8] **URL :** `https://www.sciencedirect.com/science/article/pii/S1361841524001786`

— Dexl, J., Gatidis, S., Fruh, M., Jeblick, K., Mittermeier, A., St ?ber, A. T., Schachtner, B., Topalis, J., Fabritius, M. P., Gu, S., Murugesan, G. K., VanOss, J., Ye, J., He, J., Alloula, A., Papie ?, B. W., **Mesbah, Z.**, Modzelewski, R., Hadlich, M., . . . Kustner, T. (2025). AutoPET Challenge on Fully Automated lesion Segmentation in Oncologic PET/CT Imaging, Part 2 : Domain Generalization. Journal of Nuclear Medicine, jnumed.125.270260. https ://doi.org/10.2967/jnumed.125.270260[9] **URL :**`https://jnm.snmjournals.org/content/jnumed/early/2025/12/30/jnumed.125.270260.full.pdf`

— **Mesbah, Z.**, Mottay, L., Modzelewski, R., Decazes, P., Hapdey, S., Ruan, S., & Thureau, S. (2024). AutoPET III : The tracer frontier. What frontier ? arXiv.org[10] **URL :**`https://arxiv.org/pdf/2410.02807`

— **Mesbah, Z.**, Perret, S., Mottay, L., Univ Rouen Normandie F-76000 Rouen, France, AIMS - QuantIF, & Siemens Healthineers. (2025). QuantIF : Semi-automatic PET/CT segmentation across tracers. In QuantIF : Semi-automatic PET/CT Segmentation Across Tracers.[11] **URL :** `https://www.autopet.org/docu/autopetiv/Task1/3_Zhack_AutoPETIV_GoodFormat.pdf`

— **Mesbah Z.**, Thureau S., Gouel P., Hapdey S., Rogé M., Jan O., Le Roux P., Langlais A., Lévêque É., Chaumet-Riffaud P., Amour É., Giroux-Leprieur É., Le Tinier F., Martin É.,

Pourel N., Boisselier P., Salem N., Le Pechoux C., Guillerm S., Audigier-Valette C., Zalcman G., Giraud P., Vera P., Modzelewski R. Strength of organ radiomics (organomics) for non small cell lung cancer (NSCLC) survival analysis based on mid-radiotherapy treatment PET/CT. A RTEP7 ancillary study. (Manuscript submitted in the International Journal of Radiation Oncology Biology Physics).

Chapitre 2

État de l'Art de l'Intelligence Artificielle pour l'Analyse d'Images Médicales en Oncologie

Contenu du Chapitre

Afin d'explorer les usages émergents de l'intelligence artificielle (I.A.) dans l'analyse d'images médicales en oncologie, il est nécessaire de précisément définir ce que l'on apelle l'I.A. et les concepts qui lui sont connexes, tels que le *machine learning* et *deep learning*.

Dans ce chapitre, les noms et descriptions des tâches et paradigmes communément utilisés dans la recherche en I.A. vont être présentés, ainsi que les principaux outils employés pour mener cette recherche. Une fois le paysage actuel de l'I.A. exploré, nous aborderons les utilisations et applications existantes de l"I.A. pour la recherche en imagerie médicale et en oncologie.

Contents

2.1 Avancées en Intelligence Artificielle 9
2.1.1 Intelligence Artificielle, *Machine Learning* et *Deep Learning* 9
2.1.2 Les Origines du *Deep Learning* 9
2.1.3 *Transformers* et *Large Language Modèles* 10
2.2 Paradigmes & Tâches de l'Intelligence Artificielle 11
2.2.1 Nomenclature des Tâches pour l'Intelligence Artificielle 11
2.2.2 Paradigmes d'Entraînement pour l'Intelligence Artificielle 13
2.3 Outils pour l'Analyse d'Images Médicales 14
2.3.1 Stockage des Images Médicales 14
2.3.2 Traitement des Images Médicales 14
2.3.3 Outils pour l'Intelligence Artificielle 15
2.4 Grands Jeux de Données en Oncologie 15
2.4.1 Facteurs Limitants 15
2.4.2 Exemples de Grands Jeux de Données 16
2.5 Conclusion 16

2.1 Avancées en Intelligence Artificielle

2.1.1 Intelligence Artificielle, *Machine Learning* et *Deep Learning*

L'Intelligence Artificielle (I.A.) est définie par Oxford comme *"la théorie et le développement de systèmes informatiques capables d'exécuter des tâches nécessitant normalement l'intelligence humaine"*.

Dans le langage courant, le terme I.A. a été utilisé pour désigner des concepts divers, allant des algorithmes contrôlant des personnages non-joueurs dans les jeux vidéo, aux agents conversationnels (*chatbots*) tels que ChatGPT. Dans le domaine de la recherche, l'I.A. regroupe tout algorithme utilisant la logique pour accomplir une tâche.

Elle englobe le *Machine Learning* (ML), c'est à dire toutes les méthodes d'I.A. qui utilisent des exemples pour apprendre une logique selon un des multiples paradigmes d'apprentissage (voir 2.2.2).

Le *Deep Learning* (DL) est un sous-ensemble du ML, qui comprend exclusivement les méthodes utilisant des réseaux de neurones profonds oud *Deep Neural Network* (DNN). Ce champ de recherche fait partie de l'I.A. presque depuis ses débuts, principalement en raison de la volonté de reproduire le fonctionnement des neurones biologiques. Grâce à l'augmentation de la puissance de calcul disponible, le DL est aujourd'hui le principal moteur de la recherche et des applications commerciales de l'I.A..

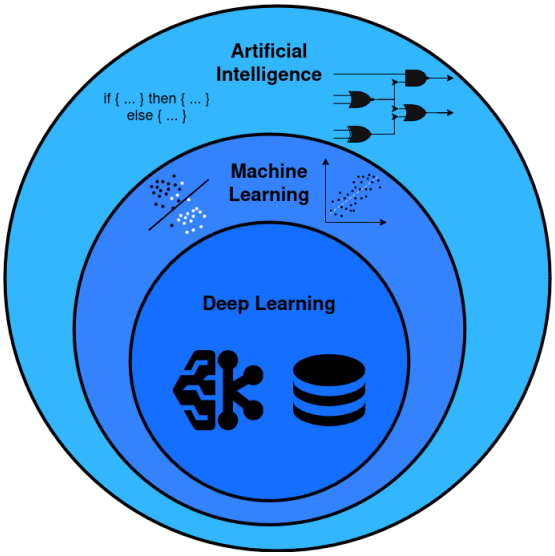

Figure 2.1 – L'Intelligence Artificielle consiste à appliquer une logique pour traiter des entrées. L'apprentissage automatique (ML) désigne spécifiquement les méthodes qui apprennent à partir d'exemples, en cherchant à généraliser à tous les cas possibles. L'apprentissage profond (DL) utilise des réseaux de neurones profonds pour cet apprentissage et nécessite souvent de larges jeux de données pour l'entraînement.

2.1.2 Les Origines du *Deep Learning*

Les fondations de l'apprentissage profond ont été posées entre les années 1960 et 1980 avec l'introduction de plusieurs concepts fondamentaux :

- **Les réseaux de neurones : communément appelés DNN**, ils sont constitués de multiples neurones artificiels, qui sont des fonctions mathématiques appliquées aux données d'entrée. Les bases mathématiques des réseaux de neurones ont été établies sur une longue période, certains concepts remontant aux travaux d'Alan Turing. La percée la plus importante fut le perceptron multicouche (MLP), en 1958 [12]. Les MLP furent les premiers DNNs, car ils contenaient une couche d'entrée, une couche cachée et une couche de sortie.
- **La rétropropagation (*backpropagation*)** : ce concept consiste à attribuer des fractions de l'erreur d'un algorithme à chaque opération de celui-ci. Introduite par Seppo Linnainmaa en 1970 (republiée en 1976)[13], cette méthode permet d'attribuer des parts de l'erreur (calculée entre la sortie d'un réseau de neurones et la vérité terrain) à chaque neurone du DNN. Une fois qu'une erreur locale est obtenue pour chaque neurone, la valeur de ses paramètres peut être modifiée afind de réduire cette erreur.
- **Les couches de convolution**[14] : elles permettent de traiter les entrées non pas comme des valeurs indépendantes, mais comme des vecteurs, avec un ordre. Cela est utile pour le traitement des séries temporelles (valeurs ordonnées dans le temps), ainsi que pour le traitement d'images et de volumes (valeurs ordonnées spatialement).

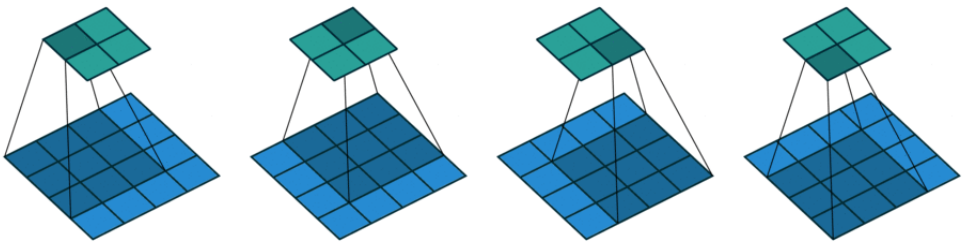

Figure 2.2 – Une convolution $3*3$ appliquée à une matrice de pixels $4*4$ produit une matrice $2*2$. Par exemple : le pixel supérieur gauche de la matrice résultante est le produit de la combinaison des valeurs des 9 values du coin supérieur gauche $3*3$ de la matrice originale, par les poids du noyau de convolution.

- Les réseaux de neurones basés sur la convolution, ou Convolutional Neural Network (CNN), avec la rétropropagation a lancé l'utilisation des réseaux de neurones pour le traitement des images. Parmi les premiers exemples figure LeNet [15], pour la classification de chiffres dans des images $32*32$ pixels (1998) ; AlexNet [16] pour la classification de 1000 objets (2012) ; UNet [17] pour la segmentation d'images (2015) et ResNet[18], un réseau de classification avec des connexions résiduelles, encore utilisé aujourd'hui.

2.1.3 *Transformers* et *Large Language Modèles*

La rupture suivante dans le DL a eu lieu en 2017, avec la publication de *Attention is all you need* [19], un article introduisant l'architecture Transformer. Cette architecture, reposant sur un nouveau mécanisme appelé "l'attention", a bouleversé la communauté académique du traitement du langage naturel, *Natural Language Processing* (NLP). Elle a d'abord démontré des performances de pointe pour les tâches de traduction, avant d'être utilisée pour la génération de texte, au coeur des modèles *Generative Pre-trained Transformer* (GPT)[20].

Ces modèles alimentent les *Large Language Model* (LLM) qui ont déclenché une révolution de l'I.A., avec des produits commerciaux comme ChatGPT qui se sont imposés massivement dans les usages professionnels et personnels. Une catégorie spécifique, les *Vision Language Model* (VLM), intègre désormais des capacités de traitement d'images.

Bien que ces modèles fassent preuve de grandes capacités de génération de texte et de performances notables dans des tâches logiques (compréhension de texte, mathématiques, codage, "raisonnement"...), ils restent des modèles génératifs qui ne peuvent fournir de garanties sur leur contenu de leur génération. Des générations incorrectes, souvent appelées "hallucinations", affectent ces modèles et sont souvent présentées de manière assertive.

La profondeur de ces modèles, soulignée par le qualificatif *Large*, est aussi source de problèmes : leur fonctionnement est presque complètement opaque. La recherche sur l'explicabilité des LLM se limite à :

- L'examen d'un groupe de quelques centaines de neurones [21] (sur des milliards !)
- L'étude du comportement du modèle[22], sans certitude quant à la reproductibilité des comportements observés dans des contextes différents [23].

2.2 Paradigmes & Tâches de l'Intelligence Artificielle

2.2.1 Nomenclature des Tâches pour l'Intelligence Artificielle

Les modèles d'intelligence artificielle restent, à ce jour, spécialisés pour une tâche précise. Cette section présente les tâches les plus courantes que les modèles d'I.A. peuvent résoudre, accompagnées d'exemples en oncologie et en imagerie médicale.

2.2.1.1 Génération

Comme expliqué précédemment, les chatbots commerciaux reposent sur des modèles génératifs. Cette classe de modèles exécute une tâche simple : générer du texte à partir d'une entrée donnée. Des modèles génératifs existent également pour les images [24], la vidéo [25] et l'audio [26].

Cependant, les modèles entraînés sur de grandes quantités de données synthétiques ont tendance à se comporter de manière erratique et échouent à se généraliser aux données réelles [27]. Cela signifie que les données générées par les modèles d'I.A. générative ne doit pas être considérés comme de vrais échantillons de données.

En imagerie médicale et en oncologie, l'I.A. générative est encore émergente, avec des défis spécifiques (images complexes, nécessité de prendre en compte des cas rares...) qui placent le seuil d'adoption plus haut que pour les chatbots ou les générateurs d'images grand public. Ils ont tout de même été appliqués à la génération d'images médicales, notamment pour créer des ensembles de données synthétiques plus vastes lorsque la rareté des données posait problème. Par exemple, Kebaili et al. [28] ont entraîné un modèle de diffusion pour générer des Imagerie par Résonance Magnétique (IRM) de patients atteints de tumeurs cérébrales, afin d'améliorer les performances des modèles de segmentation en aval.

2.2.1.2 Classification

Les signaux (texte, audio, image, vidéo) sont souvent la représentation d'un concept. Une utilisation courante de l'I.A. est de reconnaître automatiquement un concept présent dans ce signal. Lorsqu'un modèle est conçu pour cette tâche, il effectue une tâche de classification.

La classification consiste à analyser un signal et à lui attribuer une classe. Pour une image, il peut s'agir d'une étiquette "chien" ou "chat", pour un texte une étiquette "positif" ou "négatif" (analyse de sentiments) et pour une vidéo, "mouvement" ou "pas de mouvement" pour détecter un mouvement de foule.

Les données d'entraînement doivent être au moins partiellement étiquetées. Le modèle apprend alors, à partir de paires signal/étiquette, à attribuer la classe correcte à tout nouveau signal.

La classification d'images médicales est utile pour le diagnostic de pathologies, mais peut aussi être utilisée pourr clssifier classant des nodules ou des tumeurs comme bénignes ou malignes. Utiliser la classification au niveau du patient est un problème de supervision faible (voir 2.2.2.2), car une faible partie du signal sera utilisée pour la tâche. Pour éviter cela, on peut décomposer la tâche de classification en plusieurs étapes mieux supervisées (ex : segmentation des nodules, puis classification de chaque nodule, puis décision sur le stade du cancer).

La détection du cancer du sein par DNN à partir d'images mammaires complètes (échographie ou mammographie) fait l'objet de recherches [29], tout comme ou la classification de nodules pulmonaires à partir d'images de TomoDensitoMétrie (TDM), en utilisant un DNN pour l'extraction de caractéristiques et un modèle de ML pour la classification[30].

2.2.1.3 Segmentation

La segmentation consiste à effectuer une version dense de la classification, en attribuant une classe à chaque pixel (resp. voxel) d'une image (resp. volume), ce qui produit une carte de segmentation de la même taille que le signal d'entrée. La segmentation est utilisée dans de nombreux domaines, comme la conduite autonome pour localiser les autres véhicules [31].

En imagerie médicale, elle sert à localiser et délimiter les organes, les tissus et les lésions. En raison de la densité de supervision disponible (supervision par voxel), cette tâche est souvent effectuée par des modèles de DL.

2.2.1.4 Régression

Les tâches de régression sont utilisées pour prédire une variable continue à partir d'un signal d'entrée. Son fonctionnement interne est similaire à celui de la classification : les tâches de classification consistent à attribuer un score à chaque classe, puis à choisir la classe ayant le score le plus élevé.

La différence réside dans le fait que, pour la classification, les scores (pseudo-probabilités, généralement entre 0 et 1) ne sont pas calibrés, ce qui signifie qu'ils ne peuvent être utilisés que de manière comparative. Au contraire, la régression produit une valeur continue dont les variations doivent refléter celles de la valeur cible réelle.

En oncologie, une sous-catégorie de régression est particulièrement utilisée : la prédiction de survie.

Qu'il s'agisse de la survie globale, sans progression ou sans récidive, une régression peut être apprise en utilisant la durée de survie comme valeur continue cible. Cependant, lorsque l'objectif est

simplement de classer les patients selon leur risque, un score de risque peut être prédit. Ce score, inversement proportionnel au temps de survie, est plus facile à prédire qu'un nombre exact de jours, car un objectif de classement peut être (*ranking*) plutôt qu'un objectif basé sur l'erreur quadratique (MSE, MAE...).

2.2.2 Paradigmes d'Entraînement pour l'Intelligence Artificielle

Quel que soit le type de modèle utilisé, ou la tâche visée, l'entraînement d'un modèle d'I.A. peut s'effectuer selon différents paradigmes d'apprentissage.

2.2.2.1 Apprentissage Supervisé

Lorsque la vérité terrain est disponible pour chaque échantillon, l'apprentissage supervisé peut être utilisé. C'est le paradigme le plus robuste, car il garantit que les modèles soient entraînés sur une vérité terrain, souvent établie par des experts. Néanmoins, des erreurs ou du bruit peuvent apparaître si la vérité terrain est incorrecte ou incomplète.

En pratique, cela consiste à fournir un signal à un modèle (ML, DNN), à le laisser produire une sortie, puis à comparer cette sortie à la vérité terrain attendue via une fonction d'erreur (*loss function*). Cette erreur est utilisée pour modifier les paramètres du modèle (par rétropropagation dans les DNN). Pour éviter que le modèle n'apprenne "par coeur" (surapprentissage), les données sont séparées en ensembles d'entraînement et d'évaluation. Aucune chevauchement ne doit exister entre les deux ensembles : par exemple, deux scanners d'un même patient doivent impérativement rester dans le même ensemble.

Le processus est répété sur plusieurs epochs (itérations complètes du jeu de données) jusqu'à ce qu'aucune amélioration ne soit observée sur l'ensemble de test. L'utilisation de l'ensemble de test pour choisir le modèle peut cependant induire un biais de sélection [32].

2.2.2.2 Apprentissage Faiblement Supervisé

L'apprentissage faiblement supervisé (WSL) concerne les situations où l'information de supervision est pauvre par rapport à l'information d'entrée. La classification d'images en est un bon exemple : une image peut contenir des millions de pixels, tandis que la supervision n'est qu'une seule annotation.

Une approche populaire pour résoudre ce problème est l'apprentissage multi-tâches ou *Multi-Task Learning* (MTL). Le MTL consiste à accompagner la tâche faiblement supervisée (ex. : classification) par d'autres tâches à supervision plus forte (ex. : reconstruction, segmentation).

Aymyar et al. [33] ont montré que la détection du COVID sur des scanners était grandement améliorée en utilisant le MTL, avec une tâche d'auto-encodage et de une autre de segmentation de lésions.

2.2.2.3 Apprentissage Semi-supervisé

Parfois, les annotations sont manquantes ou incomplètes.

En imagerie médicale, on utilise par exemple la supervision par "gribouillis" (*scribbles*) : pour éviter l'annotation manuelle complète des volumes 3D, des chercheurs utilisent les *scribbles* comme annotations plus rapides, bien que partielles, de ces volumes.

Si les annotations sont disponibles uniquement pour une partie des données, on peut générer des pseudo-annotations en entraînant un premier modèle sur les données annotées, pour annoter le reste [34].

2.2.2.4 Apprentissage Auto-supervisé

L'apprentissage auto-supervisé (SSL) est utilisé lorsque les données ne comportent aucune annotation. Une méthode populaire est la modélisation masquée (*masked modeling*), utilisée pour entraîner les LLM : on apprend au réseau à prédire les parties manquantes d'une phrase [35]. Cela existe aussi pour les images [36], l'audio [37] et la vidéo [38].

L'apprentissage contrastif est une autre approche où l'on présente plusieurs versions modifiées d'une même image au réseau pour qu'il apprenne une représentation similaire pour toutes.

Le SSL est souvent utilisé pour le pré-entraînement de réseaux, dont les poids servent ensuite d'initialisation pour des tâches supervisées sur de plus petits jeux de données. Ce processus est appelé apprentissage par transfert (*transfer learning*).

2.3 Outils pour l'Analyse d'Images Médicales

Bien avant l'avènement des méthodes d'I.A. les plus récentes, une communauté de chercheurs et de médecins a bâti un écosystème d'outils et de standards pour le stockage et le traitement des images médicales.

2.3.1 Stockage des Images Médicales

Les images médicales contiennent des métadonnées essentielles : position dans l'espace (origine, espacement, direction), informations sur le patient, la machine ou l'examen réalisé. Le standard le plus utilisé pour stocker les images médicales est le DICOM (Digital Communications in Medicine), proposé en 1983 par l'ACR et la NEMA pour permettre l'accès aux images en dehors des formats propriétaires des fabricants des machines.

Adopté massivement grâce à son intégration avec le protocole TCP/IP, il s'est étendu de la TDM et l'IRM à l'échographie, puis à l'imagerie nucléaire et à la radiothérapie [39]. Bien qu'exhaustif, le format DICOM est lourd et stocke souvent les volumes sous forme de coupes 2D, alors que l'analyse d'images nécessite une représentation matricielle 3D.

Le format NIfTI, conçu initialement pour la neuro-imagerie IRM, est devenu un standard pour le traitement d'images médicales. Il stocke les volumes (2D, 3D ou 4D) dans un seul fichier (souvent compressé en *.nii.gz*). S'il est plus rapide à charger, il abandonne certaines métadonnées DICOM. Ainsi, toute traitement des données nécessitant de connaître ces champs (ex : calcul du SUV en TEP) doit être effectuée avant la conversion en NIfTI.

2.3.2 Traitement des Images Médicales

Des bibliothèques comme PyDICOM et Nibabel gèrent des formats spécifiques. Cependant, pour l'interopérabilité, il est recommandé d'utiliser des outils polyvalents.

