Validating and optimizing the specificity of imaging biomarkers for personalized medicine

Oreiller Valentin

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Service de médecine nucléaire et imagerie moléculaire

Validating and optimizing the specificity of imaging biomarkers for personalized medicine

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Valentin Oreiller

Master en Science en bioingénieurie, EPFL - Ecole Polytechnique Fédérale de Lausanne, Suisse

Jury

Prof. Zoltan Kutalik, Président
Prof. John O. Prior, Directeur de thèse
Prof. Adrien Depeursinge, Co-directeur de thèse (Institute of Information Systems, University of Applied Sciences Western Switzerland)
Prof. Dimitri Van De Ville, Expert
Prof. Clarisse Dromain, Experte
Prof. Ender Konukoglu, Expert

Lausanne
(2022)

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Zoltan	**Kutalik**
Directeur·trice de thèse	Monsieur	Prof.	John	**Prior**
Co-directeur·trice	Monsieur	Prof.	Adrien	**Depeursinge**
Expert·e·s	Monsieur	Prof.	Dimitri	**Van De Ville**
	Madame	Prof.	Clarisse	**Dromain**
	Monsieur	Prof.	Ender	**Konukoglu**

le Conseil de Faculté autorise l'impression de la thèse de

Valentin Oreiller

Master en Science en bioingénieurie, EPFL - Ecole Polytechnique Fédérale de Lausanne, Suisse

intitulée

Validating and optimizing the specificity of imaging biomarkers for personalized medicine

Lausanne, le 4 novembre 2022

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Zoltan Kutalik

This thesis was conducted jointly at the Institute of Information Systems of the University of Applied Sciences and Arts of Western Switzerland (HES-SO) in Sierre and the Service of Nuclear Medicine and Molecular Imaging of the Lausanne University Hospital (CHUV). This work was carried out as part of the *The future of radiomics: understanding, Validating and Increasing the Specificity of Imaging Biomarkers for personaLizEd medicine (VISIBLE)* project, led by Prof. Adrien Depeursinge and project partner Prof. John O. Prior, and funded by the Swiss National Science Foundation (SNSF) under grant number 205320_179069.

Acknowledgements

Firstly, I am extremely grateful to my supervisors, Prof. Adrien Depeursinge and Prof. John O. Prior, for their invaluable guidance, support, and patience during my Ph.D. study. Their extensive knowledge and experience were a great source of inspiration to me in my academic research and daily life.

I also want to express my appreciation to Dr. Vincent Andrearczyk for his unwavering support and the experiences we shared. Dr. Mara Graziani's constructive feedback and advice were immensely helpful, and I am grateful for her contributions.

Mario Jreige's expertise in medicine and the knowledge he shared with me have been invaluable. I also acknowledge Dr. Julien Fageot for his mathematical advice and feedback, which have greatly enhanced the quality of my work.

Roger Schaer's technical support was greatly appreciated, and I want to thank Daniel Abler for his kindness and availability.

I am indebted to all the members of the nuclear medicine department of the CHUV for their outstanding expertise. Their contributions have been instrumental in shaping my research.

I would also like to thank all the members of the Medgift Group for their kind help, support, and expertise, which have made my study and life in Sierre a wonderful time despite the pandemic.

I express my gratitude to the members of my thesis committee, namely Zoltan Kutalik, Dimitri Van De Ville, Clarisse Dromain, and Ender Konukoglu, for generously dedicating their time and valuable expertise towards my research and offering insightful feedback.

Lastly, I want to express my heartfelt gratitude to my parents, sister, and partner. Their understanding and encouragement have been invaluable to me over the past few years, and I could not have completed my thesis without their support.

Monthey, November 5th 2022 V. O.

Abstract

In the last decade, biomedical image analysis has advanced significantly, thanks to radiomics and Convolutional Neural Networks (CNNs) being used for constructing predictive models in personalized medicine. These models are built using image features that can be divided into three main categories: intensity, shape, and texture. Although intensity and shape features are essential, texture features have the potential to reveal complex relationships between tumor architecture and patient outcomes. Therefore, the advancement and comprehension of texture features hold potential as effective strategies for clinicians to enhance disease characterization and facilitate personalized medicine. However, common texture features suffer from several limitations. In this thesis, we reviewed the most common texture features, explaining their advantages and disadvantages, and focusing on their robustness to rotations of the images and structures of interest. We proposed a novel method for designing directional image operators that are Locally Rotation Invariant (LRI), which we implemented based on the power spectrum and bispectrum of the circular harmonics expansion for 2D images or the spherical harmonics expansion for 3D images. We further integrated these LRI operators into a convolutional layer and used them in a CNN to obtain various LRI CNNs. We tested several shallow 3D LRI CNNs to classify benign or malignant lung nodules and demonstrated the advantages of bispectral LRI CNNs in terms of accuracy and data efficiency. Additionally, we evaluated our bispectral LRI layer in a 2D U-Net to segment nuclei in histopathological images and obtained comparable performance between the LRI U-Net and a standard U-Net. Furthermore, we showed that the LRI U-Net was more resilient to input rotations than the standard U-Net.

The development of machine learning in biomedical imaging requires large datasets and benchmarks to create robust predictive models. Therefore, the second contribution of this thesis was to participate in organizing the HEad and neCK tumOR segmentation and outcome prediction in PET/CT images (HECKTOR) challenge. The challenge aimed to benchmark automatic head and neck tumor segmentation methods and prognosis radiomics models. We obtained satisfying participation and scientific outcomes during the previous editions, and the third edition of the challenge is currently ongoing. A notable finding was that fully automatic prognosis methods could be effectively tested on large datasets without requiring human-made segmentation. This could open the door to more comprehensive benchmarking efforts in the field.

Key words: machine learning, personalized medicine, radiomics, texture, medical image analysis, image processing, CNN, rotation invariance, head and neck cancer, image segmentation.

Résumé

Au cours de la dernière décennie, l'analyse d'images biomédicales a considérablement progressé, grâce à l'utilisation de la radiomique et des réseaux de neurones convolutionnels (CNN) pour construire des modèles prédictifs en médecine personnalisée. Ces modèles sont construits à partir de caractéristiques d'images qui peuvent être divisées en trois catégories principales : l'intensité, la forme et la texture. Bien que les caractéristiques d'intensité et de forme soient essentielles, les caractéristiques de texture ont le potentiel de révéler des relations complexes entre l'architecture tumorale et l'évolution clinique des patients. Par conséquent, le développement et la compréhension des caractéristiques de texture sont des approches prometteuses pour aider les cliniciens à mieux caractériser les maladies et à permettre une médecine plus personnalisée. Cependant, les caractéristiques de texture communément utilisées présentent plusieurs limitations. Dans cette thèse, nous avons examiné les caractéristiques de texture les plus courantes, en expliquant leurs avantages et leurs inconvénients, et en mettant l'accent sur leur robustesse aux rotations des images et des structures d'intérêt. Nous avons proposé une nouvelle méthode pour concevoir des opérateurs d'image directionnels qui sont localement invariants à la rotation (LRI), que nous avons implémentée en nous basant sur le spectre de puissance et le bispectre de l'expansion harmonique circulaire pour les images 2D ou l'expansion harmonique sphérique pour les images 3D. Nous avons ensuite intégré ces opérateurs LRI dans une couche de convolution et les avons utilisés dans un CNN pour obtenir différents CNN LRI. Nous avons testé plusieurs CNN LRI peu profonds en 3D pour classifier les nodules pulmonaires bénins ou malins et avons démontré les avantages des CNN LRI bispectraux en termes d'exactitude et d'efficacité des données. De plus, nous avons évalué notre couche LRI bispectrale dans un U-Net 2D pour segmenter des noyaux dans des images histopathologiques et avons obtenu des performances comparables entre l'U-Net LRI et un U-Net standard. De plus, nous avons montré que l'U-Net LRI était plus résilient aux rotations de l'image que l'U-Net standard.

Le développement de l'apprentissage automatique en imagerie biomédicale nécessite des ensembles de données volumineux et des benchmarks pour créer des modèles prédictifs robustes. Par conséquent, la deuxième contribution de cette thèse a été de participer à l'organisation du challenge HEad and neCK tumOR segmentation and outcome prediction in PET/CT images (HECKTOR). Le challenge visait à évaluer les méthodes automatiques de segmentation de tumeurs de la tête et du cou et les modèles de radiomique de pronostic. Nous avons obtenu une participation et des résultats scientifiques satisfaisants lors des éditions précédentes, et la troisième édition du challenge est actuellement en cours. Une découverte notable était que les méthodes de pronostic entièrement automatiques pourraient être testées de manière efficace sur de grands ensembles de données sans nécessiter de segmentation humaine. Cela pourrait ouvrir la voie à des efforts d'évaluation plus complets dans le domaine.

Mots clefs : apprentissage automatique, médecine personnalisée, radiomique, texture, analyse d'images médicales, traitement d'images, CNN, invariance de rotation, cancer de la tête et du cou, segmentation d'images.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Biomedical Texture Analysis

1.1.1 Notations

We consider 2D or 3D gray-scale images as square-integrable function with bounded support $I \in L^2(\mathbb{R}^D)$ where $D = 2, 3$ depending on whether it is a 2D or 3D image. Multi-channel images are the concatenation of multiple gray-scale images. Actual images are stored as a collection of pixels, for 3D images, the term voxel is often used, but for simplicity we will use the word pixel for any D-dimensional image. The conversion from a continous image I to its pixel values is done by sampling, *i.e.* the image is evaluated on a discrete set of positions. The discrete image is written as $I[\boldsymbol{k}]$ with $\boldsymbol{k} = (k_1, \dots, k_D)$ the vector of pixel indices. The continuous image $I(\boldsymbol{x})$ defined on the continuous positions $\boldsymbol{x} = (x_1, \dots, x_D) \in \mathbb{R}^D$ is sampled as $I[k_1, \dots, k_D] = I(k_1 \cdot \Delta\xi_1 + c_1, \dots, k_D \cdot \Delta\xi_D + c_D)$ with $\Delta\xi_i$ the sampling spacing for the i^{th} dimension, and c_i the dimension-wise offset. Without loss of generality the offset c_i are chosen to obtain indices $k_i \in \{0, \dots, K_i - 1\}$. Since the image I is defined on a bounded domain, we can always find a rectangular domain containing the image domain written as $\boldsymbol{F} \in \mathbb{R}^D$ and $\boldsymbol{F} = F_1 \times \ldots \times F_D$ the Cartesian product of the intervals $F_i = [c_i, (K_i - 1) \cdot \Delta\xi_i + c_i]$.

The Regions Of Interest (ROIs) are bounded subset $\boldsymbol{V} \subset \mathbb{R}^D$ of the spatial domain of the images. The formalism developed in (Depeursinge, Fageot, & Al-Kadi, 2017a) is used to describe the extraction of radiomics features and more precisely of texture features. We refer to the extraction of the scalar measurements from an image I within a ROI $\boldsymbol{V}$ by the letter $\eta(I, \boldsymbol{V})$.

The Fourier transform of a continuous integrable image $I \in L^1(\mathbb{R}^D)$ is written as $\hat{I}(\boldsymbol{\omega}) = \mathcal{F}\{I\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^D} I(\boldsymbol{x}) e^{-\mathrm{j}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \mathrm{d}\boldsymbol{x}$. The extension of the Fourier transform to a square-integrable function $I \in L^2(\mathbb{R}^D)$ is done in the standard way. We also use the notation $I(\boldsymbol{x}) \xleftrightarrow{\mathcal{F}} \hat{I}(\boldsymbol{\omega})$ to indicate a change of from the spatial domain to the Fourier domain.

The sets of 2D and 3D rotation matrices are denoted by $SO(2)$ and $SO(3)$. For D-dimensional rotation matrices, the set is written as $SO(D)$. In 2D, we write a rotation matrix by an angle θ as R_θ.

1.1.2 Radiomics Features

Some authors (Gillies et al., 2016a) use a broad definition of radiomics and refer to it as the conversion of digital medical images into mineable high-dimensional data. In this work, we consider a more restrained definition, and we define radiomics as the process of extracting scalar measurements, referred to as *radiomics features*, from a D-dimensional image I and an associated ROI $\boldsymbol{V} \subset \mathbb{R}^D$. These radiomics features are then used in statistical modeling and Machine learning (ML) for diagnosis or prognosis purposes. A typical example would be the extraction of scalar measurement from contoured lung nodules in Computed Tomography (CT) scanners in order to diagnose malignancy.

Radiomics features can be classified in three main categories: intensity, texture and shape. Intensity features are derived from the intensity distribution of the image pixels without taking into account spatial relationships. On the contrary, texture features are designed to capture spatial relationships. The shape features are computed from the contours of the ROI $\boldsymbol{V}$ and are quite different in nature to the other ones, since the intensity values of the image are not directly used in their computation. However, the pixel values most often play an important role in their computation since they are used in the delineation processed. Shape features are often very informative in cancer characterization and are widely used for cancer staging. For instance, as part of the Response Evaluation Criteria in Solid Tumors (RECIST) (Eisenhauer et al., 2009) the maximum diameter of the ROI is one of the main information used to assess the evolution in tumor burden during treatment. Other shape features, such as the ROI volume, are widely used in clinical practice to assess the tumor burden in a patient and thus, this category of features must be considered in standard radiomics analyses since their prognosis value have been demonstrated many times(M.-K. Chen et al., 2004; Iliadis et al., 2009; Miller & Grigsby, 2002).

In this work, we will focus on texture features, and since a rigorous definition of texture can be complicated, we will define texture features as quantities that are not invariant to pixel shuffling. Meaning that if pixels of an image are shuffled, the corresponding extracted features will be different. We note that it is not the case for intensity features since they are derived from the intensity distribution, which is invariant to pixel shuffling, as illustrated in Figure 1.1. Texture features have the potential to carry important complementary information about the tumor architecture and micro-environment. Relying only on shape and intensity does not allow for the characterization of complex structures such as necrotic parts, highly heterogeneous tumor sub-compartment, and tumor micro-environment that may correlate with tumor aggressiveness or therapy outcome (Junttila & De Sauvage,

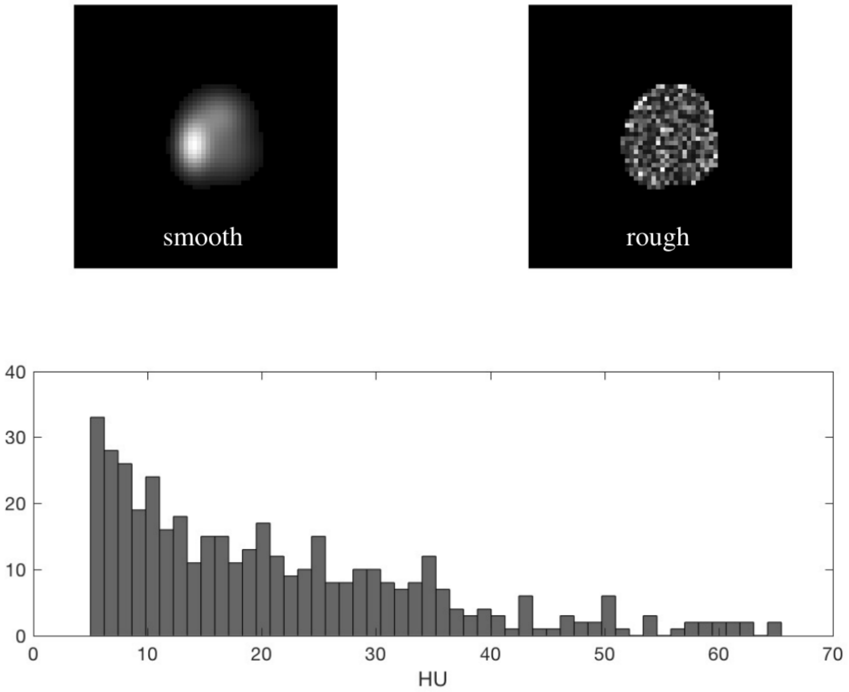


Figure 1.1: Example of two simulated tumors with the same distribution of pixel intensities in Hounsfield Unit (HU). The tumor on the right was generated by shuffling the pixels of the tumor on the left yielding very different looking images while maintaining the same intensity distribution. This picture illustrates what type of transformation intensity features are insensitive to and what texture features are focusing on. Reproduced from (Depeursinge, Fageot, & Al-Kadi, 2017b).

2013; Tang et al., 2018).

For these reasons, we think that texture features have a high potential in personalized medicine. However, texture features suffer from stability issues and are often less reproducible than intensity and shape features (Reiazi et al., 2021; Traverso et al., 2018). Different factors are responsible for this lack of reproducibility. First, the implementation of texture features involves many steps that can be interpreted in several ways leading to a discrepancy in the definitions of these features. Initiatives such as The Image Biomarker Standardisation Initiative (IBSI) (Zwanenburg et al., 2020) propose a framework to precisely define and test radiomics feature implementations.

Second, as texture features measure spatial differences, they are more sensitive to noise, and the lack of standard protocol for image acquisition dedicated to radiomics. It has been shown that radiomics features are very sensitive to imaging protocol (Mackin et al., 2015)

Third, one source of non-robustness is also due to feature design. The following sections review the most popular texture features implemented in standard radiomics toolboxes. We will highlight their strengths and weaknesses with a focus on robustness towards rotations of the input image.

1.1.3 Texture Features and Image Operators

Without loss of generality, any texture feature can be expressed as the result of the application of an image operator $\mathcal{G}$ and an aggregation function. The image operator captures local structures in the image at different positions, scales, and orientations. Image operators are not limited to linear operators, especially since we are interested in operators invariant to rotations. Rotation invariant operators are generally not linear, with the exception of isotropic filters, as discussed in Section 1.1.7.

A local image operator is defined as follow.

Definition 1 *A local image operator $\mathcal{G}$ is a mapping $\mathcal{G} : L^2(\mathbb{R}^D) \rightarrow L^2(\mathbb{R}^D)$ satisfying the following properties:*

- Locality*: there exists $r_0 > 0$ such that, for every $\boldsymbol{x} \in \mathbb{R}^D$ and every image $I(\boldsymbol{x}) \in L^2(\mathbb{R}^D)$, the quantity $\mathcal{G}\{I\}(\boldsymbol{x})$ only depends on local image values $I(\boldsymbol{y})$ for $\|\boldsymbol{y}-\boldsymbol{x}\| \leq r_0$.*

- Global equivariance to translations: *For any $I \in L^2(\mathbb{R}^D)$,*

$$\mathcal{G}\{I(\boldsymbol{x}-\boldsymbol{x}_0)\} = \mathcal{G}\{I\}(\boldsymbol{x}-\boldsymbol{x}_0), \quad \textit{for all} \quad \boldsymbol{x}, \boldsymbol{x}_0 \in \mathbb{R}^D.$$

The resulting image $g(\boldsymbol{x}) = \mathcal{G}\{I\}(\boldsymbol{x})$ is referred to as a *feature map*. The operator $\mathcal{G}$ can be seen as the application of a local descriptor $\mathcal{G}\{\cdot\}(\boldsymbol{x}) : L^2(\mathbb{R}^D) \rightarrow \mathbb{R}$ defined on a bounded support $\boldsymbol{G} + \boldsymbol{x} \subset \mathbb{R}^D$. This local operator extracts value from the neighborhood of the position $\boldsymbol{x}$ and maps them to a scalar value $g(\boldsymbol{x})$. Applying this local operator to all positions of the image domain leads to translation equivariance. This operation is equivalent to sliding the local operator as well as its support to the position $\boldsymbol{x}$ and computing the value of $g(\boldsymbol{x})$.

1.1.4 Aggregation Functions

In order to extract scalar measurements, *i.e.* a unique feature for each ROI, the values of the feature map g within the ROI $\boldsymbol{V}$ are summarized with an aggregation function. Aggregation functions can be diverse, but their goal is to summarize in one scalar the content of the feature map within the ROI. A particular family of aggregation is called *integrative* and regroup aggregations of the form:

$$\eta = \frac{1}{|\boldsymbol{V}|} \int_{\boldsymbol{V}} f(g(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \tag{1.1}$$

where $|\boldsymbol{V}| = \int_{\boldsymbol{V}} \mathrm{d}\boldsymbol{x}$ is the volume of the ROI $\boldsymbol{V}$, f is a real-valued function which would typically be of the form $f(\cdot) = |\cdot|^2$ to avoid computing vanishing integral.

Some other standard aggregation functions are the ones that extract statistics from the distribution of the pixel values of the feature maps g such as the minimum, maximum, or other quantile-based statistics. Of interesting note, the application of these types of aggregation functions directly to the image I instead of the feature map g results in most of the standard intensity features (such as SUV_{max}, SUV_{mean} for Positron Emission Tomography (PET) images).

1.1.5 Invariance and Equivariance to Geometric Transformations

Depending on the task, geometric transformations such as rotation or scaling do not change the semantic information of the image content. In other words, it is often required that such transformations of the input image do not change the decision of the prediction model. The class of transformations that should not impact the model's output can vary depending on the application. For instance, scale invariance would be required if a model is designed for self-driving cars since objects far from the camera will appear smaller in the image, but the semantic content would remain the same.

In the context of biomedical image analysis, the patient's position with respect to the imaging device is well controlled, and the physical pixel size is known. Hence, scale invariance is usually not required. On the contrary, the scale often includes semantic information about the imaged pathology and, in general, should not be discarded. A typical example is the tumor size, which holds crucial prognosis information. However, small rotations or translations of the patient or of small structures should not impact the final prediction of the model. Therefore, models that are invariant to these transformations are expected to be more robust.

A model that does not vary against a given transformation of its inputs is said to be *invariant* to this transformation. In the following, we will define the notion of transformation more precisely and see how we obtain radiomics features invariant to a given one. This invariance is induced by applying the aggregation function on an *equivariant* image operator. The difference between invariance and equivariance will be elaborated on in the following.

We consider geometric transformation operators $\mathcal{T}$ that can be written as $\mathcal{T}\{I\}(\boldsymbol{x}) = I(T(\boldsymbol{x}))$ with $T : \mathbb{R}^D \to \mathbb{R}^D$ a bijective mapping that associates a spatial coordinate with another one. This mapping can be seen as a deformation field. A well-known class of such transformation is the one where $T(\boldsymbol{x}) = \mathrm{A}\boldsymbol{x} + \boldsymbol{c}$ with A a $D \times D$ invertible matrix and $\boldsymbol{c} \in \mathbb{R}^D$ which forms the group of *affine* transformations. The subgroup of affine transformations with $\det(\mathrm{A}) = \pm 1$ is called the *rigid* transformation group and includes rotations, translations, and reflections. The subgroup of transformations with a positive determinant is called the *proper rigid* transformation group and does not include reflections.

Formally, an operator $\mathcal{G}$ is said to be *invariant* to the transformation $\mathcal{T} : L^2(\mathbb{R}^D) \to L^2(\mathbb{R}^D)$ if

$$\mathcal{G}\{\mathcal{T}\{I\}\} = \mathcal{G}\{I\}, \quad \text{for any} \quad I \in L^2(\mathbb{R}^D). \tag{1.2}$$

Imposing this condition on an image operator is very constraining and can even be inconsistent with the definition of the local image operator. Indeed, if we require that the operator is invariant to translations, it directly contradicts the second point of Definition 1 (equivariance to translation). Thus, in the context of image operator we relax the notion of invariance to the one of equivariance.

An operator $\mathcal{G}$ is *equivariant* to the transformation $\mathcal{T} : L^2(\mathbb{R}^D) \to L^2(\mathbb{R}^D)$ if

$$\mathcal{G}\{\mathcal{T}\{I\}\} = \mathcal{T}\{\mathcal{G}\{I\}\}, \quad \text{for any} \quad I \in L^2(\mathbb{R}^D). \tag{1.3}$$

In other words, the operator $\mathcal{G}$ is equivariant to the transformation $\mathcal{T}$ if the two operations commute.

We observe that a particular notion of invariance can be obtained for scalar measurements if we use an equivariant image operator and an appropriate aggregation function. It may not hold for any aggregation function, but it does for the integrative ones and the quantile-based statistics. We must also consider that the ROI $\boldsymbol{V}$ associated with an image I undergoes the opposite transformation applied to the argument of I. If we denote by $\boldsymbol{V}'$ the ROI associated with the transformed image $I'(\boldsymbol{x}) = I(T(\boldsymbol{x}))$ by composition of functions, we have $\boldsymbol{V}' = T^{-1}(\boldsymbol{V})$.

We will first discuss the case of integrative aggregation functions and assume the operator $\mathcal{G}$ to be equivariant to the affine transformation $T(\boldsymbol{x}) = \mathrm{A}\boldsymbol{x} + \boldsymbol{c}$. Hence, we have $\mathcal{G}\{I(\mathrm{A}\cdot +\boldsymbol{c})\}(\boldsymbol{x}) = \mathcal{G}\{I(\cdot)\}(\mathrm{A}\boldsymbol{x} + \boldsymbol{c})$. In addition, we observe that the determinant of the Jacobian of the inverse affine transformation is equal to $|\det(\mathrm{A}^{-1})|$. Applying the change of variable $\boldsymbol{x}' = \mathrm{A}\boldsymbol{x} + \boldsymbol{c}$ in Equation (1.1) leads to:

$$\begin{aligned}\frac{1}{|\boldsymbol{V}'|}\int_{\boldsymbol{V}'} f(\mathcal{G}\{I(\mathrm{A}\cdot +c\}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} &= \frac{1}{|\boldsymbol{V}'|}\int_{\boldsymbol{V}'} f(\mathcal{G}\{I\}(\mathrm{A}\boldsymbol{x}+c))\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{|\boldsymbol{V}||\det(A)^{-1}|}\int_{T(\boldsymbol{V}')} f(\mathcal{G}\{I\}(\boldsymbol{x}'))|\det(A^{-1})|\mathrm{d}\boldsymbol{x}'\\ &= \frac{1}{|\boldsymbol{V}|}\int_{\boldsymbol{V}} f(\mathcal{G}\{I\}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\end{aligned}$$

$$.$$

Since the ROIs $\boldsymbol{V}$ and $\boldsymbol{V}'$ are linked by $|\boldsymbol{V}'| = |\det(\mathrm{A}^{-1})||\boldsymbol{V}|$. We can conclude that the integrative aggregation function of an equivariant image operator yields an invariant scalar measurement for affine transformations.

This result coincides with intuition since the ROI is transformed accordingly to the transformation applied to the image, and we expect aggregated features over this ROI to remain the same. For instance, if an image is rotated as well as its associated ROI, the extracted features should remain the same. It might be less evident when the transformation affects the scale of the image. However, since the integrative aggregation is divided by the volume of the ROI, it discards this dependency.

The proof for the equivariance of the maximum operator as an aggregation function follows the same logic. We consider the same equivariant operator $\mathcal{G}$ and by substituting $\boldsymbol{x}' = \mathrm{A}\boldsymbol{x} + c$ we obtain:

$$\max_{\boldsymbol{x}\in\boldsymbol{V}'} \mathcal{G}\{I(\boldsymbol{A}\cdot +\boldsymbol{c})\}(\boldsymbol{x}) = \max_{\boldsymbol{x}\in T^{-1}(\boldsymbol{V})} \mathcal{G}\{I\}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{c}) = \max_{T^{-1}(\boldsymbol{x}')\in T^{-1}(\boldsymbol{V})} \mathcal{G}\{I\}(\boldsymbol{x}') = \max_{\boldsymbol{x}'\in\boldsymbol{V}} \mathcal{G}\{I\}(\boldsymbol{x}') \tag{1.4}$$

The same holds for the minimum aggregation function and other quantile-based aggregation functions.

In general, other aggregation functions derived from the statistics of the intensity values

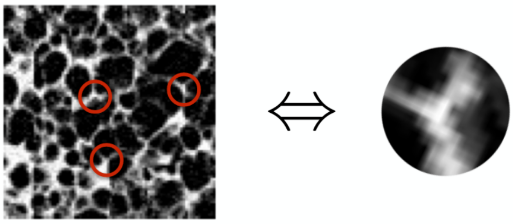


Figure 1.2: Illustration of collagen junctions in the lung, imaged by CT. We observe that approximately the same y-shaped junction appears at multiple positions and orientations.

do not yield invariant representation for operator equivariant to affine transforms. However, considering only rigid transformations, these aggregations produce invariant scalar measurements as rigid transformations conserve the distance between any pair of points.

In the next section, we will discuss the general strategy to obtain operators equivariant to any element of the rotation group $SO(D)$.

1.1.6 Local Rotation Invariance and Directional Sensitivity

This work mainly focused on designing image operators equivariant to image rotations. We referred to this type of operator as Locally Rotation Invariant (LRI) operators. The primary motivation for using and designing such operators is that biomedical textures are composed of local patterns appearing at random positions and orientations, as illustrated in Figure 1.2. The information in the local orientation of patterns is usually not informative for the task at hand. Therefore, being invariant with this information can be helpful in the design of robust texture operators.

Definition 2 *An image operator $\mathcal{G}$ is said to be LRI if, in addition to the two properties of Definition 1, it satisfies the following one:*

- Global equivariance to rotations: *For any $I \in L^2(\mathbb{R}^D)$,*

$$\mathcal{G}\{I(\mathrm{R}_0\cdot)\} = \mathcal{G}\{I\}(\mathrm{R}_0\cdot) \quad \textit{for any } \mathrm{R}_0 \in SO(D).$$

The global equivariance property implies that the local descriptor of the LRI operator is rotation invariant which can be highlighted if we consider an image I' rotated around the position $\boldsymbol{x}_0$ by a rotation $\mathrm{R} \in SO(D)$. We can write $I'(\boldsymbol{x}) = I(\mathrm{R}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{x}_0)$. The application of a LRI operator $\mathcal{G}$ to I' yields $\mathcal{G}\{I'\}(\boldsymbol{x}) = \mathcal{G}\{I(\mathrm{R}(\cdot - \boldsymbol{x}_0) + \boldsymbol{x}_0\}(\boldsymbol{x}) = \mathcal{G}\{I(\cdot)\}(\mathrm{R}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{x}_0)$ since LRI operators are both rotation and translation equivariant. Therefore, evaluating the application of the LRI operator $\mathcal{G}$ to the image I' at the position $\boldsymbol{x}_0$ trivially leads to $[\mathcal{G}\{I\}(\mathrm{R}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{x}_0)]_{\boldsymbol{x}=\boldsymbol{x}_0} = \mathcal{G}\{I\}(\boldsymbol{x}_0)$, indicating that the rotation around the position $\boldsymbol{x}_0$ has no effect on the value of the operator at that position. Therefore, we refer to this class of operators as LRI since local rotations, *i.e.* rotations around a specific position, do not impact the value of the operator at that position.

The design of such operators relies on finding local descriptors that are rotation invariant. In practice, two main strategies are used. The first strategy employs local descriptors that can be efficiently rotated and aligned to a specific orientation at every position. This alignment requires some criteria to define the orientation of the local descriptor. For instance, it can be the orientation that maximizes the response of the descriptor as it was implemented in (Andrearczyk & Depeursinge, 2018; Andrearczyk, Fageot, et al., 2019), or the direction of the local gradient as in the scale-invariant feature transform (Lowe, 2004) or as in the Riesz transform based features (Dicente Cid et al., 2017b). As explained in Section 1.1.7, local binary patterns (Ojala, Pietikainen, et al., 2002) also use this strategy to yield LRI image operators.