2.3.2.1 SimpleITK : L'interface d'imagerie interopérable

SimpleITK [40] sert d'interface au Insight ToolKit (ITK) et gère de nombreux formats (NIfTI, DICOM, NRRD, PNG, etc.). Disponible pour C++ et Python, il s'interface nativement avec NumPy [41].

SimpleITK propose des outils de rééchantillonnage, normalisation, recadrage et recalage d'images (via SimpleElastix [42]). Sa rapidité, son accessibilité et sa large communauté en font un outil de référence[43].

2.3.2.2 ANTs et TorchIO

Advanced Normalization Tools (ANTs [44]) est une bibliothèque C++ pour le recalage et la segmentation d'images médicales.

Basée sur SimpleITK, la bibliothèque TorchIO [45] propose quant à elle des outils de traitement d'images de haut niveau pour les chercheurs en DL. Elle introduit les concepts de *Subject* (regroupant images et étiquettes d'un patient) et de *Dataset*.

2.3.3 Outils pour l'Intelligence Artificielle

2.3.3.1 Scikit-Learn et Scikit-Survival

Scikit-Learn [46] implémente les algorithmes classiques de ML (forêts aléatoires, SVM, K-Means). Cette bibliothèque privilégie la stabilité et les méthodes éprouvées. Scikit-Survival [47] étend les capacités de Scikit-Learn en gérant nativement les données censurées et en proposant des modèles et métriques pour la prédiction de survie.

2.3.3.2 PyTorch et MONAI

PyTorch [48], développé par Meta, est devenu le framework dominant en recherche grâce à sa flexibilité et son support GPU (CUDA).

MONAI est une extension de PyTorch dédiée aux images médicales, proposant des architectures spécifiques (UNet, SwinUNETR) et des modèles pré-entraînés sur ImageNet[49] ou MedicalNet[50].

2.4 Grands Jeux de Données en Oncologie

Afin d'obtenir des performances similaires à celles des LLMs commerciaux, de grands jeux de données sont nécessaires. Cependant, à l'instar du texte, la collection de grands jeux de données en imagerie oncologique est beaucoup plus difficile.

2.4.1 Facteurs Limitants

Le premier frein est la rareté des données pour certaines pathologies. La collaboration entre centres d'imagerie oncologique est nécessaire, mais complexe en raison des réglementations sur la protection des données. L'anonymisation complète (suppression des noms, dates, identifiants) ou la pseudo-anonymisation est requise. Le format NIfTI facilite cela en ne conservant aucune métadonnée identifiante.

Toutefois, le risque de ré-identification par reconstruction 3D existe [51]. Certains jeux de données utilisent le *defacing* (floutage du visage) [52] pour protéger les patients, bien que cela puisse gêner l'analyse de la région ORL.

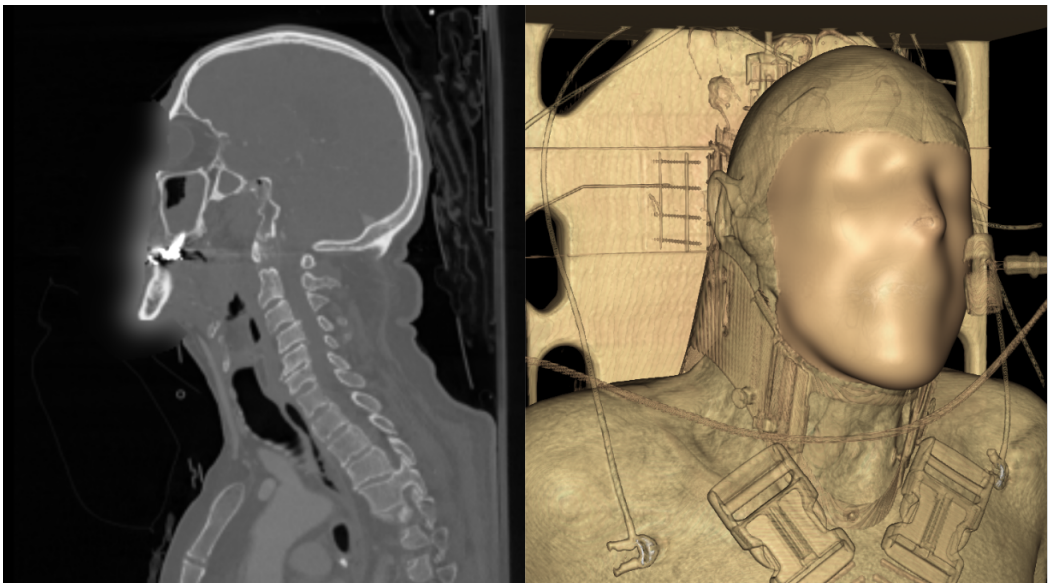

Figure 2.3 – Exemple de scanner (*s0591* de TotalSegmentator) ayant subi un floutage facial. À gauche : rendu volumique 3D. À droite : coupe coronale 2D montrant le floutage appliqué.

2.4.2 Exemples de Grands Jeux de Données

Malgré ces limites, des jeux de données massifs ont pû émerger.

Le *CT Foundation dataset* de Google (500 000 volumes) a permis de pré-entraîner le modèle Gemini [53]. Plus récemment, *CT-FM* (148 000 volumes) a été rendu public, ce qui favorise la reproductibilité.

Pour la segmentation, les chercheurs derrière *TotalSegmentator* (1228 volumes, 117 structures) ont fourni un outil **et** un dataset de référence[6]. *AbdomenAtlas* [54] propose plus de 8 000 scanners segmentés.

Pour atteindre ces échelles, des frameworks comme ScaleMAI [55] utilisent l'I.A. pour pré-annoter les images, laissant aux experts humains le soin de corriger les prédictions les plus incertaines.

2.5 Conclusion

Depuis 2010, les outils d'I.A. en imagerie médicale se sont considérablement développés. Le standard NIfTI et la bibliothèque SimpleITK forment la base du traitement d'images. L'apprentissage profond, porté par PyTorch, a révolutionné le domaine grâce à la constitution de jeux de données massifs segmentés à l'aide de méthodes semi-automatiques. La segmentation est aujourd'hui la seule tâche supervisée où l'I.A. atteint des performances comparables à celles des experts humains.

Dans le prochain chapitre, nous détaillerons l'application du DL à la segmentation des *Organes À Risque* (OAR) ORL.

Chapitre 3

Segmentation d'Images Médicales : État-de-l'art et application à la segmentation des OAR de la tête et du cou

Chapter Overview

La segmentation d'images médicales est une application de la segmentation aux images obtenues par les différentes modalités d'imagerie médicale. Elle présente des difficultés spécifiques, dues à la nature des image médicales (3D, diversité de machines et constructeurs, quantité limité).

It comes with unique challenges to be overcome, due to the nature of medical images. These are often acquired in three dimensions, using different imaging modalities, on machines provided by different vendors and have only recently (and rarely) become available in amounts large enough to use deep learning methods efficiently. In this chapter, the evolution of the state-of-the-art for medical image segmentation will be presented, starting from the theoretical concepts, towards existing tools and works in the domain. Finally, we present our application of this concept to the segmentation of organs at risk in the head and neck area for radiotherapy planning, using CT/MRI images.

Contents

3.1 Segmentation d'images 19
3.1.1 Segmentation sémantique vs segmentation d'instance 19
3.1.2 Segmentation d'images par apprentissage profond 19
3.1.3 Fonctions de perte et métriques 21
3.2 Segmentation d'images médicales 21
3.2.1 Défis spécifiques aux images médicales 21
3.2.2 Avancées en segmentation médicale automatique 21
3.2.3 Adaptation des Transformers et de Mamba 22
3.2.4 Segmentation semi-automatisée 22
3.3 Segmentation des organes à risque de la tête et du cou 22
3.3.1 Contexte technique 22
3.3.2 Le challenge HaN-Seg 23
3.3.3 Algorithme proposé 23
3.3.4 Résultats et Conclusion 23

3.1 Segmentation d'images

3.1.1 Segmentation sémantique vs segmentation d'instance

La segmentation d'images se divise en deux grandes catégories d'objectifs : la segmentation sémantique et la segmentation d'instance. Ces deux approches sont présentées ci-dessous :

— **Segmentation sémantique** : une tâche de segmentation dense qui attribue à chaque pixel (ou voxel) une classe décrivant le type d'objet auquel il appartient. Dans la segmentation d'images naturelles, cela peut consister à attribuer la classe *voiture* à tout pixel faisant partie d'un véhicule. En imagerie médicale, cela revient par exemple à attribuer la valeur *poumon gauche* à chaque voxel appartenant au poumon gauche d'un patient sur un scanner (TDM). Ce type de segmentation est optimal pour les objets dits *indénombrables*, tels que le ciel, les feuilles ou, dans un contexte médical, l'emphysème ou les lésions fibrotiques. Bien que plusieurs instances d'un objet dénombrable puissent être segmentées par cette méthode, elles seront regroupées au sein d'une classe unique.

— **Segmentation d'instances** : cette tâche attribue également une valeur à chaque voxel décrivant l'objet, mais elle distingue les différentes instances d'un même objet. Pour des images naturelles, cela signifie attribuer une valeur distincte aux voxels de chaque voiture unique dans une scène. On utilise la segmentation d'instance pour segmenter des objets dénombrables, comme plusieurs personnes dans une foule ou, en médecine, différentes cellules sur des lames d'anatomo-pathologie [56] ou de multiples nodules pulmonaires. Sa limite réside dans sa segmentation sans concept : il n'est pas possible d'attribuer des classes d'objets différentes aux instances segmentées en utilisant uniquement la segmentation d'instances.

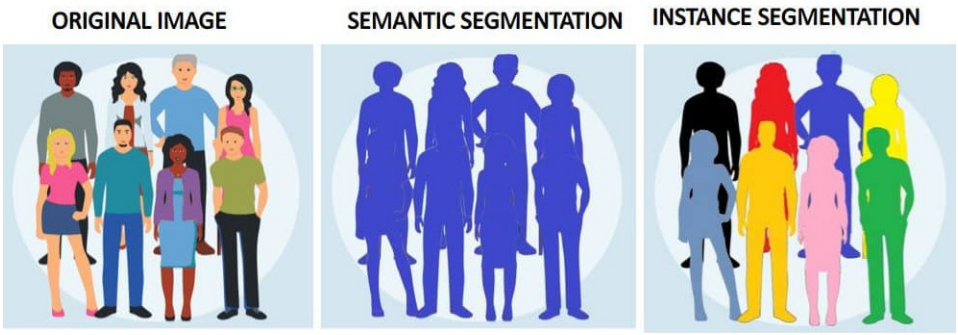

Figure 3.1 – À partir d'une image, la segmentation consiste à classifier les pixels (ou voxels) selon l'objet auquel ils appartiennent. La segmentation sémantique regroupe toutes les instances d'un objet dans la même classe (ex. : personne), tandis que la segmentation d'instance attribue des étiquettes différentes à chaque instance (ex. : chaque personne individuellement). *Source : learnopencv.com*

3.1.2 Segmentation d'images par apprentissage profond

Ces dernières années, la recherche en segmentation s'est largement détournée des méthodes traditionnelles de vision par ordinateur, comme le seuillage d'Otsu [57], au profit de méthodes reposant sur l'apprentissage profond (*deep learning*) et les réseaux de neurones profonds (DNN). À cette fin,

des architectures spécifiques, des fonctions de perte (*losses*) et des métriques d'évaluation ont été conçues.

3.1.2.1 Architectures

L'architecture la plus couramment utilisée est *UNet* [17]. Proposé en 2015, ce réseau de neurones convolutif (CNN) tire son nom de sa forme en U. Il se compose d'un encodeur convolutif de sous-échantillonnage, qui extrait des caractéristiques de haut niveau de l'image, et d'un décodeur convolutif, qui reconstruit la carte de segmentation à partir de ces caractéristiques. Entre l'encodeur et le décodeur se trouvent des connexions résiduelles (appelées *skip connections*), qui transfèrent des caractéristiques spatiales de plus bas niveau de l'encodeur vers le décodeur. Cela permet au réseau de mieux préserver l'information spatiale et les détails fins, tels que les contours. Ces connexions aident également à stabiliser les gradients dans les architectures les plus profondes [58].

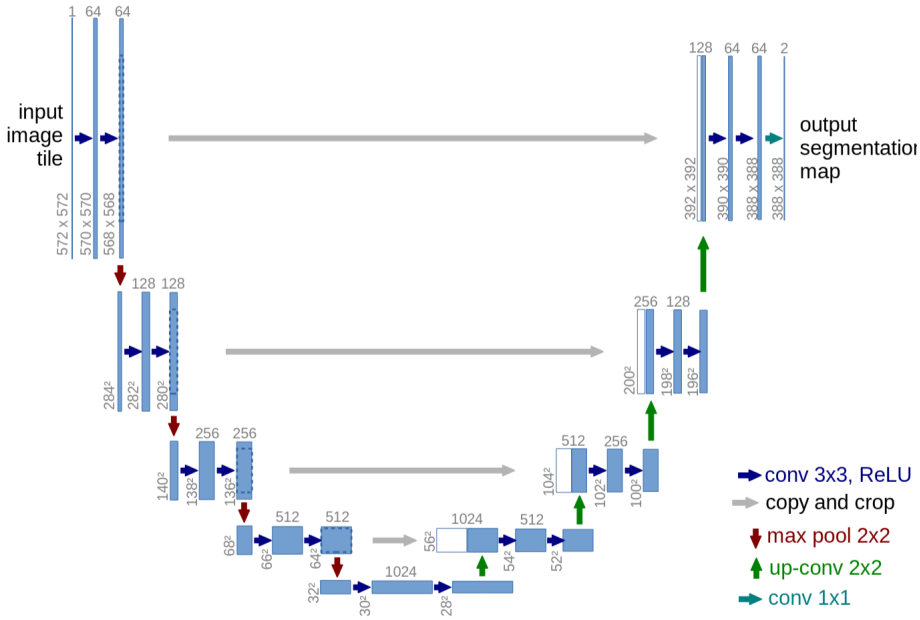

Figure 3.2 – Architecture originale de l'UNet. *Source : Ronneberger et al.*[17]

Après 2020 et l'adaptation de l'architecture *Transformer* aux images avec le *Visual Transformer* (ViT) [59], des réseaux de segmentation basés sur les Transformers ont été proposés. On note particulièrement l'UNETR [60], utilisant un ViT comme encodeur, ainsi que sa variante plus économe, le Swin-UNetR [61]. D'autres équipes ont proposé d'entrelacer les couches de Transformers et de convolutions, comme dans l'architecture nnFormer [62].

Le dernier changement architectural majeur a eu lieu en 2024 avec l'arrivée des modèles d'espace d'états (*state-space models*), notamment l'architecture *Mamba* [63], adaptée à la segmentation médicale avec U-Mamba [64]. Bien que les blocs élémentaires évoluent, l'architecture en U avec encodeur, décodeur et connexions résiduelles reste le coeur de la plupart des conceptions actuelles.

3.1.3 Fonctions de perte et métriques

Pour entraîner un DNN supervisé, il faut définir une fonction d'erreur mathématique quantifiant l'écart entre la sortie du réseau et la vérité terrain (*Ground Truth*, GT) : c'est la **fonction de perte** (*loss function*).

Le premier groupe est celui des **pertes d'information**, dont fait partie l'entropie croisée (*Cross-Entropy*, CE). Ces fonctions calculent l'entropie entre les classes prédites (sous forme de pseudo-probabilités obtenues via une fonction softmax ou sigmoïde) et la vérité terrain.

Les **pertes géométriques** comparent la superposition spatiale. La perte de Dice, basée sur le *Dice-Sorensen Similarity Coefficient*[65] (DSC), est la plus utilisée. Elle mesure la superposition entre la prédiction et la vérité terrain par rapport à leur taille totale.

DSC et *Intersection over Union* (IoU) sont utilisées comme métriques pour évaluer la qualité des segmentations produites. Des variantes ont été créées pour des cas spécifiques, comme la distance de surface normalisée [66] pour les contours, ou le *centerline Dice* [67] pour les structures tubulaires.

Ces métriques de surface ne capturant pas les distances réelles, elles sont complétées par la **distance de Hausdorff**, qui représente l'écart maximal entre deux points des segmentations. Le 95eme percentile de cette distance (HD^{95}) est préféré car il est plus robuste aux valeurs aberrantes.

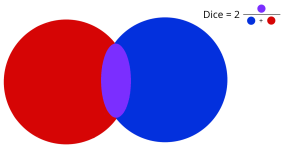

Figure 3.3 – Le DSC compare l'intersection de deux formes géométriques avec la somme de leurs surfaces ou volumes.

3.2 Segmentation d'images médicales

3.2.1 Défis spécifiques aux images médicales

L'entraînement d'un DNN pour la segmentation médicale nécessite des annotations denses et précises, dont le coût est élevé. Cela requiert l'expertise d'un spécialiste capable d'identifier des structures complexes ou des pathologies, et représente un temps de travail considérable.[68].

De plus, la variabilité inter-observateur est significative : deux experts produisent rarement des contours identiques [69].

Enfin, l'accès aux données est limité par les contraintes réglementaires et cliniques, les essais cliniques dépassant rarement le millier de patients.

3.2.2 Avancées en segmentation médicale automatique

3.2.2.1 Le framework nnUNet

Pour pallier à la prolifération de variantes d'architectures peu généralisables, Isensee et al. ont publié le framework **nnUNet**[70].

Leur thèse était que la méthode d'entraînement importe plus que l'architecture elle-même.

Le framework nnUNet automatise la planification, le pré-traitement, l'entraînement et l'évaluation des modèles de segmentation.

Il intègre des mécanismes robustes :

— **Augmentation de données** pour améliorer la généralisation

— **Optimisations mathématiques** : *gradient clipping* pour la stabilité et *weight decay* contre le surapprentissage

— **Perte combinée** : (e.g. : Dice + CE) et supervision profonde (*deep supervision*) à différents niveaux du décodeur

En 2026, nnUNet reste la référence absolue en imagerie médicale, ayant remporté la quasi-totalité des défis internationaux (MICCAI) ces dernières années.

3.2.3 Adaptation des Transformers et de Mamba

L'introduction du **Visual Transformer** (ViT) en 2021 a ouvert l'ère des Transformers en segmentation. En découpant les images en *patchs*, ces modèles visent à capturer des dépendances à longue distance. Cependant, malgré des résultats prometteurs avec le pré-entraînement auto-supervisé (UNETR, SwinUNETR), ils peinent souvent à surpasser les CNN classiques comme nnU-Net lorsqu'ils sont entraînés à partir de zéro.

L'architecture **Mamba** (2024) propose une alternative plus efficace. Contrairement aux Transformers dont le coût de calcul croît de manière quadratique avec la taille de l'image, Mamba propose un coût linéaire. Le modèle **U-Mamba** combine cette efficacité pour le contexte global avec les capacités d'analyse locale des convolutions.

3.2.4 Segmentation semi-automatisée

Malgré les progrès, la segmentation totalement automatique n'est pas infaillible. La segmentation guidée (*promptable segmentation*) permet à un utilisateur d'interagir avec le modèle via des points, des boîtes englobantes ou du texte.

Le modèle SAM (*Segment Anything Model*) de Meta a révolutionné ce domaine, puis a été adapté à la médecine avec MedSAM puis MedSAM2.

Plus récemment, nnInteractive a intégré ces capacités de correction en temps réel directement dans des logiciels de visualisation médicale (3D Slicer, ITK-SNAP), réduisant le temps d'annotation de plus de 85%.

3.3 Segmentation des organes à risque de la tête et du cou

Cette section présente une application concrète : la segmentation des OAR pour la planification en radiothérapie ORL, utilisant des images TDM et IRM non recalées.

3.3.1 Contexte technique

La TDM fournit des informations anatomiques basées sur l'atténuation des rayons X, mesurée en unités Hounsfield (HU). C'est l'examen de référence pour le calcul de dose en radiothérapie.

L'IRM, quant à elle, offre un contraste supérieur pour les tissus mous (cerveau, muscles) en exploitant les propriétés des noyaux d'hydrogène. Combiner les deux permet une vision complète de l'anatomie du patient.

En radiothérapie de la tête et du cou, il est crucial de détourer précisément les OAR (tronc cérébral, nerfs optiques, moelle épinière) pour détruire la tumeur tout en épargnant les tissus sains. Certains organes, peu visibles au scanner, pourraient bénéficier de l'apport de l'IRM.

3.3.2 Le challenge HaN-Seg

Le défi HaN-Seg [8] porte sur 42 patients atteints d'un cancer ORL, pour lesquels 30 OAR ont été contourés.

La difficulté majeure réside dans le fait que les images TDM et IRM ne sont pas acquises simultanément et présentent donc un décalage spatial.

3.3.3 Algorithme proposé

3.3.3.1 Recalage

Notre solution utilise un recalage affine, implémenté *via* la bibliothèque ANTsPy. Le scanner sert de référence anatomique tandis que l'IRM est déplacée pour s'aligner sur le scanner, créant un nouveau volume IRM recalé (*regMRI*).

3.3.3.2 Architecture et entraînement

Nous avons utilisé STU-Net, un U-Net pré-entraîné sur une grande base de données (TotalSegmentator), que nous avons intégré dans le framework nnUNet.

Nous avons apporté deux modifications clés :

1. **Perte Dice adaptée aux classes (ca-Dice) :** elle augmente l'importance des petits organes souvent négligés
2. **Suppression des symétries miroir :** pour permettre au réseau de distinguer la gauche de la droite (essentiel pour les glandes parotides ou les yeux)

3.3.3.3 Inférence adaptative

Pour respecter la contrainte de temps (15 min) et éviter les erreurs de mémoire (*Out-of-Memory*), nous avons implémenté une inférence adaptative. Selon le nombre de voxels de l'image, l'algorithme choisit soit un modèle unique, rapide, soit un ensemble de cinq modèles, plus robuste mais plus lent.

3.3.4 Résultats et Conclusion

Notre modèle a obtenu la 3eme place du challenge avec un score de Dice moyen de 0,750. Il a surpassé les autres méthodes sur plusieurs organes, comme la cavité buccale et le larynx. Ces résultats démontrent que les méthodes d'I.A. atteignent désormais une précision comparable à la variabilité entre experts humains, ouvrant la voie à une intégration clinique pour assister les oncologues radiothérapeutes dans leur pratique quotidienne.

Chapitre 4

Automated PET/CT Lesion Segmentation : four years of improvements with AutoPET

The works presented in this chapter have contributed to three papers :

— Dexl J., Gatidis S., Früh M., ..., Mesbah Z., Modzelewski R. et al. (2025) *AutoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging, Part 2 : Domain Generalization* Published in Journal of Nuclear Medicine
URL : `https://jnm.snmjournals.org/content/jnumed/early/2025/12/30/jnumed.125.270260`

— Mesbah Z., Mottay L. et al. (2025) *AutoPET III : The Tracer Frontier*
URL : `https://arxiv.org/pdf/2410.02807`

— Mesbah Z., Perret S., Mottay L. (2025) *Semi-automatic PET/CT segmentation across tracers*
URL : `https://autopet.org/docu/autopetiv/Task1/3_Zhack_AutoPETIV_GoodFormat.pdf`

Chapter Overview

PET/CT is a multi-modality imaging technique, which combines functional and anatomical images. The segmentation of lesions on PET/CT volumes is key to the computation of prognostic factors, but remains a time-consuming task for nuclear physicians. Medical imaging researchers set out to solve this issue, using automated segmentation algorithms.

This is the motivation behind the AutoPET challenge series, which has taken place for 4 consecutive years from 2022 to 2025. AutoPET aimed to improve the state-of-the-art for automated PET/CT lesion segmentation, by providing large datasets and involving the medical imaging and nuclear medicine community.

In this chapter, we will describe how these challenges took place, which novel methods were proposed in our contributions, how AutoPET improved the PET/CT segmentation SoTA and which were the key factors to foster this improvement.

Contents

4.1 Contexte 27
4.1.1 L'imagerie TEP/TDM 27
4.1.2 Quantification en TEP : le Standardized Uptake Value (SUV) 29
4.1.3 Méthodes d'acquisition TEP 30
4.1.4 Acquisition multimodale TEP/TDM 30
4.1.5 Analyse TEP/TDM 30
4.2 AutoPET : données et historique des challenges 31
4.2.1 Jeux de données 31
4.2.2 Challenges 31
4.3 Nos méthodes 31
4.3.1 AutoPET II 31
4.3.2 AutoPET III 31
4.3.3 AutoPET IV 32
4.4 Discussion et Conclusion 32
4.4.1 Enseignements d'AutoPET 32
4.4.2 Conclusion 33

4.1 Contexte

4.1.1 L'imagerie TEP/TDM

La TEP/TDM est une technique d'imagerie multimodalité qui combine la tomographie par émission de positrons (TEP) et la tomodensitométrie (TDM ou scanner, voir 3.3.1.3). La physique de l'imagerie TEP repose sur la radioactivité et sera présentée, ainsi que les mécanismes biologiques qui lui sont liés, dans cette section.

4.1.1.1 Physique et biologie de la tomographie par émission de positrons

La TEP est une technique d'imagerie fonctionnelle volumétrique qui repose sur l'injection d'un radiotraceur dans le corps du patient. Un radiotraceur est la combinaison d'une molécule appelée ligand (ex. : FluoroDoxyGlucose (FDG), eau, choline, PSMA) et d'un isotope radioactif émetteur de positrons (ex. : ^{18}Fluor, ^{68}Gallium, ^{15}Oxygène, ^{64}Cuivre, ^{11}Carbone).

Le rôle du ligand est de se lier à, ou d'être consommé par, une partie spécifique du corps, des cellules particulières ou des tissus présentant un comportement spécifique. Les ligands ciblant l'antigène membranaire spécifique de la prostate (PSMA), comme le gozétotide[71], ont par exemple été conçus pour se lier à un antigène principalement présent sur la membrane des cellules prostatiques bénignes et malignes.

Le FDG, le ligand le plus utilisé en oncologie, est un analogue du sucre consommé par les cellules de tous les organes utilisant le glucose comme source d'énergie. D'autres ligands, tels que le fluoromisonidazole (FMISO), s'accumulent dans les cellules privées d'oxygène (cellules hypoxiques). Une certaine quantité de ligand s'accumule dans les organes d'élimination (reins, vessie), mais une activité intense sera observée dans les zones contenant le composé cible, là où le radiotraceur s'est lié en grande quantité.

Une fois que le ligand est lié à sa cible, sa position est révélée par l'émission de rayonnements de haute énergie. Pour déclencher l'émission de ces rayonnements, un isotope émetteur de positrons est fixé au ligand.

Cet isotope peut soit remplacer un atome stable du ligand (par exemple, un isotope ^{18}F peut remplacer un atome de fluor stable dans le FDG pour obtenir du ^{18}F-FDG), soit y être attaché par un chélateur, comme c'est le cas pour le ^{68}Ga-PSMA. L'isotope est un atome instable, dont le nombre de neutrons est supérieur à son numéro atomique. Pour cette raison, il se désintègre en un élément de numéro atomique inférieur en perdant un proton, gagnant un neutron et en émettant un positron (une particule d'anti-électron chargée positivement) ainsi qu'un neutrino.

Ce positron parcourt environ 2 à 3 millimètres dans la matière avant de rencontrer un électron, point auquel ces deux particules s'annihilent[72]. L'énergie de ces particules est principalement convertie en deux rayons gamma de 511 keV chacun, émis dans des directions opposées.

4.1.1.2 Acquisition de l'image TEP

Ces rayons gamma sont ensuite détectés à l'aide de cristaux de scintillation, qui émettent de la lumière lorsqu'ils sont frappés par un rayonnement ionisant. Cette lumière est détectée par des photo-détecteurs et un impact est enregistré par le scanner TEP. Lorsque deux impacts sont détectés simultanément par deux détecteurs de la machine, l'événement est marqué comme une coïncidence.

La lumière se déplaçant en ligne droite, cet événement signifie qu'une annihilation positron-électron a probablement eu lieu quelque part sur la ligne reliant les deux détecteurs, appelée *line of response* (LOR). Certains scanners TEP récents sont équipés de capteurs capables de mesurer précisément le temps de collision d'un photon (généralement inférieur à la nanoseconde), ce qui permet de localiser l'annihilation sur un segment plus court de la LOR en calculant la (très faible) différence de temps entre les deux collisions coïncidentes[73].

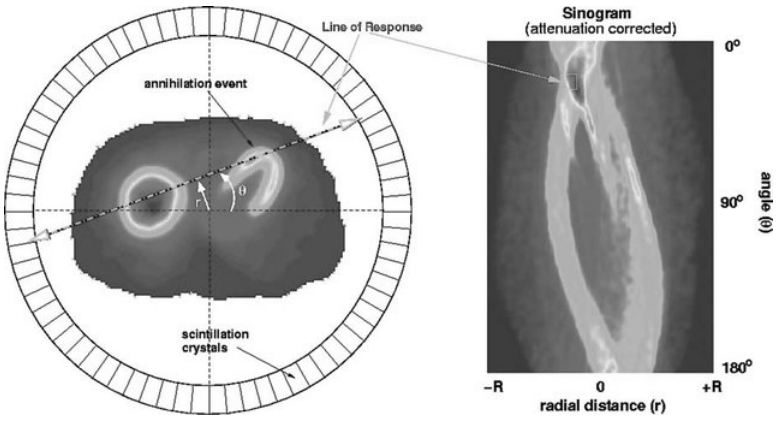

Figure 4.1 – Line Of Response (LOR) et le sinogramme correspondant, résultat de l'addition des LORs. *Source : Volkow et al.[74]*

Après quelques minutes, à l'aide de la liste des coïncidences détectées (regroupées dans un sinogramme), une carte 3D peut être reconstruite, représentant la localisation probable de chaque annihilation.

À partir de cette carte de probabilité, l'activité générée par le radiotraceur peut être quantifiée spatialement, montrant ainsi la quantité de traceur accumulée dans chaque partie du corps. Cette carte est toutefois souvent bruitée, en partie à cause de la diffusion (*scatter*), une interaction où un photon est dévié de sa trajectoire rectiligne initiale.

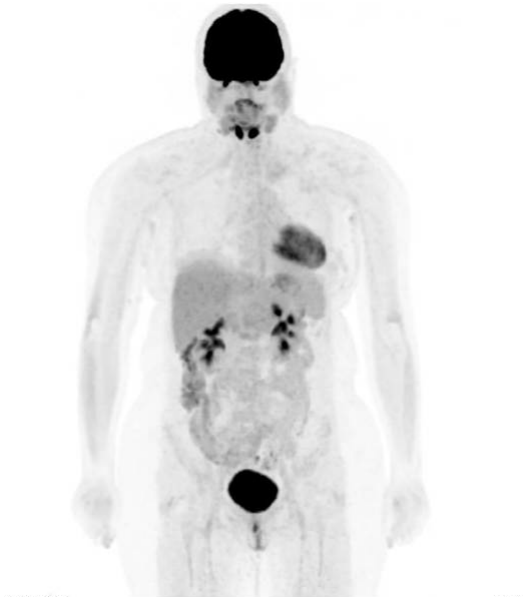

Figure 4.2 – Maximum Intensity Projection (MIP) d'une TEP au 18F-FDG normale (bénigne)[75]. La fixation du radiotraceur est visible dans le cerveau et le myocarde, qui consomment constamment du glucose, ainsi que dans les reins et la vessie. Une fixation thyroïdienne peut également être observée, signe possible d'une pathologie non cancéreuse[76]. *Source : radiopaedia.org*

4.1.2 Quantification en TEP : le Standardized Uptake Value (SUV)

L'activité est influencée par plusieurs facteurs : l'activité injectée (en MBq), le délai entre l'injection et l'acquisition, le poids et la composition corporelle du patient, ainsi que la demi-vie de l'isotope.

Pour pouvoir comparer d'examens TEP à différents moments ou entre différents patients, les images doivent être quantifiées à l'aide du *Standardized Uptake Value* (SUV). Cette valeur est calculée selon la formule suivante :

$$SUV = \frac{Activité}{Activité_{Injecteé} * Facteur\,de\,Décroissance} * Poids \tag{4.1}$$

L'activité est la radioactivité $\beta+$ mesurée en MBq/mL, le poids est en grammes, l'activité injectée en MBq et le facteur de décroissance est défini par :

$$Facteur\,de\,Décroissance = \exp \frac{-\Delta t * \log 2}{\lambda} \tag{4.2}$$

où λ est la demi-vie de l'isotope et Δt le temps écoulé entre l'injection et l'acquisition.

Les volumes TEP en SUV permettent d'extraire des caractéristiques telles que le SUV_{max}, le SUV_{mean} ou le SUV_{peak}, facilitant les études sur l'impact de ces paramètres sur la survie ou la réponse au traitement. Des variantes comme le SUL (SUV normalisé par la masse maigre) ont été développées pour réduire la variabilité inter-patients[77].

4.1.3 Méthodes d'acquisition TEP

La méthode classique consiste en une acquisition 3D unique, réalisée une à deux heures après l'injection. Cependant, de nouvelles modalités sont apparues, comme la TEP paramétrique et la TEP 4D.

La **TEP paramétrique** (souvent corps entier) repose sur des acquisitions successives rapprochées. Elle permet de suivre l'évolution temporelle de la fixation du traceur. Ses applications incluent la séparation de multi-traceurs par courbes activité-temps[78] ou l'analyse de la perfusion tumorale.

La **TEP 4D** repose sur une acquisition synchronisée avec la respiration du patient. Elle réduit les artefacts de mouvement, particulièrement au niveau des poumons, et permet une délimitation métabolique précise de la tumeur pour chaque phase du cycle respiratoire, améliorant ainsi la précision de la radiothérapie.

4.1.4 Acquisition multimodale TEP/TDM

Si la TEP permet une imagerie fonctionnelle, elle manque de précision anatomique. Pour y remédier, un scanner TDM est acquis simultanément. Contrairement à l'IRM, il n'y a pas de limitation physique à l'intégration d'un scanner dans une machine TEP. Cette acquisition combinée assure un recalage intrinsèque des deux examens, servant deux objectifs principaux :

— **Correction d'atténuation :** Le scanner TDM est utilisé pour compenser l'atténuation des photons à l'intérieur du corps, permettant d'obtenir une carte 3D plus fidèle à l'émission réelle des positrons.

— **Localisation anatomique :** Les médecins nucléaires peuvent confronter l'imagerie fonctionnelle aux repères anatomiques pour mieux localiser les lésions et évaluer l'extension de la maladie.

La résolution spatiale de la TEP reste limitée à environ 3 mm (due au parcours du positron), tandis que le scanner atteint une résolution sub-millimétrique. Cette différence nécessite souvent un compromis de résolution lors des analyses automatisées.

4.1.5 Analyse TEP/TDM

Pour évaluer la progression de la maladie, les cliniciens s'appuient sur des mesures pronostiques comme le volume tumoral métabolique total (TMTV)[79]. Obtenir ces valeurs nécessite de détourer toutes les lésions, une tâche extrêmement chronophage, surtout pour les lymphomes ou les cancers métastatiques.

Pour alléger cette charge, des algorithmes de segmentation automatique peuvent assister les praticiens de deux manières :

— **En effectuant un premier passage** que le médecin pourra ensuite corriger

— **En interagissant avec le médecin** via des méthodes de segmentation semi-automatiques qui propagent intelligemment ses annotations[80].

4.2 AutoPET : données et historique des challenges

4.2.1 Jeux de données

Pour la première édition (AutoPET I), un ensemble de 1014 examens TEP/TDM corps entier, avec des segmentations manuelles au voxel près, a été rendu public[81]. Il comprend des patients sains et des patients atteints de mélanome, lymphome ou cancer du poumon. Les éditions suivantes ont introduit des données provenant d'autres centres, des cas pédiatriques et de nouveaux traceurs comme le PSMA.

4.2.2 Challenges

AutoPET I & II L'objectif initial était la délimitation entièrement automatique des lésions au FDG. La deuxième édition a mis l'accent sur la **généralisation** en incluant des données "hors distribution" (différents hôpitaux, pathologies variées, pédiatrie) sans en informer les participants au préalable.

AutoPET III Ce challenge a introduit la problématique du **multi-traceur** (FDG et PSMA). Les algorithmes devaient fonctionner sans connaître le type de traceur *a priori*. Deux catégories ont été créées : *model-centric* (utilisation libre de données/modèles externes) et *data-centric* (focus sur le pré-traitement et l'augmentation de données).

AutoPET IV Cette édition a déplacé le curseur vers la **segmentation semi-automatique** avec l'intégration de "clics" d'experts simulant une interaction humaine (*human-in-the-loop*). Une tâche de segmentation TDM longitudinale a également été introduite.

4.3 Nos méthodes

J'ai participé aux éditions II, III et IV d'AutoPET. Voici les méthodes que j'ai participé à développer.

4.3.1 AutoPET II

J'ai utilisé l'architecture **ResidualEncoderUNet**, plus complexe que l'UNet classique, pour capturer des relations complexes dans les données. J'ai adapté le cycle d'entraînement de la bibliothèque `nnUNet` en augmentant considérablement le nombre d'itérations par époque pour maximiser l'apprentissage sur ce large jeu de données.

4.3.2 AutoPET III

Notre approche reposait sur trois points :

1. **Discriminateur de traceur :** Un CNN 2D analysant les projections MIP pour identifier le ligand (FDG vs PSMA) en moins de 3 secondes avec une précision de 99,64 %.
2. **Segmentation experte :** Deux réseaux distincts (spécialisés par traceur) recevant en entrée les images TEP/TDM ainsi que des versions « fenêtrées » pour améliorer le contraste des tissus mous et des lésions.

3. **Supervision par organes :** Nous avons ajouté des étiquettes d'organes (segmentées via *TotalSegmentator*) pour apprendre au réseau à ne pas confondre la fixation physiologique (cerveau, vessie) avec des lésions malignes.

4.3.3 AutoPET IV

Pour l'interaction semi-automatique, nous avons introduit des **cartes de probabilité**. Contrairement aux approches des autres équipes, nos cartes prennent en compte les valeurs de SUV de la TEP : un clic "lésion" ne diffuse son influence que sur les voxels ayant une fixation significative du traceur.

Nous avons également fusionné les clics positifs et négatifs dans une représentation jointe pour gérer les zones conflictuelles.

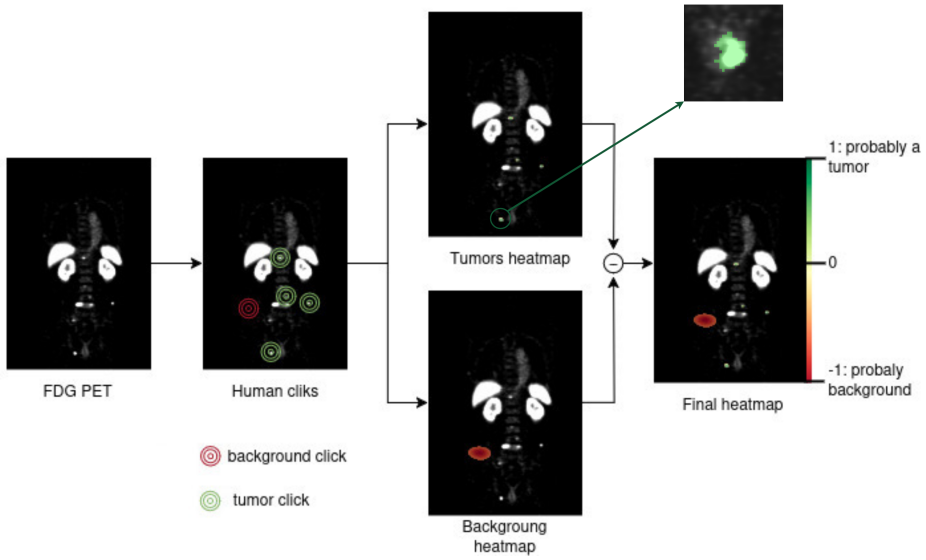

Figure 4.3 – L'algorithme de construction des cartes de probabilité. *Crédit : Solène Perret.*

—

4.4 Discussion et Conclusion

4.4.1 Enseignements d'AutoPET

— **nnUNet reste l'état de l'art :** Les architectures basées sur les CNNs entraînées avec nnUNet surpassent encore les Transformers pour les tâches de segmentation d'images médicales.

— **L'importance du pré-entraînement :** L'utilisation de poids pré-entraînés sur de vastes bases de données améliore significativement la précision et la robustesse.

— **La segmentation semi-automatique est l'avenir :** Elle permet d'intégrer l'expertise clinique tout en réduisant le temps de contourage.

4.4.2 Conclusion

Ces quatre années de challenges AutoPET ont permis des avancées majeures dans la segmentation TEP/TDM en favorisant l'*open-source* et la reproductibilité. Nos contributions, notamment le discriminateur de traceur par MIP et les cartes de probabilité sensibles au SUV, s'inscrivent dans cette dynamique. Bien que des études de validation clinique soient encore nécessaires, la technologie est désormais mature pour assister les cliniciens, que ce soit pour le calcul automatique de la charge tumorale (TMTV) ou la planification en radiothérapie.

Chapitre 5

Radiomics for survival prediction using PET/CT : state of the art and improved reliability using data augmentations

Chapter Overview

Outside of radiotherapy planning, lesion contours obtained from PET/CT imaging are often used to compute lesion descriptors. These descriptors are then used for tasks such as patient classification and survival prediction.

In this chapter, we discuss radiomic features, a high dimensional class of descriptors, how they are obtained from medical images and how they can be used in machine learning to predict survival and patient response to treatment. Additionally, we present how we applied data augmentation to radiomics-based models to improve the performance of ML models on the task of survival prediction.

Contents

5.1 Radiomiques : descripteurs d'imagerie médicale de haute dimension 37
5.1.1 Descripteurs d'imagerie médicale . . . 37
5.1.2 Radiomiques : descripteurs de haute dimension . . . 37
5.1.3 Extraction des caractéristiques radiomiques . . . 38
5.2 Prédiction de la survie à l'aide de la radiomique . . . 39
5.2.1 Sélection de caractéristiques . . . 39
5.2.2 Apprentissage automatique pour la prédiction de survie . . . 40
5.2.3 Métriques pour l'évaluation de la prédiction de survie . . . 41
5.3 *Radiomics Augmentation* : augmentations de données pour combattre le surapprentissage dans les études radiomiques . . . 42
5.3.1 Surapprentissage et augmentations de données . . . 42
5.3.2 Augmentations de données et radiomique . . . 43
5.4 Matériel et méthodes . . . 43
5.4.1 Le jeu de données HECKTOR 2022 . . . 43
5.4.2 Augmentations de données . . . 44
5.4.3 Radiomique . . . 47
5.4.4 Apprentissage automatique . . . 47
5.5 Résultats . . . 47
5.5.1 Distribution des données . . . 47
5.5.2 Impact sur la performance des modèles . . . 48
5.5.3 Impact sur la sélection de caractéristiques et le surapprentissage . . . 48
5.6 Discussion . . . 49
5.6.1 Enseignements . . . 49
5.6.2 Limites . . . 49
5.7 Conclusion . . . 50

5.1 Radiomiques : descripteurs d'imagerie médicale de haute dimension

5.1.1 Descripteurs d'imagerie médicale

Une fois les images médicales traitées et les contours des lésions obtenus, des descripteurs peuvent être calculés pour extraire des informations sur l'image à l'intérieur de ces contours.