The second strategy is to define descriptors that are intrinsically rotation invariant. Circularly/spherically symmetric filters have that property, which comes at the price of being non-directional. These filters are non-directional since their local operator performs a local averaging in every direction. Directional sensitive invariant operators based on spherical harmonics were used in (Depeursinge et al., 2018), and how we can obtain them is explained in Section 1.1.9. The rotation invariant version of the Grey Level Co-occurence Matrices (GLCM) also falls into this category since rotation invariance is approximated by averaging the output of the local descriptor on different orientations (see Section 1.1.7).

1.1.7 Common Texture Analysis Approaches

In the following, we review the most commonly used texture features that are typically implemented in popular radiomics toolbox such as (Nioche et al., 2018; Van Griethuysen et al., 2017) and others.

We will cover the three main categories of texture features: gray level matrices, local binary patterns, and convolutional-based. We will not cover all texture feature families, such as the one based on fractal theory, since they are less prevalent in the radiomics community.

Gray Level Matrices (GLM)

GLMs form a large family of non-linear texture operators. Among the most popular ones are the Gray Level Co-occurrence Matrices (GLCM), the Gray Level Size Zone Matrix (GLSZM), and the Gray Level Run Length Matrix (GLRLM). These three types of GLM are described below. The core idea behind GLM is to characterize the texture by a probability distribution over the gray levels and pixel positions that constitute the image. In other words, the GLM-based methods count the occurrences of gray levels according to a given condition and organize them in a matrix denoted by P. How the occurrences are counted and organized in the matrix defines the specific family of GLM. Scalar measurements are computed from the GLM P and are quantities derived from the statistics framework, such as autocorrelation, joint energy, or joint entropy. These measures are used to characterize the randomness of P.

An essential step common to most GLMs is the quantization of the image. This step is required in order to avoid computing sparse matrices. This quantization relies on the binning of the image's gray level and thus reduces the image's dynamic range to a smaller number of bins. We denote the quantized version of an image $I[\boldsymbol{k}]$ on a number of bins Λ as $I_\Lambda[\boldsymbol{k}]$ and the typical number of bins ranges from $\Lambda = 8, 16, 32, 64, 128$. Image quantization can be implemented in many different ways, and the choice of implementation can dramatically impact the final feature value (Leijenaar et al., 2015). IBSI recommends the following implementation:

$$I_\Lambda[\boldsymbol{k}] = \begin{cases} \lfloor \Lambda \frac{(I[\boldsymbol{k}] - \min(I)}{\max(I) - \min(I)} \rfloor + 1 & \text{for} \quad I[\boldsymbol{k}] < \max(I)) \\ \Lambda & \text{for} \quad I[\boldsymbol{k}] = \max(I) \end{cases} \tag{1.5}$$

Quantization with a fixed number of bins has a regularization effect. For instance, image modalities such as Magnetic Resonance Imagery (MRI), where the dynamic range can significantly vary from image to image benefits from a fixed number of bins quantization as it ensures to compare images with similar contrast. In the case of images with well-calibrated dynamic range, such as CT, it is preferred to use a binning scheme that conserves the information of the pixel values. It is usually implemented by imposing a fixed width for the bins, which depends on the imaging modality[I].

While being one of the most popular families of texture features in radiomics, GLMs suffer from several limitations. First, the quantization step potentially discards a lot of information since biomedical pixels are usually stored in 16 bits which amounts to a total of 65535 gray levels. Hence, a reduction to 128 bins is a drastic loss of information. Moreover, GLMs have poor scale coverage since the support of their local operators is small. Finally, GLM-based methods lead to an anisotropic representation of the spatial domain since every displacement of one pixel to any of its neighbors is considered a distance 1. Thus, displacements to diagonal pixels are considered equal to a displacement to adjacent pixels.

[I] For more details please refer to https://pyradiomics.readthedocs.io/en/v3.0.1/radiomics.html?highlight=getbinedges#radiomics.imageoperations.getBinEdgesas of November 2022

Gray Level Co-occurence Matrices (GLCM) First proposed by (Haralick et al., 1973), the GLCM can be seen as the application of a collection of operators defined by:

$$\mathcal{G}_{\Delta \boldsymbol{k}}^{\lambda_1,\lambda_2}\{I_\Lambda\}[\boldsymbol{k}_0] = \begin{cases} 1 \text{ if } I[\boldsymbol{k}_0] = \lambda_1 \text{ and } I[\boldsymbol{k}_0 + \Delta \boldsymbol{k}] = \lambda_2, \\ 0 \text{ otherwise,} \end{cases} \tag{1.6}$$

where $\Delta \boldsymbol{k} \in \mathbb{Z}^D$ is a discrete vector representing the direction and the scale of the operator, $\lambda_1, \lambda_2 \in [1, \ldots, \Lambda]$ are the gray levels to which the operator is sensitive. A GLCM matrix is the spatial aggregation of the operator $\mathcal{G}_{\Delta \boldsymbol{k}}^{\lambda_1,\lambda_2}$ over the ROI and, thus, can be written as

$$\mathrm{P}(\lambda_1, \lambda_2 \mid \Delta \boldsymbol{k}) = \sum_{\boldsymbol{k}_0 \in \boldsymbol{V}} \mathcal{G}_{\Delta \boldsymbol{k}}^{\lambda_1,\lambda_2}\{I_\Lambda\}[\boldsymbol{k}_0] \tag{1.7}$$

with $\boldsymbol{V} \subset \mathbb{Z}^D$ the set of the pixels belonging to the ROI. This matrix is then normalized as to obtain a probability density function of gray-level co-occurrences $p(\lambda_1, \lambda_2 \mid \Delta \boldsymbol{k})$.

GLCM matrices are directional as they are computed along the direction of the vector $\Delta \boldsymbol{k}$. However, they are not LRI. The invariance to rotation is commonly approximated by averaging the matrices over different orientations. It can be understood as marginalizing out the direction of $\Delta \boldsymbol{k}$ to obtain $p(\lambda_1, \lambda_2 \mid ||\Delta \boldsymbol{k}|| = \delta)$. In practice, since the image is discretized and takes values on the discrete grid $\mathbb{Z}^D$, the infinity norm is used to approximate the distance $\delta = ||\Delta \boldsymbol{k}||$. To illustrate this, let us consider the 2D case for $\delta = 1$. In order to cover all the neighborhoods of a central pixel, one has to average the GLCM matrices resulting from the operators $\mathcal{G}_{\Delta \boldsymbol{k}}^{\lambda_1,\lambda_2}$ with $||\Delta \boldsymbol{k}||_\infty = 1$ which results in extracting the 8 neighbors of a central pixels. As a reminder, the infinity norm is the maximum value of the vector components, resulting in an anisotropic norm, unlike the Euclidean norm. This induces features that are not invariant to local rotations, as it was pointed out and experimentally demonstrated in (Depeursinge et al., 2018).

Another way to approximate the invariance to local rotation is by averaging the final computed features in all directions, which remains anisotropic.

Gray Level Size Zone Matrices (GLSZM) The GLSZM (Thibault et al., 2009) features are based on the matrix $\mathrm{P}(\lambda, n)$ with entries being the count of gray level zones. A *gray level zone* is defined as an area of connected pixels that share the same pixel intensity or gray level. Therefore, the matrix $\mathrm{P}(\lambda, n)$ outputs the number of times a gray level zone with n pixels and with gray level λ appear in the image. The GLSZM features cover all directions and are isotropic with respect to the infinity norm. However, they are not sensitive to direction as differently shaped zones with the same gray level λ containing the same number of pixels can potentially appear within the same image. In this case, the gray level zone will increment the same entry in $\mathrm{P}(\lambda, n)$ and thus will not be discriminated.

Gray Level Run Length Matrices (GLRLM) In 2D, the GLRLM (Galloway, 1975) features are based on the statistics of the matrix $\mathrm{P}(\lambda, \gamma \mid \theta)$. This matrix counts the gray-level λ appearing for a consecutive length γ along the axis θ. In 2D, the features are commonly computed for each direction $\theta \in [0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}]$ and then averaged. The *run-length* is also derived from the infinity norm, leading to an anisotropic image representation. Furthermore, the orientation coverage is limited since only multiple of $\frac{\pi}{4}$ are considered.

Local Binary Patterns (LBPs)

LBPs were first introduced by (Ojala, Pietikainen, et al., 2002). The basic idea behind this method is to attribute to each pixel a binary sequence which is then converted to a decimal number. The binary sequence is constructed as follows. First, a sequence of N equally spaced neighboring pixels are extracted on a disk of radius r around the position $\boldsymbol{x}_0$. The neighboring pixels can be expressed as $I(\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k)$ with $\Delta\boldsymbol{x}_k = (r\cos(\theta_k), r\sin(\theta_k))$ for $\theta_k = \frac{2\pi}{N}k$. The neighbors that do not fall on the spatial grid are interpolated. Second, a value of 0 is attributed to the neighbor at position $\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k$ if its pixel value is lesser than $I(\boldsymbol{x}_0)$. Conversely a value of 1 is attributed to the neighbor at position $\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k$ if its pixel value is greater than the $I(\boldsymbol{x}_0)$. The LBP operator for a radius r, a number of neighbors γ, and a 2D continuous image I can be expressed as

$$\mathcal{G}_{N,r}\{I\}(\boldsymbol{x}_0) = \sum_{k=0}^{N-1} H(I(\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k) - I(\boldsymbol{x}_0))2^k = \sum_{k=0}^{N-1} b_k 2^k, \tag{1.8}$$

with H the Heaviside step function, and the *binary sequence* $b_k = H(I(\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k) - I(\boldsymbol{x}_0))$. Note that the term 2^k ensures that the operator $\mathcal{G}_{N,r}$ is in decimal number.

The LBP framework is very interesting since it allows to implement the two strategies mentioned in Section 1.1.6 to obtain LRI operators. First, let us consider how aligning the local descriptor would be implemented for the LBP. It suffices to note that shifting the binary sequence b_k is equivalent to rotating the local descriptor since $b_{k-k_0} = H(I(\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k - \Delta\boldsymbol{x}_{k_0} - I(\boldsymbol{x}_0))$ with $\Delta\boldsymbol{x}_{k_0} = \frac{2\pi}{N}k_0$ which expresses a rotation on the extracted circle. Thus, computing every shifts of the binary sequence b_k at position x_0 and keeping the one that minimizes the decimal code $\sum_{k=0}^{N-1} H(I(\boldsymbol{x}_0 + \Delta\boldsymbol{x}_k - \Delta\boldsymbol{x}_{k_0}) - I(\boldsymbol{x}_0))2^k$ $b_{k-k_0}2^k$ leads to an invariant representation, which is equivalent to locally aligning the LBP to the direction which minimizes the decimal code.

Second, using an invariant representation can be done, similarly to what is explained in Section 1.1.8, by using the property of the Fourier transform and applying the fast Fourier transform to the binary sequence b_k (G. Zhao et al., 2011). The power spectrum, *i.e.* the modulus, of this Fourier representation yields an invariant operator since the power spectrum is shift-invariant.

LBP is a well explored framework (Pietikäinen & Zhao, 2015) and its extension to 3D is done similarly to what is explained in Section 1.1.8 and is based on spherical harmonics (Banerjee et al., 2012). The power spectrum of the spherical harmonics representation is used to obtain 3D LRI LBP as well as the kurtosis of this representation.

Convolution-based

The convolution framework constitutes the basis of a very rich family of texture features. In the following, we will take a few examples and explain the main concepts to obtain LRI operators derived from the convolution framework. The *convolution* for an image $I \in L^2(\mathbb{R}^D)$ and a filter $h \in L^\infty(\mathbb{R}^D)$ is defined as

$$(I * h)(\boldsymbol{x}_0) = \int_{\mathbb{R}^D} I(\boldsymbol{x}) h(\boldsymbol{x}_0 - \boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{1.9}$$

If the filter h is defined on a bounded support, *i.e.* there exists a $r_0 > 0$ such that $h(\boldsymbol{x}) = 0$ for any $||\boldsymbol{x}|| > r_0$, the operator formed by the convolution with the filter h is an image operator satisfying Definition 1. Furthermore, this operator is linear.

The convolution operation is, in general, not equivariant to rotation as the convolution with a rotated version of the image I yields:

$$\begin{aligned}(I(\mathrm{R}\cdot) * h)(\boldsymbol{x}_0) &= \int_{\mathbb{R}^D} I(\mathrm{R}\boldsymbol{x}) h(\boldsymbol{x}_0 - \boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\mathrm{R}^D} I(\boldsymbol{x}') h(\mathrm{R}^{-1}\mathrm{R}\boldsymbol{x}_0 - \mathrm{R}^{-1}\boldsymbol{x}')) \mathrm{d}\boldsymbol{x}' \\ &= (I * h(\mathrm{R}^{-1}\cdot))(\mathrm{R}\boldsymbol{x}_0), \quad \text{for any } \mathrm{R} \in SO(D).\end{aligned} \tag{1.10}$$

This is the result of performing a change of variable and noting that the determinant of the Jacobian of a rotation matrix is 1. This relation is central since all the efforts to construct LRI convolutional operator are focused on countering or annihilating the local rotation R^{-1} of the filter h.

Isotropic filters

The simplest way to achieve LRI convolutional operators is by using an isotropic filter $h(\boldsymbol{x}) = h_r(||\boldsymbol{x}||)$. Such filters lead to LRI operator since the filter is radially symmetric *i.e.* $h_r(||\mathrm{R}^{-1}\boldsymbol{x}||) = h_r(||\boldsymbol{x}||)$. This relation plugged into Equation (1.10) leads to global rotation equivariance. Furthermore, obtaining rotation equivariance with purely convolutional operators can only be achieved with circularly/spherically symmetric filters (Andrearczyk, Fageot, et al., 2020, Appendix A). However, isotropic filters are non-directional and, thus, quite limited in characterizing complex textures.

In this category of filters, the Laplacian of Gaussian (LoG) is widely used, and it consists of the second order derivative of D-dimensional Gaussians defined by:

$$h(\boldsymbol{x}) = -\frac{1}{\pi\sigma^2}\left(1 - \frac{||\boldsymbol{x}||^2}{2\sigma^2}\right)e^{-\frac{||\boldsymbol{x}||^2}{2\sigma^2}} \tag{1.11}$$

The parameter σ controls the size of the Gaussian and is used to analyze various scales of the image. In practice, a range of values of σ are used, and the most informative ones are selected during the final modeling. The scalar measurements are derived from aggregated statistics of the intensity distribution of the feature maps.

Wavelets

Another commonly implemented family of filters in radiomics is the wavelet filters. This family of operators forms a rich and successful mathematical framework (Mallat, 1999) that is still widely used today as for instance in the JPEG2000 coding format (Rabbani, 2002). The wavelet framework is best known for the decimated Discrete Wavelet Transform (DWT), which is implemented by the convolution of an image $I[\boldsymbol{k}]$ with two one-dimensional filters: a high-pass filter h_H and a corresponding low-pass filter h_L. For $D-$dimensional images, this process is performed for each dimension separately. For instance, in 2D, it results in the four response maps $g_{LL}, g_{LH}, g_{HL}, g_{HH}$, which are obtained after the subsequent application of h_H and h_L to each dimension. As an example, the computation of g_{LH} is performed by first applying one pass of the one-dimensional filter h_L over the first dimension of the image I:

$$g_L[k_1, k_2] = \sum_{k_1'} I[k_1', k_2]h_L[k_1 - k_1']. \tag{1.12}$$

Then, the final g_{LH} is obtained by applying the filter h_H to the feature map g_L:

$$g_{LH}[k_1, k_2] = \sum_{k_2'} g_L[k_1, k_2']h_H[k_2 - k_2']. \tag{1.13}$$

These two one-dimensional convolutions taken together are referred to as *separable convolution* since they are equivalent to a 2D convolution with the *separable* filter $h_{LH}[k_1, k_2] = h_L[k_1]h_H[k_2]$, but it is computationally more efficient to apply them separately.

These feature maps are then downsampled by a factor of 2 in every direction, which yields the first iterations of the DWT and the feature maps $g^1_{LL}, g^1_{LH}, g^1_{HL}, g^1_{HH}$. The subsequent iterations of the transform are obtained by applying the same scheme (first filtering by h_L and h_H, then downsampling) to the low-pass feature maps g^j_{HH}. This process is usually carried out j times such as $2^j = d$ for a $d \times d$ image, which leads to a final decomposition with no low-pass component g^j_{LL} left. Figure 1.3 illustrates the

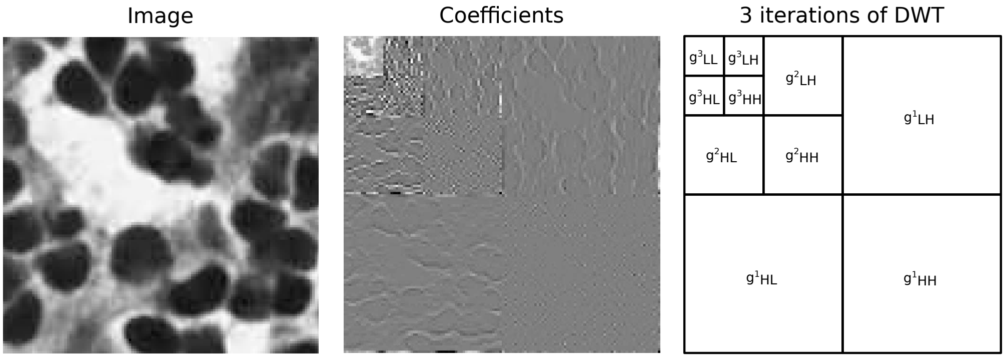


Figure 1.3: Example of the decimated DWT for three iterations on a grayscale image of cell nuclei.

decimated DWT for three iterations.

The decimated DWT is very efficient in encoding images in sparse representation and was designed for image compression. The downsampling step of the decimated DWT deteriorates the translation equivariance of the convolution. For this reason, it is preferred to use the stationary, also called undecimated, DWT (Holschneider et al., 1990) for image analysis purposes. This approach is very similar to the decimated DWT, but instead of downsampling the feature maps, the filters h_L and h_H are upscaled between each iteration to cover all the image scales. The resulting feature maps have the same size as the input image, and the translation equivariance is conserved. However, the undecimated DWT is redundant because the number of coefficients in the decimated DWT is sufficient for perfect image reconstruction. Thus, the undecimated DWT includes redundant coefficients and is suboptimal for image compression. Nevertheless, since the goal is image characterization, it is more appropriate to use the undecimated DWT for radiomics. For instance, the wavelet operators used in the pyradiomics (Van Griethuysen et al., 2017) toolbox are based on the implementation of the undecimated DWT of PyWavelets (G. Lee et al., 2019).

One major drawback of both the decimated and undecimated DWT is their directional sensitivity since they both rely on separable convolutions. In general, separable filters produce directionally sensitive operators, with the only exception being the Gaussian filters. Therefore, the generated feature maps are not LRI since linear directional sensitive operators cannot be LRI, as explained in Section 1.1.6.

As an alternative, non-separable wavelets can be used (Unser et al., 2011). Various approaches exist to design LRI and directionally sensitive image operators based on non-separable wavelets, which include steerable wavelets (Chenouard & Unser, 2012b), the Riesz transform, which is explained in Section 1.1.7, and CHs/SHs modulated by a non-separable wavelet radial profile (Depeursinge et al., 2018).

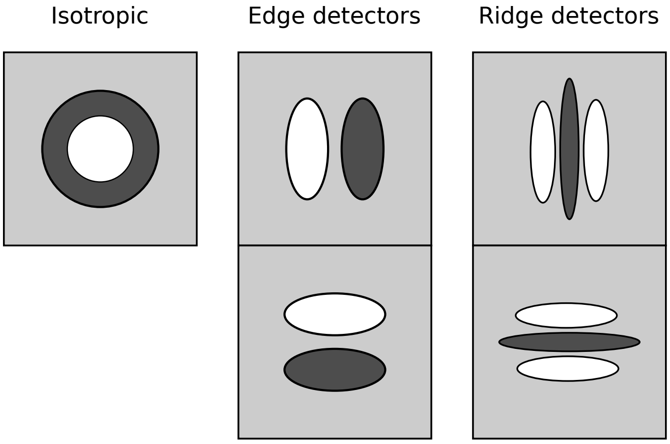


Figure 1.4: Hypothetic filterbank containing one isotropic filer, one edge detectors at two different orientations, and a ridge detector at two different orientations.

Filterbank of Rotated Filters

A straightforward and convenient way to obtain collections of relevant feature maps is using filterbanks. A filterbank is a collection of filters that are chosen according to the task at hand. The feature maps are generated by convolving the image $I(\boldsymbol{x})$ with the element of the filterbank.

Building feature maps that are LRI is straightforward with this approach since it suffices to use the same filters at different orientations and combining the resulting feature maps with a pixel-wise operation such as the maximum. Taking the pixel-wise maximum on a set of feature maps is often referred to as *orientation max-pooling*. More formally, we can compute an LRI feature map g_{LRI} as:

$$g_{LRI}(\boldsymbol{x}_0) = \max_{\mathrm{R}\in SO(D)} (I * h(\mathrm{R}\cdot))(\boldsymbol{x}_0), \tag{1.14}$$

for each position $\boldsymbol{x}_0$ of the image domain. In practice, though, since we cannot use infinite filterbanks, the set $SO(D)$ must be discretized. For instance, in 2D, a discretization of $SO(2)$ could be the set of rotations by an angle $\theta = 0, \frac{\pi}{4}, \frac{\pi}{2}, \ldots$. For this reasons, we can only compute approximate LRI operators from filterbanks. Figure 1.4 illustrates a schematic representation of a simple filterbank with five filters, including one isotropic, an edge detector at two different orientations, and a ridge detector at two different orientations.

An example of implementation of such approaches is the Maximum Response 8 (MR8) filterbank (Varma & Zisserman, 2005b) which contains 38 different filters, including two circularly symmetric filters, one edge detector, and one ridge detector, both of which are expressed on six orientations and three scales. After orientation max-pooling, the method yields a total of eight feature maps, which are approximately LRI. The feature can then be extracted with the usual aggregation function. However, in the original publication they used pixel-wise clustering.

An important class of filters that are often used in filterbanks are the *steerable* filters (Freeman & Adelson, 1991). This class of filters allows for the implementation of the operator defined by Equation (1.14) more efficiently. The solid circular and spherical harmonics are an example of steerable filters, as explained in Section 1.1.9.

Riesz Features

A well-known approach to describing variations in images is using image derivatives. This approach leads to interpretable quantities. For instance, first-order derivatives, *i.e.* the gradient, express local gray-level variations and second-order derivatives, *i.e.* the Hessian, describe local curvature in terms of gray-level variations. In computer vision, since images are composed of spatial transitions at different scales and also noise, it is useful to consider image derivatives at a given scale, and this is implemented thanks to the following property of the convolution:

$$I * \frac{\partial h}{\partial x_d} = \frac{\partial I}{\partial x_d} * h, \tag{1.15}$$

with $1 \leq d \leq D$. Thus, convolving the image I with the derivative of circularly/spherically symmetric filter h leads to a directionally sensitive image operator that approximates the image derivative at the scale of the filter h. Various choices of h can be used. For instance, the first-order derivatives of Gaussian filters are often employed to compute the local gradient of an image. The variance of the Gaussian used defines the scale at which the gradient is computed. In practice, it is advantageous to compute the gradient of an image by the convolution with derivatives of radially symmetric filters since it allows to focus on a given scale of the image and increases the operator's robustness toward the noise.

Filters derivatives can be computed directly in the image domain by finite difference. However, computing them in the Fourier domain is more elegant since it can be done very efficiently for any order by the following formula:

$$\mathcal{F}\left\{\frac{\partial^l h}{\partial x_d^l}\right\}(\boldsymbol{\omega}) = (\mathrm{j}\omega_d)^l \hat{h}(\boldsymbol{\omega}), \tag{1.16}$$

with l the derivative order. Therefore, differentiating a filter along the d^{th} dimension in the Fourier domain amounts to multiply its Fourier transform by $\mathrm{j}\omega_d$. Computing derivative as (1.16) leads to high-pass filters and accentuates high frequencies along the direction x_d. For this reason, it is preferable to use an all-pass version of Equation (1.16) which is implemented by the real Riesz transform. The first order Riesz transform of a filter h is given by

$$\boldsymbol{\mathcal{R}}\{h\}(\boldsymbol{x}) = \begin{pmatrix} \mathcal{R}_1\{h\}(\boldsymbol{x}) \\ \vdots \\ \mathcal{R}_D\{h\}(\boldsymbol{x}) \end{pmatrix} \xleftrightarrow{\mathcal{F}} -\mathrm{j}\frac{\boldsymbol{\omega}}{||\boldsymbol{\omega}||}\hat{h}(\boldsymbol{\omega}). \tag{1.17}$$

The first order Riesz transform $\boldsymbol{\mathcal{R}}\{h\}$ defines D all-pass image operators that can be understood as an all-pass version of the gradient of h. Similarly, higher order Riesz transform can be considered to compute all-pass surrogates of Hessian matrices.

In practice, first orders and second orders Riesz transform of circularly/spherically symmetric filter h are computed to generate a filterbank. LRI is achieved by locally aligning each element of the filterbank to the local direction of the image gradient. This operation is computationally expensive, but the Riesz transform's steerability helps reduce this cost. Several studies have shown the benefits of the Riesz ransform (Depeursinge et al., 2015; Dicente Cid et al., 2017b).

Summary

In the previous sections, we described the most commonly implemented and used texture features in radiomics. Despite their limitations, GLM-based texture operators remain the most prevalent in radiomics toolboxes. We saw that LBP is an exciting framework for designing LRI operators. However, their computation involves a binarization step that discards much of the information contained in the image dynamic range. Alternatively, convolutional-based operators offer a wide range of implementations of LRI operators. However, highly directional filters without being LRI are also commonly used, such as the undecimated DWT. We do not recommend using this type of features in radiomics analysis since robustness to image rotation is a must.

The Riesz transform of non-separable wavelets offers a solid approach to obtaining LRI operators. Nonetheless, this approach requires defining an arbitrary criterion to align the filterbank, which is computationally expensive.

In the following sections, we describe a framework based on Fourier descriptors that allow us to implement efficiently the two strategies explained in Section 1.1.6 to obtain LRI operators.

Table 1.1 summarizes the strengths and weaknesses of each discussed method and the approach discussed in the following sections.

Method	**Directional sensitivity**	**Directional coverage**	**LRI**	**Continuous rotations**	**Gray level quantization**
Isotropic filters (*e.g.* LoG)	✗	complete	✓	✓	✗
GLCM	✓	complete	approx.	✗	✓
GLSZM	✗	complete	✗	-	✓
GLRLM	✓	incomplete	approx.	✗	✓
LBP	✓	complete	approx.	✗	✓
Undecimated DWT	✓	incomplete	✗	-	✗
Riesz transform	✓	complete	✓	✗	✗
circular/spherical Fourier invariants	✓	complete	✓	✓	✗

Table 1.1: Qualitative comparison of the common radiomics features.

1.1.8 Fourier-based Rotation Invariants

One of the goals of this thesis was to design image operators that are LRI while being sensitive to directions. In this section, we show how such operators can be derived from the well-known Fourier transform.

The critical property of the Fourier transform that allows building invariant representations is its behavior toward shifts of the argument of a function f. Let us consider the case of functions defined on the real line $f : \mathbb{R} \to \mathbb{R}$. We denote by $\hat{f}(\omega) = \int_{\mathbb{R}} f(t) e^{-\mathrm{j}\omega t} \mathrm{d}t$ its Fourier transform, and we observe that the Fourier transform of a shifted function is $\mathcal{F}\{f(\cdot - t_0)\}(\omega) = \hat{f}(\omega) e^{-\mathrm{j}\omega t_0}$. Naturally, it follows that any representation that discards the phase shift $e^{-\mathrm{j}wt_0}$ induced by the shift in the argument of f will result in a shift-invariant representation of f.

One simple quantity that discards this dependency is the power spectrum $|\hat{f}(\omega)|^2$. While invariant to shifts, the power spectrum discards all the information in the phase of the function. This information is, in general, crucial as it informs on how the Fourier expansion's sinusoids interact to form important patterns such as edges and ridges. To tackle this issue, another quantity was introduced, the bispectrum (Kakarala & Mao, 2010), defined by $b_f(\omega_1, \omega_2) = \hat{f}(\omega_1)\hat{f}(\omega_2)\hat{f}^*(\omega_1 + \omega_2)$. The bispectrum has been shown to be complete, meaning that if two functions f_1 and f_2 have the same bispectrum, *i.e.* $b_{f_1}(\omega_1, \omega_2) = b_{f_2}(\omega_1, \omega_2)$ for any $\omega_1, \omega_2 \in \mathbb{R}$ then the functions f_1 and f_2 are shifted versions of each other. The completeness of the bispectrum makes it an interesting candidate for the design of invariant representation since all the information contained in the phase is conserved, except for the one that encodes the relative position of f.

In the following sections, we show how we can generalize this intuition to rotation invariance and introduce this into the framework of image operators. Section 1.1.8 introduces the Circular Harmonics (CHs) as they are very close to the 1D dimensional Fourier series of 2π-periodic functions since they are the Fourier basis of functions defined on the circle $\mathbb{S}^1$. We then extend the theory to Spherical Harmonics (SHs), which represent the Fourier basis for functions defined on the sphere $\mathbb{S}^2$. Finally, in Section 1.1.9, we show

how the rotation invariance for functions defined on the circle or the sphere is introduced in a convolutional framework to obtain LRI image operators.

Circular Harmonics (CHs)

This section describes how we can derive rotation invariant representation for functions defined on the circle.

The unit circle is defined as

$$\mathbb{S}^1 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}. \tag{1.18}$$

The most natural choice of coordinates to describe the unit circle are the polar coordinates, and we will use the standard definition:

$$\boldsymbol{x} = (x_1, x_2) = (r\cos(\theta), r\sin(\theta)), \tag{1.19}$$

with $r \geq 0$ and $\theta \in [0, 2\pi)$ referred to as the radius and the polar angle.

The parametrization of the unit circle in polar coordinates is given by:

$$\mathbb{S}^1 = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} = (\cos(\theta), \sin(\theta)) \text{ and } 0 \leq \theta < 2\pi\}. \tag{1.20}$$

Thus, any function defined on the circle $f : \mathbb{S}^1 \to \mathbb{R}$ can be represented by a 2π-periodic function of the polar angle θ. It follows that any square-integrable function defined on the circle, *i.e.* $f \in L^2(\mathbb{S}^1)$, can be expanded according to the standard Fourier series as

$$f(\theta) = \sum_{n=-\infty}^{\infty} \hat{f}_n e^{\mathrm{j}n\theta}, \tag{1.21}$$

where $\hat{f}_n = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-\mathrm{j}n\theta}\mathrm{d}\theta$ is the Fourier coefficient of f. The family of functions $[e^{\mathrm{j}n\theta}]_{n\in\mathbb{Z}}$ are the so-called Circular Harmonics (CHs) and are simply sinusoids defined on the circle. Figure 1.5 depicts representative of the CH family for degree $n = 0, \ldots, 4$ modulated by a Gaussian radial profile for illustration purpose. This figure shows that for increasing degrees, the patterns that the CHs expansion can express becomes finer.