En médecine nucléaire, il a été démontré que les descripteurs TEP, parfois aussi simples que le volume tumoral métabolique (MTV), la valeur SUV_{max} ou la valeur SUV_{mean}, sont cliniquement pertinents pour le pronostic du patient [82],[83]. Des descripteurs combinés tels que le volume tumoral métabolique total (TMTV), produit du MTV et du SUV_{mean}, ont également montré des performances intéressantes pour identifier les patients à haut risque, par exemple dans des pathologies à forte charge tumorale comme le lymphome diffus à grandes cellules B [84].

Cependant, même ces descripteurs simples peuvent être sensibles aux variations de la texture de l'image et de la forme des contours. Ces variations proviennent de multiples sources et sont difficiles à contrôler ou à éliminer totalement. Les variations de l'image peuvent être causées par des acquisitions réalisées sur des machines différentes, dans différents centres d'imagerie [85], avec des paramètres de reconstruction distincts [86], ou même en raison de la variabilité intrinsèque du métabolisme d'un patient [87]. Cette variabilité peut conduire à considérer certains descripteurs, tels que ceux de texture, comme insuffisamment robustes pour être utilisés de manière fiable dans le pronostic des patients.

Les variations de contours, causées par la variabilité inter-observateur sont un phénomène bien documenté où les contours manuels obtenus par deux ou plusieurs observateurs diffèrent en partie. Ce problème, évoqué précédemment pour le contourage d'organes sur les scanners, existe également pour le contourage des lésions sur les images TEP [88], bien que certaines méthodes s'appuyant sur les spécificités de la TEP puissent être utilisées pour limiter cette variabilité. La plus populaire de ces méthodes est le contourage semi-automatique, utilisant un volume d'intérêt (VOI) défini manuellement, puis un seuillage de SUV [89].

5.1.2 Radiomiques : descripteurs de haute dimension

Pour tirer parti des modèle récents et plus complexes, les chercheurs ont commencé à produire de plus en plus de descripteurs, extrayant des informations de forme et de texture à partir du VOI de l'image. Ces descripteurs ont été nommés radiomiques [90] (*radiomic features*).

Il existe différents types de caractéristiques, qui peuvent être classées principalement dans les catégories suivantes :

— **Descripteurs simples**, tels que la valeur moyenne, les quartiles ou les caractéristiques issues de l'histogramme de premier ordre de l'image. Ils sont similaires aux descripteurs plus anciens décrits précédemment.

— **Descripteurs de forme**, tels que le volume, la surface ou le diamètre transversal des structures détourées. Ces caractéristiques décrivent la géométrie d'un contour et, bien que certains descripteurs de forme aient été utilisés avant l'apparition du terme radiomique, la plupart étaient liés au volume.

— **Descripteurs de texture**, qui décrivent la texture, c'est à dire les relations entre les valeurs des pixels voisins. Ces caractéristiques peuvent décrire des motifs plus complexes présents dans les données, mais leur complexité les rend moins intuitives à comprendre.

Certaines caractéristiques radiomiques ont été proposées, sans entrer dans ces catégories, comme les distances entre les lésions[91] et le nombre de lésions [92]. Bien qu'elles aient montré des résultats prometteurs, elles n'ont pas encore été aussi largement utilisées et évaluées que les caractéristiques des trois catégories présentées ci-dessus.

Même si un grand nombre de caractéristiques radiomiques peuvent être calculées pour une paire *image/VOI*, la radiomique reste une technique de réduction de dimensionnalité, car elle n'extrait qu'un nombre limité de caractéristiques à partir d'images contenant une grande quantité de voxels. Ces descripteurs augmentent également la robustesse et la sémantique de l'information, en l'élevant au-delà du simple voxel.

Ce changement d'échelle est utile pour toutes les tâches en aval, car l'analyse de tous les voxels d'une image 3D prend beaucoup de temps et nécessite des modèles profonds et lourds, qui exigeraient à leur tour de grands ensembles de données pour l'entraînement. De plus, certains voxels ont plus d'importance que d'autres, tandis que d'autres n'ont aucun intérêt (voxels à l'extérieur du corps du patient). En utilisant les contours, l'information disponible à l'intérieur des parties de l'image jugées importantes est priorisée.

Les caractéristiques radiomiques ont été utilisées pour la classification des maladies [93], la survie des patients [94][95] et la prédiction de la réponse au traitement [96].

5.1.3 Extraction des caractéristiques radiomiques

Afin de rendre les caractéristiques radiomiques accessibles à la communauté de l'imagerie médicale, ainsi que pour établir une norme pour des définitions et des calculs reproductibles, PyRadiomics [97] a été publiée en 2017 sous la forme d'une bibliothèque Python. Cette bibliothèque a offert un moyen d'extraire automatiquement une grande quantité de caractéristiques radiomiques, sans avoir besoin de les redéfinir à chaque fois.

À ce jour, PyRadiomics est le framework d'extraction radiomique le plus utilisé, avec 6 700 citations de son article de référence, 1 300 étoiles [98] sur son dépôt GitHub et des centaines de téléchargements quotidiens[99] via PyPi, le gestionnaire de paquets de Python.

PyRadiomics calcule 107 descripteurs dans sa configuration par défaut, qui peut être modifiée pour exclure certaines caractéristiques ou certains types de caractéristiques. Des descripteurs supplémentaires, exclus par défaut, peuvent toujours être obtenus en forçant PyRadiomics à les inclure. Les raisons de l'exclusion d'une caractéristique de la configuration par défaut peuvent être : elle est égale ou proportionnelle à une caractéristique existante, elle est toujours égale à la même valeur, ou elle fait partie du groupe de caractéristiques de forme calculées en 2D.

En plus des descripteurs, PyRadiomics inclut des outils de prétraitement, dont certains ont été développés pour aider à la standardisation des caractéristiques (IBSI, voir ci-dessous). D'autres outils de pré-traitement peuvent être utilisés pour appliquer des filtres sur les images d'entrée. Ces filtres peuvent être activés par l'utilisateur afin de faire ressortir certaines parties de l'image.

En 2020, des chercheurs de plusieurs laboratoires ont collaboré pour proposer IBSI [100],[101] : ***Imaging Biomarker Standardization Initiative***. Cet article définit des règles pour harmoniser les pratiques de calcul radiomique entre les différentes études, équipes et logiciels (LifeX [102], MITK

[103], PyRadiomics [97]). Un manuel accompagnant le standard a été publié, avec des recommandations pour des études radiomiques reproductibles. PyRadiomics a participé à cette initiative de standardisation et certaines recommandations ont été intégrées dans son code et sa documentation.

Comme mentionné plus haut, d'autres logiciels d'extraction radiomique existent, bien que moins populaires que PyRadiomics. LifeX, MITK, CERR [104], MIRP [105] ou RaCat [106] sont des exemples de frameworks incluant un pipeline d'extraction radiomique. Certains de ces frameworks ont été développés et sont maintenus par des universités ou des laboratoires de recherche, ce qui rend les initiatives comme IBSI nécessaires pour maintenir la cohérence des valeurs radiomiques d'un framework à l'autre.

Cependant, comme le montre le rapport IBSI, aucun framework n'a implémenté exactement le même ensemble de caractéristiques, ce qui laisse subsister une différence dans l'information contenue dans le vecteur radiomique selon le logiciel utilisé.

5.2 Prédiction de la survie à l'aide de la radiomique

L'imagerie médicale est utilisée par les médecins pour choisir la meilleure stratégie thérapeutique, qu'il s'agisse d'une surveillance accrue ou d'un traitement. Dans ce dernier cas, le traitement peut également être adapté en fonction de l'imagerie.

Pour assister les cliniciens dans leur décision, l'analyse d'images via des modèles de ML est un domaine de recherche prometteur. En apprenant de larges cohortes de patients et en observant des variations subtiles dans les données d'imagerie, ces modèles peuvent offrir un éclairage quantitatif complémentaire à l'évaluation qualitative du médecin.

Les radiomique ont été utilisées dans de nombreuses études de recherche et ont démontré une capacité à différencier, au moins partiellement, les patients en fonction de leur future réponse au traitement, ou du risque de décès ou de rechute.

Dans cette section, un cadre de base pour l'utilisation des méthodes de ML radiomiques pour la prédiction de la survie sera présenté.

5.2.1 Sélection de caractéristiques

Une fois les caractéristiques radiomiques extraites comme décrit précédemment, le vecteur résultant reste souvent de haute dimension.

La première étape de toute étude radiomique doit être la sélection de caractéristiques ou *Feature Selection* (FS), c'est-à-dire sélectionner les variables portant le plus d'informations, tout en écartant les autres.

Les caractéristiques radiomiques présentent souvent une corrélation inter-caractéristiques, ce qui signifie qu'il existe une redondance entre elles. Une première étape de réduction de la redondance est recommandée [107]. Elle peut consister soit à supprimer certaines caractéristiques égales ou fortement corrélées (par exemple, un coefficient de corrélation de Spearman $> 0,90$) au profit d'une seule variable conservée, soit à créer des méta-caractéristiques qui sont une combinaison de plusieurs variables [108] (par exemple via une analyse en composantes principales). Bien que cette dernière méthode réduise la dimensionnalité tout en conservant une partie de l'information, elle rend le modèle ML plus difficile à interpréter, à cause des méta-caractéristiques.

Une fois les redondances traitées, une réduction supplémentaire peut être obtenue en sélectionnant les caractéristiques les plus importantes. Il peut s'agir de celles présentant la plus forte variabilité inter-patients, ou de celles corrélant le plus avec la cible à prédire [107]. Lorsque cette cible est entièrement disponible (pas de censure), des coefficients de corrélation tels que ceux de Spearman [109] ou de Pearson [110] peuvent être utilisés.

Si une partie de la cible à prédir est manquante (par exemple, des données censurées à droite, très fréquent en analyse de survie), des modèles univariés peuvent être entraînés, caractéristique par caractéristique, pour attribuer un score d'importance à chacune selon sa capacité à résoudre le problème seule.

Cette méthode a été utilisée dans plusieurs articles de recherche, souvent comme *baseline* pour de nouvelles méthodes [111],[112].

Il est intéressant de noter que les implémentations d'algorithmes de FS pour les données de survie se trouvent parfois dans des packages R [113], mais sont absentes de Python. Cela montre que la prédiction de survie est souvent considérée, dans la littérature, sous un angle statistique plutôt que sous celui de l'apprentissage automatique.

5.2.2 Apprentissage automatique pour la prédiction de survie

La plupart des modèles de ML, tels que les machines à vecteurs de support [114] (SVM), les régressions logistiques [115], les arbres de décision [116] ou les forêts aléatoires [117], sont conçus pour résoudre des tâches de classification.

Les tâches de survie sont donc souvent réduites à des tâches de classification [118],[119] (par exemple, la survie globale devient une survie à 6 mois, avec une tâche de classification binaire *survie/décès*).

Grâce à ce changement de définition, les modèles de ML classiques peuvent être utilisés directement. Cependant, cela entraîne une perte de précision en passant d'une cible continue à une cible discrète plus large. Les patients ne peuvent alors être stratifiés qu'en classes et, une fois le seuil de temps passé, le modèle n'a plus d'informations supplémentaires pour apprendre.

Pour conserver la nature continue de l'information de survie, le choix logique suivant est d'utiliser des modèles de régression. Ces modèles utilisent une valeur continue comme cible et sortie, ce qui permet de modéliser des variables telles que le temps jusqu'à l'événement (décès, rechute, progression de la maladie) ou le risque qu'un événement survienne. Le problème des modèles de régression est qu'ils ne gèrent pas nativement un aspect crucial des données de survie : la **censure**.

En effet, les ensembles de données médicales sont souvent collectés pendant une durée limitée, après quoi la base de données est figée et les patients n'ayant pas connu l'événement suivi sont censurés à droite. La valeur obtenue n'est alors pas la valeur réelle, mais une borne inférieure : par exemple, une étude avec un suivi de trois ans fixera le temps de survie à trois ans, pour tous les patients encore en vie après trois ans, mais ces patients seront marqués comme censurés. C'est ce qu'on appelle la censure de type I. Parfois, le contact peut être perdu avec un patient sans qu'il ne quitte officiellement l'étude. Dans ce cas, le patient est marqué comme censuré et son temps jusqu'à l'événement est fixé à la date de sa dernière visite.

Pour gérer la censure, les fonctions de vraisemblance sont adaptées pour inclure uniquement les cas avec événement via la fonction de densité de probabilité, tandis que les cas censurés sont inclus directement via la fonction de survie, au moment de leur censure.

5.2.3 Métriques pour l'évaluation de la prédiction de survie

Pour évaluer la performance des modèles de prédiction de survie, une métrique appropriée, capable de gérer des valeurs continues, doit être utilisée.

Différents types de performances peuvent être évalués selon le modèle :

— Les métriques de classement évaluent la capacité du modèle à classer un ensemble de cas en fonction de leur risque ou de leur temps de survie ; elles sont cruciales pour les modèles sans régression qui attribuent des rangs aux patients.

— Les métriques de précision, en revanche, doivent être utilisées pour les modèles fournissant une prédiction temporelle réelle, car elles comparent le résultat observé avec la valeur prédite.

Ci-dessous, une métrique est présentée pour chacune de ces catégories.

5.2.3.1 Les indices de concordance

Les indices de concordance sont des métriques qui évaluent la qualité du classement des patients selon leur temps de survie. Le *C-Index* de Harrell [120] a été démontré comme étant une extension de la courbe ROC pour les données dépendantes du temps [121]. Cette métrique compare, pour chaque paire de patients, le rang relatif de leur temps de survie et de leur survie prédite par le modèle.

Pour tenir compte de la censure, les paires de patients i et j sont considérées comme concordantes si

$$y_i < y_j \wedge \hat{y}_i < \hat{y}_j \tag{5.1}$$

ou si

$$y_i > y_j \wedge \hat{y}_i > \hat{y}_j \tag{5.2}$$

Elles sont discordantes seulement si

$$y_i < y_j \wedge \hat{y}_i > \hat{y}_j \tag{5.3}$$

si la survie du patient i n'est pas censurée, ou si

$$y_i > y_j \wedge \hat{y}_i < \hat{y}_j \tag{5.4}$$

si la survie du patient j n'est pas censurée.

L'indice C de Harrell est alors calculé par la formule :

$$CI = \frac{nb_{concordant}}{nb_{concordant} + nb_{discordant}} \tag{5.5}$$

En tant que métrique basée sur le rang, le C-Index ne peut être calculé que pour plusieurs patients, avec une robustesse croissante selon le nombre de cas. Différentes versions de le C-Index ont été proposées pour tenir compte de la censure et des égalités dans le classement.

L'utilisation de poids de censure inversement proportionnels (IPCW), comme dans l'indice C d'Uno [122], donne plus de poids aux échantillons non censurés pour lesquels des échantillons ayant un temps similaire sont censurés. Les égalités peuvent également être traités de différentes manières

[123], soit en considérant les échantillons avec des résultats ou des prédictions identiques comme non comparables, soit totalement comparables, soit partiellement, avec un poids inférieur à un.

Cette métrique est la plus utilisée pour évaluer les modèles de survie [124], bien qu'elle soit limitée car elle ne mesure pas la précision des prédictions temporelles, mais seulement leur ordre, et n'évalue pas la calibration du modèle.

5.2.3.2 Le Score de Brier

Le score de Brier [125] (ou indice de Brier) calcule la différence quadratique moyenne entre la distribution des résultats et la distribution des temps de survie prédits. Cette différence est calculée, dans la version originale du score de Brier, à un ensemble donné de points temporels, ce qui la rend dépendante des temps d'évaluation choisis. Pour résoudre ce problème, le score de Brier intégré (IBS) considère théoriquement l'intégrale du score de Brier, soit le score de Brier à tous les points temporels de l'étude. Dans les expériences réelles, l'IBS consiste à calculer le score de Brier pour tous les points disponibles.

Ces deux métriques mesurent la précision d'un modèle en calculant ces différences entre les temps de survie observés et prédits. En raison de cela, ils ne peuvent être utilisés que pour évaluer des modèles qui produisent un temps de survie et non une unité arbitraire, comme un score de risque.

Si le nombre de patients dans l'ensemble de test est suffisamment important pour ne considérer que les patients non censurés, l'erreur quadratique moyenne (MSE) et l'erreur absolue moyenne (MAE) peuvent être utilisées pour évaluer la précision des modèles de survie.

5.3 *Radiomics Augmentation* : augmentations de données pour combattre le surapprentissage dans les études radiomiques

5.3.1 Surapprentissage et augmentations de données

Les ensembles de données d'imagerie médicale ont souvent été de taille limitée, ce qui a par la suite limité les performances des méthodes de ML appliquées à des tâches médicales spécifiques [126]. En particulier, les modèles entraînés avec une quantité limitée de données sont sujets au surapprentissage (*overfitting*)[127], un phénomène où un modèle obtient de bons résultats sur ses données d'entraînement mais échoue à généraliser sur de nouvelles données utilisées pour l'évaluer.

De plus, le nombre de caractéristiques d'entrée présentées au modèle de ML peut parfois être plus important que le nombre d'échantillons dans l'ensemble de données, ce qui crée une situation avec plus d'inconnues (l'importance de chaque caractéristique radiomique) que d'équations (chaque patient de l'ensemble d'entraînement).

Les modèles d'apprentissage profond (*Deep Learning*) ont toujours été sensibles au surapprentissage lorsqu'ils sont entraînés sur de petits ensembles de données, en raison du grand nombre de paramètres entraînables dans leur architecture. Ce problème a été atténué en partie grâce à l'utilisation d'augmentations de données (*data augmentation*, DA). Les DA consistent en des transformations appliquées aux données pour créer des échantillons légèrement différents, agissant comme de nouveaux échantillons. Ces DA doivent être réalistes et représenter la variabilité trouvée dans la distribution réelle des données.

L'utilisation de DA permet donc au modèle d'apprendre à partir de données qui correspondent mieux à la distribution réelle, tout en diminuant l'importance des spécificités propres aux données d'entraînement. Cela aide le modèle à généraliser, augmentant ses performances lorsqu'il est confronté à des données qu'il n'a jamais vues auparavant. Cette méthode est aussi ancienne que le DL, l'inclusion d'images déformées ayant montré une réduction de l'erreur de test pour la reconstruction de chiffres avec LeNet [15].

Différentes DA sont disponibles selon le type de données [128] et de nouvelles augmentations sont créées pour des cas d'utilisation spécifiques, comme l'imagerie médicale [129], la conduite autonome [130] et la reconnaissance vocale [131]. Les DA génériques incluent l'ajout de bruit (généralement gaussien, mais peut être d'autres types, comme le bruit de Poisson pour les sinogrammes TEP [132]), les transformations géométriques (traductions, rotations et mises à l'échelle), les modifications de valeurs (modification du contraste, exponentialisation).

5.3.2 Augmentations de données et radiomique

Dans le domaine des radiomiques, les DA ont été principalement appliquées directement sur les caractéristiques, après leur calcul et stockage sous forme tabulaire. Par exemple, la technique de sur-échantillonnage synthétique des minorités (SMOTE) [133] et sa variante haute dimensionnelle [134] sélectionnent des échantillons des classes minoritaires avec une probabilité plus élevée pour traiter le déséquilibre des classes.

D'autres méthodes ont été présentées pour générer des données radiomiques synthétiques, en utilisant des auto-encodeurs [135], des *Generative Adversarial Networks*[136](GAN) et des GAN tabulaires conditionnels[137] (CTGAN). Les DA ont également été appliquées aux images, soit pour mesurer la sensibilité d'une caractéristique à une perturbation [138], soit pour améliorer la robustesse dans des régimes à faible données, par exemple pour la classification du score de Gleason dans le cancer de la prostate [139].

Dans cette étude, nous présentons le framework *Radiomics Augmentation*, dans lequel les DA sont utilisées sur les images et les masques de segmentation avant le calcul des caractéristiques radiomiques, afin de mesurer si elles conduisent à une performance accrue et à un surapprentissage réduit.

À cette fin, un benchmark de ML a été construit pour évaluer huit modèles différents sur la tâche de prédiction de la survie sans récidive (Relapse Free Survival, RFS), avec et sans *Radiomics Augmentations*.

5.4 Matériel et méthodes

5.4.1 Le jeu de données HECKTOR 2022

L'ensemble de données fourni pour la seconde tâche du challenge HECKTOR22 [140] a été utilisé dans cette étude. Dans ce jeu de données, les scanners TEP/TDM de 489 patients souffrant d'un cancer de la tête et du cou et traités par radiothérapie ont été collectés. Pour chacun de ces patients, les contours du *Gross Tumor Volume* (GTV) de la tumeur primitive (GTV_p) et des ganglions lymphatiques affectés (GTV_n) ont été délimités. Ce contourage a été réalisée par un premier annotateur, puis vérifié par un second, les deux disposant de *guidelines* de contourage.

Pour les besoins de notre étude, seules les contours manuelles ont été utilisées, afin de ne pas introduire de biais via l'implémentation d'une méthode de segmentation automatisée spécifique.

La tâche consistait à prédire la RFS, soit le temps (en jours) entre la fin du traitement de radiothérapie et le décès (ou la rechute) du patient. Pour utiliser uniquement les contours manuels lors de l'extraction des radiomiques, seul l'ensemble d'entraînement a été utilisé dans cette étude, et la performance de généralisation est évaluée grâce à une validation croisée. Ce choix est également motivé par des contraintes expérimentales : l'ensemble de test étant privé et limité à trois soumissions par mois, il n'était pas possible de l'utiliser pour évaluer plusieurs modèles et effectuer plusieurs cycles pour chaque modèle.

Pour chaque patient, les scanners TEP et TDM sont fournis sous forme de volumes 3D, avec les contours des régions d'intérêt (ROI), le tout stocké au format NIfTI. Comme ils ont été acquis simultanément, les volumes TEP et TDM sont supposés recalés l'un par rapport à l'autre. Les contours sont fournis sous forme de masques d'entiers multi-classes (0 : arrière-plan, 1 : GTV_p, 2 : GTV_n).

En plus des données d'imagerie, des informations cliniques ont été fournies (sexe, âge, chimiothérapie, etc.). Cependant, dans cette étude, les données cliniques n'ont pas été incluses, car nous nous sommes concentrés l'effet des augmentations de données sur les caractéristiques radiomiques.

La RFS, critère d'évaluation de notre étude, a été fournie pour chaque patient. En l'absence d'événement, le patient est considéré comme **censuré** et la valeur de RFS est fixée au temps écoulé entre la radiothérapie et la date de dernières nouvelles.

Les modèles de ML ont été évalués sur la tâche consistant à attribuer un score de risque à chaque patient. Ce score devait permettre de classer correctement les patients du plus au moins susceptible de subir un événement après N jours.

5.4.2 Augmentations de données

Dans notre étude, plusieur DA ont été appliquées aux images (TEP et TDM) et aux masques de contours (GTV_p et GTV_n).

Les DA appliquées aux images sont les suivantes :

— **Flou (Blur) :** application d'un filtre gaussien sur l'image, avec une taille de noyau (en millimètres) choisie selon une distribution uniforme entre 0 et 4.

— **Bruit (Noise) :** ajout d'un bruit gaussien à l'image (moyenne 0, écart-type entre -0,1 et 0,1).

— **Gamma :** modification du contraste de l'image en élevant ses valeurs à la puissance γ (γ choisi entre $-\exp 0.9$ et $\exp 0.9$).

— **Décalage de Contraste (ContrastShift) :** modification du contraste en normalisant l'image par Z-Score et en multipliant par un multiplicateur (0,9 ou 1,1) avant de restaurer les paramètres originaux.

— **Luminosité Multiplicative (MultiplicativeBrightness) :** multiplication des valeurs de l'image par un nombre (0,8 ou 1,2).

— **Simulation de Basse Résolution (SimulateLowResolution) :** sous-échantillonnage de l'image par interpolation au plus proche voisin, puis ré-échantillonnage vers les dimensions d'origine.

Des DA ont également été appliquées aux contours GTV_p et GTV_n, traités séparément comme des masques binaires :

— **Dilatation :** application d'une opération de dilatation sur le masque avec un noyau de 4 mm.

— **Seuillage Relatif SUV (SUVRelativeThreshold) :** re-segmentation du contour en utilisant les valeurs SUV du scanner TEP. Les contours sont définis avec un seuil fixé à 40% de la valeur SUVmax à l'intérieur du contour original.

— **Seuillage Absolu SUV (SUVAbsoluteThreshold) :** re-segmentation à l'aide de seuils absolus (2,5 et 4 SUV).

Au total, 8 transformations d'image et 4 transformations de masque ont été utilisées, ce qui signifie qu'en incluant les versions non modifiées, il y avait $9 \times 5 = 45$ fois plus d'échantillons dans l'ensemble augmenté que dans l'original.

22 005 échantillons (dont 21 516 augmentés) ont donc été obtenus à partir des 489 patients d'origine.

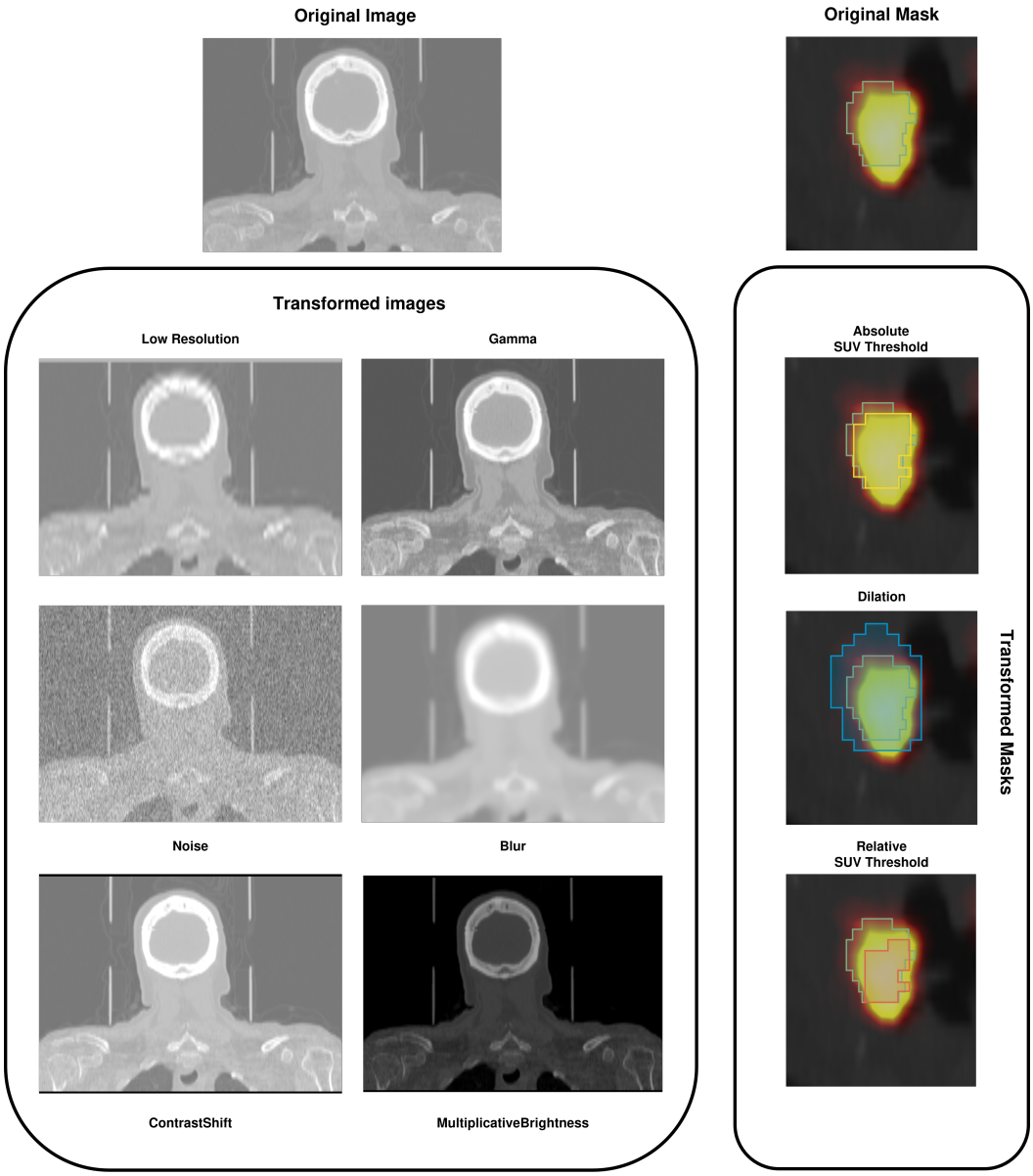

Figure 5.1 – Représentation visuelle des transformations appliquées sur les images et les masques. Le contour vert est le contour manuel d'origine, tandis que les jaunes, bleus et rouges sont les contours modifiés par les DA.

5.4.2.1 Visualisation de l'impact des DA sur la distribution des données

Notre première expérience a consisté à visualiser comment la distribution des échantillons augmentés se comparait à la distribution originale, via une projection UMAP (*Uniform Manifold Approximation Projection*). UMAP est un algorithme de réduction de dimensionnalité qui préserve mieux la structure que t-SNE ou PCA.

Les vecteurs projetés en 2D ont été affichés avec la bibliothèque Matplotlib, avec un code couleur représentant la valeur RFS en jours. Les échantillons originaux étaient représentés par des points de

plus grand diamètre. Cette représentation a permis d'évaluer si le processus d'augmentation comblait les lacunes de la distribution des données sans trop s'éloigner de la distribution originale.

5.4.3 Radiomique

107 descripteurs radiomiques ont été utilisés, extraits via PyRadiomics. Ces caractéristiques incluaient 18 caractéristiques de premier ordre, 14 descripteurs de forme 3D et 75 caractéristiques de texture. Les radiomiques ont été obtenues séparément pour les masques GTV_p et GTV_n, ainsi que pour les images TEP et TDM, pour un total de 4 paires image/masque, soit 428 caractéristiques radiomiques par échantillon.

Après suppression des variables égales, 221 caractéristiques radiomiques uniques par échantillon ont été conservées pour l'entraînement des modèles de ML.

5.4.4 Apprentissage automatique

Modèles de ML Plusieurs modèles ont été intégrés au benchmark :

- FS-SVM (SVM rapide pour la survie)
- CoxSA (réseau Cox avec régularisation ElasticNet)
- Bagged Icare (modèle d'ensemble limitant le surapprentissage)
- Random Survival Forest (RSF)
- Gradient Boosting Survival Analysis (GB-SA).

Chaque modèle a été entraîné pour classer les patients par score de risque.

Sélection de caractéristiques Le meilleur modèle a été choisi via une sélection exhaustive univariée basée sur l'indice C (UCI) : 1) classement des variables par importance via un modèle CoxNet univarié, 2) entraînement de modèles multivariés avec k variables, 3) sélection du nombre de variables optimal (maximisant le score sur l'ensemble d'entraînement.

Évaluation L'indice de concordance de Harrell a été utilisé pour évaluer les performances. Une validation croisée à 5 folds, répétée 10 fois a été mise en place. **Il est crucial de noter** que pendant la validation croisée, les folds de test ne contenaient que des échantillons originaux et non modifiés, afin d'évaluer la généralisation sur des données réelles uniquement.

5.5 Résultats

5.5.1 Distribution des données

La distribution des échantillons augmentés ressemble à la distribution originale, avec un accent mis sur certaines structures existantes. Dans certains cas, les vecteurs augmentés ont comblé des espaces vides, suggérant qu'ils ont complété des parties manquantes de la distribution ou se sont légèrement éloignés de la distribution initiale. Le fait que les zones vides de la distribution originale soient restées majoritairement vides suggère que la DA n'a pas déplacé les données de manière significative hors distribution, mais a plutôt généré des échantillons complétant la distribution initiale.

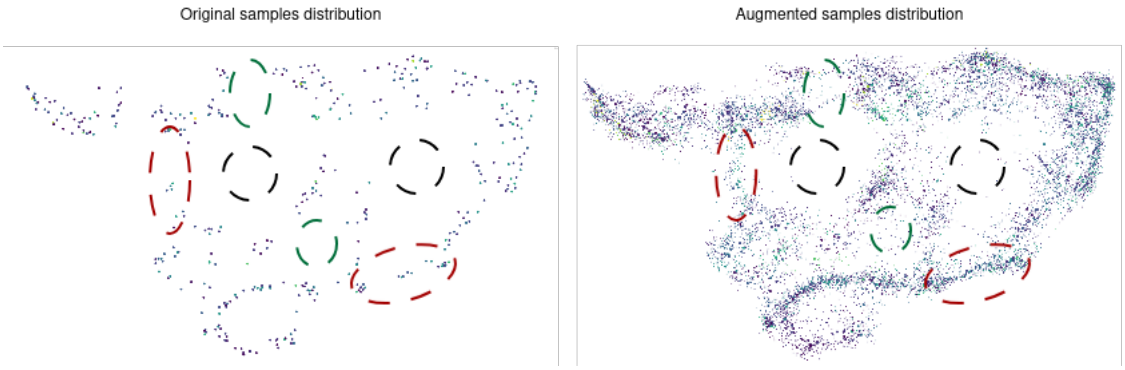

Figure 5.2 – Résultats de la projection UMAP appliquée aux caractéristiques radiomiques originales et augmentées de l'ensemble de données HECKTOR22. L'image de gauche présente les projections 2D des vecteurs radiomiques pour tous les cas originaux. Celle de droite présente les projections pour les cas augmentés. Le code couleur des points avec la RFS vise principalement à améliorer la lisibilité.

5.5.2 Impact sur la performance des modèles

Model	Basic Radiomics	Augmented Radiomics	Performance Change
FS-SVM	0.571	**0.576**	+0.8%
CoxSA	0.580	**0.595**	+2.5%
Bagged Icare (10 estimators)	0.569	**0.618**	+8.6%
Bagged Icare (100 estimators)	0.593	**0.614**	+3.5%
RSF (10 estimators)	0.606	**0.607**	+0.2%
RSF (100 estimators)	0.647	**0.652**	+0.8%
GB-SA (10 estimators)	**0.621**	0.614	-1.1%
GB-SA (100 estimators)	0.632	**0.636**	+0.6%

Table 5.1 – C-Index for each model, using basic and augmented radiomics. In the last column the percent change in performance is added.

Pour 7 modèles sur 8, une meilleure prédiction de la RFS a été obtenue lorsque les radiomiques augmentées étaient incluses dans l'ensemble d'entraînement. La seule exception est le modèle GB-SA avec 10 estimateurs. Les améliorations de performance vont de **+0,2 %** pour la plus petite RSF à **+8,6 %** pour le plus grand Bagged Icare. Le RSF avec 100 estimateurs entraîné sur les données augmentées a obtenu la meilleure performance globale avec un **indice C de 0,652**.

5.5.3 Impact sur la sélection de caractéristiques et le surapprentissage

L'utilisation d'augmentations radiomiques a permis de réduire l'écart entre CI_{train} et CI_{test}. Pour FS-SVM, CoxSA et Bagged Icare, les valeurs maximales de performance sur l'entraînement et le test tendent à coïncider davantage lorsque les données augmentées sont utilisées. Cela démontre

une meilleure robustesse : même si le nombre de variables sélectionnées était différent, la baisse de performance sur le test serait probablement plus limitée avec les radiomiques augmentées.

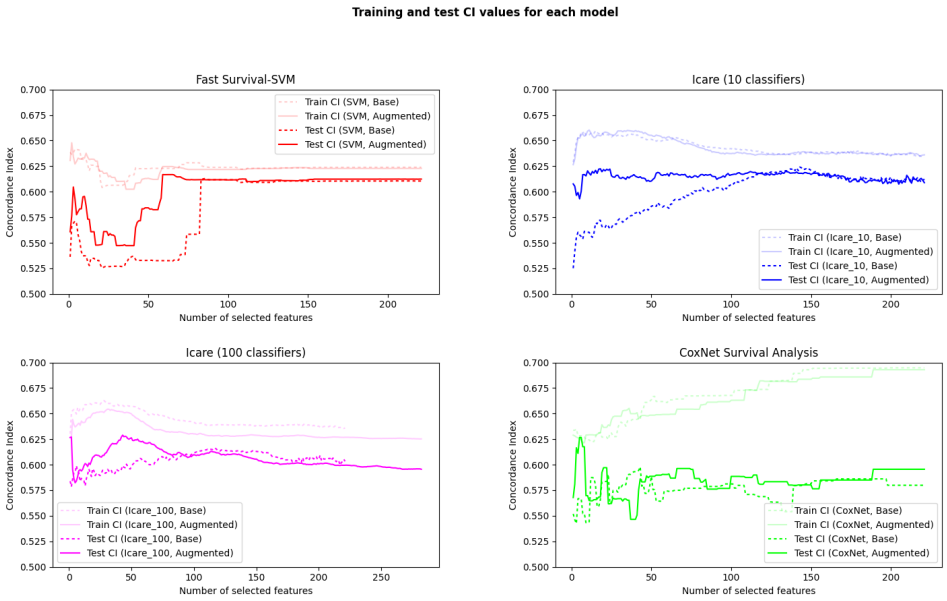

Figure 5.3 – Les valeurs de CI_{train} et CI_{test} sont présentées, avec le nombre de variablesélectionnées en abscisse. Les lignes en pointillées et complètes représentent les performances des radiomiques orioginales et augmentées respectivement

5.6 Discussion

5.6.1 Enseignements

L'augmentation radiomique a montré un fort potentiel pour améliorer la performance des modèles ML entraînés sur des caractéristiques radiomiques. En appliquant des transformations d'images et de masques couramment utilisées en DL avant l'extraction des caractéristiques, nous avons obtenu jusqu'à 8,6% d'amélioration du C-Index.

Notre hypothèse initiale selon laquelle ce gain de performance est dû à une réduction du sur-apprentissage semble renforcée par les résultats. Le rétrécissement de l'écart entre les métriques d'entraînement et de test au point optimal de sélection des caractéristiques montre qu'une meilleure généralisation peut être obtenue via l'utilisation de DA.

5.6.2 Limites

Une étude plus complète pourrait inclure d'autres transformations (affine, élastique, érosion) ou des combinaisons de transformations. Les paramètres de transformation pourraient également être optimisés par pathologie. De plus, une sélection rigoureuse des échantillons augmentés (pour exclure ceux trop éloignés de la distribution réelle) pourrait encore améliorer les résultats. Enfin, bien que validée sur un large ensemble de données TEP/TDM pour le cancer de la tête et du cou, cette

méthode devrait être testée sur d'autres modalités (IRM, échographie 3D) et d'autres tâches (survie globale, réponse au traitement).

5.7 Conclusion

Cette étude a démontré que les augmentations de données peuvent être appliquées aux images avant l'extraction des caractéristiques radiomiques pour créer des échantillons augmentés suivant une distribution similaire à l'originale. Sept modèles sur les huit évalués ont montré une amélioration de leurs performances. De plus, nous avons observé une réduction du risque de surapprentissage pour quatre modèles.

Ces résultats constituent une première validation méthodologique de l'augmentation radiomique, ouvrant la voie à une application plus large pour améliorer la robustesse des modèles de ML en imagerie médicale.

Le code de *Radiomics Augmentation* est disponible à l'adresse suivante : `https://www.github.com/Zhack47/radiomics-augmentation`.

Chapitre 6

Organomics : Extracting radiographic information from organs imaging

Chapter Overview

Radiomics derived from manually delineated lesion contours are widely used in medical imaging studies, but this approach has many limitations. First, the human contouring is not always reproducible[141] from one observer to another. Even though the availability of precise guidelines improves contours reproducibility, manual delineation remains a bottleneck in many studies, due to its time-intensive nature. Additionally, only a limited amount of information is obtained when only the lesions, such as the primary tumor and malignant lymph nodes, are analyzed.

In 2024, a team of researchers explored the possibility to use automated organ segmentation tools in conjunction with radiomics, to extract the information contained in the organs instead of the lesions. They introduced the term organomics (organ-radiomics) to describe their method[7]. However, to the best of our knowledge, no publicly available end-to-end pipeline has automated the extraction of organ-based radiomics extraction from medical imaging datasets.

To fill this gap, we have developed and made publicly available **Organ-omics**, an automated, configurable Python pipeline. This pipeline (1) generates organ contours (2) groups organs according to a user-defined configuration and (3) extracts reproducible organ-level radiomic features.

To validate the usability and the performances of our pipeline, Organ-omics was applied to the task of survival prediction using PET/CT imaging from a lung cancer trial : RTEP7 IFCT-1402. Specifically, we compared the performance of Organ-omics, lesion radiomics and clinical data for this task.

Contents

6.1 Vue d'ensemble 53
6.2 Description détaillée 53
6.2.1 Structure et Configuration 53
6.2.2 Segmentation des organes 54
6.2.3 Extraction de la radiomique 54
6.3 Applications et Limites 55
6.3.1 Applications 55
6.3.2 Limites et Améliorations futures 55
6.4 Application des organomiques à la prédiction de la survie des patients atteints de CBNPC 55
6.4.1 Introduction 55
6.4.2 Matériel et Méthodes 55
6.4.3 Résultats 57
6.4.4 Discussion 57
6.4.5 Conclusion 58

6.1 Vue d'ensemble

Organ-omics est une bibliothèque Python conçue pour extraire et stocker des caractéristiques radiomiques à l'échelle de l'organe. Cette extraction peut être configurée *via* un fichier JSON (voir ci-dessous).

Une fois ce fichier de configuration analysé, TotalSegmentator (TS) est exécuté, segmente les organes, puis les annotations sont assemblées selon la configuration. Les masques de segmentation multi-labels sont ensuite stockés au format NIfTI dans un dossier choisi par l'utilisateur. Organ-omics privilégie le format NIfTI pour le stockage et la lecture des images médicales, facilitant ainsi son intégration dans des flux de prétraitement plus larges, bien qu'il puisse être utilisé comme un outil autonome pour des études de radiomique organique.

La seconde partie utilise automatiquement les contours obtenus précédemment, parallèlement aux images, pour calculer les caractéristiques radiomiques de chaque groupe d'organes et pour chaque modalité d'imagerie via PyRadiomics. Ces caractéristiques sont ensuite stockées dans un unique fichier CSV. Cette extraction radiomique accepte les paramètres définis dans la configuration JSON.

6.2 Description détaillée

Cette section détaille chaque composant d'Organ-omics : la structure initiale des données et le fichier de configuration, l'algorithme de segmentation des organes et le calcul de l'organomique. Dans la suite de ce chapitre, *Organ-omics* fait référence à notre pipeline, tandis qu' *organomique* désigne les caractéristiques radiomiques des organes.

6.2.1 Structure et Configuration

Organ-omics a été intentionnellement divisé en deux parties : la segmentation, *puis* le calcul organomique. Cette conception rend le contourage et le calcul des radiomiques indépendants, permettant d'utiliser différents contours pour l'analyse : contours manuels, contours issus d'un autre algorithme ou contours de TS révisés par un radiologue expert.

Pour le bon fonctionnement d'Organ-omics, seules deux exigences doivent être respectées : une organisation spécifique des données au format NIfTI et un fichier de configuration JSON.

Structure La structure des données suit une arborescence de dossiers similaire à celle utilisée dans le framework nnUNet[70] :

```
Dossier Racine
├─ imagesTr
│  ├─ P01_0000.nii.gz
│  ├─ P01_0001.nii.gz
│  ├─ P02_0000.nii.gz
│  └─ P02_0001.nii.gz
└─ labelsTr
   └─ <vide>
```

Configuration Pour permettre l'adaptation d'Organ-omics, sans code, un fichier de configuration JSON est défini par l'utilisateur comprenant plusieurs champs requis :

— **channel_names** : Un dictionnaire des modalités d'imagerie avec leur index correspondant (0, 1, 2 ..., *nb_modalities*). L'une d'entre elles doit être nommée ***CT*** et servira à la délinéation des organes via TS. Les autres noms de modalités sont conservés pour l'extraction radiomique (un ensemble de caractéristiques est extrait par paire modalité/masque).