A shift in the argument of the function $f(\cdot - \theta_0)$ results in rotation of the function on the circle. We introduce the following notation for rotations applied to a function $f \in L^2(\mathbb{S}^1)$:

$$\mathcal{R}_{\theta_0}\{f\} = f(\cdot - \theta_0). \tag{1.22}$$

This notation can seem a little excessive in the context of CHs but will make the extension to the sphere more evident.

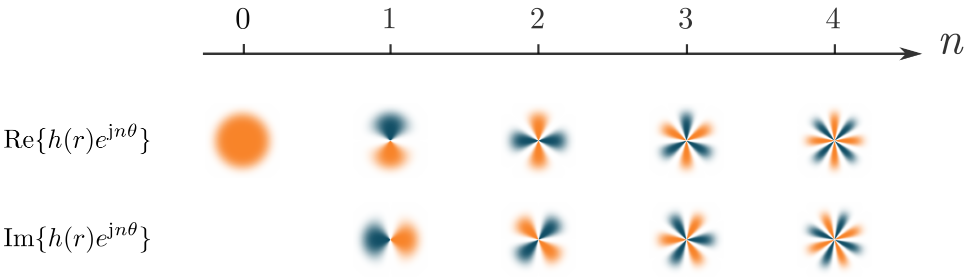

0
1
2
3
4
n
$\mathrm{Re}\{h(r)e^{\mathrm{j}n\theta}\}$
$\mathrm{Im}\{h(r)e^{\mathrm{j}n\theta}\}$

Figure 1.5: Illustration of CHs for $n = 1, \ldots, 4$. The CHs are modulated by a Gaussian profile $h(r)$ for illustration purposes.

One key aspect of the Fourier expansion, as stated previously, is its shifting property:

$$\mathcal{R}_{\theta_0}\{f\}(\theta) = f(\theta - \theta_0) = \sum_{n=-\infty}^{\infty} \left(\hat{f}[n]e^{-\mathrm{j}n\theta_0}\right)e^{\mathrm{j}n\theta} \tag{1.23}$$

for any $\theta_0 \in [0, 2\pi)$. By identifying the terms of Equation (1.23) and (1.21), we see that the rotation of f translates to a modulation with a sinusoid in the Fourier expansion. Thus, any representation that discards the sinusoid $e^{\mathrm{j}n\theta_0}$ is rotation invariant. For instance, the power spectrum is defined by:

$$s_n(f) = \hat{f}_n\hat{f}_n^* = |\hat{f}_n|^2 \tag{1.24}$$

satisfies this requirement. One can readily verify that $s_n(\mathcal{R}_{\theta_0}\{f\}) = s_n(f)$ since the power spectrum discards completely the phase of the Fourier expansion. However, this results in the loss of the phase between the CHs which contains important information about how they interact to form complex patterns. An illustration of 2D patterns that are not discriminated by the spectrum is shown in Figure 1.6.

A quantity that is phase-sensitive while being invariant to rotation is the bispectrum and is defined as follows:

$$b_{n,n'}(f) = \hat{f}_n\hat{f}_{n'}\hat{f}_{n+n'}^*. \tag{1.25}$$

One can verify its invariance by replacing $\hat{f}_n$ by $\hat{f}_n e^{-\mathrm{j}n\theta_0}$ in Equation (1.25). Furthermore, the bispectrum is complete (Kakarala, 2012) in the sense of Proposition 1.

Proposition 1 *For any pair of functions $f, f' \in L^2(\mathbb{S}^1)$ whose bispectral coefficients are equal $b_{n,n'}(f) = b_{n,n'}(f')$ for all $n, n' \in \mathbb{Z}$, there exists a rotation $\mathcal{R}$ such that $\mathcal{R}\{f'\} = f$.*

In other words, the bispectral decomposition of a function completely characterizes

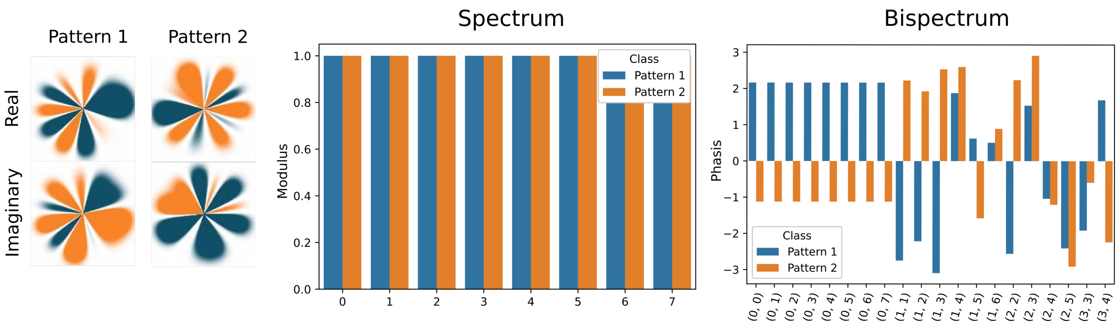


Figure 1.6: Example of two distinct patterns expressed on the circle having the same power spectrum while exhibiting a very different bispectral decomposition.

this function up to a rotation. Figure 1.6 shows two patterns built from CHs that are not discriminated by the power spectrum, while the bispectrum captures the difference. Proposition 1 can be proved by using a recursive algorithm such as the one proposed in (Giannakis, 1989). The proof can be quite technical and is out of the scope of this thesis. Another proof was proposed in (Kakarala et al., 2011) in the context of the bispectrum defined on homogenous groups. This result can be applied to Proposition 1, since it is less general.

Spherical Harmonics (SHs)

The unit sphere is defined as

$$\mathbb{S}^2 = \{(x_1, x_2, x_3) \in \mathbb{R}^2 : x_1^2 + x_2^2 + x_3^2 = 1\}. \tag{1.26}$$

The usual spherical coordinate mapping is given by

$$\boldsymbol{x} = (x_1, x_2, x_3) = (r\cos(\phi)\sin(\theta), r\sin(\phi)\sin(\theta), r\cos(\theta)), \tag{1.27}$$

with $r \geq 0$ the radius, $\phi \in [0, 2\pi)$ the azimuthal angle, and $\theta \in [0, \pi]$ the polar angle.

The canonical parametrization of the sphere is given by setting the radius to $r = 1$ and letting the azimuthal and polar angle as free parameters. Thus, functions that take values on the sphere can be expressed as functions of these two angles. Similarly to functions defined on the circle, any square-integrable function defined on the sphere $f \in L^2(\mathbb{S}^2)$ can be expanded in the SHs basis as

$$f(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \hat{f}_{n,m} Y_n^m(\theta, \phi), \tag{1.28}$$

where $\hat{f}_{n,m} = \langle f, Y_n^m \rangle_{L^2(\mathbb{S}^2)} = \int_0^{\pi} \int_0^{2\pi} f(\theta, \phi)\overline{Y_n^m(\theta, \phi)} \sin(\theta) \mathrm{d}\phi \mathrm{d}\theta$ are the Fourier coefficients of f, also referred to as the Fourier transform of f, and Y_n^m are the so-called SHs.

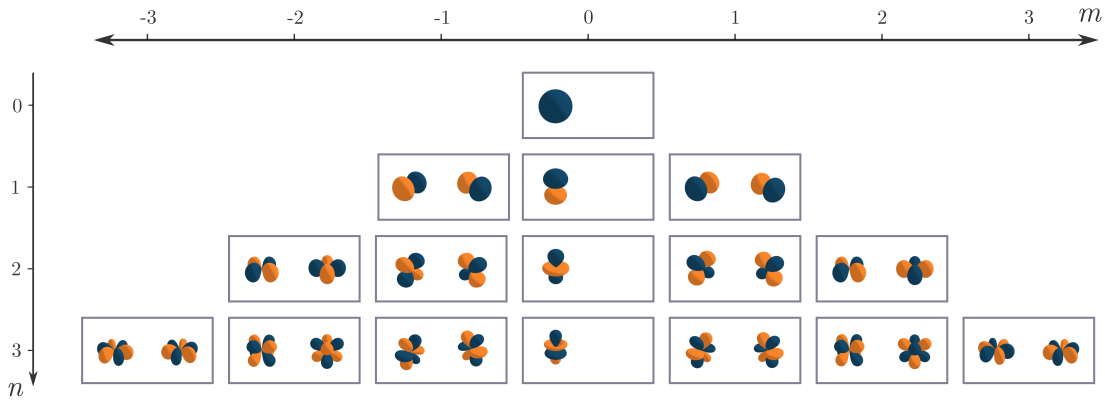


Figure 1.7: Representative of the SHs family for $n = 0, \ldots, 3$. The real and imaginary parts are represented on the left and right of the grey rectangles. The orange and blue colors represent positive and negative values, respectively.

Figure 1.7 illustrates the representative of the SH family for $n = 0, \ldots, 3$. The SHs are defined as (Driscoll & Healy, 1994)

$$Y_n^m(\theta, \phi) = A_n^m P_n^{|m|}(\cos(\theta)) e^{\mathrm{j}m\phi} \tag{1.29}$$

with $A_n^m = (-1)^{(m+|m|)/2} \left(\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}\right)^{(1/2)}$ a normalization constant and $P_n^{|m|}$ the associated Legendre polynomial given for $0 \leq m \leq n$ by

$$P_n^m(x) = \frac{(-1)^m}{2^n n!}(1-x^2)^{m/2} \frac{\mathrm{d}^{n+m}}{\mathrm{d}x^{n+m}}(x^2-1)^n, \tag{1.30}$$

The indices n and m are referred to as the SH's degree and order, respectively.

The Fourier coefficients $\hat{f}_{n,m}$ are grouped together to form the $(1 \times 2n+1)$ vector:

$$\boldsymbol{\mathcal{F}}_n = [\hat{f}_{n,-n} \ldots \hat{f}_{n,0} \ldots \hat{f}_{n,n}]. \tag{1.31}$$

This quantity is the spherical equivalent of the coefficients $\hat{f}_n$ in Equation (1.21) and are accordingly referred to as the spherical Fourier transform of $f(\theta, \phi)$.

Rotations on the sphere are more complex than on the circle since three degrees of freedom are required to express them. Indeed, any 3D rotation can be expressed as the rotation by an angle α_0 around an axis defined by the two spherical coordinate θ_0 and ϕ_0. Several conventions exist to represent rotations in a 3D space. However, we do not need to compute them explicitly in this work. Hence, we will use an abstract notation $\mathcal{R}_0\{f\}$ to represent a 3D rotations applied to a function $f \in L^2(\mathbb{S}^2)$.

The spherical Fourier vector $\boldsymbol{\mathcal{F}}'_n$ of a function $f' = \mathcal{R}_0\{f\}$ is given by (Kakarala & Mao, 2010, Equation (5))

$$\boldsymbol{\mathcal{F}}'_n = \boldsymbol{\mathcal{F}}_n \mathrm{D}_n(\mathcal{R}_0) \tag{1.32}$$

with $\mathrm{D}_n(\mathcal{R}_0)$ a $(2n+1) \times (2n+1)$ unitary matrix commonly referred as the Wigner D-matrix (Varshalovich et al., 1988, Chapter 4). The Wigner D-matrices can intuitively be understood as the sinusoids in Equation (1.23).

Similarly to the power spectrum of the CHs expansion, we can define the power spectrum of the SHs expansion as follow:

$$s_n(f) = \frac{1}{2n+1}\boldsymbol{\mathcal{F}}_n\boldsymbol{\mathcal{F}}_n^\dagger = \frac{1}{2n+1}\sum_{m=-n}^{n} |\hat{f}_{n,m}|^2, \tag{1.33}$$

where the symbol † represents the Hermitian transpose. The quantity $s_n(f)$ is as well rotation invariant which can be shown using the fact that the Wigner D-matrices are unitary, meaning that $\mathrm{D}_n(\mathcal{R}_0)\mathrm{D}_n^\dagger(\mathcal{R}_0) = \mathrm{I}$ for any $\mathcal{R}_0$, and replacing $\boldsymbol{\mathcal{F}}_n$ by $\boldsymbol{\mathcal{F}}_n\mathrm{D}_n(\mathcal{R}_0)$.

The circular bispectrum also has its spherical counterpart. The spherical bispectrum is defined as (Kakarala & Mao, 2010, Equation (24)):

$$b^\ell_{n,n'}(f) = [\boldsymbol{\mathcal{F}}_n \otimes \boldsymbol{\mathcal{F}}_{n'}]\mathrm{C}_{nn'}\widetilde{\boldsymbol{\mathcal{F}}_\ell}^\dagger, \tag{1.34}$$

with $\otimes$ the Kronecker product, $\mathrm{C}_{nn'}$ the $(2n+1)(2n'+1)\times(2n+1)(2n'+1)$ Clebsh-Gordan matrix containing the Clebsh-Gordan coefficients (detailed definition and description can be found in Section A.1), and $\widetilde{\boldsymbol{\mathcal{F}}_\ell} = [0,\dots,0,\boldsymbol{\mathcal{F}}_\ell,0,\dots,0]$ is a zero-padded vector of size $1 \times (2n+1)(2n'+1)$ containing the spherical Fourier vector of degree ℓ with $|n-n'| \leq \ell \leq n+n'$. The zero-padding is performed to match the size of $\mathrm{C}_{nn'}$ and to select only the rows corresponding to the ℓ^{th} degree.

Summary

We described the structure shared by the circular and spherical Fourier expansions in the previous sections. The behavior of these expansions with respect to rotations can be exploited to build rotation invariant representations. In the circular expansion, a linear phasis factor appears in response to rotations. In the spherical case, the unitary Wigner matrix appears in response to rotations. Thus, the strategy to build rotation invariant representations is to use quantities that discard this dependency. Table 1.2 compares the circular and spherical Fourier expansion, the effect of rotations, and rotations invariants built from the power spectrum and bispectrum.

			Circle $\mathbb{S}^1$	Sphere $\mathbb{S}^2$
Expansion	f		$\sum_{n=-\infty}^{\infty} \hat{f}_n e^{\mathrm{j}n\theta}$	$\sum_{n=0}^{\infty} \sum_{m=-n}^{n} \hat{f}_{n,m} Y_n^m(\theta, \phi)$
Fourier transform	f	$\xleftrightarrow{\mathcal{F}}$	$\hat{f}_n = \langle f \,, e^{\mathrm{j}n\cdot} \rangle_{L^2(\mathbb{S}^1)}$	$\hat{f}_{n,m} = \langle f \,, Y_n^m \rangle_{L^2(\mathbb{S}^2)}$
Rotation	$\mathcal{R}_0\{f\}$	$\xleftrightarrow{\mathcal{F}}$	$\hat{f}_n e^{-\mathrm{j}n\theta_0}$	$\boldsymbol{\mathcal{F}}_n \mathrm{D}_n^{\dagger}(\mathcal{R}_0)$
Spectrum			$s_n(f) = \lvert \hat{f}_n \rvert^2$	$s_n(f) = \frac{1}{2n+1} \sum_{m=-n}^{n} \lvert \hat{f}_{n,m} \rvert^2$
Bispectrum			$b_{n,n'}(f) = \hat{f}_n \hat{f}_{n'} \hat{f}_{n+n'}^*$	$b_{n,n'}^{\ell}(f) = [\boldsymbol{\mathcal{F}}_n \otimes \boldsymbol{\mathcal{F}}_{n'}] \mathrm{C}_{nn'} \widetilde{\boldsymbol{\mathcal{F}}_{\ell}}^{\dagger}$

Table 1.2: Summary and comparison of the Fourier transform properties and rotation invariants on the circle and the sphere.

1.1.9 Building LRI from Fourier

We saw in Section 1.1.6 that there are mainly two approaches to obtaining LRI image operators. First, the descriptor of the image operator can be aligned at every position of the image. Second, a local descriptor that is invariant to rotations can be used.

The CHs and SHs provide a framework that allows for the implementation of both strategies. In this section, we will develop the theory for the CHs (2D) as it is simpler than the SHs (3D). Nevertheless, the extension to SHs is straightforward, given the similar structure that they share with the CHs.

Aligning the Local Descriptor

As described earlier, in this approach, we need to define a criterion to select the local orientation of the descriptor. Here, we consider the orientation that maximizes the response of the filter at position $\boldsymbol{x}_0$. Other criteria can also be used, such as maximizing the local gradient of the image(Chenouard & Unser, 2012a). In (Andrearczyk, Fageot, et al., 2019, 2020), we used as a criterion the direction that maximizes the absolute value of the convolution.

The image operator that we will discuss here is formulated as

$$\mathcal{G}_h^{\text{steer}}\{I\}(\boldsymbol{x}_0) = \max_{\mathrm{R} \in SO(2)} (I * h(\mathrm{R}\cdot))(\boldsymbol{x}_0). \tag{1.35}$$

In practice, the operator G_h^{steer} is approximated by using a finite subset $S \subset SO(2)$ of the 2D rotations. We use the notation $g^{\theta_0} = I * h(\mathrm{R}_{\theta_0}\cdot)$ to mean that the feature map is obtained by the application of filter with orientation θ_0. Aligning the local descriptor for each position can be implemented efficiently with filters that are expressed as

$$h(r,\theta) = \sum_{n=-N}^{N} h_n(r)e^{\mathrm{j}n\theta}. \tag{1.36}$$

These filters are steerable, meaning that their rotation $h(\mathrm{R}\cdot)$ can be expressed as a linear combination of themselves, obviating the need to reconvolve the rotated filters with the input. As an aside, one might be concerned that the filters defined by (1.36) are complex, but they can be made real by imposing the Hermitian symmetry $h_{-n}(r) = h_n^*(r)$.

The steerability for such filters takes the form

$$h(\mathrm{R}_{\theta_0}\boldsymbol{x}) = \sum_{n=-N}^{N} h_n(r)e^{\mathrm{j}n\theta}e^{-\mathrm{j}n\theta_0} \tag{1.37}$$

and is direct consequence of Equation (1.40). Computing the feature map g^{θ_0} becomes:

$$g^{\theta_0} = I * \sum_{n=-N}^{N} h_n(r)e^{\mathrm{j}n\theta}e^{-\mathrm{j}n\theta_0} = \sum_{n=-N}^{N} (I * h_n(r)e^{\mathrm{j}n\theta})e^{-\mathrm{j}n\theta_0}. \tag{1.38}$$

This equation shows that with a limited number of convolutions, here $2N+1$, we can express the feature maps h^{θ_0} for any orientation θ_0.

The implementation of the LRI operator G_h^{steer} consists of computing the set of feature maps $[I * h_n(r)e^{\mathrm{j}n\theta}]_{n=-N}^{n=N}$, then recombining them according to Equation (1.38) to obtain another set of feature maps $[g^{\theta_0}(\boldsymbol{x}_0)]_{\theta_0 \in S}$. The final operator is obtained by applying the maximum operator among all the rotated feature maps at every position $\boldsymbol{x}_0$.

This approach was implemented in a Convolutional Neural Network (CNN) framework in (Andrearczyk, Fageot, et al., 2019, 2020) for 3D CT images by using the SHs expansion. The extension to SHs relies on choosing filters expressed as

$$h(\boldsymbol{x}) = \sum_{n=0}^{N}\sum_{m=-n}^{n} h_{n,m}(r)Y_n^m(\theta,\phi). \tag{1.39}$$

These filters are also steerable, which is the consequence of the behaviour of the SHs with respect to rotation. The steerabiliy of SHs filters is given by

$$h(\mathrm{R}\boldsymbol{x}) = \sum_{n=0}^{N}\sum_{m=-n}^{n}\sum_{m'=-n}^{n} [\mathrm{D}_n(\mathrm{R})]_{m',m} h_{n,m'}(r)Y_n^m(\theta,\phi) \tag{1.40}$$

The implementation follows the same logic and consist of computing the feature maps with the SHs, then computing the feature maps for a subset of $SO(3)$.

The benefit of steerability may not seem evident in 2D as we could use a filterbank of rotated filters. However, since 3D rotations have more degrees of freedom than 2D rotations, steerability can dramatically decrease the computational cost in that case.

The main drawback of the steering approach is that many orientations must be sampled to obtain a reasonable LRI operator, which leads to an approximation of the LRI operator. Furthermore, this method requires an arbitrary criterion that may be sub-optimal.

The following section shows how we can use CHs and SHs to derive rotation invariant operators that do not suffer from these limitations.

Using Invariant Descriptor

The first step of this approach is the same as the one described in the previous Section. The image is first convolved with the *solid* CHs $[h_n(r)e^{\mathrm{j}n\theta}]_{n=-N}^{n=N}$. The term *solid* CHs is used to mean that they are expressed on the 2D domain instead of the circle, which is achieved by multiplying the CHs by a radial profile $h_n(r)$. The feature maps are then recombined according to the power spectrum or bispectrum formula to obtain LRI operators.

The *power spectral* operator takes the form

$$\mathcal{G}_n^{\mathrm{sp}}\{I\}(\boldsymbol{x}_0) = g_n(\boldsymbol{x}_0)\overline{g_n(\boldsymbol{x}_0)}, \tag{1.41}$$

and the *bispectral* one:

$$\mathcal{G}_{n,n'}^{\mathrm{bisp}}\{I\}(\boldsymbol{x}_0) = g_n(\boldsymbol{x}_0)g_{n'}(\boldsymbol{x}_0)\overline{g_{n+n'}(\boldsymbol{x}_0)}. \tag{1.42}$$

With $g_n(\boldsymbol{x}_0) = (I(\boldsymbol{x}) * h_n(r(\boldsymbol{x}))e^{\mathrm{j}n\theta(\boldsymbol{x})})(\boldsymbol{x}_0)$, and $r(\boldsymbol{x}), \theta(\boldsymbol{x})$ the mapping from Cartesian coordinate to polar coordinates. In Section B.1, we showed that operators defined as in Equations (1.41) and (1.42) are LRI. The intuition behind the proof is to observe that $g_n(\boldsymbol{x}_0)$ implements the n^{th} CHs expansion of a local projection of the image at the position $\boldsymbol{x}_0$. Thus, by applying the spectral or bispectral formula, we obtain a quantity invariant to the descriptor's local rotation.

The extension to 3D images is also done by mean of the SHs and the computed feature maps have the form $g_{n,m} = I * h_n(r)Y_n^m(\theta, \phi)$. Since the spherical power spectrum and bispectrum formula are defined for vectors, the feature maps $g_{n,m}$ are grouped to satisfy that structure. More details can be found in Section 2.3.6.

Summary

We described how Fourier theory could be used to build LRI operators. Since CHs and SHs share a similar structure, the same strategies can be used in both cases, making an extension to 3D images easier.

Figure 1.8 summarizes the two main strategies, namely using steerable filters or invariant representations. This Figure illustrates the implementation for the 2D case, but a similar picture could be done for the 3D case (Andrearczyk, Fageot, et al., 2019, Figure 2).

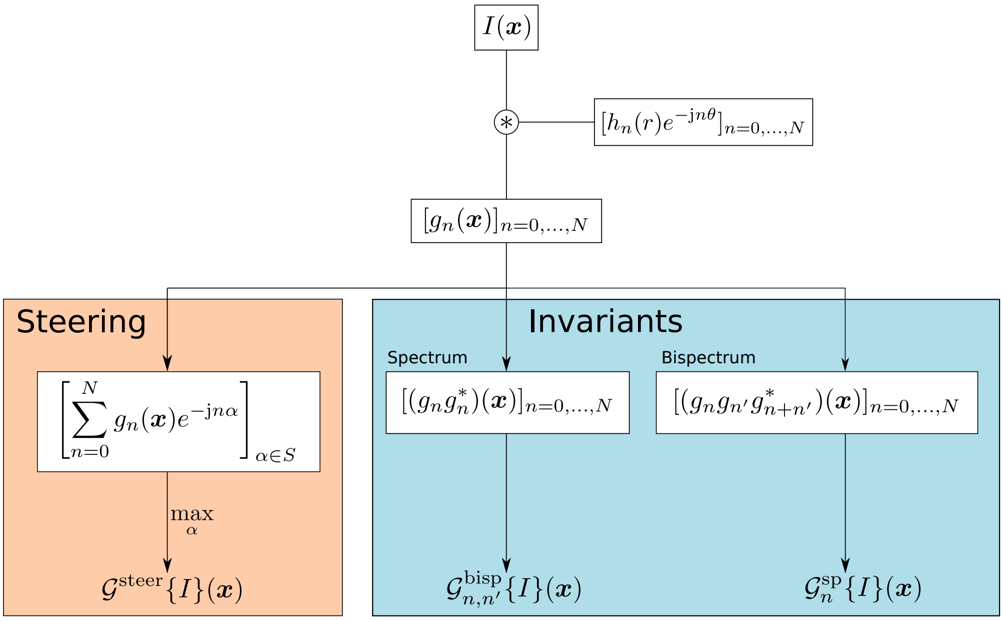

$I(\boldsymbol{x})$
$*$
$[h_n(r)e^{-\mathrm{j}n\theta}]_{n=0,\dots,N}$
$[g_n(\boldsymbol{x})]_{n=0,\dots,N}$
Steering
Invariants
Spectrum
Bispectrum
$\left[\sum_{n=0}^{N} g_n(\boldsymbol{x})e^{-\mathrm{j}n\alpha}\right]_{\alpha\in S}$
$[(g_n g_n^*)(\boldsymbol{x})]_{n=0,\dots,N}$
$[(g_n g_{n'} g_{n+n'}^*)(\boldsymbol{x})]_{n=0,\dots,N}$
$\max_{\alpha}$
$\mathcal{G}^{\mathrm{steer}}\{I\}(\boldsymbol{x})$
$\mathcal{G}^{\mathrm{bisp}}_{n,n'}\{I\}(\boldsymbol{x})$
$\mathcal{G}^{\mathrm{sp}}_{n}\{I\}(\boldsymbol{x})$

Figure 1.8: The two strategies used to obtain LRI operators are derived from CHs. The image is first convolved with a filterbank of CHs modulated by radially symmetric filter $h_n(r)$. The feature maps are then recombined according to the different methods to obtain the three different LRI operators: $\mathcal{G}^{\mathrm{steer}}$ (Equation (1.35)) for the operator coming from the steering strategy, $\mathcal{G}^{\mathrm{sp}}$ (Equation (1.41)) for the operator coming from the spectral invariant, and $\mathcal{G}^{\mathrm{bisp}}$ (Equation (1.42)) for the operator coming from the bispectral invariant.

1.1.10 Convolutional Neural Networks (CNNs)

CNNs have revolutionized the field of computer vision since the breakthrough of AlexNet (Krizhevsky et al., 2012) on the ImageNet challenge. CNNs are usually large models that contain billions of parameters. For instance, the commonly used ResNet50 architecture amounts to 26 million learnable parameters. This success was possible mainly thanks to the GPU's parallelization of the convolution operation. CNNs are complex models but created from fully differentiable elements. Therefore, CNNs optimization can and is mostly done by stochastic gradient descent.

The essence of CNNs is the convolutional layer which is often implemented as a correlation. Similarly to the convolution of Equation (1.9), the correlation of an image $I \in L^2(\mathbb{R}^D)$ and a filter $h \in L^\infty(\mathbb{R}^D)$ is defined by:

$$(f \star h)(\boldsymbol{x}_0) = \int_{\mathbb{R}^D} f(\boldsymbol{x}) h(\boldsymbol{x} - \boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}. \tag{1.43}$$

The only difference with the standard convolution is the sign of the argument of h. Thus, the correlation $f \star h$ is equivalent to the convolution with the filter $\tilde{h}(\boldsymbol{x}) = h(-\boldsymbol{x})$.

Each convolutional layer of a CNN compute a multi-channel correlation with a number C_{in} and C_{out} of input and output channels. This operation is given by

$$\overline{g}_o^{(l)}[\boldsymbol{k}] = \sum_{i=1}^{C_{\text{in}}} (g_i^{(l-1)} \star \kappa_o^i)[\boldsymbol{k}] \tag{1.44}$$

with $i = 1, \ldots, C_{\text{in}}$, $o = 1, \cdots, C_{\text{out}}$, $g_i^{(l-1)} : \mathbb{Z}^D \to \mathbb{R}$ the feature maps from the previous layer, and $\kappa_o^i : \mathbb{Z}^D \to \mathbb{R}$ the set of filters of the current layer. This operation is followed by the application of a pixel-wise non-linearity $g_o^{(l)}[\boldsymbol{k}] = \sigma(\overline{g}_o^{(l)}[\boldsymbol{k}])$ to obtain the final feature map $g_o^{(l)}$, where σ is a usually a Rectified Linear Unit (ReLU).

The filters $\kappa_o^i[\boldsymbol{k}]$, often referred to as *kernels*, are stored in a multi-dimensional array of size $K^D \times C_{\text{in}} \times C_{\text{out}}$ where K is the spatial size of the kernel. A typical value for the kernel size in modern CNNs is $K = 3$. All the entries of this multi-dimensional array are initialized randomly and learned during training time.

Current CNNs contain dozens of convolutional layers, which amount to many free parameters. Training these models usually requires enormous datasets. In the following section, we describe how the quantity of data is artificially augmented by data augmentation and how it is used to enforce invariance to pre-defined sets of transformations.

Invariance through Data Augmentation

One of the earliest methods used to enforce invariance to geometric transformations, and more generally, to any implementable transformations of the input image is data augmentation (Cireşan et al., 2011; Krizhevsky et al., 2012; Simard et al., 2003). It is also a good approach to artificially increase the size of the dataset, since deep learning models usually requires many data points for learning.

Data augmentation relies on an applying transformations of the input images, which are expected to happen in the natural distribution of the dataset. It is also applied to increase the robustness of the networks towards a group of transformations. As described in Section 1.1.5, the group of transformations can vary depending on the task. For instance, in the self-driving car case example, if we wish to train robust models towards scaling of the input image, we will train the networks on zoomed versions of the input images.

The main advantage of data augmentation is its versatility, as any transformation can be used as long as it is implementable. Complex geometric transformations, such as elastic transformations, can be implemented quickly. JPEG artifacts and Gaussian noise can also be added to the data augmentation pipeline to increase the robustness of machine learning methods. The Albumentations toolbox (Buslaev et al., 2020) is an excellent example of the implementation of these augmentations.

Furthermore, recent deep learning frameworks use implementations of data augmentation on the fly without having to store an extra augmented dataset. With a few lines of code, complex pipelines can be implemented that apply custom data transformation on the CPU and feed it to the model on the GPU, all those happening in parallel without impacting the time required for training.

Data augmentation can be used during the training phase, referred to as *train-time* augmentation (the one described in the previous paragraph), and during the inference phase, referred to as *test-time* augmentation. Test-time augmentation is a little different and generally works as follows. The input image is transformed according to a predefined set of fixed transformations. The output for the different augmented images is ensembled to obtain one final and more robust prediction. Any method for ensembling can be used, such as averaging the output or a majority voting.