— **labels** : Un dictionnaire contenant les noms de tous les groupes d'organes avec leur numéro d'annotation label(le fond doit correspondre à l'étiquette 0). Ces étiquettes finales peuvent correspondre à un ou plusieurs volumes d'intérêt (VOI) de TS regroupés.

— **tasks** : Une liste des tâches de TS à exécuter pour obtenir les organes inclus dans les VOI définis. Chaque élément doit contenir :

  — **name** : Le nom de la tâche (ex : *total*, qui délinée 117 organes).

  — **correspondence** : Un dictionnaire associant les **labels** définis précédemment aux listes d'organes de la tâche TS correspondante.

6.2.2 Segmentation des organes

Une fois Organ-omics installé et les données configurées, deux commandes sont disponibles. La première, **Organomics_contour_dataset**, effectue la segmentation des organes selon les étapes suivantes :

1. **Initialisation des tâches** : Les tâches TS définies sont initialisées, en utilisant *roi_subset* pour réduire le temps de calcul.
2. **Inférence de segmentation** : Chaque patient est traité par les tâches TS.
3. **Regroupement des étiquettes** : Les contours finaux sont obtenues par combinaisondes résultats des tâches TS.
4. **Sauvegarde** : Les masques finaux sont sauvegardés dans un fichier NIfTI unique par patient.

6.2.3 Extraction de la radiomique

La seconde commande, **Organomics_extract_radiomics**, calcule les caractéristiques organomiques pour chaque paire organe/image et les stocke dans un fichier CSV. Le processus suit trois étapes par patient :

1. **Chargement** : Les images (CT et autres modalités recalées) et les masques sont chargés via SimpleITK[142].
2. **Rééchantillonnage (optionnel)** : Si un espacement est défini, les volumes sont ré-échantillonnés vers cette résolution.
3. **Calcul des radiomiques** : Via PyRadiomics, les 107 caractéristiques radiomiques standard sont calculées pour chaque paire modalité/masque.

Le code source et les instructions d'installation sont disponibles sur : `https://www.github.com/Zhack47/Organ-omics`.

6.3 Applications et Limites

6.3.1 Applications

Organ-omics permet d'explorer les liens entre l'imagerie des organes et des observations cliniques tels que la survie des patients ou la réponse au traitement. En favorisant la découverte de nouveaux biomarqueurs universels (applicables même aux patients sans lésions visibles), il vise à faciliter une médecine personnalisée. L'imagerie oncologique, riche en modalités quantitatives (TEP, TDM, IRMq), est le domaine d'application privilégié, mais l'outil peut s'étendre aux maladies musculo-squelettiques (analyse des os et des muscles).

6.3.2 Limites et Améliorations futures

La principale contrainte actuelle est la nécessité d'un scanner CT comme référence de segmentation. L'utilisation de modèles basés sur l'IRM[143] pourrait lever cette restriction. Une autre limite concerne la qualité des contours, dépendante du champ de vue de l'imagerie : un organe partiellement hors champ faussera les caractéristiques calculées.

Enfin, l'absence de filtres d'image (LoG, Wavelet) dans la version actuelle limite le nombre de caractéristiques, bien que leur ajout risque d'accentuer le déséquilibre entre le nombre de variables et le nombre d'échantillons.

6.4 Application des organomiques à la prédiction de la survie des patients atteints de CBNPC

Pour démontrer l'utilité d'Organ-omics, nous l'avons appliqué à un jeu de données TEP/TDM (deux points temporels) pour prédire la survie globale (Overall Survival, OS) de 138 patients atteints de cancer bronchique non à petites cellules (CBNPC) et traités par radio-chimiothérapie.

6.4.1 Introduction

Le CBNPC est le cancer le plus meurtrier au monde. La TEP au ^{18}F-FDG permet de mesurer le métabolisme du glucose et d'évaluer l'agressivité tumorale. Récemment, l'extension de la radiomique aux organes sains (organomique) a montré un fort potentiel prédictif pour la survie[144].

En parallèle, l'essai RTEP7 IFCT 14.02, qui étudiait l'escalade de dose via une TEP de mi-traitement, a servi de base à notre étude.

6.4.2 Matériel et Méthodes

L'essai multicentrique RTEP7 incluait deux examens TEP/TDM : un avant traitement (TEP/TDM1) et un après 42 Gy (TEP/TDM2).

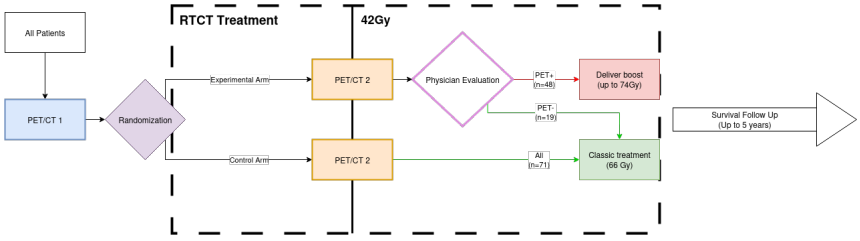

Figure 6.1 – Le protocole RTEP7 IFCT-1402 : tous les patients inclus ont un scan TEP/TDM avant d'être assignés à une des deux branches : expérimentale ou contrôle, puis d'avoir un deuxième scan TEP/TDM, après 42 Gray de traitement. Dans le bras de contrôle, le traitement classique est suivi (66 Gray), dans le bras expérimental le traitement est adapté en fonction de ce deuxième examen (66 ou 74 Gray, en fonction de la réponse).

Trois médecins nucléaires ont segmenté manuellement les lésions (Volume Tumoral Biologique, VTB). En parallèle, 7 groupes d'organes thoraciques (trachée, oesophage, coeur, aorte, moelle épinière, vertèbres thoraciques et poumons) ont été segmentés automatiquement par TS.

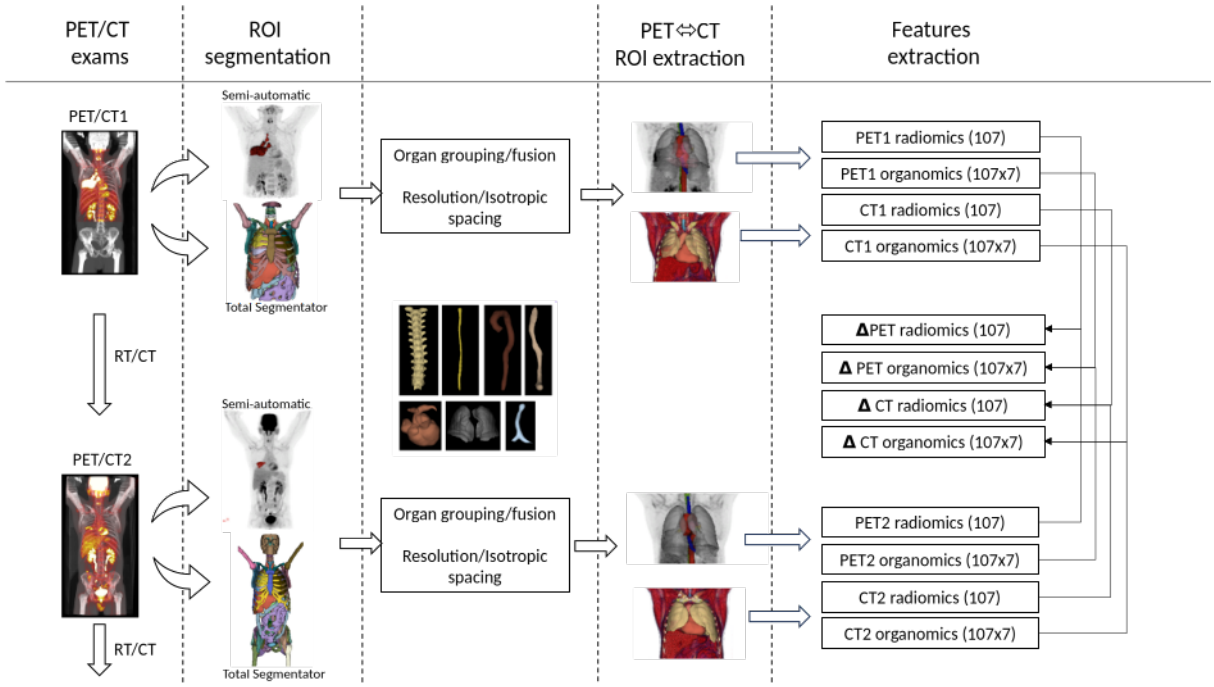

Figure 6.2 – Vue d'ensemble de la pipeline d'extraction des caractéristiques. Les contours d'organes ont été extraits des scans TDM avec TS. Nous obtenons deux ensembles de caractéristiques, un pour chaque point temporel, et un troisième, Δ F, la différence entre les caractéristiques à ces deux points.

6.4.2.1 Benchmark d'Apprentissage Automatique (Machine Learning)

Pour comparer la puissance prédictive de l'organomique et de la radiomique lésionnelle, un benchmark fixe a été construit :

1. **Sélection de caractéristiques** : Utilisation de l'indice de concordance univarié (UCI) pour

			Radiomics		Organomics		Results	
Model	Nb Feat.	Clin.	PET	CT	PET	CT	CI ($\pm Cint$) ↑	cdAUC ($\pm Cint$) ↑
Clinical	6	✓(3)					0.573 (±0.008)	0.592 (±0.012)
Radiomics	327	✓(3)	✓(20)				0.517 (±0.012)	0.515 (±0.017)
	327	✓		✓(29)			0.591 (±0.012)	0.612 (±0.017)
	648	✓	✓(8)	✓(29)			0.584 (±0.012)	0.602 (±0.016)
Organomics	2253	✓			✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2253	✓				✓(2)	**0.663** (±0.008)	0.688 (±0.012)
	4500	✓			✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
Combined (3)	2574	✓	✓		✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2574	✓	✓			✓(2)	**0.663** (±0.008)	0.688 (±0.012)
	2574	✓		✓	✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2574	✓		✓		✓(2)	**0.663** (±0.008)	0.688 (±0.012)
Combined (4)	2895	✓	✓	✓	✓(35)		0.648 (±0.008)	0.682 (±0.01)
	2895	✓	✓	✓		✓(2)	**0.663** (±0.008)	0.688 (±0.012)
	4821	✓	✓		✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
	4821	✓		✓	✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)
All	5142	✓	✓	✓	✓(35)	✓(15)	0.656 (±0.011)	**0.689** (±0.011)

Figure 6.3 – Toutes les combinaisons de données d'entrée sont utilisées et comparées. Une coche- signifie que le type de donnée a été utilisé : rouge ou vert, selon si il a été rejeté ou inclus par la FS

entraîner des modèles CoxNet.

2. **Validation croisée** : Une validation croisée à 6 plis, répétée 100 fois, utilisant un modèle *Fast Survival SVM*.
3. **Métriques** : Indice C de Harrell et Aire sous la courbe dynamique cumulative (cdAUC).

6.4.3 Résultats

L'analyse a porté sur 138 patients (âge moyen 62,5 ans).

Le modèle basé uniquement sur les données cliniques a obtenu un C-Index de $0,573 \pm 0,008$. L'ajout des radiomiques TEP a paradoxalement diminué les performances $(0,513)$.

Les modèles utilisant uniquement les organomiques ont montré les meilleures performances. Le modèle basé sur les organomiques de la TDM a atteint un C-Index de $0,663 \pm 0,008$ et une cdAUC de $0,688 \pm 0,012$. De manière notable, lorsque les caractéristiques organomiques étaient incluses, les variables cliniques et radiomiques lésionnelles n'étaient plus sélectionnées par l'algorithme, soulignant la dominance pronostique des descripteurs d'organes sains.

6.4.4 Discussion

Nos résultats suggèrent que les organomiques des organes thoraciques sains, en particulier les descripteurs issus de la TDM, fournissent des informations plus solides pour le pronostic de la survie globale que les radiomique tumorales classiques ($p \ll 0,001$). Cela pourrait s'expliquer par le fait que la texture des organes sains capture indirectement la réponse inflammatoire systémique du patient.

D'un point de vue clinique, l'impact des organomiques extraits de la TDM est particulièrement intéressant, car les TDM sont largement plus disponibles et moins coûteuses que les TEP. Parmi les caractéristiques clés, les variations de volume de l'oesophage entre les deux scans sont apparues comme des prédicteurs majeurs, suggérant une détection précoce de signaux de toxicité.

6.4.5 Conclusion

En déplaçant le regard de la tumeur vers les organes environnants, les organomiques semblent capturer la tolérance au traitement et la réponse systémique du patient. Cette approche ouvre de nouvelles voies pour la médecine personnalisée dans le traitement du cancer du poumon. Le code est disponible à l'adresse suivante : `https://www.github.com/Zhack47/Organ-omics`.

Chapitre 7

Conclusion and perspectives

7.1 Conclusion

Cette thèse a présenté l'état actuel des technologies d'IA pour l'imagerie multimodale en oncologie.

La segmentation entièrement automatisée et semi-automatisée est à l'avant-garde de l'intégration de l'IA dans la pratique clinique, tandis que les descripteurs d'imagerie demeurent un domaine de recherche ouvert.

La combinaison d'outils de segmentation automatisés avec ces descripteurs offre une nouvelle voie pour la prédiction de la survie et l'évaluation de l'efficacité d'un traitement.

Bien que les techniques de réseaux de neurones profonds (DNN) et d'apprentissage profond (DL) soient disponibles et présentent les meilleures performances pour la segmentation, la prédiction de la survie utilise encore majoritairement des méthodes d'apprentissage automatique (ML) classiques.

Nous avons démontré comment un pipeline de bout en bout pouvait néanmoins être créé, en utilisant des outils de segmentation à l'état de l'art (SoTA) pour détourer les volumes d'intérêt (VOI), en extrayant des descripteurs radiomiques de ces contours, puis en utilisant des modèles ML pour prédire la survie à partir de ces descripteurs tout en obtenant des caractéristiques explicatives significatives.

Voici un résumé des travaux les plus importants contribués au cours de cette thèse :

— **Algorithmes de segmentation automatisée TEP/TDM** : Dans le chapitre 4, les contributions à l'état de l'art en segmentation TEP/TDM ont été présentées.

  En commençant par un modèle simple de segmentation TEP/TDM au ^{18}F-FDG, entraîné avec une version modifiée de l'algorithme nnU-Net, qui a surpassé le modèle de référence et s'est classé 4ème au défi AutoPET II **(2023)**.

  La stratégie d'entraînement a ensuite été mise à jour pour inclure la supervision d'organes supplémentaires, une architecture U-Net résiduelle ainsi qu'une segmentation multi-traceurs (FDG et PSMA).

  Un CNN de classification des traceurs TEP, qui s'est révélé presque infaillible (précision de 99,64 %), a été introduit pour orienter la segmentation vers l'un des deux modèles experts par traceur, pour AutoPET III **2024**.

Enfin, l'intervention humaine a été intégrée au U-Net via la construction de cartes de chaleur (heatmaps) tenant compte des spécificités de la TEP, servant d' *a priori* informé par le domaine pour la segmentation.

La combinaison de toutes ces améliorations a culminé dans une contribution d'équipe pour AutoPET IV **2025**.

L'ensemble de ce travail est accessible publiquement, tant à travers des publications[9][10][11] que des dépôts contenant le code et les poids nécessaires pour entraîner et exécuter ces modèles sur des images TEP/TDM[145][146][147].

— **Le framework Radiomics Augmentation** : Un cadre de travail permettant d'appliquer des augmentations de données sur les images et les masques avant l'extraction des caractéristiques radiomiques.

Ce framework vise à transposer le concept d'augmentation de données, classique en DL, vers le ML pour les données médicales.

Une étude comparative mesurant la performance de 8 modèles de ML sur la tâche de prédiction de la survie sans récidive (RFS) à l'aide du jeu de données HECKTOR 2022 a révélé que la *Radiomics Augmentation* améliorait les performances pour 7 de ces 8 modèles et réduisait le surapprentissage (overfitting) lors de l'étape de sélection des caractéristiques pour 4 d'entre eux.

Le code utilisé pour appliquer la *Radiomics Augmentation* sur le jeu de données Hecktor2022 a été rendu public[148].

— **Le pipeline Organ-omics** : Un pipeline extrayant des descripteurs basés sur les organes à partir d'images médicales.

Ce pipeline de bout en bout incarne la philosophie de cette thèse : d'abord, la segmentation basée sur le DL est utilisée pour obtenir les contours des VOI pertinents, puis des descripteurs de haut niveau sont extraits de ces VOI et utilisés pour entraîner des modèles de ML pour des tâches moins densément supervisées.

Organ-omics a été conçu pour rendre le concept d'organomique[7] accessible aux chercheurs, afin de favoriser de nouvelles recherches utilisant ces descripteurs plus holistiques.

Le code et les instructions pour exécuter Organ-omics sur n'importe quel jeu de données sont disponibles en ligne[149].

Ce pipeline a été validé sur un jeu de données privé, collecté rétrospectivement à partir d'un essai clinique, contenant des données d'imagerie TEP/TDM à deux points temporels différents.

Le pouvoir prédictif de l'organomique, de la radiomique et des données cliniques a été comparé à l'aide de ces données et du critère d'évaluation recueilli : la survie globale (OS) censurée à droite.

Cette expérience a démontré la nette supériorité de l'organomique sur les données cliniques (0,663 contre 0,573) et la radiomique (0,663 contre 0,591).

De plus, il a été noté que l'organomique calculée sur le scanner (TDM) offrait les meilleures performances, ce qui a des implications cliniques intéressantes : la survie globale semble être liée à des caractéristiques d'imagerie visibles sur un simple scanner, plutôt que sur une TEP.

Les caractéristiques organomiques ayant la plus grande importance dans le modèle (extraites de ses poids) ont montré que les poumons et l'œsophage étaient les organes contenant les informations les plus pertinentes.

Les travaux produits au cours de cette thèse ont été mis à la disposition de la communauté de recherche, sous forme d'articles, de code et d'outils.

Ils offrent aux chercheurs une plateforme de départ pour mener de nouvelles études avec des concepts à l'état de l'art, sans le fardeau de l'implémentation technique initiale.

7.2 Perspectives

7.2.1 Vers des modèles fondationnels pour la segmentation d'images médicales

À partir des nombreux petits jeux de données de segmentation médicale, la tendance semble évoluer vers des jeux de données à grande échelle, avec des annotations pour un grand nombre de classes.

Cette tendance est corrélée à l'augmentation des performances des réseaux de neurones profonds (DNN) dans le domaine de la segmentation d'images médicales :

— La performance des DNN croît avec l'augmentation des données d'entraînement ; la segmentation d'images médicales ne fait pas exception à cette règle[150].

Bien que la qualité des images, des annotations et du protocole d'imagerie puisse influencer la performance, les données du monde réel sont hétérogènes et nécessitent des algorithmes de segmentation dotés de grandes capacités de généralisation.

Il est impossible d'inclure l'intégralité des images possibles dans un ensemble d'entraînement ; par conséquent, l'obtention d'un modèle de segmentation adapté à la pratique clinique repose sur les capacités de généralisation des DNN.

— Les DNN entraînés pour la segmentation sur de petits jeux de données peuvent être utilisés en apprentissage itératif ou en apprentissage actif (*active learning*), pour annoter de grands jeux de données à moindre coût.

Un DNN entraîné sur un petit jeu de données de haute qualité peut inférer des masques de segmentation sur un grand nombre d'images médicales.

Ces masques peuvent ensuite être utilisés comme pseudo-annotations pour entraîner un DNN de grande taille sur toutes les images précédemment non annotées.

Les pseudo-labels peuvent également être corrigés par des experts humains, permettant ainsi une annotation efficace de grandes bases de données.

Cette relation mutuellement bénéfique entre les jeux de données à grande échelle et les modèles de segmentation DNN est de plus en plus exploitée, le projet BodyMaps étant à l'avant-garde de cet effort.

Cependant, les grands jeux de données ne font qu'émerger pour la tâche de segmentation, laissant les autres tâches dispersées entre de multiples petits jeux de données, souvent privés.

7.2.2 Améliorer l'état de la prédiction de la survie

La prochaine étape pour améliorer l'utilisabilité clinique de l'IA sera de rassembler et de rendre publics des jeux de données destinés aux tâches de prédiction de la survie.

Cela serait nécessaire pour permettre le développement et la validation de modèles de prédiction de la survie fiables, robustes et ouverts.

En effet, nous avons vu ce processus se dérouler pour la segmentation d'images médicales : de la première percée (U-Net, 2015) aux premiers grands jeux de données publics (AutoPET (2022/2025), TotalSegmentator (2022)), aux modèles et frameworks d'entraînement de pointe (nnU-Net et MedNeXt, 2019/2023), à un cadre de validation rigoureux[151] et, enfin, à de très grands jeux de données publics (AbdomenAtlas 3.0[152], 2026) et des modèles de segmentation surpassant l'humain (segmentation de tumeurs pancréatiques, 2025[55]).

En ce qui concerne la prédiction de la survie, des modèles ont été mis à disposition (Scikit-Survival[47], Icare[92]), mais un consensus clair sur l'état de l'art et les meilleures pratiques d'entraînement et d'évaluation n'a pas encore été établi.

Ce manque d'un cadre d'entraînement et d'évaluation strict, parallèlement à l'absence de référence (*baseline*), conduit à une prolifération de méthodes, ce qui complique davantage l'élaboration d'une validation rigoureuse, car chaque nouvelle méthode devrait être évaluée comparativement.

De plus, ces méthodes sont principalement développées et évaluées sur des jeux de données restreints et privés, ce qui ne garantit pas leur capacité de généralisation à de nouvelles données et empêche la reproduction des travaux par d'autres chercheurs.

Cette absence de grands jeux de données ouverts entrave l'émergence d'une méthode de référence pour la prédiction de la survie.

Néanmoins, la construction de grands jeux de données de survie n'est pas hors de portée, surtout pour les équipes de recherche travaillant dans un environnement clinique.

En effet, certains pays, comme la France, maintiennent un service ouvert qui enregistre les dates de naissance et de décès des personnes décédées[153].

En utilisant ces bases de données, les cliniciens pourraient rassembler les examens d'imagerie de patients décédés et y associer le temps de survie en jours, à compter de cet examen.

Une fois les métadonnées d'imagerie supprimées (et le *defacing* appliqué, si nécessaire), ces paires examen/temps-jusqu'au-décès pourraient même être rendues publiques, conformément aux lois sur la protection de la vie privée.

Au vu des résultats présentés au chapitre 6, cette méthode pourrait être appliquée aux scanners (TDM), qui sont réalisés en quantités importantes et croissantes, tant dans les centres de lutte contre le cancer que dans les cabinets d'imagerie.

Une fois que des jeux de données publics à grande échelle seront disponibles, les efforts de recherche devront se concentrer sur l'évaluation des méthodes de prédiction de la survie disponibles afin d'établir un état de l'art robuste.

Certaines tentatives ont été faites avec des données cliniques issues de dossiers de santé électroniques (*Electronic Health Records* (EHR)), mais à notre connaissance, pas encore avec des données d'imagerie.

7.2.3 Quantifier le corps humain dans sa globalité

Notre motivation à inclure des données d'imagerie dans les prédictions n'est pas dogmatique ; elle découle d'un constat : l'imagerie, en particulier l'imagerie multimodale, offre au médecin la vision et la compréhension la plus complète du corps humain.

Table des figures

1.1 Une amélioration structurelle peut être obtenue en utilisant l'I.A. pour assister les médecins dans les tâches chronophages. Des tâches telles que la détection de lésions, la segmentation d'organes ou même la génération de comptes rendus d'imagerie, pourront ainsi être automatisées, réservant le temps du médecin aux décisions médicales les plus complexes. 3

2.1 L'Intelligence Artificielle consiste à appliquer une logique pour traiter des entrées. L'apprentissage automatique (ML) désigne spécifiquement les méthodes qui apprennent à partir d'exemples, en cherchant à généraliser à tous les cas possibles. L'apprentissage profond (DL) utilise des réseaux de neurones profonds pour cet apprentissage et nécessite souvent de larges jeux de données pour l'entraînement. 9

2.2 Une convolution $3 * 3$ appliquée à une matrice de pixels $4 * 4$ produit une matrice $2 * 2$. Par exemple : le pixel supérieur gauche de la matrice résultante est le produit de la combinaison des valeurs des 9 values du coin supérieur gauche $3 * 3$ de la matrice originale, par les poids du noyau de convolution. 10

2.3 Exemple de scanner (*s0591* de TotalSegmentator) ayant subi un floutage facial. À gauche : rendu volumique 3D. À droite : coupe coronale 2D montrant le floutage appliqué. 16

3.1 À partir d'une image, la segmentation consiste à classifier les pixels (ou voxels) selon l'objet auquel ils appartiennent. La segmentation sémantique regroupe toutes les instances d'un objet dans la même classe (ex. : personne), tandis que la segmentation d'instance attribue des étiquettes différentes à chaque instance (ex. : chaque personne individuellement). *Source : learnopencv.com* 19

3.2 Architecture originale de l'UNet. *Source : Ronneberger et al.*[17] 20

3.3 Le DSC compare l'intersection de deux formes géométriques avec la somme de leurs surfaces ou volumes. 21

4.1 Line Of Response (LOR) et le sinogramme correspondant, résultat de l'addition des LORs. *Source : Volkow et al.[74]* 28

4.2 Maximum Intensity Projection (MIP) d'une TEP au 18F-FDG normale (bénigne)[75]. La fixation du radiotraceur est visible dans le cerveau et le myocarde, qui consomment constamment du glucose, ainsi que dans les reins et la vessie. Une fixation thyroïdienne peut également être observée, signe possible d'une pathologie non cancéreuse[76]. *Source : radiopaedia.org* 29
4.3 L'algorithme de construction des cartes de probabilité. *Crédit : Solène Perret.* 32

5.1 Représentation visuelle des transformations appliquées sur les images et les masques. Le contour vert est le contour manuel d'origine, tandis que les jaunes, bleus et rouges sont les contours modifiés par les DA. 46
5.2 Résultats de la projection UMAP appliquée aux caractéristiques radiomiques originales et augmentées de l'ensemble de données HECKTOR22. L'image de gauche présente les projections 2D des vecteurs radiomiques pour tous les cas originaux. Celle de droite présente les projections pour les cas augmentés. Le code couleur des points avec la RFS vise principalement à améliorer la lisibilité. 48
5.3 Les valeurs de CI_{train} et CI_{test} sont présentées, avec le nombre de variablesélectionnées en abscisse. Les lignes en pointillées et complètes représentent les performances des radiomiques orioginales et augmentées respectivement 49

6.1 Le protocole RTEP7 IFCT-1402 : tous les patients inclus ont un scan TEP/TDM avant d'être assignés à une des deux branches : expérimentale ou contrôle, puis d'avoir un deuxième scan TEP/TDM, après 42 Gray de traitement. Dans le bras de contrôle, le traitement classique est suivi (66 Gray), dans le bras expérimental le traitement est adapté en fonction de ce deuxième examen (66 ou 74 Gray, en fonction de la réponse). 56
6.2 Vue d'ensemble de la pipeline d'extraction des caractéristiques. Les contours d'organes ont été extraits des scans TDM avec TS. Nous obtenons deux ensembles de caractéristiques, un pour chaque point temporel, et un troisième, Δ F, la différence entre les caractéristiques à ces deux points. 56
6.3 Toutes les combinaisons de données d'entrée sont utilisées et comparées. Une coche- signifie que le type de donnée a été utilisé : rouge ou vert, selon si il a été rejeté ou inclus par la FS 57

1 Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for all of the benchmarked models. Dashed lines represent the performance using basic radiomics and solid lines represent the performance using augmented radiomics. CI_{train} values are displayed with more transparency. 68

Liste des tableaux

5.1 C-Index for each model, using basic and augmented radiomics. In the last column the percent change in performance is added. 48

Appendix 1

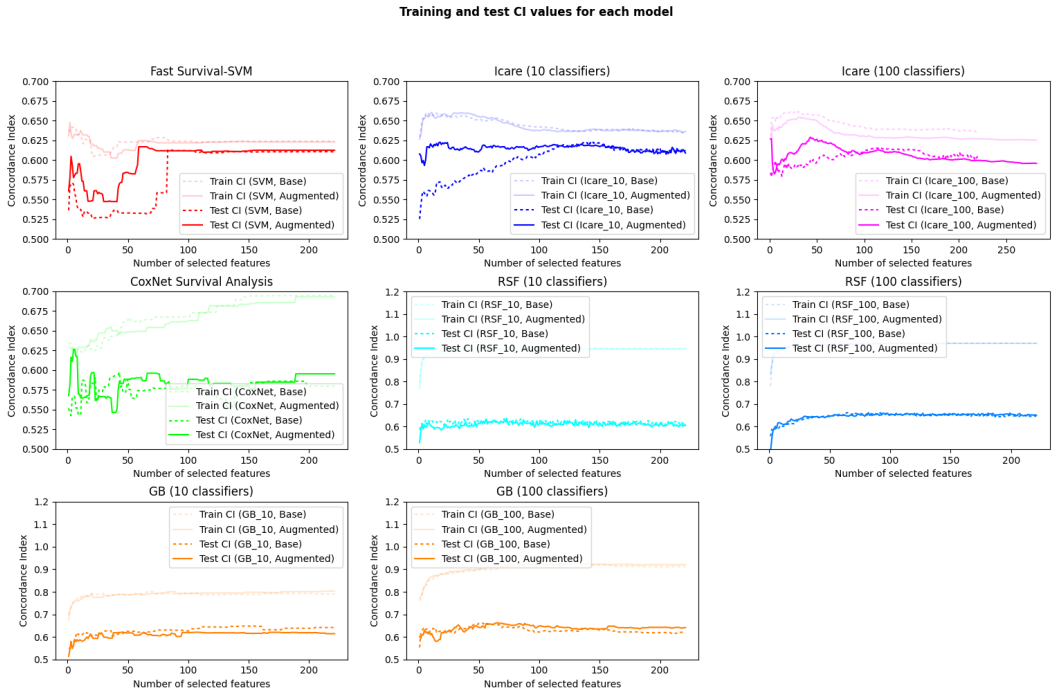

Figure 1 – Values of CI_{train} and CI_{test} with respect to the number of selected input variables, for all of the benchmarked models. Dashed lines represent the performance using basic radiomics and solid lines represent the performance using augmented radiomics. CI_{train} values are displayed with more transparency.

Bibliographie

[1] Mariana Fernandez. High-end Global Computed Tomography Purchases to Propel the High-end CT Segment Revenue, September 2021.

[2] The National Lung Screening Trial : Overview and Study DesignRadiology. https ://pubs.rsna.org/doi/abs/10.1148/radiol.10091808.

[3] Jens Vogel-Claussen, Florian Lasch, Benjamin-Alexander Bollmann, Katharina May, Alexander Kuhlmann, Gerald Schmid-Bindert, Rudolf Kaaks, Jörg Barkhausen, Sabine Bohnet, and Martin Reck. Design and Rationale of the HANSE Study : A Holistic German Lung Cancer Screening Trial Using Low-Dose Computed Tomography. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 194(12) :1333–1345, December 2022. ISSN 1438-9029, 1438-9010. doi : 10.1055/a-1853-8291.

[4] Alain Bernard and Hani Shennib. Resection of pulmonary nodules using video-assisted thoracic surgery. *The Annals of Thoracic Surgery*, 61(1) :202–205, January 1996. ISSN 0003-4975. doi : 10.1016/0003-4975(95)01014-9.

[5] Cancer Tomorrow. https ://gco.iarc.who.int/today/.

[6] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator : Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology : Artificial Intelligence*, 5(5) :e230024, September 2023. doi : 10.1148/ryai.230024.

[7] Yazdan Salimi, Ghasem Hajianfar, Zahra Mansouri, Amirhosein Sanaat, Mehdi Amini, Isaac Shiri, and Habib Zaidi. Organomics : A Concept Reflecting the Importance of PET/CT Healthy Organ Radiomics in Non–Small Cell Lung Cancer Prognosis Prediction Using Machine Learning. *Clinical Nuclear Medicine*, 49(10) :899, October 2024. ISSN 0363-9762. doi : 10.1097/RLU.0000000000005400.

[8] Gašper Podobnik, Bulat Ibragimov, Elias Tappeiner, Chanwoong Lee, Jin Sung Kim, Zacharia Mesbah, Romain Modzelewski, Yihao Ma, Fan Yang, Mikołaj Rudecki, Marek Wodziński, Primož Peterlin, Primož Strojan, and Tomaž Vrtovec. HaN-Seg : The head and neck organ-at-risk CT and MR segmentation challenge. *Radiotherapy and Oncology*, 198 :110410, September 2024. ISSN 0167-8140. doi : 10.1016/j.radonc.2024.110410.

[9] Jakob Dexl, Sergios Gatidis, Marcel Früh, Katharina Jeblick, Andreas Mittermeier, Anna Theresa Stüber, Balthasar Schachtner, Johanna Topalis, Matthias P. Fabritius, Sijing Gu, Gowtham Krishnan Murugesan, Jeff VanOss, Jin Ye, Junjun He, Anissa Alloula, Bartłomiej W. Papież, Zacharia Mesbah, Romain Modzelewski, Matthias Hadlich, Zdravko Marinov, Rainer Stiefelhagen, Fabian Isensee, Klaus H. Maier-Hein, Adrian Galdran, Konstantin Nikolaou, Christian la Fougère, Moon Kim, Nico Kallenberg, Jens Kleesiek, Ken Herrmann, Rudolf Werner, Michael Ingrisch, Clemens C. Cyran, and Thomas Küstner. AutoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging, Part 2 : Domain Generalization. *Journal of Nuclear Medicine*, December 2025. ISSN 0161-5505, 2159-662X. doi : 10.2967/jnumed.125.270260.

[10] Zacharia Mesbah, Léo Mottay, Romain Modzelewski, Pierre Decazes, Sébastien Hapdey, Su Ruan, and Sébastien Thureau. AutoPETIII : The Tracer Frontier. What Frontier ?, September 2025.

[11] Zacharia Mesbah and Solène Perret. QuantIF : Semi-automatic PET/CT segmentation across tracers.

[12] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958. ISSN 1939-1471. doi : 10.1037/h0042519.

[13] Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2) :146–160, June 1976. ISSN 1572-9125. doi : 10.1007/BF01931367.

[14] Kunihiko Fukushima. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4) :193–202, April 1980. ISSN 1432-0770. doi : 10.1007/BF00344251.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, November 1998. ISSN 1558-2256. doi : 10.1109/5.726791.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi : 10.1007/978-3-319-24574-4_28.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

[21] Decomposing Language Models Into Understandable Components. https ://www.anthropic.com/news/decomposing-language-models-into-understandable-components.

[22] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don't Always Say What They Think.

[23] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, December 2024.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[25] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic Video Generation with Diffusion Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 393–411, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72986-7. doi : 10.1007/978-3-031-72986-7_23.

[26] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen : Textually Guided Audio Generation, March 2023.

[27] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The Curse of Recursion : Training on Generated Data Makes Models Forget, April 2024.

[28] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. 3D MRI Synthesis with Slice-Based Latent Diffusion Models : Improving Tumor Segmentation Tasks in Data-Scarce Regimes. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. doi : 10.1109/ISBI56570.2024.10635533.

[29] Adyasha Sahu, Pradeep Kumar Das, and Sukadev Meher. An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. *Biomedical Signal Processing and Control*, 87 :105377, January 2024. ISSN 1746-8094. doi : 10.1016/j.bspc.2023.105377.

[30] Devinder Kumar, Alexander Wong, and David A. Clausi. Lung Nodule Classification Using Deep Features in CT Images. In *2015 12th Conference on Computer and Robot Vision*, pages 133–138, June 2015. doi : 10.1109/CRV.2015.25.

[31] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving : Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3) :1341–1360, March 2021. ISSN 1558-0016. doi : 10.1109/TITS.2020.2972974.

[32] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of Significance : Model selection and overfitting. *Nature Methods*, 13(9) :703–705, September 2016. ISSN 15487091.

[33] Amine Amyar, Romain Modzelewski, Hua Li, and Su Ruan. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia : Classification and segmentation. *Computers in Biology and Medicine*, 126 :104037, November 2020. ISSN 0010-4825. doi : 10.1016/j.compbiomed.2020.104037.

[34] Dong-Hyun Lee. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-1423.

[36] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked Image Modeling : A Survey. *International Journal of Computer Vision*, 133(10) : 7154–7200, October 2025. ISSN 1573-1405. doi : 10.1007/s11263-025-02524-1.

[37] Yifei Xin, Xiulian Peng, and Yan Lu. Masked Audio Modeling with CLAP and Multi-Objective Learning, January 2024.

[38] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE : Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems*, 35 :10078–10093, December 2022.

[39] History. https ://www.dicomstandard.org/history.

[40] SimpleITK/SimpleITK : SimpleITK : A layer built on top of the Insight Toolkit (ITK), intended to simplify and facilitate ITK's use in rapid prototyping, education and interpreted languages. https ://github.com/SimpleITK/SimpleITK.

[41] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane,

Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825) :357–362, September 2020. ISSN 1476-4687. doi : 10.1038/s41586-020-2649-2.

[42] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix : A User-Friendly, Multi-Lingual Library for Medical Image Registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 134–142, 2016.

[43] Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimpleITK Image-Analysis Notebooks : A Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3) :290–303, June 2018. ISSN 1618-727X. doi : 10.1007/s10278-017-0037-8.

[44] Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11(1) :9068, April 2021. ISSN 2045-2322. doi : 10.1038/s41598-021-87564-6.

[45] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO : A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208 :106236, September 2021. ISSN 0169-2607. doi : 10.1016/j.cmpb.2021.106236.

[46] Fabian Pedregosa, Fabian Pedregosa, Gael Varoquaux, Gael Varoquaux, Normalesup Org, Alexandre Gramfort, Alexandre Gramfort, Vincent Michel, Vincent Michel, Logilab Fr, Bertrand Thirion, Bertrand Thirion, Olivier Grisel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, Alexandre Tp, and David Cournapeau. Scikit-learn : Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

[47] Sebastian Pölsterl. Scikit-survival : A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212) :1–6, 2020. ISSN 1533-7928.

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch : An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[49] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet : A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi : 10.1109/CVPR.2009.5206848.

[50] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D : Transfer Learning for 3D Medical Image Analysis, July 2019.

[51] Connie L. Parks and Keith L. Monson. Automated Facial Recognition of Computed Tomography-Derived Facial Images : Patient Privacy Implications. *Journal of Digital Imaging*, 30(2) :204–214, April 2017. ISSN 1618-727X. doi : 10.1007/s10278-016-9932-7.

[52] M. Lindholz, R. Ruppel, S. Schulze-Weddige, G. L. Baumgärtner, I. Schobert, A. Panten, R. Schmidt, T. A. Auer, J. Nawabi, A. M. Haack, L. Stepansky, L. Poggi, R. Hosch, C. A. Hamm, and T. Penzkofer. Analyzing the TotalSegmentator for facial feature removal in head CT scans. *Radiography*, 31(1) :372–378, January 2025. ISSN 1078-8174. doi : 10.1016/j.radi.2024.12.018.

[53] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing Multimodal Medical Capabilities of Gemini, May 2024.

[54] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan L. Yuille, and Zongwei Zhou. AbdomenAtlas-8K : Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks. *Advances in Neural Information Processing Systems*, 36 :36620–36636, December 2023.

[55] Wenxuan Li, Pedro R. A. S. Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, Xiaoxi Chen, Lizhou Wu, Qilong Wu, Yannick Kirchhoff, Maximilian Rokuss, Saikat Roy, Yuxuan Zhao, Dexin Yu, Kai Ding, Constantin Ulrich, Klaus Maier-Hein, Yang Yang, Alan L. Yuille, and Zongwei Zhou. ScaleMAI : Accelerating the Development of Trusted Datasets and AI Models, January 2025.

[56] Antonella Santone, Francesco Mercaldo, Luca Brunese, Antonella Santone, Francesco Mercaldo, and Luca Brunese. A Method for Real-Time Lung Nodule Instance Segmentation Using Deep Learning. *Life*, 14(9), September 2024. ISSN 2075-1729. doi : 10.3390/life14091192.

[57] Nobuyuki Otsu. A Tlreshold Selection Method from Gray-Level Histograms.

[58] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem : If resnets are the answer, then what is the question ?

[59] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale, June 2021.

[60] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR : Transformers for 3D Medical Image

Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[61] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR : Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08999-2. doi : 10.1007/978-3-031-08999-2_22.

[62] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer : Interleaved Transformer for Volumetric Segmentation, February 2022.

[63] Albert Gu and Tri Dao. Mamba : Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling*, August 2024.

[64] Jun Ma, Feifei Li, and Bo Wang. U-Mamba : Enhancing Long-range Dependency for Biomedical Image Segmentation, January 2024.

[65] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26 (3) :297–302, 1945. ISSN 1939-9170. doi : 10.2307/1932409.

[66] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R. Ledsam, and Olaf Ronneberger. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy : Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research*, 23(7) :e26151, July 2021. doi : 10.2196/26151.

[67] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16555–16564, June 2021. doi : 10.1109/CVPR46437.2021.01629.

[68] Woorim Choi, Chul-Ho Kim, Hyein Yoo, Hee Rim Yun, Da-Wit Kim, and Ji Wan Kim. Development and validation of a reliable method for automated measurements of psoas muscle volume in CT scans using deep learning-based segmentation : A cross-sectional study. *BMJ Open*, 14(5) :e079417, May 2024. ISSN 2044-6055, 2044-6055. doi : 10.1136/bmjopen-2023-079417.

[69] Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3) :1391–1399, March 2019. ISSN 1432-1084. doi : 10.1007/s00330-018-5695-5.

[70] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net : A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2) :203–211, February 2021. ISSN 1548-7105. doi : 10.1038/s41592-020-01008-z.

[71] Gozétotide : substance active à effet thérapeutique. https ://www.vidal.fr/medicaments/substances/gozetotide-100484.html.

[72] Electron–positron annihilation. *Wikipedia*, January 2026.

[73] Artem Zatcepin and Sibylle I. Ziegler. Detectors in positron emission tomography. *Zeitschrift für Medizinische Physik*, 33(1) :4–12, February 2023. ISSN 0939-3889. doi : 10.1016/j.zemedi.2022.08.004.

[74] Nora Volkow, Nizar Mullani, and Bernard Bendriem. Positron emission tomography instrumentation : An overview. *American journal of physiologic imaging*, 3 :142–53, February 1988.

[75] David Little. Normal F-18 FDG PET-CT | Radiology Case | Radiopaedia.org. https ://radiopaedia.org/cases/normal-f-18-fdg-pet-ct.

[76] K. Agrawal, J. Weaver, R. Ngu, and H. Krishnamurthy Mohan. Clinical significance of patterns of incidental thyroid uptake at 18F-FDG PET/CT. *Clinical Radiology*, 70(5) :536–543, May 2015. ISSN 0009-9260. doi : 10.1016/j.crad.2014.12.020.