Contrary to train-time augmentation, which does not increase the training compute (just by adding the computation of the transformation), the test-time augmentation requires multiple passes on the input images to compute one prediction, thus increasing the computation by the number of transformations.

Furthermore, it is not satisfying to use data augmentation, and hoping the network will learn the best way to encode robustness toward the given transformations. Better approaches that would profit from the transformation's structure to build efficient networks would be more elegant and interpretable.

Finally, it can be a waste of parameters in some contexts, as illustrated in Figure 1.9. We generated this figure training an AlexNet (Krizhevsky et al., 2012) on the CIFAR10 (Krizhevsky, Hinton, et al., 2009) dataset with data augmentation. We can observe that the same edge detector is learned multiple times but at different orientations. Hence, the following question arises: can we design networks that are intrinsically robust toward the rotation of the input image? We will discuss in the following sections how other researchers addressed this question. Finally, we will describe our solution to this problem.

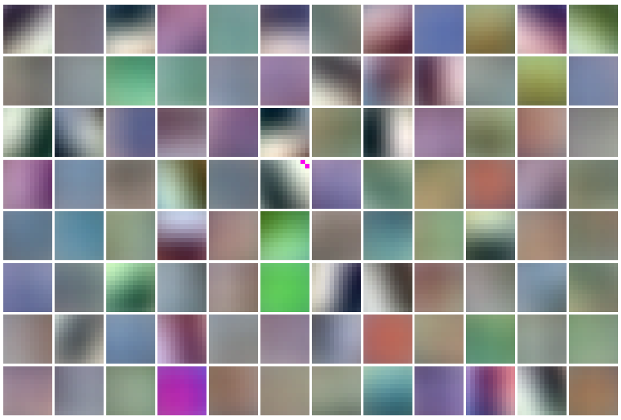


Figure 1.9: Kernels of the first layer of an AlexNet (Krizhevsky et al., 2012) model. This CNN was trained for 250 epochs on CIFAR10 (Krizhevsky, Hinton, et al., 2009). Similar learned kernels at different orientation can be observed.

Group Equivariant CNN

A fair amount of research has been made on designing CNNs that are rotation equivariant in addition to the translation equivariance inherited from spatial convolution. The key idea to designing this type of Group convolution CNNs (G-CNN) is to generalize the definition of convolution 1.9 to the more general convolution for functions defined on a group, *e.g.* besides the group of translations.

In this section, we will recall the theory about group convolution and consider images and feature maps with only one channel. In real applications, the convolutions are applied for muti-channels images and feature maps, as explained in Section 1.1.10. We also omit the non-linearities, but their usage in the G-CNN is straightforward (T. Cohen & Welling, 2016b). For simplicity and coherence with the pioneering work of (T. Cohen & Welling, 2016b), we will develop the mathematical formulation with respect to the correlation. Since correlation and convolution are very similar, the derivations of the formula for the standard convolution follow the same logic.

We will now define the mathematical notion of a group, action of group on a set, and group correlation. The definition of groups is derived from (Artin, 2018, Chap. 2).

Definition 3 *A group is defined as a set G together with a law of composition that satisfies the following properties*

1. *Associativity: $(uv)w = u(vw)$, for every $u, v, w \in G$*
2. *The existence of an identity element 1 in the set G such that $u1 = 1u = u$, for every $u \in G$*

3. Existence of an inverse, noted u^{-1}, for every $u \in G$ such that $u^{-1}u = uu^{-1} = 1$

Furthermore, we call *group action* α of G on a set χ the mapping $\alpha : G \times \chi \to \chi$ satisfying:

1. Identity: $\alpha(1, x) = x$
2. Compatibility: $\alpha(v, \alpha(u, x)) = \alpha(vu, x)$

for every $x \in \chi$ and for every $u, v \in G$. When the context is clear we use the shorthand notation $\alpha(u, x) = ux$. In the context of image analysis, the set χ will be the image domain for the first layer of the network. In downstream layers, the set χ is the group upon which the G-CNN is built, *e.g* the union of the translation group with the rotation group SO(2).

In addition, if the group G is equipped with a measure μ and if f, g are real or complex valued functions on G, we can define the group correlation as

$$(f \star g)(u) = \int_G f(v)g(u^{-1}v)\mathrm{d}\mu(v). \tag{1.45}$$

This definition of correlation is compatible with the usual spatial correlation since the group of D-dimensional translations is isomorphic to $\mathbb{R}^D$ equipped with the addition as a law of composition. For simplicity and since one of the first works made in this domain was done on discrete groups (T. Cohen & Welling, 2016b), we will develop the idea for a discrete group G. Many generalizations to continuous compact groups exist, such as (Kondor & Trivedi, 2018). Therefore, Equation (1.45) becomes for a discrete group G:

$$(f \star h)(u) = \sum_{v \in G} f(v)h(u^{-1}v) \tag{1.46}$$

Note that $u, v \in G$ and the functions are defined over elements of the group.

In addition to group correlation layers, G-CNNs also contain a lifting layer. The lifting layer is the first in the network and links input images that are expressed on the spatial domain $\mathbb{Z}^D$ to feature maps expressed on the group G. For a discrete image $I[\boldsymbol{k}]$ and a discrete kernel $\kappa[\boldsymbol{k}]$, the lifting layer is computed as

$$f(u) = \sum_{\boldsymbol{k} \in \mathbb{Z}^D} I[\boldsymbol{k}]\kappa[u^{-1}\boldsymbol{k}]. \tag{1.47}$$

Note that the feature map is now expressed on elements of the group. We also observe that if the group G is the group of translations on $\mathbb{Z}^D$, Equation (1.47) becomes the standard discrete correlation, with the feature maps expressed on translations rather than positions which is equivalent in this case.

The attractive property of the group correlation is its behavior towards the transformation of the input image by a group element. We note the transformation of a discrete image $I[\boldsymbol{k}]$ by a group element $w \in G$ by[II]

$$L_w\{I[\boldsymbol{k}]\} = I[w^{-1}\boldsymbol{k}]. \tag{1.48}$$

By plugging this into Equation (1.47) we obtain and applying the substitution $\boldsymbol{k} \to w\boldsymbol{k}$:

$$\begin{aligned}\sum_{\boldsymbol{k}\in\mathbb{Z}^D} L_w\{I[\boldsymbol{k}]\}\kappa[u^{-1}\boldsymbol{k}] &= \sum_{\boldsymbol{k}\in\mathbb{Z}^D} I[w^{-1}\boldsymbol{k}]\kappa[u^{-1}\boldsymbol{k}]\\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^D} I[\boldsymbol{k}]\kappa[u^{-1}w\boldsymbol{k}]\\ &= f(w^{-1}u) = L_w\{f(u)\}.\end{aligned}$$

Therefore, the transformation of the input image is propagated to the feature map, which ensures that the layer is equivariant to any element of the group G. The same mathematical development used on the group correlation layer leads to the same conclusions. Hence, a sequence of one lifting layer and many group correlation layers yields a fully equivariant network with respect to the group considered.

In practice, to benefit from the efficiency of the spatial correlation, the group considered is separated into the translation group and the remaining group. By computing feature maps in the spatial domain and recombining them according to the group correlation structure, it is possible to implement efficient G-CNNs.

As an example, we will consider one of the groups of symmetries used in (T. Cohen & Welling, 2016b) for 2D images, which is the group of discrete translation and right-angle rotations referred to as the *p4* group. This group can be divided in the group of translations by the amount $\boldsymbol{k} \in \mathbb{Z}^2$ and the group of rotations by the angles $\theta_0 = 0, \theta_1 = \frac{\pi}{2}, \theta_2 = \pi, \theta_3 = \frac{3\pi}{2}$. In this case, the lifting layer is implemented as

$$f^{(1)}(\boldsymbol{k}_0, \theta_i) = \sum_{\boldsymbol{k}\in\mathbb{Z}^2} I[\boldsymbol{k}]\kappa^(1)[\mathrm{R}_{\theta_i}(\boldsymbol{k}_0 - \boldsymbol{k})], \tag{1.49}$$

and we observe that now the feature maps are indexed on the positions and the orientations; this is the practical way of representing the feature maps on the group *p4*. We also observe that this operation is the same as convolving the image with a filterbank containing the same filter at different orientations.

Once lifted, the feature maps can be filtered using the group correlation:

[II] Note that the operator $L_w\{\cdot\}$ is equivalent to the operator $\mathcal{T}\{\cdot\}$ introduced in Section 1.1.5 when the transformation of the image is the action of a group element on the image domain.

$$f^(2)(\boldsymbol{k}_0,\theta_j) = \sum_{\boldsymbol{k}\in\mathbb{Z}^2}\sum_{i=0}^{3} f^(1)(\boldsymbol{k},\theta_i)\kappa^(2)[\mathrm{R}_{\theta_j-\theta_i}(\boldsymbol{k}_0-\boldsymbol{k})] \tag{1.50}$$

In this section, we illustrated the basics of group correlation. However, this framework is very general and can be extended to different contexts and groups. For instance, in (Weiler et al., 2017), they used 2D steerable representations of the kernels to obtain a finer representation of the rotation group. In (Kondor & Trivedi, 2018), the authors developed the theory for any compact groups. In (Lafarge et al., 2019) a $SE(2)$ equivariant CNN is applied to several tasks in histopathology. In (Kondor et al., 2018) and (T. Cohen et al., 2018) the group equivariant convolution is used to define equivariant CNN on the 2D sphere. In (T. Cohen et al., 2019), the authors extend this concept to local gauge transformations. In (Weiler et al., 2018), they used the SHs to design 3D CNNs equivariant to rotations in $SO(3)$.

LRI Layers

In order to design CNNs robust against rotations of the input image, we propose to use the same LRI operators introduced in Section 1.1.9. Few modifications are made to embed them into a convolutional layer. The main one concerns filter parametrization.

For 2D images, the LRI layer convolves, according to Equation (1.44), the input image or feature map with a filterbank of kernels of the form $\kappa^i_{n,o}(r,\theta) = h^i_{n,o}(r)e^{-\mathrm{j}n\theta}$, where $n = 0,\dots,N$ with N the maximal degree of the CH expansion, i, o are indices running through $[1,\dots,C_{\mathrm{in}}]$ and $[1,\dots,C_{\mathrm{out}}]$, respectively, representing the indices of the input and output channels. The radial profiles $h^i_{n,o}$ are expressed as a linear combination of radially symmetric functions:

$$h^i_{n,o}(r) = \sum_{j=0}^{J} w^{i,j}_{n,o}\psi_j(r) \tag{1.51}$$

with $\psi_j(r) = 1 - r$ if $r > 0$ and $\psi_j(r) = r + 1$ if $r \leq 0$.

Once convolved, the features maps are recombined according to the formulae of Section 1.1.9 in order to obtain the $\mathcal{G}^{\mathrm{steer}}$, $\mathcal{G}^{\mathrm{sp}}$, or the $\mathcal{G}^{\mathrm{bisp}}$ layers. This layer is then used in a CNN and the parameters $w^{i,j}_{n,o}$ of Equation (1.51) are learned from the data via gradient descent.

3D LRI CNNs are designed similarly, but instead of a filterbank of modulated CHs, SHs are used.

The two approaches were developed, implemented, and evaluated in the context of

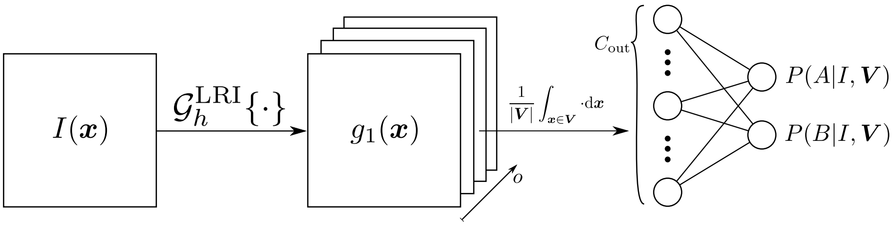

$I(\boldsymbol{x})$
$\mathcal{G}_h^{\mathrm{LRI}}\{\cdot\}$
$g_1(\boldsymbol{x})$
o
$\frac{1}{|\boldsymbol{V}|}\int_{\boldsymbol{x}\in\boldsymbol{V}}\cdot\mathrm{d}\boldsymbol{x}$
C_{out}
$P(A|I,\boldsymbol{V})$
$P(B|I,\boldsymbol{V})$

Figure 1.10: Representation of the architecture used in Chapter 2 and in (Andrearczyk, Fageot, et al., 2019). The features maps are computed by the application of a LRI layer resulting in C_{out} output channels. A Muli-Layer Perceptron (MLP) is then applied on the average aggregation of the feature maps within the pre-defined ROI $\boldsymbol{V}$. The MLP outputs the probability of ROI $\boldsymbol{V}$ of the image I belonging to class A or B.

this thesis, namely invariant and steering. In (Andrearczyk, Fageot, et al., 2019, 2020; Andrearczyk, Oreiller, Fageot, et al., 2019a), we implemented the $\mathcal{G}^{\mathrm{steer}}$ and $\mathcal{G}^{\mathrm{sp}}$ of Section 1.1.9 in a 3D LRI CNN with an architecture similar to the one illustrated in Figure 1.10. Furthermore, in Chapter 2, we implemented the $\mathcal{G}^{\mathrm{bisp}}$ with the same architecture. These LRI CNNs were evaluated on a classification task of benign versus malign lung nodules in CT images.

In Chapter 3, we implemented a 2D LRI CNN based on the U-Net architecture (Ronneberger et al., 2015), which is illustrated in Figure 1.11. In this LRI U-Net we used the $\mathcal{G}^{\mathrm{sp}}$ and $\mathcal{G}^{\mathrm{bisp}}$ layers. We evaluated these LRI U-Nets on a nuclei segmentation task of histopathological images obtained from the MoNuseg2018 (N. Kumar et al., 2019) challenge.

Compared to the G-CNN presented in Section 1.1.10, our LRI design discards the local information about the orientation of the patterns, while the G-CNNs conserve this information through equivariance. Since we are interested in biomedical images with textures composed of local patterns whose orientation is not significant for the task at hand, as mentioned in Section 1.1.6, we think discarding this information is beneficial.

1.2 Validation of Biomedical Image Analysis Models: the Need for Large and High quality Data and Benchmarks

As introduced in Section 1.1.2, radiomics has become popular this last couple of year. However, its application to clinical routine remains challenging (Guiot et al., 2022). In this Section, we review the importance of high-quality data in biomedical image analysis, the reasons that hinder the translation of radiomics model to the clinic, and why high-quality benchmarks may be a good solution to this problem. Finally, we describe the motivations of the HEad and neCK tumOR segmentation and outcome prediction in PET/CT images (HECKTOR) challenge, which constitutes teh second main contribution of this thesis.

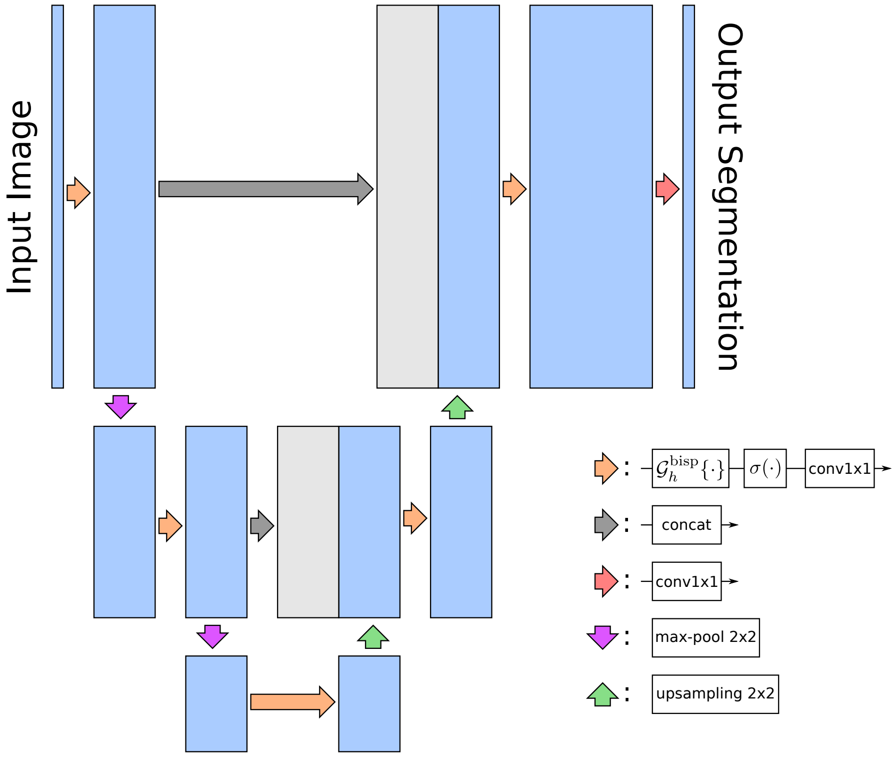


Figure 1.11: Representation of the U-Net architecture used in Chapter 3. The standard convolution is replaced by the $\mathcal{G}^{\text{bisp}}$ layer followed by a pixel-wise non-linearity $\sigma(\cdot)$ and a 1x1 convolution.

1.2.1 Need for Data in Biomedical Image Analysis

Deep learning algorithms and, more generally, machine learning methods have revolutionized many fields ranging from natural language processing to computer vision. The efficiency of these approaches is directly linked to the quality and quantity of data. For instance, in computer vision, many advances were made possible with the broad availability of large annotated datasets such as ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), CIFAR10 (Krizhevsky, Hinton, et al., 2009), and many others.

Consequently, scientists are encouraged to openly publish their data and the software used to generate and process them, which is embodied in the Open Science movement (Vicente-Saez & Martinez-Fuentes, 2018). This movement aims for more reproducible science and, at the same time, to promote discoveries. The trend is pronounced with the creation of peer-reviewed journals which value data description and sharing, such as Scientific Data published by Nature. In order to promote data sharing, the Findability, Accessibility, Interoperability, and Reuse of digital assets (FAIR) principles (Wilkinson et al., 2016) were established. These principles define a score to evaluate the reusability of data used in research to enforce data sharing and collaboration. There is also an emergence of platforms like Zenodo[III], which seek to implement the concepts of Open Science by allowing researchers to publish their data for free. Zenodo was developed in the context of the OpenAIRE[IV] consortium, a project of the Horizon 2020 funding program of the European Union.

Biomedical image analysis and radiomics research also require a large number of high-quality datasets, since they rely more and more on ML algorithms. However, the curation of large annotated biomedical datasets is hindered by different factors, such as the cost of labeling and collecting the data since it requires high expert knowledge. Moreover, imaging and clinical data are often sensitive data and cannot be shared without taking precautions such as patient consent and the approval of an ethical committee. Despite these difficulties, there are many research groups that invest much effort in sharing their data, which is also simplified with the implementation of platforms such as The Cancer Imaging Archive (TCIA) (Clark et al., 2013). TCIA was launched in 2010 and funded by the National Institute of Health (NIH). It is an open-access database of medical images that includes 109 different collections with more than $32'000$ images as of July 2022. These collections are openly accessible along with the publications describing them and often contain additional clinical information.

Although many resources to access biomedical images are available, it remains difficult for researchers not directly involved in the curation process to understand and use the data in a meaningful way. Furthermore, researchers can use the same images but for different purposes and measure the performances of their approaches with different metrics. All

[III] https://zenodo.org/ as of November 2022

[IV] https://www.openaire.eu as of November 2022 these degrees of freedom make direct comparisons between published work challenging.

In addition, and despite the Open Science movement, there is still a fair amount of research conducted on private data and a lack of good reporting guidelines to assess the studies' quality. To promote best reporting practices, the Radiomics Quality Score (RQS) (Lambin, Leijenaar, Deist, Peerlings, de Jong, et al., 2017) was published in 2017. The RQS is determined by 16 critical criteria and assigned to them points of value with a maximum of 36 points. These criteria include image acquisition protocols, statistical data processing, cohort origins, open science guidelines, and all the relevant aspects that a high-quality radiomics study must report. Even though the community well received the RQS, scientists are still reluctant to use it in their research (Guiot et al., 2022). This lack of good reporting practice, together with a lack of data sharing, delays the development and clinical impact of radiomics.

An exciting approach to tackling these problems is the organization of scientific challenges, as they offer an excellent opportunity to gather researchers from around the world to test their methods on the same clinical question, data, validation methodology, and performance metrics.

1.2.2 Challenges and Benchmarks

Successful challenges promote competition and drive innovation. For instance, the well-known AlexNet(Krizhevsky et al., 2012) network was implemented in the context of the ImageNet(Deng et al., 2009) competitions. The U-Net architecture (Ronneberger et al., 2015) was also developed and evaluated thanks to a segmentation challenge ("Segmentation of neuronal structures in EM stacks challenge - ISBI 2012", 2012) that was started at the conference ISBI 2012. More recently, the very successful nnU-Net framework (Isensee et al., 2021) was conceived in the context of the Medical Segmentation Decathlon (Antonelli et al., 2022).

The goal of a scientific challenge is to provide a control setup to evaluate methods for a given clinical question. The two main ingredients of a challenge are a training dataset shared with the participants and an evaluation platform where the proposed methods are evaluated on a held-out test set. The methods submitted on the platform are then ranked according to one or multiple metrics. In practice, challenges are organized as any competition and span weeks to several months between the release of the training data to the final evaluation when the winner is declared. After the end of the competition, it is customary to let a leaderboard open for new submissions to be made and evaluated. Nowadays, many platforms exist to organize challenges, such as Grand Challenge, AIcrowd, and Kaggle.

There is a long history of biomedical challenges, and some, like the Brain Tumor Segmentation (BRATS) challenge (Menze et al., 2014) has been running since 2012. For the first edition, the BRATS challenge contained 120 multi-modal MR images and has been updated yearly to reach 1470 multi-modal MR in 2022. This challenge is widely recognized, serves as a benchmark for brain tumor segmentation algorithms, and allows the development of reliable models for tumor segmentation.

Success stories like BRATS have inspired many researchers to organize challenges, and the number of organized challenges has increased continuously in recent years. For instance, the total number of accepted challenges at the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference was 15, 22, 25, and 38 for the years 2018, 2019, 2020, and 2021 respectively. In response to this expanding number of challenges and to raise the quality of proposed challenges, the Biomedical Image Analysis ChallengeS (BIAS) initiative (Maier-Hein et al., 2020) was created. This initiative is a set of recommended guidelines to report challenge results ranging from the definition of the cohort to the metrics employed to rank the different competing teams. These guidelines aim to improve the overall quality of challenges and ensure a better generalization of the conclusions drawn from the challenge results. Several biomedical image analysis conferences, such as MICCAI, the International Symposium on Biomedical Imaging (ISBI), and Medical Imaging with Deep Learning (MIDL), which have challenges organized as satellite events, promoted and used these guidelines in their review process for accepting or rejecting challenge propositions.

Currently, the task which is most commonly proposed in biomedical image challenges is segmentation. It is the most represented task on the Grand Challenge platform[V] with up to 40% of the challenges including this task among 309 challenges. The second most popular task is classification, with 23 % of the challenges proposing this task. This trend was also reported in a review by (Maier-Hein et al., 2018) where among the 150 biomedical challenges analyzed, 70% of the total number of tasks was segmentation.

One of the reasons that may explain the overrepresentation of segmentation challenges is that it addresses a critical clinical need. Indeed, downstream tasks like radiotherapy planning, cancer staging, and treatment response evaluation require tumor delineation, and automatizing this process can save clinicians' work hours. Furthermore, segmentation seems well adapted for CNN-based methods since fewer data points than a typical classification problem are usually required to achieve satisfying performance. As a comparison, the dataset used for image classification in the ImageNet competition contains 14 million images. On the other hand, biomedical segmentation challenges usually include hundreds of images. The differences in the number of samples required for both tasks may be explained by the fact that the learning problem is more constrained in segmentation tasks. For image classification, there is only one label per image; all the images' pixels are annotated for image segmentation. Thus, one delineated image contains more information than a classification label.

[V] https://grand-challenge.org/ as of November 2022

Even though lots of challenges in image segmentation have been proposed in the past, organizing segmentation challenges on unaddressed image modalities and types of tissue is still relevant since it allows for building extensive benchmarks with various tasks such as the one proposed in the context of the Medical Segmentation Decathlon (Antonelli et al., 2022). This challenge regrouped ten different segmentation tasks and, as mentioned above, allowed the implementation and evaluation of the well-known nnU-Net (Isensee et al., 2021). This kind of benchmark is beneficial for establishing a general framework that can be easily deployed on new tasks with few adjustments.

In the following, we describe the interest in organizing a challenge in Head and Neck (H&N) tumor segmentation for large-scale radiomics validation. One of the HECKTOR challenge's end goals is to evaluate automatic methods for prognosis in H&N cancer. Our vision for HECKTOR is to break this complex problem into smaller parts. Since obtaining high-quality contoured datasets is very expensive, we first propose to solve the segmentation of malignant tissues with a relatively small training dataset. Then, it will allow for evaluating prognosis methods on larger unannotated datasets.

1.2.3 The HECKTOR Challenge

Head and Neck (H&N) cancer is the seventh most incident cancer, with more than 930,000 new cases and 450,000 deaths worldwide, which amounts to 5% of all new cancer cases and 5% of all cancer deaths (Sung et al., 2021). 60% of presenting patients have locally advanced cancer and have a higher risk of local recurrence (15 to 40%) and a poor prognosis with a 5-year overall survival 50% (Braakhuis et al., 2012; Chow, 2020). Hence, improving cancer characterization with the help of radiomics can allow identifying patients requiring more aggressive treatments.

Fluorodeoxyglucose-PET/CT (FDG-PET/CT) is a modality of choice in H&N cancer staging, treatment planning, and assessing treatment response (Al-Ibraheem et al., 2009). At the time of the first edition of the challenge, several studies had shown the potential of radiomics on PET/CT images of H&N cancers for recurrence prediction and the prognosis value of radiomics (Bogowicz et al., 2017; Vallieres et al., 2017). However, there was still a lack of validation on huge datasets. Furthermore, radiomics studies are often performed on datasets contoured for radiotherapy planning which can be sub-optimal for radiomics modeling (Fontaine et al., 2022b).

HECKTOR aims to create a benchmark to address tumor and lymph node segmentation in FDG-PET/CT images of H&N cancers and build a robust benchmark to develop prognostic radiomics models in this context.

When we launched the HECKTOR challenge in 2020, it was the first challenge in PET/CT tumor segmentation. One prior challenge on PET tumor segmentation was proposed (Hatt, Laurent, Ouahabi, Fayad, Tan, Li, Lu, Jaouen, Tauber, Czakon, et al., 2018a). This challenge contained 19 training 3D FDG-PET examples and 157 testing images, including simulated, phantom-based, and clinical images. In total, the challenge included 25 clinical images and 151 simulated and phantom-based images. A key advantage of using simulated and phantom-based images is that it allows for perfect ground-truth definition. However, these images are often too simplistic and may not reflect the variability observed in clinical images.

In the HECKTOR challenge, we took a different approach and used only clinical images contoured by expert radiologists. This choice was motivated by one of the main goals: automatic tumor segmentation on PET/CT image fusion for further radiomics analysis.

Two editions of the challenge have already taken place at the MICCAI conference, and the third one is currently running. The dataset used in HECKTOR is multicentric. We focus on PET/CT images of oropharyngeal cancer to reduce anatomical variability and offer a more controlled setting for implementing segmentation algorithms.

Each edition of the challenge had an increasing number of data. The first edition included 201 training cases and 53 testing cases. The current edition of the challenge includes 524 training images and 362 testing images. At each iteration of the challenge, we also introduced new tasks. In HECKTOR 2020, primary tumor (GTVt) segmentation was the only task. In HECKTOR 2021, we added the prediction of Progression Free Survival (PFS). Finally, in HECKTOR 2022, we added the segmentation of malignant lymph nodes (GTVn) and updated the survival task to model Recurrence Free Survival (RFS). The main difference between the PFS and RFS outcomes is that for RFS, we only include patients with a complete response determined at the end of the treatment. In the first two editions of the challenge, the PET/CT images were provided with bounding boxes of 144 × 144 × 144 mm^3 defining a volume containing the oropharyngeal region. We made this choice to help participants. In the last edition, since tumor segmentation performance reached a plateau and to simulate conditions closer to the clinical data, we provided only the complete images.

In each edition, the participants were asked to provide a paper describing their method to be eligible for the official ranking. We organized reviews of these papers, and at least two experts reviewed each manuscript. The peer-reviewed manuscript as well as a BIAS-compliant overview of the challenge organization was published in proceedings (Andrearczyk, Oreiller, & Depeursinge, 2021). Table 1.3 summarizes the amount of data and tasks of the three editions of the HECKTOR challenge.

All the details concerning the data, the metrics used as well as challenge design can be found for HECKTOR 2020 in BIAS compliant format in (Andrearczyk, Oreiller, Vallières, Jreige, et al., 2021). Additional post-challenge analyses and discussion about the limitations of the challenge are described in Chapter 4. The BIAS compliant reporting of HECKTOR 2021 is exposed in Chapter 5. As HECKTOR 2022 is not finalized yet,

Table 1.3: Summary of the three editions of the HECKTOR challenge in terms of the amount of data, number of centers, and tasks proposed. The number of submitted papers for each team is also mentioned. The acronym BB stands for "Bounding Box", meaning we provided a fixed-size bounding box around the oropharynx region for the first two editions. The FI acronym stands for "Full Image", meaning we provided the entire image for this challenge edition.

		HECKTOR 2020	HECKTOR 2021	HECKTOR 2022
Data	# Training subjects	201	224	524
	# Test subjects	53	101	362
	# centers	5	6	9
	Inputs	BB	BB	FI
	Clinical data	✓	✓	✓
Tasks	GTVt segmentation	✓	✓	✓
	Outcome prediction	✗	PFS	RFS
	GTVn segmentation	✗	✗	✓
	Participant papers	10	31	?

the overview paper will be published by the end of 2022.

1.3 Thesis Contributions

This section describes my personal contributions to the material presented in Chapter 2, 3, 4 and 5 as well as other contributions.

1.3.1 Material presented in this manuscript

The manuscript presented in Chapter 2 is available on arXiv[VI] and was not published in a peer-reviewed journal for the reasons discussed in Section 6.1. In this work, I implemented and evaluated the bispectral CNN discussed in Section 1.1.10. Related to this work, I also had an active role in the conceptualization of the work of (Andrearczyk & Depeursinge, 2018), (Andrearczyk, Fageot, et al., 2019), and (Andrearczyk, Oreiller, Fageot, et al., 2019a). In (Andrearczyk & Depeursinge, 2018), the 3D operator $\mathcal{G}^{\text{steer}}$ was embedded in CNN, this paper won the best paper award at MIDL 2019. In (Andrearczyk, Oreiller, Fageot, et al., 2019a), the spectral operator $\mathcal{G}^{\text{sp}}$ was implemented in a 3D CNN. (Andrearczyk, Fageot, et al., 2019) comprehensively analyzes different LRI CNNs.

The article presented in Chapter 3 was accepted in the MIDL conference for a poster session. In this work, I implemented and evaluated the 2D LRI U-Net described in Section 1.1.10.

I had a leading role in the conceptualization, data curation, formal analysis, methodology,

[VI] https://arxiv.org/ as of November 2022 software, and writing of both of the works presented in Chapter 4 and Chapter 5. The work presented in Chapter 4 describes the HECKTOR 2020 challenge and additional post-challenge analyses and was published in Medical Image Analysis. The work presented in Chapter 5 is the overview paper of HECKTOR 2021 and was published in Lecture Notes in Computer Science (LNCS) (Andrearczyk et al., 2022). We published the participants' paper in LNCS proceedings (Andrearczyk, Oreiller, & Depeursinge, 2021) for both editions of the challenge. I had a leading role in the edition of these proceedings ranging from the organization of the reviews to reviewing the participants' papers.

In the context of the HECKTOR challenge, I also had a role in the conceptualization, data analysis, and writing of the following papers (Andrearczyk, Fontaine, et al., 2021; Andrearczyk, Oreiller, & Depeursinge, 2020; Andrearczyk, Oreiller, Vallières, Castelli, et al., 2020; Fontaine et al., 2022b; Fontaine et al., 2021)

1.3.2 Other contributions

During my thesis, I also contributed to other projects.

I had an active role in the implementation of the QuantImage platform[VII]. QuantImage is an open-source web-based platform allowing radiomics feature extraction and radiomics modeling for image classification or survival analysis. This platform aims to democratize radiomics and enable clinicians to test hypotheses in a radiomics framework quickly. My part was to develop a toolbox that implements an entire pipeline for image conversion (CT, MRI, PET) and preprocessing and feature extraction using in-house and external radiomics libraries. The code is available on my Github repository[VIII]. It can be used as a standalone command line interface. This toolbox is used as one of the main backend components of QuantImage. Furthermore, I also participated in elaborating the Quantimage platform and writing the manuscript (Abler et al., in preparation) describing the platform.

In (Jreige et al., 2020), I did the radiomics analysis and evaluation.

[VII] https://quantimage2.ehealth.hevs.ch/ as of November 2022
[VIII] https://github.com/voreille/okapy as of November 2022

1.3.3 Research Output

Peer-reviewed publications in international scientific journals

- **Valentin Oreiller,** Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, Andrei Iantsen, Mathieu Hatt, Yading Yuan, Jun Ma, Xiaoping Yang, Chinmay Rao, Suraj Pai, Kanchan Ghimire, Xue Feng, Mohamed A. Naser, Clifton D. Fuller, Fereshteh Yousefirizi, Arman Rahmim, Huai Chen, Lisheng Wang, John O. Prior and Adrien Depeursinge, Head and neck tumor segmentation in PET/CT: The HECKTOR challenge, in: Medical Image Analysis, 77:102336, 2022.
- Pierre Fontaine, Vincent Andrearczyk, **Valentin Oreiller**, Daniel Abler, Joel Castelli, Oscar Acosta, R. De Crevoisier, Martin Vallières, Mario Jreige, John O. Prior and Adrien Depeursinge, Cleaning radiotherapy contours for radiomics studies, is it worth it? A head and neck Cancer Study (2022), in: Clinical and Translational Radiation Oncology, 33:153-158, 2022.
- Vincent Andrearczyk, Julien Fageot, **Valentin Oreiller,** Xavier Montet and Adrien Depeursinge, Local rotation invariance in 3D CNNs, in: Medical Image Analysis, 65:101756,2020.

Peer-reviewed books/monographs

- Vincent Andrearczyk, **Valentin Oreiller**, Mathieu Hatt and Adrien Depeursinge, Head and neck tumor segmentation and outcome prediction, Springer LNCS, 1-328, 2022.
- Vincent Andrearczyk, **Valentin Oreiller** and Adrien Depeursinge, Head and neck tumor segmentation, Springer LNCS, 1-109, 2021.

Peer-reviewed conference papers

- **Valentin Oreiller**, Julien Fageot, Vincent Andrearczyk, John O. Prior and Adrien Depeursinge, Multi-organ nucleus segmentation using a locally rotation invariant bispectrum U-Net, in: Medical Imaging with Deep Learning, 2022.
- Pierre Fontaine, Vincent Andrearczyk, **Valentin Oreiller**, Joel Castelli, Mario Jreige, John O. Prior and Adrien Depeursinge, Fully automatic head and neck cancer prognosis prediction in PET/CT, in: Multimodal Learning for Clinical Decision Support, Springer LNCS, pages 59-68, 2021.
- Vincent Andrearczyk, Pierre Fontaine, **Valentin Oreiller**, Joel Castelli, Mario Jreige, John O. Prior and Adrien Depeursinge, Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer, in: 4th Workshop on PRedictive Intelligence in MEdicine, Springer LNCS, pages 147-156, 2021.
- Vincent Andrearczyk, **Valentin Oreiller**, Martin Vallières, Joel Castelli, Hesham Elhalawani, Mario Jreige, Sarah Boughdad, John O. Prior and Adrien Depeursinge, Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans, in: Medical Imaging with Deep Learning (MIDL), Montréal, Canada, 2020.
- Vincent Andrearczyk, Julien Fageot, **Valentin Oreiller**, Xavier Montet and Adrien Depeursinge, Exploring local rotation invariance in 3D CNNs with steerable filters, in: Medical Imaging with Deep Learning (MIDL), 15-26, 2019 (overall best paper award of MIDL 2019).
- Vincent Andrearczyk, **Valentin Oreiller**, Julien Fageot, Xavier Montet and Adrien Depeursinge, Solid Spherical Energy (SSE) CNNs for efficient 3D medical image analysis, in: Irish Machine Vision and Image Processing Conference, 37-44, 2019.

ArXiv paper

- **Valentin Oreiller**, Vincent Andrearczyk, Julien Fageot, John O. Prior, Adrien Depeursinge, 3D solid spherical bispectrum CNNs for biomedical texture analysis, arXiv:2004.13371, 2020.

Contributions to books

- Vincent Andrearczyk, **Valentin Oreiller**, Sarah Boughdad, Catherine Chez Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt and Adrien Depeursinge, Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images, Springer LNCS, 2022.
- Vincent Andrearczyk, **Valentin Oreiller**, Mario Jreige, Martin Vallières, 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, Springer LNCS, pages 1-21, 2021.

Oral contributions to international conferences

- Schaer, **Valentin Oreiller**, Daniel Abler, Himanshu Verma, Julien Reichenbach, Florian Evéquoz, Mario Jreige, John O. Prior and Adrien Depeursinge, QuantImage v2: A Clinician-in-the-loop Cloud Platform for Radiomics Research, in: European Congress of Radiology, 2022.
- Roger Schaer, **Valentin Oreiller**, Orfeas Aidonopoulos, John O. Prior and Adrien Depeursinge, QuantImage v2: an Open-Source and Web-Based Integrated Platform for Clinical Radiomics Research, in: Joint scientific session SSRMP/SGR-SSR, 2021.
- Vincent Andrearczyk, **Valentin Oreiller** and Adrien Depeursinge, Oropharynx Detection in PET-CT for Tumor Segmentation, in: Irish Machine Vision and Image Processing Conference, 2020.
- Sarah Boughdad, **Valentin Oreiller**, Marie Meyer, Niklaus Schaefer, Marie Nicod-Lalonde, Mario Jreige, Adrien Depeursinge and John Prior, Impact of a Gaussian filter applied to post-reconstruction PET images on radiomic features to predict complete pathological response in breast cancer, in: Journal of Nuclear Medicine, 61: supplement 1 (606-606), 2020.
- Sarah Boughdad, **Valentin Oreiller**, Marie Meyer, Niklaus Schaefer, Marie Nicod Lalonde, Mario Jreige, Adrien Depeursinge and John Prior, Impact of a Gaussian filter applied to post-reconstruction PET on radiomic features in assessing tumor heterogeneity in breast cancer, in: Journal of Nuclear Medicine, 61: supplement 1 (612-612), 2020.
- Mario Jreige, **Valentin Oreiller**, Igor Letovanec, Niklaus Schaefer, Adrien Depeursinge and John Prior, PET/CT Radiomics predict Pulmonary Lymphangitic Carcinomatosis (PLC) in Non-Small Cell Lung Cancer (NSCLC), in: Journal of Nuclear Medicine, 61: supplement 1 (1311-1311), 2020.
- **Valentin Oreiller**, Mario Jreige, John O. Prior and Adrien Depeursinge, PET/CT radiomics analysis contributes to detection of Pulmonary Lymphangitic Carcinomatosis (PLC) in Non-Small Cell Lung Cancer (NSCLC), in: Swiss Congress of Radiology, 2019.

2 3D Bispectral LRI CNN

The paper presented in this chapter is:

- Oreiller, Valentin, Vincent Andrearczyk, Julien Fageot, John O. Prior, and Adrien Depeursinge. "3D solid spherical bispectrum CNNs for biomedical texture analysis." *arXiv preprint arXiv*:2004.13371 (2020).

Abstract

Locally Rotation Invariant (LRI) operators have shown great potential in biomedical texture analysis where patterns appear at random positions and orientations. LRI operators can be obtained by computing the responses to the discrete rotation of local descriptors, such as Local Binary Patterns (LBP) or the Scale Invariant Feature Transform (SIFT). Other strategies achieve this invariance using Laplacian of Gaussian or steerable wavelets for instance, preventing the introduction of sampling errors during the discretization of the rotations. In this work, we obtain LRI operators via the local projection of the image on the spherical harmonics basis, followed by the computation of the bispectrum, which shares and extends the invariance properties of the spectrum. We investigate the benefits of using the bispectrum over the spectrum in the design of a LRI layer embedded in a shallow Convolutional Neural Network (CNN) for 3D image analysis. The performance of each design is evaluated on two datasets and compared against a standard 3D CNN. The first dataset is made of 3D volumes composed of synthetically generated rotated patterns, while the second contains malignant and benign pulmonary nodules in Computed Tomography (CT) images. The results indicate that bispectrum CNNs allows for a significantly better characterization of 3D textures than both the spectral and standard CNN. In addition, it can efficiently learn with fewer training examples and trainable parameters when compared to a standard convolutional layer.

2.1 Introduction

Convolutional Neural Networks (CNNs) have recently gained a lot of attention as they outperform classical handcrafted methods in almost every computer vision tasks where data scarcity is not an issue. In biomedical image analysis, data are abundant. However, obtaining high quality and consistently labeled images is expensive as data curation and annotation require hours of work from well-trained experts (Greenspan et al., 2016). Thus, the effective number of training examples is often low. This limitation is usually handled by transfer learning and data augmentation. Transfer learning, the process of fine-tuning a network trained on another task to the task at hand, is very common for 2D images. For 3D images, however, the lack of very large datasets hinders the availability of pre-trained models. Another approach, data augmentation, refers to the application of geometric transforms and perturbations to the training examples to make the CNN invariant to these distortions (Shorten & Khoshgoftaar, 2019). The cost of data augmentation is a substantial increase in the data size leading to a slower convergence rate and potential waste of trainable parameters.

A lot of recent research has focused on how to build CNNs that are invariant to these transforms by imposing constraints on the architecture of the network (Andrearczyk, Fageot, et al., 2020; T. Cohen & Welling, 2016a; Eickenberg et al., 2017; Weiler et al., 2017). The motivation of these approaches is to obviate the need to learn these invariances from the data and their transformation. As a result, an effective reduction of the number of trainable parameters is achieved and, potentially, a reduction of the number of training examples needed for the generalization of the network.

This work focuses on 3D biomedical texture analysis and on the design of CNNs that are invariant to local 3D rotations, *i.e.*, rotations of individual local patterns. This invariance is obtained using continuously defined Rotation Invariant (RI) descriptors of functions on the sphere. By relying on a continuous-domain formulation, we avoid the difficulties associated with rotations of discretized images (Ke & Li, 2014; Vivaldi, 2006). Neighborhoods defined by learned radial profiles are used to locally project the image on the solid sphere. These descriptors are used together with a convolution operation to obtain Locally Rotation Invariant (LRI)[I] operators in the 3D image domain as proposed in (Andrearczyk, Fageot, et al., 2019). These types of operators are relevant in biomedical texture analysis where discriminative patterns appear at random positions and orientations. The RI descriptors used in (Andrearczyk, Fageot, et al., 2019, 2020; Andrearczyk, Oreiller, Fageot, et al., 2019a; Eickenberg et al., 2017; Weiler et al., 2018) and in the present work are derived from the Spherical Harmonics (SH) decomposition of the kernels. The SHs are the generalization of the Circular Harmonics (CH) to the 2D sphere (Gallier, 2009). These two families of functions are intimately linked with Fourier theory, and both decompositions correspond to the Fourier transform of the function

[I] LRI is used for Locally Rotation Invariant and Local Rotation Invariance interchangeably defined on the sphere $\mathbb{S}^2$ for the SHs and on the circle $\mathbb{S}^1$ for the CHs.

To better apprehend the two invariants considered in this work, namely the spectrum and the bispectrum, it is useful to consider them on the circle. The CH expansion of a function $f \in L_2(\mathbb{S}^1)$ for a degree n is computed as $\widehat{f}_n = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-\mathrm{j}\theta n}\mathrm{d}\theta$, which is the Fourier series for 2π-periodic functions. For $m, n \in \mathbb{Z}$, the spectrum of the CH expansion is calculated as $s_n(f) = \widehat{f}_n\widehat{f}_n^* = |\widehat{f}_n|^2$ and the bispectrum as $b_{n,m}(f) = \widehat{f}_n\widehat{f}_m\widehat{f}_{n+m}^*$. One readily verifies that for a function $g(\theta) = f(\theta - \theta_0)$ we have for any $m, n \in \mathbb{Z}$ the equalities $s_n(f) = s_n(g)$ and $b_{n,m}(f) = b_{n,m}(g)$, since $\widehat{g}_n = \widehat{f}_n e^{-\mathrm{j}\theta_0 n}$. This means that the spectrum and bispectrum are RI, since a shift θ_0 in the parameter of f is equivalent to a rotation on the circle. The spectrum is the most simple, yet informative, Fourier-based RI quantity. However, it discards the phase between harmonics which contains all the information on how the sinusoids from the expansion add up to form edges and ridges (Smith et al., 1997, Chapter 10). The bispectrum, on the contrary, conserves the phase information (Kakarala & Mao, 2010) and constitutes a more specific pattern descriptor.

The main contributions of this paper are the introduction of a novel image operator based on the Solid Spherical Bispectrum (SSB) that is LRI and a corresponding CNN layer, resulting in a locally rotation invariant CNN. This work builds upon (Andrearczyk, Oreiller, Fageot, et al., 2019a), where a Solid Spherical Energy (SSE) layer was proposed. The radial profiles used to locally project the image on the solid sphere as well as the relative importance of the bispectrum coefficients can be learned end to end with backpropagation. We experimentally investigate the relevance of the proposed SSB layer for biomedical texture classification. Finally, we study the ability of the SSB-CNN to learn with small amounts of data and compare with a classical CNN.

This manuscript is organized as follows. In Section 2.2, we review the main related works. Sections 2.3.1 to 2.3.3 describe the nomenclature and the mathematical tools used in this work. The definitions of the spectrum and bispectrum for functions defined on the sphere are reported in Section 2.3.4 and are drawn from the work of Kakarala and Mao, 2010. We recall the theoretical benefits of the bispectrum over the spectrum in Section 2.3.5. In Section 2.3.6, we define the SSE and SSB image operators and state that they are LRI. In Section 2.3.7, we discuss the implementation details to integrate these image operators into a convolutional layer, referred to as the SSE or SSB layer. Sections 2.4 and 2.5 detail and discuss the experimental evaluation of the proposed approach. Conclusions and perspectives are provided in Section 2.6.

2.2 Related Work

2.2.1 Rotation Invariant Image Analysis

Combining LRI and directional sensitivity is not straightforward and is often antagonist in simple designs (Andrearczyk, Fageot, et al., 2020; Depeursinge et al., 2018). Several methods exist to combine both properties. Ojala, Pietikäinen, et al., 2002 proposed the Local Binary Patterns (LBP) where they compare values of pixels within a circular neighborhood to the middle pixel. Pixels of the neighborhood are thresholded based on the central pixel to generate a binary code. LRI is achieved by ordering the binary code to obtain the smallest binary number.

Several LRI filtering approaches were proposed. Varma and Zisserman, 2005a used a filter-bank including the same filters at different orientations, where LRI is achieved by max pooling over the orientations. Instead of explicitly computing responses (*i.e.* convolving) to oriented filters, steerable filters can be used to improve efficiency (Freeman & Adelson, 1991; Unser & Chenouard, 2013a). The work of Perona, 1992 shows the use of steerable filters for LRI edges and junctions analysis. Dicente Cid et al., 2017a used a filter-bank composed of steerable Riesz wavelets. LRI is obtained by locally aligning the filters to the direction maximizing the gradient of the image. Data-driven steerable filters were used in (Fageot et al., 2018) as LRI detectors of a given template within an image. Steerable Wavelet Machines (SWMs) were proposed in (Depeursinge, Püspöki, et al., 2017a), where task-specific profiles of steerable wavelets are learned using support vector machines.

Other approaches have been described to obtain invariants without explicitly rotating the descriptors. Such methods relies on moments (Flusser et al., 2009) or invariants built from the SH decomposition (Kakarala, 2012). Kakarala and Mao, 2010 introduced the bispectrum of the SH decomposition and they demonstrated the superiority of the bispectrum over the spectrum for 3D shape discrimination. Kakarala, 2012 showed that the bispectrum has better properties and contains more information than the spectrum, also proving its completeness for functions defined on compact groups. More recently, an extension of the spectral and bispectral invariants was used by Zucchelli et al., 2020 for the analysis of diffusion Magnetic Resonance Imaging data.

In (Depeursinge et al., 2018; Eickenberg et al., 2017), the authors used the spectrum of the SH expansion to compute LRI operators. Their work shares similarities with the method exposed here. However, our approach is more data-driven since we learn the radial profiles, whereas they rely on handcrafted ones.

2.2.2 Rotation Equivariance in CNNs

Recently, several research contributions focused on the explicit encoding of rotation equivariance into CNNs. One group of methods relies on the extension of the classic convolution on the group of translations to groups of symmetries including rotations and reflections. A detailed description of the generalization of the convolution to compact groups is given in (Kondor & Trivedi, 2018) and to homogeneous spaces in (T. S. Cohen et al., 2019). Regarding the application of this generalization, T. Cohen and Welling, 2016a used rotations of the filters together with recombinations of the response maps, which is performed according to the rules of group theory and allows equivariance to 2D right-angle rotations. The same strategy was extended to 3D images in (Winkels & Cohen, 2019; D. Worrall & Brostow, 2018). This 3D group CNN was applied to 3D texture classification in (Andrearczyk & Depeursinge, 2018). Bekkers et al., 2018 used the convolution on the discretized group of 2D roto-translations. Weiler et al., 2017 proposed a CH kernel representation to achieve a more efficient rotation of the filters via steerability, still in the context of the convolution on groups. T. Cohen and Welling, 2016c used the irreducible representation of the dihedral group to build CNNs that are equivariant to 2D discrete rotations.

The aforementioned methods offer the possibility to encode the equivariance to virtually any finite group. The 2D rotations group $SO(2)$ can be uniformly discretized by choosing a finite subgroup of $SO(2)$ with an arbitrary large number of elements. This is not anymore the case for 3D rotations since there is only 5 regular convex polyhedrons (Coxeter, 1961, Chapter 10). Therefore, approaches allowing for the propagation of the rotational equivariance without explicitly sampling the different orientations are crucial in 3D. Methods involving CH and SH have been introduced to address this problem. D. E. Worrall et al., 2016 used CHs representation of the kernels together with a complex convolution and complex non-linearities to achieve the rotational equivariance. The main drawback is that it generates many channels that must be disentangled to achieve rotation invariance. A SH representation of the kernels was used in (Weiler et al., 2018) to propagate the equivariance as a generalization of (D. E. Worrall et al., 2016) to 3D images. It is also possible to adapt neural networks to non-Euclidean domains, for instance, to the 2D sphere, where the invariance to rotations plays a crucial role as in (Kondor et al., 2018) and (T. Cohen et al., 2018). Finally, the group convolution can be extended to more general Lie groups as proposed by Bekkers, 2019, where CNNs equivariant to roto-translation and scale-translation were implemented.

Most of these methods focused on the propagation of the rotation equivariance throughout the network, whereas we propose lightweight networks discarding this information after each LRI layer, similarly to (Andrearczyk, Fageot, et al., 2020).

2.3 Methods

2.3.1 Notations and Terminology

We consider 3D images as functions $I \in L_2(\mathbb{R}^3)$, where the value $I(\boldsymbol{x}) \in \mathbb{R}$ corresponds to the gray level at location $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$. The set of 3D rotation matrices in the Cartesian space is denoted as $SO(3)$. The rotation of an image I is written as $I(\mathrm{R}\cdot)$, where $\mathrm{R} \in SO(3)$ is the corresponding rotation matrix.

The sphere is denoted as $\mathbb{S}^2 = \{\boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x}||_2 = 1\}$. Spherical coordinates are defined as (ρ, θ, ϕ) with radius $\rho \geq 0$, elevation angle $\theta \in [0, \pi]$, and horizontal plane angle $\phi \in [0, 2\pi)$. Functions defined on the sphere are written as $f \in L_2(\mathbb{S}^2)$ and are expressed in spherical coordinates. The inner product for $f, g \in L_2(\mathbb{S}^2)$ is defined by $\langle f\,, g\rangle_{L_2(\mathbb{S}^2)} = \int_0^{\pi} \int_0^{2\pi} f(\theta, \phi)\overline{g(\theta, \phi)} \sin(\theta)\mathrm{d}\phi\mathrm{d}\theta$. With a slight abuse of notation, the rotation of a function $f \in L_2(\mathbb{S}^2)$ is written as $f(\mathrm{R}\cdot)$, despite the fact that spherical functions are expressed in spherical coordinates.

The Kronecker delta $\delta[\cdot]$ is such that $\delta[n] = 1$ for $n = 0$ and $\delta[n] = 0$ otherwise. The Kronecker product is denoted by $\otimes$. The triangle function is referred to as $\mathrm{tri}(x)$ and is defined as $\mathrm{tri}(x) = 1 - |x|$ if $|x| < 1$ and $\mathrm{tri}(x) = 0$ otherwise. A block diagonal matrix formed by the sub-matrices A_i is written as $[\bigoplus_i \mathrm{A}_i]$. The Hermitian transpose is denoted by †.

2.3.2 LRI Operators

This work focuses on image operators $\mathcal{G}$ that are LRI as previously introduced in (Andrearczyk, Fageot, et al., 2020). An operator $\mathcal{G}$ is LRI if it satisfies the three following properties:

- *Locality*: there exists $\rho_0 > 0$ such that, for every $\boldsymbol{x} \in \mathbb{R}^3$ and every image $I \in L_2(\mathbb{R}^3)$, the quantity $\mathcal{G}\{I\}(\boldsymbol{x})$ only depends on local image values $I(\boldsymbol{y})$ for $\|\boldsymbol{y} - \boldsymbol{x}\| \leq \rho_0$.
- *Global equivariance to translations:* For any $I \in L_2(\mathbb{R}^3)$,
$$\mathcal{G}\{I(\cdot - \boldsymbol{x}_0)\} = \mathcal{G}\{I\}(\cdot - \boldsymbol{x}_0) \quad \text{for any } \boldsymbol{x}_0 \in \mathbb{R}^3.$$
- *Global equivariance to rotations:* For any $I \in L_2(\mathbb{R}^3)$,
$$\mathcal{G}\{I(\mathrm{R}_0\cdot)\} = \mathcal{G}\{I\}(\mathrm{R}_0\cdot) \quad \text{for any } \mathrm{R}_0 \in SO(3).$$

To reconcile the intuition of LRI with this definition, let us consider a simple scenario where two images I_1 and I_2 are composed of the same small template $\tau \in L_2(\mathbb{R}^3)$ appearing at random locations and orientations and distant enough to avoid overlaps between them.

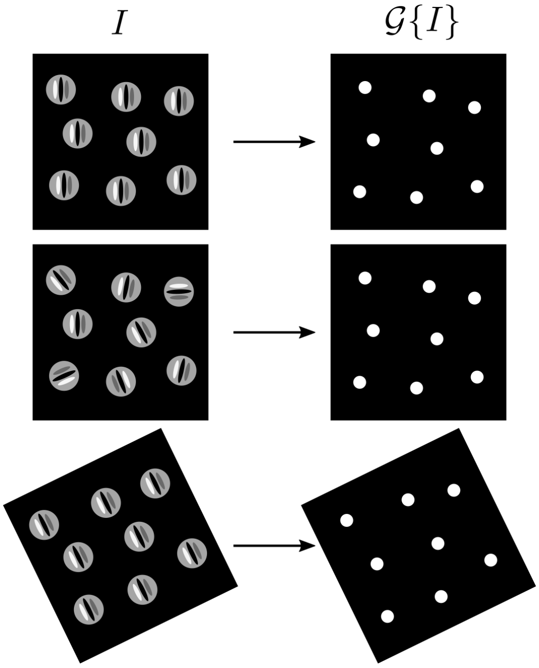


Figure 2.1: Visual representation of the output of a LRI operator (right) computed from an input image (left). For the sake of simplicity, only the output values at the template centers are represented. The top row shows an example an image composed of small directional patterns. The middle row shows the same image with local rotation applied to the patterns and the bottom rows illustrates the effect of a global rotation also highlighting the global rotation equivariance of a LRI operator.

The locations of the templates τ are identical for I_1 and I_2, the difference between the two images being in the local orientation of the templates. These images can be written as

$$I_k = \sum_{1 \leq j \leq J} \tau(\mathrm{R}_{j,k}(\cdot - \boldsymbol{x}_j)),$$

where J is the number of occurrence of the template τ and $k = 1, 2$. The local orientation and position of the j^{th} template in image k are represented by $\mathrm{R}_{j,k}$ and $\boldsymbol{x}_j$, respectively. If the operator $\mathcal{G}$ is LRI, then for any $1 \leq j \leq J$ and any rotations $\mathrm{R}_{j,1}$, $\mathrm{R}_{j,2} \in SO(3)$,

$$\mathcal{G}\{I_1\}(\boldsymbol{x}_j) = \mathcal{G}\{I_2\}(\boldsymbol{x}_j).$$

From the definition of LRI, this equality is required to hold only at the center of the templates. This example is illustrated in Fig. 2.1, where only the responses at the center of the templates are represented.

In this work, the design of LRI operators is obtained in two steps. First, the image $I \in L_2(\mathbb{R}^3)$ is convolved with SHs modulated by compactly supported radial profiles, referred to as solid SHs. The second step involves the computation of RI descriptors for each position.

2.3.3 Spherical Harmonics

Any function $f \in L_2(\mathbb{S}^2)$ can be expanded in the form of

$$f(\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} F_n^m Y_n^m(\theta,\phi), \tag{2.1}$$

where Y_n^m are the so-called SHs for a degree $n \in \mathbb{N}$ and order m with $-n \leq m \leq n$. For their formal definition, see (Depeursinge et al., 2018, Section 2.5) and for their visual representation, refer to Fig. 2.2. The SHs form an orthogonal basis of $L_2(\mathbb{S}^2)$ (Varshalovich et al., 1988, Chapter 5.6). Thus, the expansion coefficients of Eq. (2.1) can be computed by projecting f onto the SH basis using the inner product on the sphere

$$F_n^m = \langle f \,, Y_n^m \rangle_{L_2(\mathbb{S}^2)}. \tag{2.2}$$

This expansion corresponds to the Fourier transform on the sphere. We regroup the coefficients of all orders m for a given degree n as the $1 \times (2n+1)$ vector

$$\boldsymbol{\mathcal{F}}_n = [F_n^{-n} \dots F_n^0 \dots F_n^n], \tag{2.3}$$

called the spherical Fourier vector of degree n. One important property of SHs is their steerability, *i.e.* the rotation of one SH can be determined by a linear combination of the other SHs of same degree:

$$\mathrm{Y}_n^m(\mathrm{R}_0\cdot) = \sum_{m'=-n}^{n} [\mathrm{D}_n(\mathrm{R}_0)]_{m',m} Y_n^{m'}, \tag{2.4}$$

where $\mathrm{D}_n(\mathrm{R}_0)$ is a unitary matrix known as the Wigner D-matrix (Varshalovich et al., 1988, Chapter 4). Therefore, if two functions $f, f' \in L_2(\mathbb{S}^2)$ differ only by a rotation $\mathrm{R}_0 \in SO(3)$, *i.e.* $f' = f(\mathrm{R}_0\cdot)$, their spherical Fourier vectors, $\boldsymbol{\mathcal{F}}_n$ and $\boldsymbol{\mathcal{F}}'_n$, satisfy the following relation (Kakarala & Mao, 2010, Section 3, Eq. (5)):

$$\boldsymbol{\mathcal{F}}'_n = \boldsymbol{\mathcal{F}}_n \mathrm{D}_n(\mathrm{R}_0). \tag{2.5}$$

This property is similar to the shifting property of the Fourier transform on the real line. In the spherical case, instead of multiplying by a complex exponential, the transform is multiplied by the Wigner D-matrix of degree n associated with the rotation R_0.

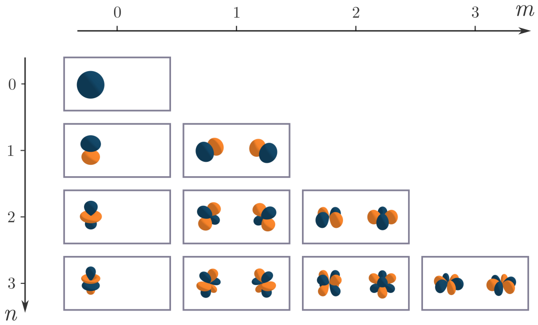


Figure 2.2: Visual representation of the real and imaginary part of the $h(r)Y_n^m(\theta,\phi)$ here h is chosen Gaussian for simplicity. Each box represents a given SH with the real part on the left and the imaginary part on the right. The blue represents positive values and orange negative values. Only the SHs for $m \geq 0$ are represented since we have the following symmetry $Y_n^{-m} = (-1)^m \overline{Y_n^m}$.

2.3.4 Spherical RI: the Spectrum and the Bispectrum

With the properties of the spherical Fourier vectors, it is possible to efficiently obtain RI operators for functions defined on the sphere. Two quantities computed from these coefficients will be of interest: the spherical spectrum and the spherical bispectrum.