[77] Andrei Gafita, Jeremie Calais, Charlott Franz, Isabel Rauscher, Hui Wang, Andrew Roberstson, Johannes Czernin, Wolfgang A. Weber, and Matthias Eiber. Evaluation of SUV normalized by lean body mass (SUL) in 68Ga-PSMA11 PET/CT : A bi-centric analysis. *EJNMMI Research*, 9 :103, December 2019. ISSN 2191-219X. doi : 10.1186/s13550-019-0572-z.

[78] Léo Mottay, Pierre Decazes, Sébastien Hapdey, and Su Ruan. In Silico Dynamic Dual-Tracer PET Image Separation for Prostate Cancer Diagnosis using Deep Learning. In *2024 IEEE 17th International Conference on Signal Processing (ICSP)*, pages 678–681, October 2024. doi : 10.1109/ICSP62129.2024.10846781.

[79] Fereshteh Yousefirizi, Ivan S. Klyuzhin, Joo Hyun O, Sara Harsini, Xin Tie, Isaac Shiri, Muheon Shin, Changhee Lee, Steve Y. Cho, Tyler J. Bradshaw, Habib Zaidi, François Bénard, Laurie H. Sehn, Kerry J. Savage, Christian Steidl, Carlos F. Uribe, and Arman Rahmim. TMTV-Net : Fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images — a multi-center generalizability analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(7) :1937–1954, June 2024. ISSN 1619-7089. doi : 10.1007/s00259-024-06616-x.

[80] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, Jonathan Deissler, Ralf Floca, and Klaus Maier-Hein. nnInteractive : Redefining 3D Promptable Segmentation, March 2025.

[81] Sergios Gatidis, Marcel Früh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Rainer Stiefelhagen, Jan Egger, Jens Kleesiek, Ludovic Sibille, Lei Xiang, Simone Bendazolli, Mehdi Astaraki, Bernhard Schölkopf, Michael Ingrisch, Clemens Cyran, and Thomas Küstner. The autoPET challenge : Towards fully automated lesion segmentation in oncologic PET/CT imaging, June 2023. ISSN 2693-5015.

[82] Ercan Kurtipek, Mustafa Çayc, Nuri Düzgün, Hdr Esme, Yüksel Terzi, Süleyman Bakdk, Murat Serhat Aygün, Yasar Unlü, Cengiz Burnik, and Taha Tahir Bekci. 18F-FDG PET/CT Mean SUV and Metabolic Tumor Volume for Mean Survival Time in Non–Small Cell Lung Cancer. *Clinical Nuclear Medicine*, 40(6) :459, June 2015. ISSN 0363-9762. doi : 10.1097/RLU.0000000000000740.

[83] Pierre Vera, Sandrine Mezzani-Saillard, Agathe Edet-Sanson, Jean-François Ménard, Romain Modzelewski, Sebastien Thureau, Marc-Etienne Meyer, Khadija Jalali, Stéphane Bardet, Delphine Lerouge, Claire Houzard, Françoise Mornex, Pierre Olivier, Guillaume Faure, Caroline Rousseau, Marc-André Mahé, Philippe Gomez, Isabelle Brenot-Rossi, Naji Salem, and Bernard Dubray. FDG PET during radiochemotherapy is predictive of outcome at 1 year in non-small-cell lung cancer patients : A prospective multicentre study (RTEP2). *European Journal of Nuclear Medicine and Molecular Imaging*, 41(6) :1057–1065, June 2014. ISSN 1619-7089. doi : 10.1007/s00259-014-2687-9.

[84] Anne-Ségolène Cottereau, Hélène Lanic, Sylvain Mareschal, Michel Meignan, Pierre Vera, Hervé Tilly, Fabrice Jardin, and Stéphanie Becker. Molecular Profile and FDG-PET/CT Total Metabolic Tumor Volume Improve Risk Classification at Diagnosis for Patients with Diffuse Large B-Cell Lymphoma. *Clinical Cancer Research*, 22(15) :3801–3809, July 2016. ISSN 1078-0432. doi : 10.1158/1078-0432.CCR-15-2825.

[85] Frederic H. Fahey, Paul E. Kinahan, Robert K. Doot, Mehmet Kocak, Harold Thurston, and Tina Young Poussaint. Variability in PET quantitation within a multicenter consortium. *Medical Physics*, 37(7Part1) :3660–3666, 2010. ISSN 2473-4209. doi : 10.1118/1.3455705.

[86] Paulina E. Galavis, Christian Hollensen, Ngoneh Jallow, Bhudatt Paliwal, and Robert Jeraj. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncologica*, 49(7) :1012–1016, October 2010. ISSN 0284-186X. doi : 10.3109/0284186X.2010.498437.

[87] Nancy Paquet, Adelin Albert, Jacqueline Foidart, and Roland Hustinx. Within-Patient Variability of 18F-FDG : Standardized Uptake Values in Normal Tissues. *Journal of Nuclear Medicine*, 45(5) :784–788, May 2004. ISSN 0161-5505, 2159-662X.

[88] F. Yang, G. Simpson, L. Young, J. Ford, N. Dogan, and L. Wang. Impact of contouring variability on oncological PET radiomics features in the lung. *Scientific Reports*, 10(1) :369, January 2020. ISSN 2045-2322. doi : 10.1038/s41598-019-57171-7.

[89] Angela van Baardwijk, Geert Bosmans, Liesbeth Boersma, Jeroen Buijsen, Stofferinus Wanders, Monique Hochstenbag, Robert-Jan van Suylen, André Dekker, Cary Dehing-Oberije, Ruud Houben, Søren M. Bentzen, Marinus van Kroonenburgh, Philippe Lambin, and Dirk De Ruysscher. PET-CT–Based Auto-Contouring in Non–Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes. *International Journal of Radiation Oncology*Biology*Physics*, 68(3) :771–778, July 2007. ISSN 0360-3016. doi : 10.1016/j.ijrobp.2006.12.067.

[90] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G. P. M. van Stiphout, Patrick Granton, Catharina M. L. Zegers, Robert Gillies, Ronald Boellard,

André Dekker, and Hugo J. W. L. Aerts. Radiomics : Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4) :441–446, March 2012. ISSN 0959-8049. doi : 10.1016/j.ejca.2011.11.036.

[91] Pierre Decazes, Vincent Camus, Elodie Bohers, Pierre-Julien Viailly, Hervé Tilly, Philippe Ruminy, Mathieu Viennot, Sébastien Hapdey, Isabelle Gardin, Stéphanie Becker, Pierre Vera, and Fabrice Jardin. Correlations between baseline 18F-FDG PET tumour parameters and circulating DNA in diffuse large B cell lymphoma and Hodgkin lymphoma. *EJNMMI Research*, 10(1) :120, October 2020. ISSN 2191-219X. doi : 10.1186/s13550-020-00717-y.

[92] Louis Rebaud, Thibault Escobar, Fahad Khalid, Kibrom Girum, and Irène Buvat. Simplicity Is All You Need : Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT. In Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation and Outcome Prediction*, pages 121–134, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-27420-6. doi : 10.1007/978-3-031-27420-6_13.

[93] Fangyang Jiao, Min Wang, Xiaoming Sun, Zizhao Ju, Jiaying Lu, Luyao Wang, Jiehui Jiang, Chuantao Zuo, Fangyang Jiao, Min Wang, Xiaoming Sun, Zizhao Ju, Jiaying Lu, Luyao Wang, Jiehui Jiang, and Chuantao Zuo. Based on Tau PET Radiomics Analysis for the Classification of Alzheimer's Disease and Mild Cognitive Impairment. *Brain Sciences*, 13(2), February 2023. ISSN 2076-3425. doi : 10.3390/brainsci13020367.

[94] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain Tumor Segmentation and Radiomics Survival Prediction : Contribution to the BRATS 2017 Challenge. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 287–297, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75238-9. doi : 10.1007/978-3-319-75238-9_25.

[95] Giulia Polverari, Francesco Ceci, Valentina Bertaglia, Maria Lucia Reale, Osvaldo Rampado, Elena Gallio, Roberto Passera, Virginia Liberini, Paola Scapoli, Vincenzo Arena, Manuela Racca, Andrea Veltri, Silvia Novello, Désirée Deandreis, Giulia Polverari, Francesco Ceci, Valentina Bertaglia, Maria Lucia Reale, Osvaldo Rampado, Elena Gallio, Roberto Passera, Virginia Liberini, Paola Scapoli, Vincenzo Arena, Manuela Racca, Andrea Veltri, Silvia Novello, and Désirée Deandreis. 18F-FDG Pet Parameters and Radiomics Features Analysis in Advanced Nsclc Treated with Immunotherapy as Predictors of Therapy Response and Survival. *Cancers*, 12(5), May 2020. ISSN 2072-6694. doi : 10.3390/cancers12051163.

[96] Madhurima R. Chetan and Fergus V. Gleeson. Radiomics in predicting treatment response in non-small-cell lung cancer : Current status, challenges and future perspectives. *European Radiology*, 31(2) :1049–1058, February 2021. ISSN 1432-1084. doi : 10.1007/s00330-020-07141-9.

[97] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21) :e104–e107, October 2017. ISSN 0008-5472. doi : 10.1158/0008-5472.CAN-17-0339.

[98] AIM-Harvard/pyradiomics. Artificial Intelligence in Medicine (AIM) Program, December 2025.

[99] PyPI Download Stats. https ://pypistats.org/packages/pyradiomics.

[100] The Image Biomarker Standardization Initiative : Standardized Quantitative Radiomics for High-Throughput Image-based PhenotypingRadiology. https ://pubs.rsna.org/doi/abs/10.1148/radiol.2020191145.

[101] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. *Radiology*, 295(2) :328–338, May 2020. ISSN 0033-8419, 1527-1315. doi : 10.1148/radiol.2020191145.

[102] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Frédérique Frouin, and Irène Buvat. LIFEx : A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Research*, 78(16) : 4786–4789, August 2018. ISSN 0008-5472. doi : 10.1158/0008-5472.CAN-18-0125.

[103] Ivo Wolf, Marcus Vetter, Ingmar Wegner, Thomas Böttger, Marco Nolden, Max Schöbinger, Mark Hastenteufel, Tobias Kunert, and Hans-Peter Meinzer. The Medical Imaging Interaction Toolkit. *Medical Image Analysis*, 9(6) :594–604, December 2005. ISSN 1361-8415. doi : 10.1016/j.media.2005.04.005.

[104] Joseph O. Deasy, Angel I. Blanco, and Vanessa H. Clark. CERR : A computational environment for radiotherapy research. *Medical Physics*, 30(5) :979–985, 2003. ISSN 2473-4209. doi : 10.1118/1.1568978.

[105] Alex Zwanenburg and Steffen Löck. MIRP : A Python package for standardised radiomics. *Journal of Open Source Software*, 9(99) :6413, July 2024. ISSN 2475-9066. doi : 10.21105/joss.06413.

[106] Elisabeth Pfaehler, Alex Zwanenburg, Johan R. de Jong, and Ronald Boellaard. RaCaT : An open source and easy to use radiomics calculator tool. *PLOS ONE*, 14(2) :e0212223, February 2019. ISSN 1932-6203. doi : 10.1371/journal.pone.0212223.

[107] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to Radiomics. *Journal of Nuclear Medicine*, 61 (4) :488–495, April 2020. ISSN 0161-5505, 2159-662X. doi : 10.2967/jnumed.118.222893.

[108] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Mechthild Krause, Michael Baumann, Esther G. C. Troost, Steffen Löck, and Christian Richter. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1) :13206, October 2017. ISSN 2045-2322. doi : 10.1038/s41598-017-13448-3.

[109] C. Spearman. *The Proof and Measurement of Association Between Two Things.* Studies in Individual Differences : The Search for Intelligence. Appleton-Century-Crofts, East Norwalk, CT, US, 1961. doi : 10.1037/11491-005.

[110] Karl Pearson. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352) :240–242, December 1895. ISSN 0370-1662. doi : 10.1098/rspl.1895.0041.

[111] Michael Laimighofer, Jan Krumsiek, Florian Buettner, and Fabian J. Theis. Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. *Journal of Computational Biology*, 23(4) :279–290, April 2016. doi : 10.1089/cmb.2015.0192.

[112] Shahrbanoo Goli, Hossein Mahjub, Javad Faradmal, Hoda Mashayekhi, and Ali-Reza Soltanian. Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression. *Computational and Mathematical Methods in Medicine*, 2016(1) :2157984, 2016. ISSN 1748-6718. doi : 10.1155/2016/2157984.

[113] Ross Ihaka and Robert Gentleman. R : A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3) :299–314, September 1996. ISSN 1061-8600. doi : 10.1080/10618600.1996.10474713.

[114] John Platt. Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft, April 1998.

[115] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society : Series B (Methodological)*, 20(2) :215–232, July 1958. ISSN 0035-9246. doi : 10.1111/j.2517-6161.1958.tb00292.x.

[116] William A. Belson. Matching and Prediction on the Principle of Biological Classification. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 8(2) :65–75, 1959. ISSN 1467-9876. doi : 10.2307/2985543.

[117] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, August 1995. doi : 10.1109/ICDAR.1995.598994.

[118] Jin-zhou Feng, Yu Wang, Jin Peng, Ming-wei Sun, Jun Zeng, and Hua Jiang. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of Critical Care*, 54 :110–116, December 2019. ISSN 0883-9441. doi : 10.1016/j.jcrc.2019.08.010.

[119] Margarita Kirienko, Luca Cozzi, Lidija Antunovic, Lisa Lozza, Antonella Fogliata, Emanuele Voulaz, Alexia Rossi, Arturo Chiti, and Martina Sollini. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. *European Journal of Nuclear Medicine and Molecular Imaging*, 45(2) :207–217, February 2018. ISSN 1619-7089. doi : 10.1007/s00259-017-3837-7.

[120] Frank E. HARRELL Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable Prognostic Models : Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4) :361–387, 1996. ISSN 1097-0258. doi : 10.1002/(SICI)1097-0258(19960229)15:4⟨361::AID-SIM168⟩3.0.CO;2-4.

[121] Patrick J. Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1) :92–105, March 2005. ISSN 0006-341X. doi : 10.1111/j.0006-341X.2005.030814.x.

[122] Alfensi Faruk, Jan Palczewski, and Georgios Aivaliotis. A new discrimination measure for assessing predictive performance of non-linear survival models, April 2025.

[123] Begoña B. Sierra, Colin McLean, Peter S. Hall, and Catalina A. Vallejos. The C-index Multiverse, August 2025.

[124] Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Stop Chasing the C-index : This Is How We Should Evaluate Our Survival Models, June 2025.

[125] Wayback Machine. https ://web.archive.org/web/20171023012737/https ://docs.lib.noaa.gov/rescue/mwr/07 078-01-0001.pdf, October 2017.

[126] Alessandro Stefano. Challenges and limitations in applying radiomics to PET imaging : Possible opportunities and avenues for research. *Computers in Biology and Medicine*, 179 :108827, September 2024. ISSN 0010-4825. doi : 10.1016/j.compbiomed.2024.108827.

[127] Xue Ying. An Overview of Overfitting and its Solutions. *Journal of Physics : Conference Series*, 1168(2) :022022, February 2019. ISSN 1742-6596. doi : 10.1088/1742-6596/1168/2/022022.

[128] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A Comprehensive Survey on Data Augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 38(1) :47–66, January 2026. ISSN 1558-2191. doi : 10.1109/TKDE.2025.3622600.

[129] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging : A systematic literature review. *Computers in Biology and Medicine*, 152 :106391, January 2023. ISSN 0010-4825. doi : 10.1016/j.compbiomed.2022.106391.

[130] Lisa Jöckel, Michael Kläs, and Silverio Martínez-Fernández. Safe Traffic Sign Recognition through Data Augmentation for Autonomous Vehicles Software. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 540–541, July 2019. doi : 10.1109/QRS-C.2019.00114.

[131] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment : A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617, September 2019. doi : 10.21437/Interspeech.2019-2680.

[132] A. Teymurazyan, T. Riauka, H.-S. Jans, and D. Robinson. Properties of Noise in Positron Emission Tomography Images Reconstructed with Filtered-Backprojection and Row-Action Maximum Likelihood Algorithm. *Journal of Digital Imaging*, 26(3) :447–456, June 2013. ISSN 0897-1889. doi : 10.1007/s10278-012-9511-5.

[133] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16 :321–357, June 2002. ISSN 1076-9757. doi : 10.1613/jair.953.

[134] Rok Blagus and Lara Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14(1) :106, March 2013. ISSN 1471-2105. doi : 10.1186/1471-2105-14-106.

[135] Muhammad Asif Munir, Reehan Ali Shah, Urooj Waheed, Muhammad Aqeel Aslam, Zeeshan Rashid, Mohammed Aman, Muhammad I. Masud, Zeeshan Ahmad Arfeen, Muhammad Asif Munir, Reehan Ali Shah, Urooj Waheed, Muhammad Aqeel Aslam, Zeeshan Rashid, Mohammed Aman, Muhammad I. Masud, and Zeeshan Ahmad Arfeen. Sparse Regularized Autoencoders-Based Radiomics Data Augmentation for Improved EGFR Mutation Prediction in NSCLC. *Future Internet*, 17(11), October 2025. ISSN 1999-5903. doi : 10.3390/fi17110495.

[136] Ye Zhang, Zhixiang Wang, Zhen Zhang, Junzhuo Liu, Ying Feng, Leonard Wee, Andre Dekker, Qiaosong Chen, and Alberto Traverso. GAN-based one dimensional medical data augmentation. *Soft Computing - A Fusion of Foundations, Methodologies & Applications*, 27(15) :10481, August 2023. ISSN 1432-7643. doi : 10.1007/s00500-023-08345-z.

[137] Haiyang WANG and Luca Mainardi. CTGAN in Augmentation of Radiomics Features Classification from Narrow Band Imaging for Laryngeal Cancer. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5, June 2024. doi : 10.1109/MeMeA60663.2024.10596771.

[138] F. Lo Iacono, R. Maragna, G. Pontone, and V. D. A. Corino. A Novel Data Augmentation Method for Radiomics Analysis Using Image Perturbations. *Journal of Imaging Informatics in Medicine*, 37(5) :2401–2414, October 2024. ISSN 2948-2933. doi : 10.1007/s10278-024-01013-0.

[139] Marcus R. Makowski, Keno K. Bressem, Luise Franz, Avan Kader, Stefan M. Niehues, Sarah Keller, Daniel Rueckert, and Lisa C. Adams. De Novo Radiomics Approach Using Image Augmentation and Features From T1 Mapping to Predict Gleason Scores in Prostate Cancer. *Investigative Radiology*, 56(10) :661, October 2021. doi : 10.1097/RLI.0000000000000788.

[140] Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Cheze Le Rest, Pierre Decazes, Ricardo Correia, Dina El-Habashy, Hesham Elhalawani, Clifton D. Fuller, Mario Jreige, Yomna Khamis, Agustina La Greca, Abdallah Mohamed, Mohamed Naser, John O. Prior, Su Ruan, Stephanie Tanadini-Lang, Olena Tankyevych, Yazdan Salimi, Martin Vallières, Pierre Vera, Dimitris Visvikis, Kareem Wahid, Habib Zaidi, Mathieu Hatt, and Adrien Depeursinge. Overview of the HECKTOR Challenge at MICCAI 2022 : Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. In Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation and Outcome Prediction*, pages 1–30, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-27420-6. doi : 10.1007/978-3-031-27420-6_1.

[141] Gašper Podobnik, Bulat Ibragimov, Primož Peterlin, Primož Strojan, and Tomaž Vrtovec. vOARiability : Interobserver and intermodality variability analysis in OAR contouring from

head and neck CT and MR images. *Medical Physics*, 51(3) :2175–2186, 2024. ISSN 2473-4209. doi : 10.1002/mp.16924.

[142] SimpleITK/SimpleITK. SimpleITK, December 2025.

[143] Tugba Akinci D'Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Daniel Boll, Hanns-Christian Breit, Andrew Phillip Nicoli, Martin Segeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. TotalSegmentator MRI : Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI. *Radiology*, 314(2) :e241613, February 2025. ISSN 0033-8419, 1527-1315. doi : 10.1148/radiol.241613.

[144] Mehdi Amini, Ghasem Hajianfar, Yazdan Salimi, Zahra Mansouri, and Habib Zaidi. MetaPredictomics : A Comprehensive Approach to Predict Post-Surgical NSCLC Recurrence Using Clinicopathological, Radiomics, and Organomics Data. *Journal of Nuclear Medicine*, 66 (supplement 1) :251261–251261, June 2025. ISSN 0161-5505, 2159-662X.

[145] Zhack47. Zhack47/autopetii, December 2023.

[146] Zhack47. Zhack47/AutopetIII_CHB, December 2025.

[147] Zhack47. Zhack47/autopetiv, February 2026.

[148] Zhack47/radiomics-augmentation : Radiomics augmentation code repository. https ://github.com/Zhack47/radiomics-augmentation.

[149] Zacharia Mesbah. Organ-omics, November 2025.

[150] Harrison C. Gottlich, Adriana V. Gregory, Vidit Sharma, Abhinav Khanna, Amr U. Moustafa, Christine M. Lohse, Theodora A. Potretzke, Panagiotis Korfiatis, Aaron M. Potretzke, Aleksandar Denic, Andrew D. Rule, Naoki Takahashi, Bradley J. Erickson, Bradley C. Leibovich, and Timothy L. Kline. Effect of Dataset Size and Medical Image Modality on Convolutional Neural Network Model Performance for Automated Segmentation : A CT and MR Renal Tumor Imaging Study. *Journal of Digital Imaging*, 36(4) :1770–1781, August 2023. ISSN 1618-727X. doi : 10.1007/s10278-023-00804-1.

[151] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnU-Net Revisited : A Call for Rigorous Validation in 3D Medical Image Segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 488–498, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72114-4. doi : 10.1007/978-3-031-72114-4_47.

[152] Zongwei Zhou. MrGiovanni/RadGPT, February 2026.

[153] matchID - Moteur de recherche des personnes décédées. https ://deces.matchid.io/search.