Spectrum

The spectrum is a ubiquitous quantity in signal processing and it is well known to provide a source of translation invariant descriptors for periodic functions and functions defined on the real line. In these cases, the spectrum corresponds to the squared modulus of the Fourier transform. Its spherical equivalent, the spherical spectrum, is defined by the averaged squared norm of the spherical Fourier vector $\boldsymbol{\mathcal{F}}_n$:

$$s_n(f) = \frac{1}{2n+1}\boldsymbol{\mathcal{F}}_n\boldsymbol{\mathcal{F}}_n^\dagger = \frac{1}{2n+1}\sum_{m=-n}^{n}|F_n^m|^2. \tag{2.6}$$

Bispectrum

The bispectrum is defined as in (Kakarala & Mao, 2010, Section 4 Eq. (24)):

$$b_{n,n'}^{\ell}(f) = [\boldsymbol{\mathcal{F}}_n \otimes \boldsymbol{\mathcal{F}}_{n'}]\mathrm{C}_{nn'}\widetilde{\boldsymbol{\mathcal{F}}_\ell}^\dagger, \tag{2.7}$$

where the term $\boldsymbol{\mathcal{F}}_n \otimes \boldsymbol{\mathcal{F}}_{n'}$ is a $1\times(2n+1)(2n'+1)$ vector, $\mathrm{C}_{nn'}$ is the $(2n+1)(2n'+1)\times(2n+1)(2n'+1)$ Clebsh-Gordan matrix containing the Clebsh-Gordan coefficients, whose definition and main properties are recalled in Appendix A.1, and $\widetilde{\boldsymbol{\mathcal{F}}_\ell} = [0,\dots,0,\boldsymbol{\mathcal{F}}_\ell, 0,\dots,0]$ is a zero-padded vector of size $1\times(2n+1)(2n'+1)$ containing the spherical Fourier vector of degree ℓ with $|n - n'| \leq \ell \leq n + n'$. The zero-padding is performed to match the size of $\mathrm{C}_{nn'}$ and to select only the rows corresponding to the ℓ^{th} degree.

The spectrum and the bispectrum are known to be RI. We recall this fundamental result that will be crucial for us thereafter.

Proposition 2 *The spectrum and the bispectrum of spherical functions are RI. This means that, for any rotation* $\mathrm{R}_0 \in SO(3)$ *and any function* $f \in L_2(\mathbb{S}^2)$*, we have, for* $f' = f(\mathrm{R}_0 \cdot)$*,*

$$s_n(f) = s_n(f') \quad \textit{and} \quad b^{\ell}_{n,n'}(f) = b^{\ell}_{n,n'}(f') \tag{2.8}$$

for any $n, n' \geq 0$ *and any* $|n - n'| \leq \ell \leq n + n'$*.*

The result that the bispectrum of a spherical function is RI is given in (Kakarala & Mao, 2010, Theorem 4.1). Besides, we introduce the following notations:

$$\mathcal{S}\{\boldsymbol{\mathcal{F}}_n\} = s_n(f) \tag{2.9}$$

and

$$\mathcal{B}\{\boldsymbol{\mathcal{F}}_n, \boldsymbol{\mathcal{F}}_{n'}, \boldsymbol{\mathcal{F}}_l\} = b^{\ell}_{n,n'}(f). \tag{2.10}$$

These notations highlight that the spectrum coefficient $s_n(f)$ only depends on the nth-order Fourier vector $\boldsymbol{\mathcal{F}}_n$, and that the bispectrum coefficient $b^{\ell}_{n,n'}(f)$ only depends on $\boldsymbol{\mathcal{F}}_n$, $\boldsymbol{\mathcal{F}}_{n'}$, and $\boldsymbol{\mathcal{F}}_\ell$. Moreover, the rotation invariance of the spectrum and bispectrum can be reformulated as

$$\mathcal{S}\{\boldsymbol{\mathcal{F}}_n \mathrm{D}_n(R)\} = \mathcal{S}\{\boldsymbol{\mathcal{F}}_n\} \tag{2.11}$$

and

$$\mathcal{B}\{\boldsymbol{\mathcal{F}}_n \mathrm{D}_n(R), \boldsymbol{\mathcal{F}}_{n'} \mathrm{D}_{n'}(R), \boldsymbol{\mathcal{F}}_\ell \mathrm{D}_\ell(R)\} = \mathcal{B}\{\boldsymbol{\mathcal{F}}_n, \boldsymbol{\mathcal{F}}_{n'}, \boldsymbol{\mathcal{F}}_\ell\}. \tag{2.12}$$

2.3.5 Advantages of the Bispectrum over the Spectrum

Despite the simplicity to compute the spherical spectrum, it can be beneficial to use the bispectrum, since spherical functions with non-vanishing SH decomposition, i.e. functions whose SH decomposition is not 0 for any degree, are characterized by the bispectrum up to rotations: it is therefore a complete set of invariants (Kakarala, 2012). The two following examples illustrate cases where the spectrum is not sufficient to represent a spherical function.

Inter-Degree Rotations

The spectrum does not take into account the inter-degree rotation. For instance, let us build a function f' from the SH expansion $\boldsymbol{\mathcal{F}} = (\boldsymbol{\mathcal{F}}_0, \boldsymbol{\mathcal{F}}_1, \cdots)$ of the function f as follows: for each degree n, we apply a different Wigner D-matrix $\mathrm{D}_n(\mathrm{R}_n)$ with at least one rotation matrix R_n different from the others. The corresponding SH expansion $\boldsymbol{\mathcal{F}}' = (\boldsymbol{\mathcal{F}}_0\mathrm{D}_0(\mathrm{R}_0), \boldsymbol{\mathcal{F}}_1\mathrm{D}_1(\mathrm{R}_1), \cdots)$ will have the same spectrum since the Wigner D-matrices are unit matrices (*i.e.*, they do not impact the norm of $\boldsymbol{\mathcal{F}}_n$).

Intra-Degree Variations

Another aspect to which the spectrum is insensitive is in the distinction of intra-degree variations. for $n_0 \geq 1$ fixed, the functions $f = Y_{n_0}^m$ have the same spectrum $s_n(f) = \frac{\delta[n-n_0]}{2n_0+1}$ but are not rotation of each other in general (see Fig. 2.2).

On the contrary, the bispectrum is harder to fool (see Section 2.4.1), since, as mentioned earlier, it is complete for functions with non-vanishing SH decomposition. Meaning that functions on S^2 with the same bispectrum differ only by a rotation. Furthermore, the spectral information is contained in the bispectrum. This can be easily seen as:

$$b_{0,n}^n(f) = \boldsymbol{\mathcal{F}}_0\boldsymbol{\mathcal{F}}_n\boldsymbol{\mathcal{F}}_n^\dagger = F_0^0 s_n(f). \tag{2.13}$$

This illustrates that, given a non-zero mean $\boldsymbol{\mathcal{F}}_0 = F_0^0 \in \mathbb{R}$, we can retrieve the spectral information from the bispectrum. This can appear as a restriction for the bispectrum. However, in practice, it is possible to add a constant to the signal ensuring that F_0^0 is non-zero. The aforementioned properties make the bispectrum a more faithful descriptor and a good substitute of the spectral decomposition.

2.3.6 LRI on the Solid Sphere $\mathbb{R}^3$

The previous sections introduced the theoretical aspects to build RI descriptors for functions defined on the sphere. In this work, we are interested in 3D images, therefore we will use the spherical spectrum and bispectrum in combination with solid SHs. Solid SHs are the multiplication of the SHs by a radial profile to extend them to a 3D volume. We introduce the following notation for solid SHs evaluated on the Cartesian grid:

$$\kappa_n^m(\boldsymbol{x}) = \kappa_n^m(\rho, \theta, \phi) = h_n(\rho) Y_n^m(\theta, \phi), \tag{2.14}$$

where h_n is a compactly supported radial profile that is shared among the SHs Y_n^m with same degree n. In the final network, the radial profiles h_n are learned from the data.

The image is convolved with the solid SHs and by regrouping the resulting feature maps for each degree, we obtain the spherical Fourier feature maps[II]:

$$\boldsymbol{\mathcal{F}}_n(\boldsymbol{x}) = [(I * \kappa_n^m)(\boldsymbol{x})]_{m=-n}^{m=n}. \tag{2.15}$$

In other terms, the m^{th} component of $\boldsymbol{\mathcal{F}}_n(\boldsymbol{x})$ is $\langle I(\boldsymbol{x} - \cdot), h_n Y_n^m \rangle$, and measures the correlation between I centered at $\boldsymbol{x}$ and the solid SH $\kappa_n^m = h_n Y_n^m$. Thanks to the Fourier feature maps, we introduce the image operators used in this paper in Definition 4.

Definition 4 *We define the* SSE image operator $\mathcal{G}_n^{SSE}$ *of degree* $n \geq 0$ *as*

$$\mathcal{G}_n^{SSE}\{I\}(\boldsymbol{x}) = \mathcal{S}\{\boldsymbol{\mathcal{F}}_n(\boldsymbol{x})\} \tag{2.16}$$

for any $I \in L_2(\mathbb{R}^3)$ *and* $\boldsymbol{x} \in \mathbb{R}^3$. *Similarly, we define the* SSB image operator $\mathcal{G}_{n,n',\ell}^{SSB}$ *associated with degrees* $n, n' \geq 0$ *and* $|n - n'| \leq \ell \leq n + n'$ *as*

$$\mathcal{G}_{n,n',\ell}^{SSB}\{I\}(\boldsymbol{x}) = \mathcal{B}\{\boldsymbol{\mathcal{F}}_n(\boldsymbol{x}), \boldsymbol{\mathcal{F}}_{n'}(\boldsymbol{x}), \boldsymbol{\mathcal{F}}_\ell(\boldsymbol{x})\}, \tag{2.17}$$

for any $I \in L_2(\mathbb{R}^3)$ *and* $\boldsymbol{x} \in \mathbb{R}^3$.

The SSE image operators have been considered in (Andrearczyk, Fageot, et al., 2020), where it was proven to be LRI in Appendix D. We recall this result and extend it to SSB image operators in the following proposition, whose proof is given in Appendix A.2.

Proposition 3 *The SSE and SSB image operators are globally equivariant to translations and rotations. When the radial profiles* h_n *are all compactly supported, these operators are therefore LRI in the sense of Section 2.3.2.*

2.3.7 Implementation of the LRI layers

In this section, we report the implementation details of our LRI design.

[II] The convolution $(I * \kappa_n^m)(\boldsymbol{x})$ with all the κ_n^m is equivalent to a local projection of the image around the position $\boldsymbol{x}$ to a function defined on the sphere followed by a projection onto the SHs basis. For that reason, we use the same notation as for the spherical Fourier vector of degree n. We distinguish the spherical Fourier feature maps by the evaluation over a position $\boldsymbol{x}$.

Parameterization of the Radial Profiles

The radial profiles are parameterized as a linear combination of radial functions

$$h_{q,n}(\rho) = \sum_{j=0}^{J} w_{q,n,j}\psi_j(\rho), \tag{2.18}$$

where the $w_{q,n,j}$ are the trainable parameters of the model. In (2.18), $h_{q,n}$ is the q^{th} radial profile associated to the degree n. The index q controls the number of output streams[III] in the layer. The index $j = 0, \ldots, J$ controls the radial components of the filter. The radial functions are chosen as $\psi_j(\rho) = \text{tri}(\rho - j)$.

Number of Feature Maps

The image is convolved with the kernels $\kappa_{q,n}^{m}$ to obtain the spherical Fourier feature maps $\{\boldsymbol{\mathcal{F}}_{q,n}(\boldsymbol{x})\}_{n=0,\ldots,N}^{q=1,\ldots,Q}$. Here, Q is the number of output streams of the layer and N is the maximal degree of the SH decomposition. These feature maps are combined according to Eq. (2.16) and (2.17) resulting in $\mathcal{G}_{q,n}^{\text{SSE}}\{I\}(\boldsymbol{x})$ or $\mathcal{G}_{q,n,n',l}^{\text{SSB}}\{I\}(\boldsymbol{x})$ respectively. In the following, we discuss the number of feature maps generated for only one output stream, thus we drop the index q.

In the case of the operator $\mathcal{G}_{n}^{\text{SSE}}$, the number of generated feature maps is $N + 1$. For the $\mathcal{G}_{n,n',\ell}^{\text{SSB}}$ operator, the total number of features maps is $\mathcal{O}(N^3)$. It is actually not necessary to compute all the bispectrum coefficients, some of them being redundant due to the following properties. First, for each n, n' and ℓ, the bispectral components $b_{n,n'}^{\ell}(f)$ and $b_{n',n}^{\ell}(f)$ are proportional independently of f (Kakarala & Mao, 2010, Theorem 4.1). Hence, we choose to compute the components only for $n \leq n'$ and $0 \leq n + n' \leq N$. Second, even though the bispectrum is complex-valued, when f is real, $b_{n,n'}^{\ell}(f)$ is either purely real or purely imaginary if $n + n' + \ell$ is even or odd respectively (Kakarala et al., 2011, Theorem 2.2). Thus, we can map it to a real-valued scalar. In our design, we take either the real or the imaginary part depending on the value of the indices n, n', ℓ.

Even with these two properties the number of feature maps for the $\mathcal{G}_{n,n',\ell}^{\text{SSB}}$ operator still follows a polynomial of degree 3 (see Table 2.1 for the first values), but for low N it still reduces greatly this number. Moreover, the maximal degree N for the SH expansion cannot be taken arbitrarily large as the kernels are discretized (Andrearczyk, Fageot, et al., 2020). The upper bound for N is given by $N \leq \frac{\pi c}{4}$, where c is the diameter of the kernel. This condition can be regarded as the Nyquist frequency for the SH expansion. As an example, $N = 7$ is the maximal value for a kernel of size 9 × 9 × 9.

[III] In a standard CNN, the number of streams and feature maps coincide. In our case, the feature maps are projected on the SH basis then recombined to form LRI feature maps, thus for a particular radial profile, or stream, we have several feature maps.

Table 2.1: Number of feature maps obtained for the $\mathcal{G}_n^{\mathrm{SSE}}$ and $\mathcal{G}_{n,n',\ell}^{\mathrm{SSB}}$ operators in function of the maximal degree N.

N	0	1	2	4	6	8	10	100
Spectrum	1	2	3	5	7	9	11	101
Bispectrum	1	2	5	14	30	55	91	48127

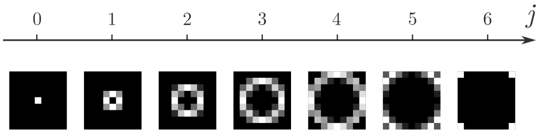


Figure 2.3: Illustration of the discretization of the radial profile ψ_j for a kernel size of 9 with $j = 0, \dots, 6$.

Discretization

The kernels $\kappa_{q,n}^m = h_{q,n} Y_n^m$ are discretized by evaluating them on a Cartesian grid. Fig. 2.3 illustrates the discretization of the radial functions $\psi(\rho)_j$. For more details about the discretization, see (Andrearczyk, Fageot, et al., 2019, Section 2.4).

Training

The learnable parameters of the LRI layer are the weights $w_{q,n,j}$ of the radial profiles $h_{q,n}(\rho)$ (Eq. (2.18)). These weights are learned via back-propagation. During the training, only an isotropic radial profile is learned in the LRI the layer. The directional information is contained in the many feature maps generated by the $\mathcal{G}_{n,n',\ell}^{\mathrm{SSB}}$ or $\mathcal{G}_n^{\mathrm{SSE}}$ operator. In order to obtain a DS network the LRI layer must be combined with a layer mixing the different feature maps. In this study, the LRI layer is directly linked to a fully connected layer.

2.4 Experiments and Results

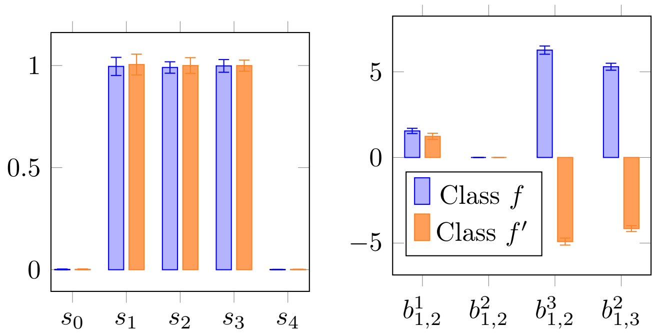
Section 2.4.1 illustrates the differences between the spherical spectrum and bispectrum with two toy experiments designed to fool the spectrum and to test if the proposed framework is applicable for functions defined on $\mathbb{R}^3$ even in the presence of noise. Then, we compare the classification performance of three different CNNs (SSE, SSB and standard) detailed in Section 2.4.3 on the two datasets described in Section 2.4.2 in terms of accuracy (Section 2.4.4) and generalization power (Section 2.4.5) *i.e.* the number of training samples required to reach the final accuracy.

2.4.1 Comparing Local Spectrum to Bispectrum Representations

Two toy experiments are conducted to highlight the differences in terms of the representation power of the spherical spectrum and bispectrum. The first experiment is designed to show that the spectrum is unable to discriminate between patterns that only differ in terms of rotations between degrees. The second experiment illustrates that the spectrum cannot capture differences within the same degree. These two experiments are done in the 3D image domain to show the applicability of the spherical spectrum and bispectrum of the solid SHs and to be as close as possible to the final application.

The images are obtained by evaluating $h(\rho)\sum_{n=0}^{N}\sum_{m=-n}^{m=n} F_n^m Y_m^n(\theta,\phi)$ on a 3D Cartesian grid of $32 \times 32 \times 32$ with h defined as an isotropic Simoncelli wavelet profile (Portilla & Simoncelli, 2000). This first experiment investigates the capability of the spectrum and bispectrum to discriminate between functions with distinct inter-degree rotations. Representatives f and f' of the two classes are obtained by summing the SHs described by their repsective spherical Fourier transform $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}}'$. $\boldsymbol{\mathcal{F}}$ is composed of $\boldsymbol{\mathcal{F}}_1 = [1, \mathrm{j}, 1]$, $\boldsymbol{\mathcal{F}}_2 = [1, -1, 1, 1, 1]$, $\boldsymbol{\mathcal{F}}_3 = [1, -1, 1, \mathrm{j}, 1, 1, 1]$ and $\boldsymbol{\mathcal{F}}_n = \mathbf{0}$ for any $n \neq 1, 2, 3$. The coefficients are chosen to ensure that the images are real and that the spherical spectrum $s_n(f) = 1$ for $n = 1, 2, 3$. The spherical decomposition $\boldsymbol{\mathcal{F}}'$ of the second class is computed as $\boldsymbol{\mathcal{F}}'_1 = \boldsymbol{\mathcal{F}}_1 \mathrm{D}_1(\mathrm{R}_1)$, $\boldsymbol{\mathcal{F}}'_2 = \boldsymbol{\mathcal{F}}_2 \mathrm{D}_2(\mathrm{R}_2)$ and $\boldsymbol{\mathcal{F}}'_3 = \boldsymbol{\mathcal{F}}_3 \mathrm{D}_3(\mathrm{R}_3)$, where R_1, R_2 and R_3 are distinct 3D rotations. This allows to combine the different degrees with different rotations resulting in a function f' with spectrum $s_n(f') = s_n(f)$ for all n but $f' \neq f$. Moreover, for each class, 50 distinct instances are created by adding Gaussian noise and randomly rotating the images. The random rotations are drawn from a uniform distribution over the 3D rotations and then we use the associated Wigner-D matrices to rotate the instances. This time, the same rotation is applied to all degrees. The bispectrum and spectrum are calculated using only the responses to the spherical filters at the origin voxel of the images and the results are presented in Fig. 2.4. Note that only a subset of distintive coefficients of the bispectrum is shown. The results indicate that the spectrum cannot detect inter-degree rotations, whereas the bispectrum can.

In the next experiment, instead of applying a distinct rotation to each degree, we choose orders that are different within the same degree. For the first class f, we use only the order $m = 0$ and for the second class f' the orders $m = n, -n$. This choice is motivated by their differences in shape as represented in Fig. 2.2. The spherical Fourier transform $\boldsymbol{\mathcal{F}}$ of the first class is chosen to be $\boldsymbol{\mathcal{F}}_1 = [0, \sqrt{3}\mathrm{j}, 0]$, $\boldsymbol{\mathcal{F}}_2 = [0, 0, \sqrt{5}, 0, 0]$, $\boldsymbol{\mathcal{F}}_3 = [0, 0, 0, \sqrt{7}\mathrm{j}, 0, 0, 0]$ and $\boldsymbol{\mathcal{F}}_n = 0$ for any $n \neq 1, 2, 3$. The spherical decomposition $\boldsymbol{\mathcal{F}}'$ of the second class is $\boldsymbol{\mathcal{F}}'_1 = [\sqrt{3/2}, 0, \sqrt{3/2}]$, $\boldsymbol{\mathcal{F}}'_2 = [\sqrt{5/2}, 0, 0, 0, \sqrt{5/2}]$, $\boldsymbol{\mathcal{F}}'_3 = [\sqrt{7/2}, 0, 0, 0, 0, 0, \sqrt{7/2}]$. The coefficients are chosen to obtain a spherical spectrum of 1 for $n = 1$, $n = 2$ and $n = 3$ and to generate real images. The results in Fig. 2.5 show that the bispectrum can discriminate between the two classes even though they have the same spectrum.



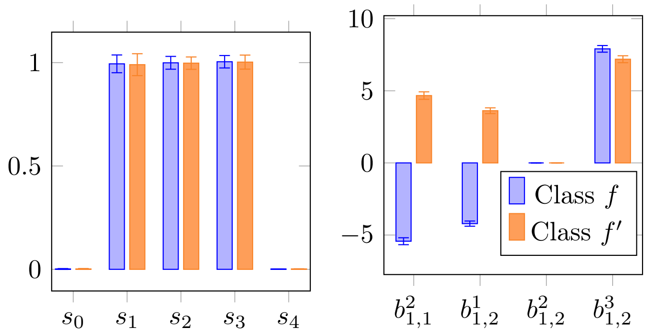
(a) Spectrum (b) Bispectrum

Figure 2.4: Experiment with distinct inter-degree rotations. Spherical spectral (left) and bispectral (right) decomposition of the two classes. The blue bars represent the decomposition for the first class f and the orange bars for the second class f' ($\boldsymbol{\mathcal{F}}'_i = \boldsymbol{\mathcal{F}}_i \mathrm{D}_i(R_i)$, $i = 1, 2, 3$). Note that only a subset of the bispectral components is displayed. It can be observed that the spectrum cannot distinguish between f and f', and that the bispectrum can.

2.4.2 Datasets

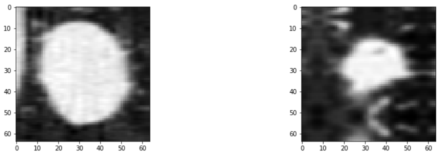
To evaluate the performance of the proposed LRI operators, we use both a synthetic and a medical dataset. The synthetic dataset constitutes a controlled experimental setup and contains two classes with 500 volumes each of size $32 \times 32 \times 32$ for each class. They are generated by placing two types of patterns with a size of $7 \times 7 \times 7$, namely a binary segment and a 2D cross with the same norm, at random 3D orientations and random locations with possible overlap. The number of patterns per volume is randomly set to $\lfloor d \cdot (\frac{s_v}{s_p})^3 \rfloor$, where s_v and s_p are the sizes of the volume and of the pattern, respectively and the density d is drawn from a uniform distribution in the range $[0.1, 0.5]$. The two classes vary by the proportion of the patterns, *i.e.* 30% segments with 70% crosses for the first class and vice versa for the second class. 800 volumes are used for training and the remaining 200 for testing. Despite the simplicity of this dataset, some variability is introduced by the overlapping patterns and the linear interpolation of the 3D rotations.

The second dataset is a subset of the American National Lung Screening Trial (NLST) that was annotated by radiologists at the University Hospitals of Geneva (HUG) (Andrearczyk, Fageot, et al., 2020). The dataset includes 485 pulmonary nodules from distinct patients in CT, among which 244 benign and 241 malignant. We pad or crop the input volumes (originally with sizes ranging from $16 \times 16 \times 16$ to $128 \times 128 \times 128$) to the size $64 \times 64 \times 64$. We use balanced training and test splits with 392 and 93 volumes respectively. Examples of 2D slices of the lung nodules are illustrated in Fig. 2.6. The Hounsfield units are clipped in the range $[-1000, 400]$, then normalized with zero mean and unit variance (using the training mean and variance).



(a) Spectrum (b) Bispectrum

Figure 2.5: Experiment with intra-degree variations. Spherical spectral (left) and bispectral (right) decomposition of the two classes. The blue bars represent the decomposition for the first class f and the orange bars for the second class f'. Note that only a subset of the bispectral components is displayed. It can be observed that the spectrum cannot distinguish between f and f', and that the bispectrum can.



(a) Benign (b) Malignant

Figure 2.6: Slices drawn from the NLST dataset showing a benign pulmonary nodule and a malignant one.

2.4.3 Network Architecture

This work uses the network architecture proposed in (Andrearczyk, Fageot, et al., 2019). The first layer consists of the LRI layer implemented as described in Section 2.3.7. The obtained responses are aggregated using spatial global average pooling, similarly to (Andrearczyk & Whelan, 2016). This pooling aggregates the LRI operator responses into a single scalar per feature map and is followed by one Fully Connected (FC) layer. For the nodule classification experiment, we average the responses inside the nodule masks instead of across the entire feature maps to remove the effect of the size allowing the network to focus on the textural content of the nodule. The final softmax FC layer is connected directly with a cross-entropy loss. Standard Rectified Linear Units (ReLU) activations are used. The two different types of LRI networks are referred to as SSE-CNN and SSB-CNN when the LRI layer uses the $\mathcal{G}^{\text{SSE}}$ or $\mathcal{G}^{\text{SSB}}$ operator respectively.

The networks are trained using an Adam optimizer with $\beta_1 = 0.99$ and $\beta_2 = 0.9999$ and a batch size of 8. Other task-specific parameters are: for the synthetic experiment (kernel size 7 × 7 × 7, stride 1, 2 filters and 50,000 iterations), for the nodule classification experiment (kernel size 9 × 9 × 9, stride 2, 4 filters and 10,000 iterations). The initial values of the trainable weights in (2.18) are drawn independently from a Gaussian distribution as $w_{q,n,j} \sim \mathcal{N}(0, 1)$ and the biases are initialized to zero. This initialization is inspired by (He et al., 2015; Weiler et al., 2017) in order to avoid vanishing and exploding activations and gradients.

We compare the proposed CNNs to a network with the same architecture but with a standard 3D convolutional layer and varying numbers of filters, referred to as Z3-CNN.

2.4.4 Classification Performance of the SSB-, SSE- and Z3-CNN

Here, we evaluate the classification performance of both the SSE-CNN and SSB-CNN on the two datasets described in Section 2.4.2. The accuracies of both designs are computed with 10 different initializations for varying maximal degrees N. Confidence Intervals (CI) at 95% and mean accuracies are reported in Fig. 2.7 and 2.8 for the synthetic and NLST datasets respectively. On both datasets, the SSB-CNN outperforms the two other networks. To exclude the possibility that this performance gain is simply due to a higher number of feature maps, we trained a SSE-CNN on the synthetic dataset with maximal degree $N = 2$ and 4 kernels in the first layer instead of 2. This amounts to a total of 12 feature maps after the first layer. This model achieves 0.9075 ± 0.006 of accuracy and is still significantly outperformed by the SSB-CNN with maximal degree 2 and 2 kernels, which has 10 feature maps after the first layer and obtains an accuracy of 0.924 ± 0.008 (Fig. 2.7). One important remark is that both LRI networks contain fewer parameters than the Z3-CNN. For instance in the NLST experiment, the SSB- and SSE-CNN have 330 and 222 parameters respectively for a maximal degree $N = 4$ against 7322 parameters for the Z3-CNN.

2.4.5 Learning Curves of the SSB-, SSE- and Z3-CNN

The SSB- and SSE-CNN are LRI networks and thus require neither additional training examples nor a large number of parameters to learn this property. In addition, they rely on compressing SH parametric representations. For these two reasons, we expect that they will learn with fewer training examples (*i.e.* steeper learning curve) than the standard Z3-CNN on data for which this property is relevant. To test this hypothesis, we compare the classification performance of each method using an increasingly large number of training examples N_s. For the synthetic dataset, we use $N_s = 16, 32, 64, 128, 200, 300, 400$ and for the nodule classification $N_s = 10, 30, 64, 128, 200, 300, 392$. For each value of N_s, 10 repetitions are made and N_s examples are randomly drawn from the same training fold as the previous experiments (Section 2.4.4). For the SSB-CNN we report the accuracy for $N = 2$ on the synthetic dataset and $N = 4$ on the NLST dataset. The accuracy of the SSE-CNN is reported for $N = 2$ on the synthetic dataset and $N = 1$ for the NLST dataset.

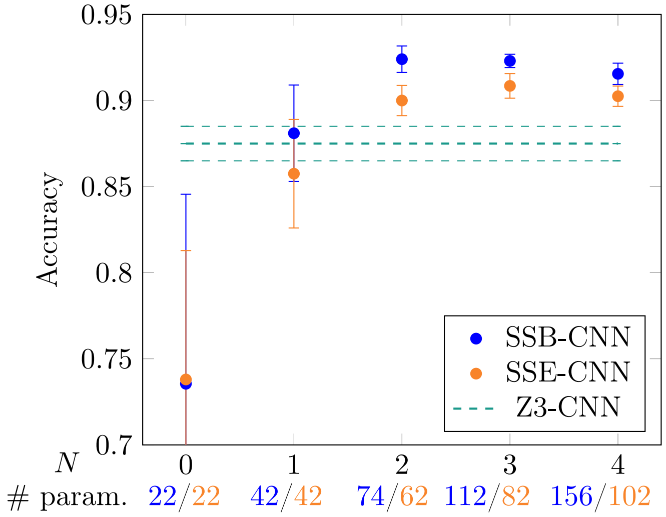


Figure 2.7: Classification accuracies and numbers of parameters on the synthetic dataset for varying maximal degrees N. The error bars represent the CIs at 95%. The accuracy of the Z3-CNN with 10 filters is 0.875 ± 0.011 with 3462 trainable parameters and is represented by the green dashed lines.

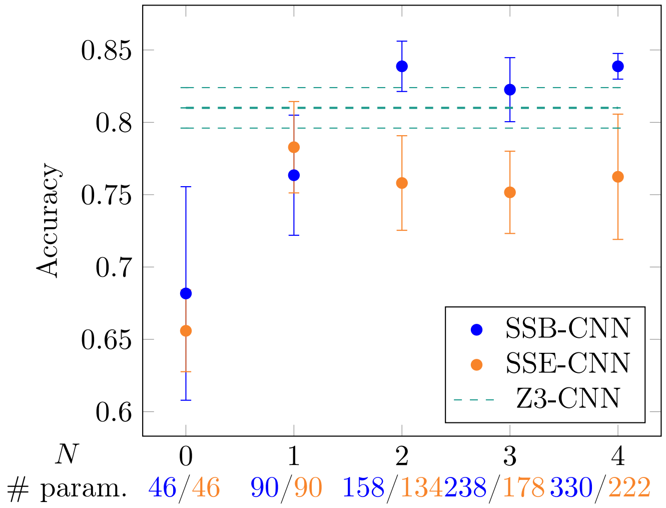


Figure 2.8: Classification accuracies and numbers of parameters on the NLST dataset for varying maximal degrees N. The error bars represent the CIs at 95%. The accuracy of the Z3-CNN with 10 filters is 0.810 ± 0.014 with 7322 trainable parameters and is represented by the green dashed lines.

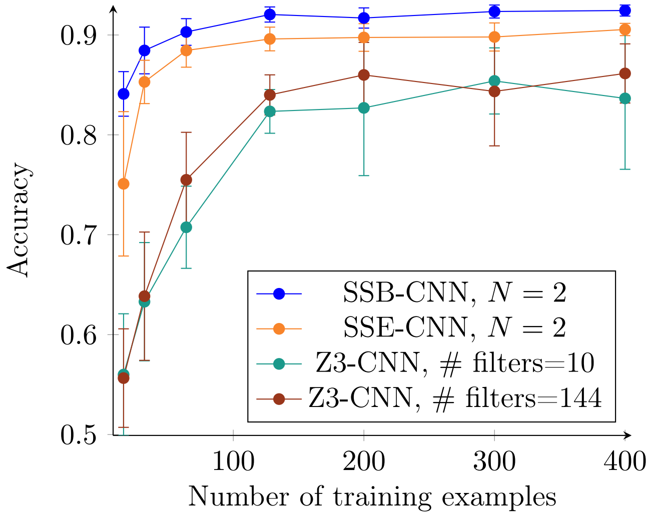


Figure 2.9: Performances on the synthetic dataset in terms of accuracy for a varying number of training examples. The error bars represent the CIs at 95%. The number of filters in the first layer for the SSB- and SSE-CNN is 2.

These parameters are chosen according to the previous experiments (Section 2.4.4, Fig. 2.7 and 2.8) as they provided the best accuracy. The experiment is also conducted with the Z3-CNN and the results are reported for both 10 and 144 filters in the convolution layer. The mean accuracy with CIs at 95% of the three models and on the two datasets is reported in Fig. 2.9 and 2.10.

2.5 Discussions

2.5.1 The Bispectrum is More Discriminative

The two experiments presented in Section 2.4.1 illustrate the types of pattern information that cannot be characterized by spectral components. In these settings, the spectrum is unable to distinguish between classes that differ either by a difference of orientation between degrees (inter-degree rotation) or by intra-degree variations. This is not the case for the bispectral coefficients that allow describing functions in $L_2(\mathbb{S}^2)$ more accurately. As expected, the cost of a more complete representation is a larger number of components. However, it is possible to compute only a subset of the bispectral components depending on the task. This larger number of components translates to a bigger number of feature maps in the CNN implementation this leads to a higher GPU RAM consumption. Furthermore, the computation of the LRI feature maps is more expensive. For instance, the training time of the SSB-, SSE- and Z3-CNN for 10000 iterations on 436 training examples 2.4.5 are 44, 17 and 7 minutes respectively[IV]. However, In the CNN implementation of these

[IV] Those numbers were obtained with a Nvidia GeForce GTX TITAN X.

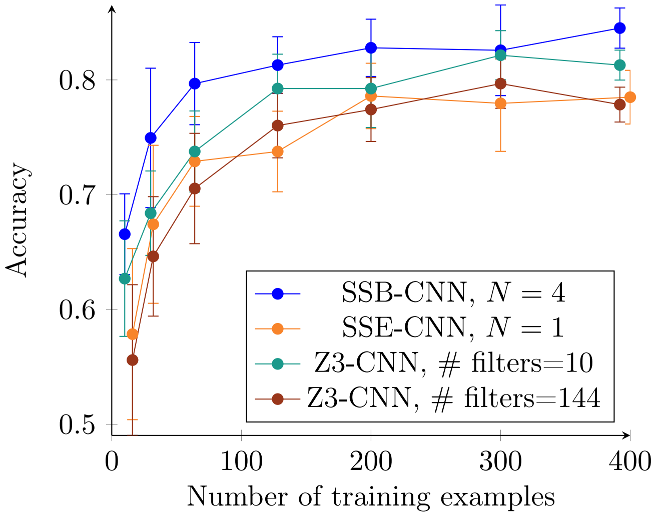


Figure 2.10: Performances on the NLST dataset in terms of accuracy for a varying number of training examples. The error bars represent the CIs at 95%. The number of filters in the first layer for the SSB- and SSE-CNN is 4.

two invariants (Section 2.4.4), we observe that the specific information captured by the SSB improves the classification performance for both datasets: as soon as the maximum degree is greater than one, the SSB-CNN outperforms the SSE-CNN (Section 2.4.4, Fig. 2.7 and 2.8).

Besides, both the SSE- and the SSB-CNN outperform the standard Z3-CNN on the synthetic data which was specifically designed to give an advantage to LRI networks. By contrast, in the nodule classification task (NLST dataset), the Z3-CNN outperforms the SSE-CNN. It seems that the simple design of the SSE-CNN fails to capture the specific signature of malignant pulmonary nodule information on these data. However, once again, the richer invariant representation of the SSB-CNN allows outperforming even the Z3-CNN with statistical significance when $N = 4$ while using approximately 22 times fewer parameters.

2.5.2 LRI Models Learn with Fewer Data

The learning curve experiment on the synthetic dataset presented in Section 2.4.5 shows that both LRI designs outperform the Z3-CNN for any number of training examples. What is more notable is the steeper learning for the two LRI networks. Both SSE- and SSB-CNNs seem to require the same number of training examples to reach their final performance level. For the Z3-CNN, two networks are compared: one with 10 filters and the other with 144 filters, accounting for 7322 and 105,410 trainable parameters, respectively. Even though the number of parameters is vastly different, the overall shape of the learning curves does not significantly change between the two Z3 networks, pointing out that the relationship between numbers of parameters and training examples is not obvious and highly depends on the architecture.

On the NLST dataset, the SSB-CNN outperforms the Z3-CNN when trained with the same number of training examples. However, the steeper learning curve of the former is less pronounced than with the synthetic dataset. We expect the gap between the two learning curves to be wider if we use deeper architecture as the difference in the number parameters will be higher. Overall, we observe that the proposed SSB-CNN requires fewer training examples than the Z3-CNN, thanks to both the LRI property and the compressing parametric SH kernel representations.

2.6 Conclusion

We showed that, by using the highly discriminative SSB RI descriptor, we are able to implement CNNs that are more accurate than the previously proposed SSE-CNN. Furthermore, we also observed that LRI networks can learn with fewer training examples than the more traditional Z3-CNN, which supports our hypothesis that the latter tends to misspend the parameter budget to learn data invariances and symmetries. The main limitation of the proposed experimental evaluation is that it relies on shallow networks that would place these approaches more at the crossroad between handcrafted methods and deep learning. In future work, the LRI layers will be implemented in a deeper architecture to leverage the fewer resources that they require in comparison with a standard convolutional layer. This is expected to constitute a major contribution to improve 3D data analysis when curated and labelled training data is scarce, which most often the case in medical image analysis. The code is available on GitHub[V].

Acknowledgment

The authors are grateful to Michael Unser, who suggested them to consider the bispectrum as a tool to capture rotation-invariant features of 3D signals. This work was supported by the Swiss National Science Foundation (SNSF grants 205320_179069 and P2ELP2_181759) and the Swiss Personalized Health Network (SPHN IMAGINE and QA4IQI projects), as well as a hardware grant from NVIDIA.

[V] https://github.com/voreille/ssbcnn, as of April 2020.

3 2D Bispectral LRI U-Net

The paper presented in this chapter is:

- Valentin Oreiller, Julien Fageot, Vincent Andrearczyk, John O. Prior, and Adrien Depeursinge. "Robust Multi-Organ Nucleus Segmentation Using a Locally Rotation Invariant Bispectral U-Net." In *Medical Imaging with Deep Learning.* 2021.

Abstract

Locally Rotation Invariant (LRI) operators have shown great potential to robustly identify biomedical textures where discriminative patterns appear at random positions and orientations. We build LRI operators through the local projection of the image on circular harmonics followed by the computation of the bispectrum, which is LRI by design. This formulation allows to avoid the discretization of the orientations and does not require any criterion to locally align the descriptors. This operator is used in a convolutional layer resulting in LRI Convolutional Neural Networks (LRI CNN). To evaluate the relevance of this approach, we use it to segment cellular nuclei in histopathological images. We compare the proposed bispectral LRI layer against a standard convolutional layer in a U-Net architecture. While they perform equally in terms of F-score, the LRI CNN provides more robust segmentation with respect to orientation, even when rotational data augmentation is used. This robustness is essential when the relevant pattern may vary in orientation, which is often the case in medical images.

3.1 Introduction

Robustness of Convolutional Neural Networks (CNNs) to changes in orientations of the input structures (*e.g.* nucleus, glands) has been little investigated and may have an important impact on the usability of the methods in practice. Biomedical textures are composed of local patterns that appear at random positions and orientations. Local Rotation Invariant (LRI) operators have been shown to be crucial to characterize such texture (Depeursinge et al., 2018). A common strategy to design LRI operators is to align local descriptors. This includes the Maximum Response 8 (MR8) filterbank (Varma & Zisserman, 2005b) and Local Binary Patterns (LBP) (Ahonen et al., 2009; Ojala, Pietikainen, et al., 2002). Other methods relying on steerability have been proposed to avoid error due to orientation sampling, such as steerable filters (Fageot et al., 2021; Unser & Chenouard, 2013b; T. Zhao & Blu, 2020), Riesz (Dicente Cid et al., 2017a), and steerable wavelets (Depeursinge, Püspöki, et al., 2017b; Puspoki et al., 2019). Another well known method is the scale-invariant feature transform (Lowe, 2004). These methods typically require discretizing orientations and an arbitrary criterion to select the dominant local orientation on which the descriptor is aligned. Built-in LRI approaches have been proposed to avoid using such arbitrary criterion. For instance, the power spectrum was used in (Andrearczyk, Oreiller, Fageot, et al., 2019b; Depeursinge et al., 2018) which allows obtaining a LRI operator continuously defined on the rotation domain. In this work, we design a LRI operator based on a similar but more evolved descriptor, *i.e.* the bispectrum, that can be embedded in a CNN.

CNNs have revolutionized the field of computer vision and biomedical image analysis. Rotation invariance in CNNs is still mainly induced via data augmentation either at training- or test-time. However, built-in rotation equivariance was shown to outperform both training- and test-time augmentation on histopathological image analysis (Lafarge et al., 2021). Rotation-equivariant networks also showed improved robustness to geometric adversarial perturbations (Dumont et al., 2018). A large body of research has focused on designing networks with built-in rotation equivariance and are mainly based on discretized rotations (*i.e.* group equivariant CNN) or steerable filters (Bekkers, 2019; T. Cohen & Welling, 2016b; T. S. Cohen et al., 2019; Kondor & Trivedi, 2018; Weiler et al., 2018; Weiler et al., 2017).

In this work, we propose a CNN design that is invariant to local rotations rather than rotation equivariant. The motivation is that such a design will provide substantial robustness to changes in pattern orientation when identifying biomedical textures. Furthermore, the proposed LRI operators are also globally rotation equivariant (Andrearczyk, Fageot, et al., 2020). While global rotation invariance may be obtained by data augmentation, invariance to local rotation can not be achieved in this way. In (Andrearczyk, Fageot, et al., 2020), the authors proposed two different designs to implement LRI CNNs, one with steerable filters and another one based on the power spectrum. In (Eickenberg et al., 2017), the power spectrum was used within the scattering transform framework (Ablowitz et al., 1974) to obtain global rotation equivariant feature maps. Those works closely relate to our design. One key difference is that we use the bispectrum rather than the power spectrum. Our design relies on the shift invariance property of the bispectrum which translates into rotational invariance for functions defined on the circle. We chose the bispectrum over the power spectrum for its completeness, *i.e.* the bispectrum completely characterizes a function up to a shift (Kakarala, 2012). We evaluated this design with a simple U-Net (Ronneberger et al., 2015) on histopathological images. The evaluation were greatly inspired by the work of Lafarge et al., 2021, in order to have an external comparison. However, our results are not directly comparable since we did not use the same training/testing/validation split.

3.2 Methods

In this section, we develop the theoretical background as well as the implementation details to design a LRI CNN. The main idea is to obtain a LRI convolutional layer that is functionally identical to a standard convolutional layer. Then, we evaluate our layer in a CNN and compare it to a CNN with the same architecture but with standard layers. This work focuses on developing 2D CNNs that are invariant to the orientation at which local patterns appear. The proposed design relies on the rotation invariance property of the bispectrum.

3.2.1 Notations

We consider 2D images as functions $I \in L_2(\mathbb{R}^2)$, where the value $I(\boldsymbol{x}) \in \mathbb{R}$ corresponds to the pixel intensity at location $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$. The rotation of an image I is written as $I(\mathrm{R}\cdot)$, where $\mathrm{R} \in SO(2)$ is the corresponding 2D rotation matrix.

The circle is denoted as $\mathbb{S}^1 = \{\boldsymbol{x} \in \mathbb{R}^2 : ||\boldsymbol{x}||_2 = 1\}$. Polar coordinates are defined as $(x_1, x_2) = (\rho\cos(\theta), \rho\sin(\theta))$ with $\rho \geq 0$ and $\theta \in [0, 2\pi)$. We use the following notation for the mapping from polar to cartesian: $\rho(\boldsymbol{x}) = ||\boldsymbol{x}||$ and $\theta(\boldsymbol{x})$ the standard mapping from $\boldsymbol{x}$ to its polar angle. For clarity purposes, we often do not disclose the dependency on $\boldsymbol{x}$ for ρ and θ. We consider square-integrable functions defined on the circle $f \in L_2(\mathbb{S}^1)$ and express them as functions of the polar angle $f(\theta)$. The inner product is defined by $\langle f\,, g\rangle_{L_2(\mathbb{S}^1)} = \int_0^{2\pi} f(\theta)\overline{g(\theta)}\mathrm{d}\theta$. The rotation of a function f by an angle θ_0 is simply the function "shifted" by that angle *i.e.* $f(\cdot - \theta_0)$. The Fourier transform of the function f is defined as $\hat{f}[n] = \int_0^{2\pi} f(\theta)e^{-jn\theta}\mathrm{d}\theta$ for any $n \in \mathbb{Z}$. The triangle function is referred to as $x \in \mathbb{R} \mapsto \mathrm{tri}(x)$ and is defined as $\mathrm{tri}(x) = 1 - |x|$ if $|x| < 1$ and $\mathrm{tri}(x) = 0$ otherwise.

3.2.2 LRI Operators

Our mathematical formalism relies on the concepts of image operators acting over continuous images, as presented in (Depeursinge, Fageot, & Al-Kadi, 2017a). This work focuses on image operators $\mathcal{G}$ that are LRI as previously introduced in (Andrearczyk, Fageot, et al., 2020). An operator $\mathcal{G}$ is LRI if it satisfies the three following properties:

- *Locality*: there exists $\rho_0 > 0$ such that, for every $\boldsymbol{x} \in \mathbb{R}^2$ and every image $I \in L_2(\mathbb{R}^2)$, the quantity $\mathcal{G}\{I\}(\boldsymbol{x})$ only depends on local image values $I(\boldsymbol{y})$ for $\|\boldsymbol{y}-\boldsymbol{x}\| \leq \rho_0$.

- *Global equivariance to translations:* For any $I \in L_2(\mathbb{R}^2)$,

$$\mathcal{G}\{I(\cdot - \boldsymbol{x}_0)\} = \mathcal{G}\{I\}(\cdot - \boldsymbol{x}_0) \quad \text{for any } \boldsymbol{x}_0 \in \mathbb{R}^2.$$

- *Global equivariance to rotations:* For any $I \in L_2(\mathbb{R}^2)$,

$$\mathcal{G}\{I(\mathrm{R}_0\cdot)\} = \mathcal{G}\{I\}(\mathrm{R}_0\cdot) \quad \text{for any } \mathrm{R}_0 \in SO(2).$$

The simplest example of a LRI operator is the convolution with filter

$$\mathcal{G}\{I\}(\boldsymbol{x}) = (I * h)(\boldsymbol{x}), \tag{3.1}$$

where h is isotropic with finite support, *i.e.* $h(x_1, x_2) = h(\rho)$ is purely radial and vanishes for $\rho > \rho_0$ for some fixed $\rho > 0$ (Andrearczyk, Fageot, et al., 2020, Prop. 1). However, isotropic filters are limited since they discard local directional information. We will now see how we can extend this notion of equivariance to directional sensitive operators using the bispectrum.

3.2.3 Bispectral LRI Operators and Layers

We introduce the theoretical background and methodology to implement our bispectral image operators. They have the advantage of being LRI and sensitive to directional information. We also detail how these operators can be embedded into a convolutional layer.

The Bispectrum: A Complete Set of Rotation Invariant Features

We first focus on features $\mathcal{H} : L_2(\mathbb{S}^1) \to \mathbb{R}$ of circular functions that are rotation invariant, *i.e.* such that

$$\mathcal{H}\{f(\theta)\} = \mathcal{H}\{f(\theta - \theta_0)\} \tag{3.2}$$

for any function f and any angle $\theta_0 \in [0, 2\pi)$. The typical example is the power spectrum of f defined from its Fourier series coefficients $\hat{f}[n]$. For any fixed discrete frequency $n_0 \in \mathbb{Z}$, the Fourier feature $f \mapsto |\hat{f}[n_0]|^2$ is easily shown to be rotation invariant in the sense of (3.2). However, the power spectrum discards the phase information of $\hat{f}$ and thus, does not allow for the complete characterization of a polar function.

For this reason, we consider a more elaborated Fourier-based feature, named the bispectrum, that retains the rotation invariance while keeping the phase information (Bartelt et al., 1984; Kakarala, 2012). The bispectrum of a function $f \colon L^2(\mathbb{S}^1) \to \mathbb{R}$ is defined for any $n_1, n_2 \in \mathbb{Z}$ as

$$b_f[n, n'] = \hat{f}[n]\hat{f}[n']\overline{\hat{f}[n+n']}. \tag{3.3}$$

One readily verifies that, for any $n, n' \in \mathbb{Z}$, the feature $f \mapsto b_f[n, n']$ is rotation invariant. The bispectrum is *complete* (Kakarala, 2012) in the sense that it discriminates between two functions up to a rotation which motivates our choice of using it over the power spectrum to design LRI layers.

Bispectral LRI Operators

The bispectrum is defined for circular functions in $L^2(\mathbb{S}^1)$ and can be used to build image operators. In this section, we fix a radial function $h(\rho) \in L_2(\mathbb{R}^2)$ and consider the steerable kernel $\kappa_n(\rho, \theta) = h(\rho)\mathrm{e}^{\mathrm{j}n\theta}$ associated to the discrete frequency $n \in \mathbb{Z}$. We moreover introduce the convolution operator $\mathcal{C}_n\{I\}(\boldsymbol{x}) = (I * \kappa_n)(\boldsymbol{x})$. We observe that we can write, for any fixed position $\boldsymbol{x}_0 \in \mathbb{R}^2$,

$$\mathcal{C}_n\{I\}(\boldsymbol{x}_0) = \int_0^{2\pi} \left(\int_0^{+\infty} \left(I(\boldsymbol{x}_0 - \cdot)\right)(\rho, \theta) h(\rho) \rho \mathrm{d}\rho \right) e^{-\mathrm{j}n\theta} \mathrm{d}\theta.$$

We can interpret the circular function $\theta \mapsto I^h_{\boldsymbol{x}_0}(\theta) := \int_0^{+\infty} \left(I(\boldsymbol{x}_0 - \cdot)\right)(\rho, \theta) h(\rho) \rho \mathrm{d}\rho$ as the radial projection of the shifted image $I(\boldsymbol{x}_0 - \cdot)$ against the radial profile h. Hence, $\mathcal{C}_n\{I\}(\boldsymbol{x}_0)$ performs the n^{th} Fourier coefficient of the periodic function $I^h_{\boldsymbol{x}_0}$.

For any $n, n' \in \mathbb{Z}$, we define the image operator $\mathcal{G}_{n,n'}$ as

$$\mathcal{G}_{n,n'}\{I\}(\boldsymbol{x}) = \mathcal{C}_n\{I\}(\boldsymbol{x}) \mathcal{C}_{n'}\{I\}(\boldsymbol{x}) \overline{\mathcal{C}_{n+n'}\{I\}(\boldsymbol{x})}. \tag{3.4}$$

Then, we see by comparing (3.4) with (3.3) that, for any fixed position $\boldsymbol{x}_0$, $\mathcal{G}_{n,n'}\{I\}(\boldsymbol{x}_0)$ is the bispectrum of the projection $I^h_{\boldsymbol{x}_0} \in L_2(\mathbb{S}^1)$. We call $\mathcal{G}_{n,n'}$ the *bispectral operator of frequencies n, n'*. Its main invariance properties are summarized in Theorem 1, whose proof is provided in Appendix B.1.

Theorem 1 *Let $n, n' \in \mathbb{Z}$ and $\rho \mapsto h(\rho)$ a radial profile with finite support. Then, the bispectrum operator $\mathcal{G}_{n,n'}$ is LRI.*

Implementation of the LRI Layer

We fix a maximal order $N \geq 0$ and consider the bispectral operators $\mathcal{G}_{n,n'}$ for any $n, n' \geq 0$ such that $n + n' \leq N$. By doing so, we only consider angular frequencies smaller than N. Moreover, we only compute non-repeating pairs as $\mathcal{G}_{n,n'} = \mathcal{G}_{n',n}$. We define L as the number of combinations of n, n' such that $n \leq n'$ and $n + n' \leq N$.

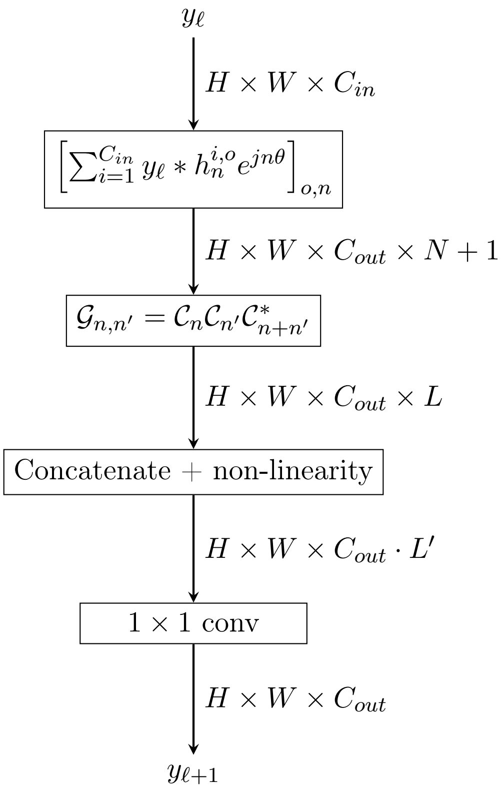

y_ℓ
$H \times W \times C_{in}$
$\left[\sum_{i=1}^{C_{in}} y_\ell * h_n^{i,o} e^{jn\theta}\right]_{o,n}$
$H \times W \times C_{out} \times N+1$
$\mathcal{G}_{n,n'} = \mathcal{C}_n \mathcal{C}_{n'} \mathcal{C}^*_{n+n'}$
$H \times W \times C_{out} \times L$
Concatenate + non-linearity
$H \times W \times C_{out} \cdot L'$
1×1 conv
$H \times W \times C_{out}$
$y_{\ell+1}$

Figure 3.1: The proposed bispectrum-based LRI convolutional layer.

For any n, n', the application of the bispectral operator $\mathcal{G}_{n,n'}$ at each layer of the bispectral LRI network is implemented in four steps as depicted in Fig. 3.1. First, the feature maps are computed as a complex convolution $\mathcal{C}_n^o(\boldsymbol{x}_0) = \sum_{i=1}^{C}(y_i(\boldsymbol{x}) * h_n^{i,o}(\boldsymbol{x})e^{-\mathrm{j}n\theta(\boldsymbol{x})})(\boldsymbol{x}_0)$, with y_i the i^{th} channel of the previous feature maps, $h_n^{i,o}$ the filters that are learned by gradient descent. The parametrization of the filters $h_n^{i,o}$ is detailed in Section 3.2.3. The indices i and o run through $[1, \ldots, C_{in}]$ and $[1, \ldots, C_{out}]$ respectively and represent the input and output channels of the layer.

The second step consists in applying (3.4) to the feature maps $\mathcal{C}_n^o(\boldsymbol{x})$, yielding the desired operator $\mathcal{G}_{n,n'}$.

The third step is the concatenation of the real and imaginary part of $\mathcal{G}_{n,n'}$, which is followed by a point-wise non-linearity of the following form $\sigma(x) = \text{sign}(x)\log(1 + |x|)$. This choice is made to avoid vanishing and exploding gradients with the cubic nature of the bispectral feature maps (see Eq. (3.4)). Learned biases are added to the resulting feature maps and Rectified Linear Unit (ReLU) is applied.

In the last step, the number of features maps is reduced by a 1 × 1 convolution to obtain C_{out} output channels. This whole process results in a layer that takes as an input C_{in} feature maps and outputs C_{out} feature maps like a standard convolutional layer.

Parametrization of the Radial Profiles

The radial profiles $h_n^{i,o}$ are parametrized as follows:

$$h_n^{i,o}(\rho) = \sum_{j=0}^{J} w_{n,j}^{i,o}\psi_j(\rho), \tag{3.5}$$

where the $w_{n,j}^{i,o}$ are the learnable parameters of the layer, $i \in [1, \cdots, C_{in}]$, $o \in [1, \cdots, C_{out}]$ and $0 \leq n \leq N$. The radial functions ψ_j are chosen as $\psi_j(\rho) = \text{tri}(\rho - j)$.

3.2.4 Dataset

We tested our design on a subset of the dataset proposed in the MoNuSeg 2018 challenge (N. Kumar et al., 2019) which consists of 24 Hematoxylin and Eosin (H&E) stained images selected from whole slice images acquired at the commonly used 40× magnification provided by The Cancer Genome Atlas (Koboldt et al., 2012). This subset contains 6 1000 × 1000 images per tissue type for a total of four different tissue types (breast, liver, kidney, and prostate). Nuclei instance segmentation is available for these 24 images. We followed a similar splitting scheme than proposed by Lafarge et al., 2019, namely 4 × 3 images for training, 4 × 1 for validation and 4 × 2 for testing. We repeated ten random splits to evaluate the variation of the models. As preprocessing, we used the method described in (Macenko et al., 2009) to normalize the H&E images. We adapted the code from https://github.com/schaugf/HEnorm_python to fit our needs.

3.2.5 Network Architecture and Training

The networks used in this work were based on the U-Net architecture (Ronneberger et al., 2015). However, we used a lighter version of the U-Net, as proposed in (Lafarge et al., 2021), which contains only two levels of down-sampling. All the convolutional layers have a kernel size of 5 × 5 and are connected to a batch normalization followed by a ReLU. The encoder path includes three convolutions layers with max-pooling to reduce the spatial dimension. The number of feature maps for each layer respectively are 8, 16, and 32. The decoder path contains 2 layers preceded by a bi-linear upsampling. The final prediction is modeled as a three classes probability, *i.e.* nucleus core, nucleus border, and background. The prediction is computed with a 1 × 1 convolution with a softmax activation. The final output is post-processed to obtain instance segmentation of each nucleus. This post-processing consists in binarizing the prediction with a threshold of 0.5, then the core and border prediction are respectively used as seed and landscape for a watershed algorithm[I] (Falcão et al., 2004).

The networks were trained by minimizing the class-balanced cross-entropy with an Adam optimizer and a learning rate of 10^{-3}. The models were trained on patches of 60 × 60 randomly drawn from the training set with a batch size of 16. We applied multiple of 90° rotation as data augmentation[II] as well as random brightness shift. The training was run for a maximum number of epochs of 200 and we applied an early stop monitoring the F-score. The experiments were performed on an Nvidia Tesla K80. The code for the implementation is available on our GitHub repository[III].

3.2.6 Metrics and Evaluation

Two types of experiments were conducted. We first evaluated the performance of our bispectral LRI U-Net against a standard U-Net. The metric used for this experiment is the F-score and we considered a match when the predicted nucleus had more than 50% overlap with the ground-truth nucleus. Since the radial profiles of the proposed LRI layer do not cover the entire 5 × 5 kernel, we used masked kernels in the standard U-Net to ensure a fair comparison. The masked kernel consists of removing the four pixels in the corners. All models were trained and tested on the same 10 train/validation/testing splits to assess performance variation.

In the second experiment, we evaluated the robustness of the predictions made by the two

[I] docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.watershed_ift.html, March 2022.

[II] We also evaluated the training without this augmentation, but did not observe any significant changes.

[III] github.com/voreille/2d_bispectrum_cnn, March 2022.

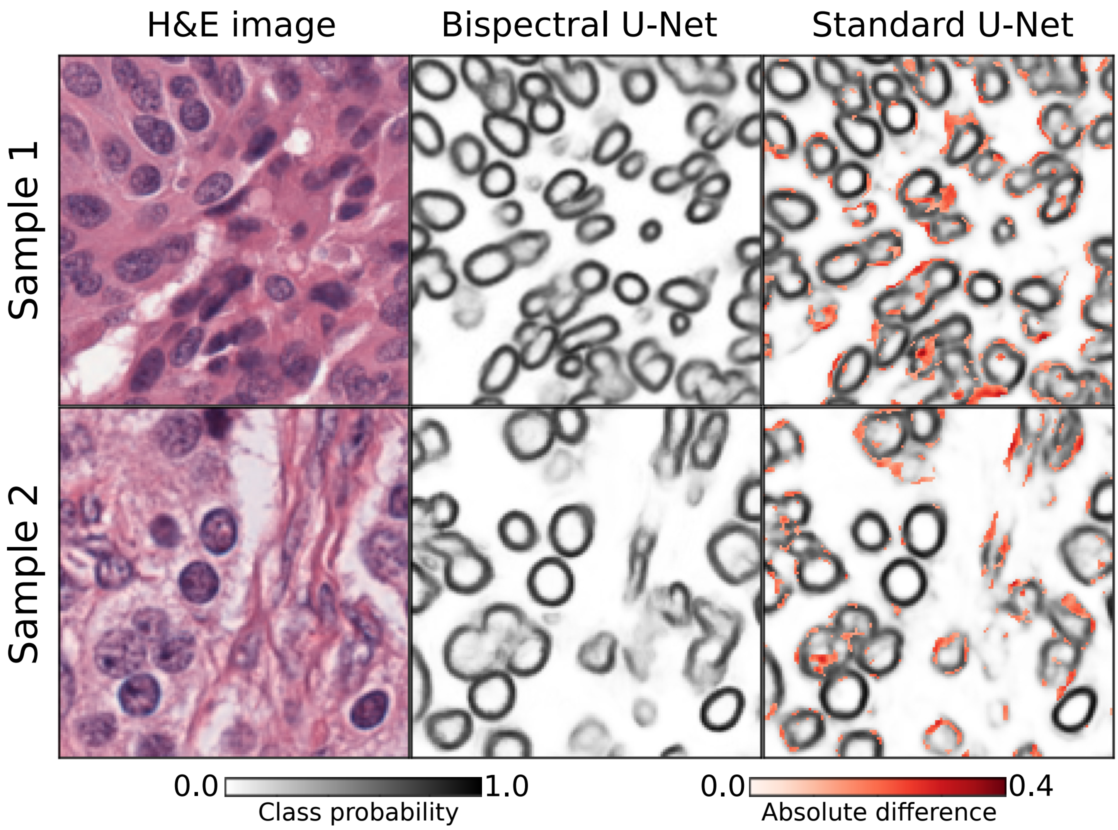


Figure 3.2: Illustration of the prediction robustness with respect to input orientation. The middle and right columns depict the probability prediction of the nucleus border class for the bispectral and standard U-Nets, respectively. The red color map indicates the mean pixel-wise differences averaged across the six pairs of 90° rotations. These differences are almost null for the bispectral U-Net.

designs in terms of 90°-rotation equivariance. We fed the networks with the same image rotated at different orientations. Then, we applied the inverse rotation to the output maps and compare the difference in pixel-wise prediction. We used the Root Mean Square Error (RMSE) on the three classes of raw probabilities prediction as well as the Dice Similarity Coefficient (DSC) on the post-processed probability to measure the overlap of predictions for each orientation, indicating the robustness of prediction with respect to input orientation.

3.3 Results

Fig. 3.3 relates the F-score for varying maximum degree N of the bispectral U-Net and standard U-Net. The best performing bispectral U-Net had an F-score of 0.7157 ± 0.0328 with a maximum degree $N = 7$ and 136,147 parameters (standard U-Net 71,571 parameters). Thus, we trained a standard U-Net with more feature maps to increase the number of parameters (134,709 parameters) which obtained an average F-score of 0.7324

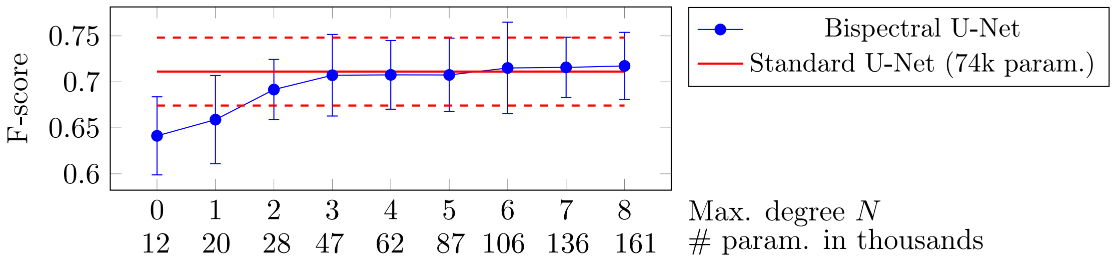


Figure 3.3: Performance of the different networks. Average F-scores are reported across ten repetitions of the proposed bispectral U-Net evaluated at different maximum degrees N (blue) and standard U-Net (red). Error bars and dashed lines indicate the standard deviation.

± 0.0326.

To account for the discrepancy in the number of parameters at different maximal degrees N, we trained a bispectral U-Net with $N = 0$ and more feature maps to obtain a network of 45,779 parameters (comparable to the number of parameters of the network with N=3). The resulting network achieved an average F-score of 0.6592 ± 0.0286.

Table 3.1 summarizes the average RMSE and DSC between predictions of rotated images. The average was calculated on all the images from the testing set of one split and all pairs of 90° rotations. Fig. 3.2 illustrates the pixel-wise variation between predictions when fed to the networks at different orientations.

Table 3.1: Quantitative evaluation of segmentation robustness. It is worth noting that rotational data augmentation was used during training for both approaches.

Model	**RMSE border**	**RMSE core**	**RMSE background**	**DSC**
Standard U-Net	8.49 ± 1.35 %	7.66 ± 1.72 %	9.15 ± 1.43 %	0.9153 ± 0.0205
Bispectral U-Net ($N = 7$)	2.50e-5 ±0.78e-5 %	2.26e-5 ±0.77e-5 %	2.53e-5 ±0.87e-5 %	0.9876 ± 0.0044

3.4 Discussions and Conclusion

We proposed a novel 2D LRI layer based on the bispectrum that can be integrated into any CNN architecture. This design aims to improve the robustness of the predictions when inputs do not have a standardized orientation, or when local structures of interest can appear at any orientation.

We first presented the bispectral operators and demonstrated their LRI property in Section 3.2.3 and Appendix B.1. We also detailed how to use them in a convolutional layer in Section 3.2.3. Second, we incorporated the LRI layer into a U-Net to allow robust segmentation of multi-organ nuclei in histopathology images. We observed that the segmentation performance of the LRI U-Net is on par with a standard U-Net (see Fig. 3.3). While the bispectral U-Net was slightly outperformed by the standard U-Net, it is difficult to evaluate to which extent the post-processing had a role in this difference (see Appendix B.2).

However, an important gain was obtained in terms of robustness with respect to the orientation of the input (see Table 3.1) thanks to the rotation equivariance property of the used image operators. This robustness is crucial for most medical image analysis tasks where structures of interest often appear at various orientations. We observed that standard methods lack robustness, even when rotational data augmentation is used (see Table 3.1 and Fig. 3.2). While most studies focused on classification or segmentation performance alone, robustness to changes in input orientation was little investigated and may have important consequences on the usability of the models.

Our work recognizes several limitations. The segmentation performance presented in Fig. 3.3 is not at the level of the state of the art on this dataset. Our goal was to compare with standard baseline methods such as the U-Net without using refinements *e.g.* postprocessing of the segmentation maps or ensembling. Future work includes the extension of the bispectral operator to 3D and extensive comparisons with group-equivariant approaches.

Acknowledgments

This work was partially supported by the SNSF grants 205320_179069 and P400P2_194364, the SPHN IMAGINE project, and the Hasler Foundation.

4 Overview and Post-Challenge Analyses of HECKTOR 2020

The paper presented in this chapter is:

- Valentin Oreiller*, Vincent Andrearczyk*, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, Andrei Iantsen, Mathieu Hatt, Yading Yuan, Jun Ma, Xiaoping Yang, Chinmay Rao, Suraj Pai, Kanchan Ghimire, Xue Feng, Mohamed A. Naser, Clifton D. Fuller, Fereshteh Yousefirizi, Arman Rahmim, Huai Chen, Lisheng Wang, John O. Prior, Adrien Depeursinge, "Head and neck tumor segmentation in PET/CT: the HECKTOR challenge." *Medical image analysis* 77 (2022): 102336.

* both authors contributed equally.

Abstract

This paper relates the post-analysis of the first edition of the HEad and neCK TumOR (HECKTOR) challenge. This challenge was held as a satellite event of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2020, and was the first of its kind focusing on lesion segmentation in combined FDG-PET and CT image modalities. The challenge's task is the automatic segmentation of the Gross Tumor Volume (GTV) of Head and Neck (H&N) oropharyngeal primary tumors in FDG-PET/CT images. To this end, the participants were given a training set of 201 cases from four different centers and their methods were tested on a held-out set of 53 cases from a fifth center. The methods were ranked according to the Dice Score Coefficient (DSC) averaged across all test cases. An additional inter-observer agreement study was organized to assess the difficulty of the task from a human perspective. 64 teams registered to the challenge, among which 10 provided a paper detailing their approach. The best method obtained an average DSC of 0.7591, showing a large improvement over

our proposed baseline method and the inter-observer agreement, associated with DSCs of 0.6610 and 0.61, respectively. The automatic methods proved to successfully leverage the wealth of metabolic and structural properties of combined PET and CT modalities, significantly outperforming human inter-observer agreement level, semi-automatic thresholding based on PET images as well as other single modality-based methods. This promising performance is one step forward towards large-scale radiomics studies in H&N cancer, obviating the need for error-prone and time-consuming manual delineation of GTVs.

4.1 Introduction

High-throughput medical image analysis, often referred to as radiomics, has shown its potential in unveiling relationships between quantitative image biomarkers and cancer prognosis, including in the context of Head and Neck (H&N) cancer (Bogowicz et al., 2017; Vallieres et al., 2017). H&N cancer is the 5$^{\text{th}}$ leading cancer by incidence (Parkin et al., 2005) and its treatment is generally based on a combination of radiotherapy with systemic treatment (e.g. Cetuximab) (Bonner et al., 2010). However, treating this cancer remains challenging since local failure occurs in about 40% of patients in the first two years after the treatment (Chajon et al., 2013). The development of non-invasive and personalized approaches (e.g. radiomics) is critical for improving disease characterization and will, hopefully, lead to more targeted therapies based on phenotypic tumor characteristics. 2-[18F]fluoro-2-deoxyglucose positron-emission tomography (FDG-PET) and Computed Tomography (CT) hold a special place for disease characterization since they contain complementary information about the metabolism and the anatomy of cancer. Furthermore, they are used for initial staging and follow-up of H&N cancer. These modalities are therefore readily available for the creation and evaluation of radiomics models based on these clinically acquired images. Typical radiomics analyses rely on localized feature extraction inside delineated lesions or Volumes Of Interest (VOI) (Gillies et al., 2016b; Lambin, Leijenaar, Deist, Peerlings, De Jong, et al., 2017). One of the reasons that impede the development of robust models is the time-consuming and error-prone manual delineation of these VOIs. To this end, the automatic segmentation of H&N Gross Tumor Volume of the primary tumor (GTVt) and the lymph nodes (GTVn) constitutes a highly promising approach to annotate and analyze very large cohorts, which is critically needed to enable robust and reproducible validation of radiomics models. Moreover, automatic segmentation also has the potential to allow radiation oncologists to improve treatment planning efficiency by reducing the time needed for tumor delineation as well as improving inter-observer reproducibility.

The goal of the HEad and neCK TumOR (HECKTOR) challenge is to establish and benchmark the best-performing methods for H&N lesions segmentation while exploiting the rich bi-modal information of combined PET/CT. In this first edition of the challenge, the participants were asked to develop automatic methods for the segmentation of the

GTVt[I] on FDG-PET/CT images of patients suffering from oropharyngeal cancer. It is worth noting that to be part of the official ranking, the participants had to provide a paper describing their methods. Furthermore, participants had to disclose the use of external training data and were in this case not eligible for the official ranking. None of the participants reported using external data. This manuscript summarizes the methods and presents the associated segmentation results of the different teams who participated in this 2020 edition of the HECKTOR challenge. It also includes several additional extensive qualitative and quantitative analyses. This paper extends the material presented in (Andrearczyk, Oreiller, Vallières, Jreige, et al., 2021) with the following:

- an extensive review of the prior work;
- an analysis of the inter-observer agreement organized with four different observers on a subset of 21 cases;
- an evaluation of a super-ensemble segmentation based on the submitted contours of the ten ranked teams;
- an addition of new participants' results from runs submitted after the end of the challenge;
- a semi-automatic segmentation based on PET thresholding as an additional baseline; and
- additional extensive qualitative and quantitative analyses of the results.

The paper is organized as follows. Section 4.2 presents the related work. Section 4.3 describes the challenge setup including the dataset, annotations, participation, and ranking. The presentation and in-depth analysis of the participants' results are provided in Section 4.4 and are discussed in Section 4.5. Finally, Section 4.6 concludes the paper.

4.2 Prior Work

4.2.1 Related Tumor Segmentation Algorithms

An abundance of works has been proposed to automatically segment tumors in PET and PET/CT images ranging from thresholding to unsupervised and supervised machine learning methods. Making an exhaustive review of all these approaches is out of the scope of this manuscript and is proposed in (Foster et al., 2014; Hatt et al., 2017). Among these different strategies, the simplest ones are based on the thresholding of the Standardized Uptake Values (SUV) in PET images. These methods are difficult to

[I] For the first and second edition of the challenge, the GTVn segmentation is not part of the tasks but will be asked in further editions.

automatize completely since the SUV is a semi-quantitative measure that highly depends on the time between the injection and the image acquisition, the device, the reconstruction algorithm, the shape of the tumor, and even the patient (Wahl et al., 2009).

More refined approaches have been proposed to further automatize this process. Most of them are relying on the distribution of SUV values or other handcrafted quantitative image features in PET only. For instance, algorithms based on Gaussian Mixtures (Aristophanous et al., 2007) or fuzzy C-means modeling (Hatt et al., 2009; Lapuyade-Lahorgue et al., 2015) were proposed. Others formulated the segmentation problem as a minimization of a Markov random field (Song et al., 2013). In the context of H&N tumors delineation, a decision-tree-based K-nearest-neighbor classifier trained with regional texture features in PET and CT images was used in (Yu et al., 2009).

Recent work was inspired by the success of deep Convolutional Neural Networks (CNN), and more precisely of the U-Net (Ronneberger et al., 2015) applied to multi-modal biomedical image segmentation (Zhou et al., 2019). PET/CT tumor segmentation has also benefited from the advancement of this field. For instance, Blanc-Durand et al., 2018 applied a 3D U-Net to segment brain tumors in O-(2-[18F]fluoroethyl)-L-tyrosine PET/CT images. Deep CNNs was also used several times in the context of lung tumor segmentation (Fu et al., 2021; Li et al., 2019; Wu et al., 2020; X. Zhao et al., 2018; Zhong et al., 2018). A 3D U-Net was used by Jemaa et al., 2020 to lung cancer and lymphoma, which was trained on 2540 volumes and tested 1124 volumes. Iantsen, Ferreira, et al., 2021 used a U-Net architecture for the automatic segmentation of cervical tumors in PET only.

The deep learning-based approaches were also specifically applied to tumor segmentation in H&N cancers. A comparison of different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation is presented in (Ren et al., 2021). In a study including 22 patients from two different centers, Huang et al., 2018 used a 2D U-Net to segment the GTV, *i.e.* the union of GTVt and GTVn. Moe et al., 2019 used a 2D U-Net for the segmentation of GTV on a dataset of 55 patients. In another study, Guo et al., 2019 applied a 3D U-Net to segment the GTVt, which was evaluated on a cohort of 250 patients. The authors showed that multimodal networks outperform networks based on a single modality. More recently, Groendahl et al., 2021 performed an analysis of the different types of automatic segmentation based on thresholding, classification at the pixel level using a shallow classifier, and deep CNN methods. They did this comparison on a mono-centric cohort of 197 patients and concluded that deep learning models outperform the others.

Identifying the best performing method among all these different strategies requires a standardized evaluation. This was already highlighted by Hatt et al., 2017 and challenges constitute a suitable way to systematically evaluate and compare state-of-the-art algorithms against the same test set and with highly controlled conditions.

4.2.2 Medical Image Segmentation Challenges

The growing interest in biomedical image analysis challenges is illustrated by and an increasing number of new challenges organized every year, which can be partly explained by the growing community. For instance at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2018, 2019, and 2020 there were 15, 22, and 25 accepted challenges, respectively. In the past three MICCAI editions, 52 out of 125 tasks (42%) were related to segmentation[II]. Several other challenges are organized as satellite events of other conferences including the International Symposium on Biomedical Imaging (ISBI), the international conference on Medical Imaging with Deep Learning (MIDL), and the annual meeting of the Radiological Society of North America (RSNA), as well as independently organized challenges (e.g. on Kaggle[III]). Remarkably successful challenges in medical image segmentation include the Brain Tumor Segmentation (BraTS) challenge (Menze et al., 2014), Kidney Tumor Segmentation (KiTS) (Heller et al., 2021) challenge and the Visual Concept Extraction Challenge in Radiology (VISCERAL) (del Toro et al., n.d.) challenge. Surprisingly, as of 2021, only one challenge was organized on PET segmentation (Hatt, Laurent, Ouahabi, Fayad, Tan, Li, Lu, Jaouen, Tauber, Czakon, et al., 2018b) and, to the best of our knowledge, none on PET/CT segmentation.

4.3 HECKTOR 2020 Challenge Set-Up

The challenge took place in 2020 and was associated with the 23rd MICCAI conference as a satellite event the same year. It was hosted on the AIcrowd platform[IV]. The training and test data were released on the 10th of June and the 1st of August, respectively. The participants were asked to submit their results before the 10th of September. The challenge's results were communicated the 15th of September, and the MICCAI associated event was held the 4th of October. The data of the challenge are currently available on the AIcrowd platform after signing an end-user agreement and the leaderboard submission was open until the 10th of September 2021[V].

The following section summarizes the challenge's set-up. A thorough and BIAS (Maier-Hein et al., 2020) compliant description of the challenge organization is provided in (Andrearczyk, Oreiller, Vallières, Jreige, et al., 2021).

[II] https://www.biomedical-challenges.org/miccai2021/Statistics, as of October 2021.

[III] https://www.kaggle.com/, as of October 2021.

[IV] https://www.aicrowd.com/challenges/miccai-2020-hecktor, as of October 2021.

[V] The leaderboard was replaced by the 2021 edition after this date: https://www.aicrowd.com/challenges/miccai-2021-hecktor/leaderboards.

Table 4.1: List of scanners used in the different centers.

Center	Device
HGJ	hybrid PET/CT scanner (Discovery ST, GE Healthcare)
CHUS	hybrid PET/CT scanner (GeminiGXL 16, Philips)
HMR	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUM	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUV	hybrid PET/CT scanner (Discovery D690 TOF, GE Healthcare)

4.3.1 Dataset

The dataset used in this challenge includes PET and CT images as well as patient information including age, sex, and acquisition center. The patients selected for this dataset suffered from H&N cancer, which was histologically proven, and they underwent radiotherapy treatment often combined with chemotherapy. The data were acquired from five centers:

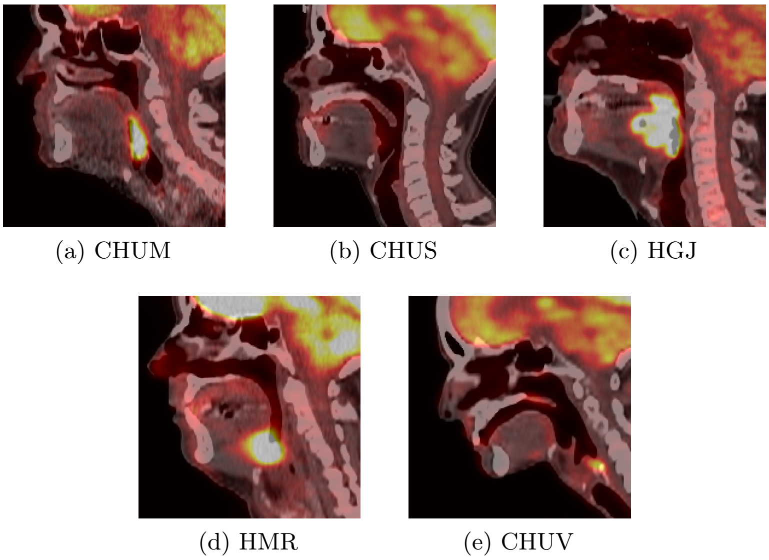
1. Hôpital Général Juif (HGJ), Montréal, CA ($n = 55$)
2. Centre Hospitalier Universitaire de Sherbooke (CHUS), Sherbrooke, CA ($n = 72$)
3. Hôpital Maisonneuve-Rosemont (HMR), Montréal, CA ($n = 18$)
4. Centre Hospitalier de l'Université de Montréal (CHUM), Montréal ($n = 56$)
5. Centre Hospitalier Universitaire Vaudois (CHUV), CH ($n = 53$)

The four centers HGJ, CHUS, HMR, and CHUM were used for the training set, which amounts to 201 cases. This training data constitute a subset of (Vallieres et al., 2017) which contains 298 cases including H&N cancers originating from various anatomical regions. For this initial edition of the HECKTOR challenge, we decided to focus on patients suffering from oropharyngeal cancer to reduce anatomical variations and provide more controlled conditions for the algorithms. The CHUV center was used for the test set, totaling a number of 53 test cases.

An example of fused PET/CT images for each of the five centers is depicted in Fig. 4.1. The list of scanners used in each center for image acquisition can be found in Table 4.1. Additional information concerning image protocols are described in (Andrearczyk, Oreiller, Vallières, Jreige, et al., 2021).

The Digital Imaging and Communications in Medicine (DICOM) files were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format. The CT and PET images were stored in Hounsfield Units (HU) and SUVs, respectively. The code used for the conversion is available on the challenge's repository[VI]. Each case comprises

[VI] github.com/voreille/hecktor/blob/hecktor2020/src/data/dicom_conversion.py, as of October 2021.


(a) CHUM (b) CHUS (c) HGJ

(d) HMR (e) CHUV

Figure 4.1: Case examples of 2D sagittal slices of fused PET/CT images from each of the five centers. These images are obtained after resampling the PET image and the CT image to 1x1x1 mm^3 with a tricubic interpolation. The CT window in Hounsfield unit is $[-140, 260]$ and the PET window in SUV is $[0, 12]$.

NIfTI files for the CT image, the PET image, and the GTVt mask (for the training cases), as well as patient information (age, sex) and center. A bounding box locating the oropharyngeal region was also provided (details of the automatic region detection can be found in (Andrearczyk, Oreiller, & Depeursinge, 2020)). The choice of preprocessing (*e.g.* resampling, image standardization) was left to the participants. Therefore, no further preprocessing was performed to mimic a clinical use of the segmentation methods. However, we provided some routines to crop, resample, and also train a baseline CNN (using NiftyNet (Gibson et al., 2018)). This code was made available on the challenge's repository[VII] to help the participants and to maximize transparency, but the participants were free to use their methods.

4.3.2 Contours

The GTVts from the original dataset were drawn by expert radiation oncologists from multiple centers for radiotherapy treatment planning. In most cases, the contours used for treatment planning are larger than the actual tumor and are presumably not optimized for radiomics with sometimes the inclusion of surrounding tissue or even air cavities. Furthermore, only 40% (80 cases) of the training set were delineated on the CT of the PET/CT scans. The remaining 60% were drawn on a dedicated CT scan for the treatment planning and were registered to the PET/CT scans using intensity-based

[VII] github.com/voreille/hecktor/tree/hecktor2020, as of October 2021.

free-form deformable registration with the software MIM (MIM Software Inc., Cleveland, OH). For more information about the original training set, please refer to (Vallieres et al., 2017). The original contours of the test set were all drawn on the fused PET/CT scans.

To homogenize the data *i.e.* to obtain delineations closer to the true tumoral volume and to remove variability due to the annotators and the registration step, each contour was controlled by an expert who is both a radiologist and a nuclear physician. Two non-experts annotators made an initial cleaning to facilitate the expert's work. During this control, multiple contours were rectified to follow the true border of the tumor as close as possible. Many original contours included air as well as various tissues around the tumor. In some cases, the registration between the dedicated CT planning and the PET/CT introduced artifacts that did not belong to the GTVt. In many cases, the GTVt and GTVn were stored under the same label and had to be separated. Three annotations were corrupted and could not be loaded, requiring the contours to be drawn from scratch. Among the 53 test cases, 11 images were contoured from scratch with the help of the radiological report. Despite the high inter-observer variability (see Section 4.4.4), and with a slight misuse of language, we refer to these “controlled" reference annotations as ground truth.

Finally, the same VOI quality control process was performed for the GTVn contours. These contours were not directly used for the HECKTOR 2020 challenge but we used them in post-analysis of the results (see Section 4.4.8). We also plan on using these annotations in future editions as an auxiliary task of lymph node segmentation. Radiomics studies including lymph nodes may carry important information about patient prognosis and response to treatment.

4.3.3 Ranking and Assement Method

Participants were given access to the test cases without the ground truth annotations and were asked to submit the results of their algorithms on these cases on the AIcrowd platform. We only accepted binary segmentations in the NIfTI file format.

Results were ranked using the 3D Dice Similarity Coefficient (DSC) computed on images cropped using the provided bounding boxes (see Section 4.3.1) in the original CT resolution as:

$$DSC = \frac{2TP}{2TP + FP + FN} \,, \tag{4.1}$$

where TP, FP, and FN are the number of True Positive, False Positive, and False Negative at the voxel level, respectively. Prior to the challenge opening, we decided to handle missing predictions by attributing a DSC of 0 to them. However, this never happened during the submission phase. If the submitted results were in a resolution different from the CT resolution, we applied nearest-neighbor interpolation before evaluation. We also computed other metrics for comparison, namely precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) to investigate whether the methods were rather providing a large FP or FN rate. The evaluation implementation can be found on our GitHub repository[VIII] and was provided to the participants to maximize transparency.

Each participating team had the opportunity to submit up to five valid runs, in case of formatting errors the participant was informed by an error message and the run was not counted. No immediate feedback was displayed on how their run was performing to avoid iterative overfit. The best result of each team was used in the final ranking, which is detailed in Section 4.4 and discussed in Section 4.5.

4.4 Results

This section regroups results in terms of challenge participation, algorithms used, segmentation performance, inter-observer agreement, ensembling “super-algorithm", simple PET thresholding, the relation between tumor size and segmentation performance, false-positive analysis, and alternative ranking of the methods.

4.4.1 Participation

The number of registered teams, as of September 10, 2020 (submission deadline), was 64. At the same date, we had also received and approved 85 signed end-user agreements, received 83 results submissions, including valid and invalid submissions. For the first iteration of the challenge, these numbers are high and show an important interest in the task.

4.4.2 Algorithms Summary

Baselines
We trained several baseline models using standard 3D and 2D U-Nets as in our preliminary results in (Andrearczyk, Oreiller, Vallières, Castelli, et al., 2020). It is worth noting that Andrearczyk, Oreiller, Vallières, Castelli, et al., 2020 used a dataset that was different from HECKTOR 2020, and that the same algorithms were re-trained and evaluated using the HECKTOR 2020 data. We trained on multi-modal PET/CT as well as individual modalities with a combination of non-weighted Dice and cross-entropy losses and without data augmentation.

Participants’ Methods
In Table 4.2, we summarize some of the main components of the participants’ algorithms,

[VIII] github.com/voreille/hecktor/tree/hecktor2020/src/evaluation, as of October 2021.

Table 4.2: Summary of the algorithms in terms of main components used: 2D or 3D U-Net, resampling, preprocessing, training or testing data augmentation, loss used for optimization, an ensemble of multiple models for test prediction and postprocessing of the results. We use the following abbreviations for the preprocessing: Clipping (C), Standardization (S), and if it is applied only to one modality, it is specified in parentheses. For the image resampling, we specify whether the algorithms use Isotropic (I) or Anisotropic (A) resampling and Nearest Neighbor (NN), Linear (L), or Cubic (Cu) interpolation. We use the following abbreviation for the losses: Cross-Entropy (CE), Mumford-Shah (MS), and Mean Absolute Error (MAE). More details can be found in the respective participants' publications.

Team	2D/3D	preproc.	resampling	augm.	loss	ensemble	postproc.
`andrei.iantsen` (Iantsen, Visvikis, et al., 2021)	3D	C+S	I/L	✓	soft Dice+Focal	✓	✗
`junma` (J. Ma & Yang, 2021)	3D	S(PET)	I/Cu	✗	Dice+Top-K	✓	✓
`badger` (J. Xie & Peng, 2021)	3D	C(CT)+S(PET)	A/Cu	✓	Dice+CE	✗	✗
`deepX` (Yuan, 2021)	3D	C(CT)+S	I/L	✓	Jaccard distance	✓	✗
`AIView_sjtu` (H. Chen et al., 2021)	3D	C+S	A/NN	✓	Dice	✗	✗
`xuefeng` (Ghimire et al., 2021)	3D	C(CT)+S	A/L	✓	Dice+CE	✓	✓
`QuritLab` (Yousefirizi & Rahmim, 2021)	3D	S	I/L	✗	MS+MAE	✗	✗
`HFHSegTeam` (Zhu et al., 2021)	2D	C+S	I/L	✓	soft Dice	✗	✗
`Fuller_MDA_Lab` (M. Naser et al., 2021)	3D	C+S	A/Cu	✓	Dice+CE	✗	✗
`Maastro-Deep-Learning` (Rao et al., 2021)	2D/3D	C	A/Cu	✗	Top-K	✓	✓
`Our baseline 3D PET/CT`	3D	C+S	I/Cu	✗	Dice+CE	✗	✗
`Our baseline 2D PET/CT`	2D	C+S	I/Cu	✗	Dice+CE	✗	✗

including model architecture, preprocessing, training scheme and postprocessing. We only report the methods of the participants with an associated publication, which was crucial to ensure the scientific relevance of the challenge. More details on the individual methods can be found in Appendix C.1 as well as in the corresponding participants' papers (H. Chen et al., 2021; Ghimire et al., 2021; Iantsen, Visvikis, et al., 2021; J. Ma & Yang, 2021; M. Naser et al., 2021; Rao et al., 2021; J. Xie & Peng, 2021; Yousefirizi & Rahmim, 2021; Yuan, 2021; Zhu et al., 2021). In the results section (4.4), we also include results of the participants without publication for comparison.

All the participants used a U-Net-based architecture. Eight used 3D architectures, one used a 2D architecture and one used a combination of the two. All participants used some sort of preprocessing prior to training their model, generally with standard data augmentation (except for three participants), using various combinations of losses, most often including the Dice loss. The participants used various cross-validation schemes to optimize the generalization performance of their models. Half of the participants used an ensemble of multiple models.

4.4.3 Segmentation Performance

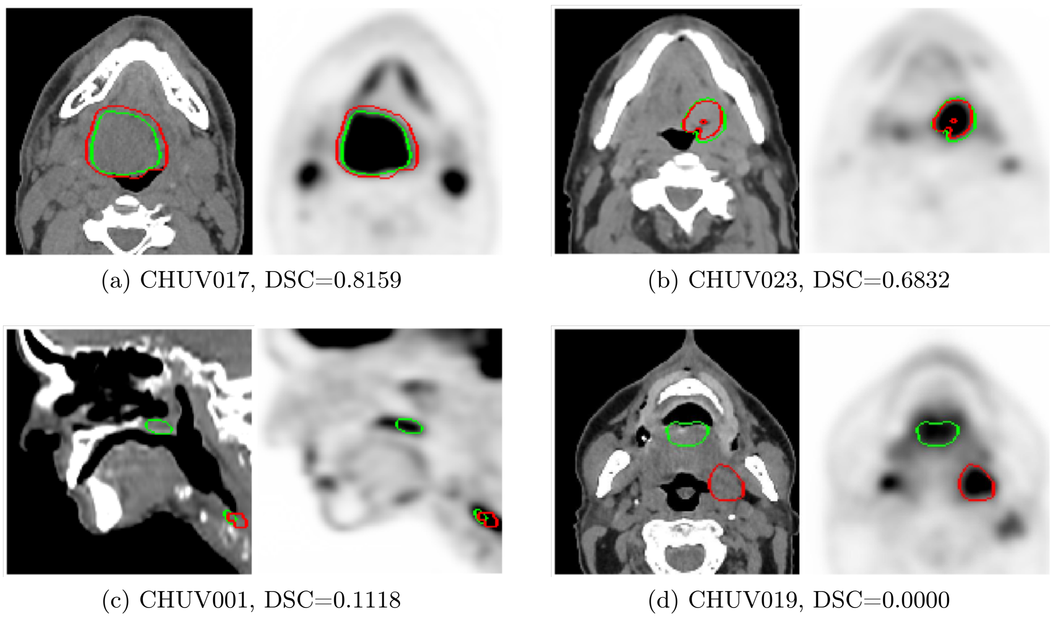
The results, including average DSC, precision, recall, and challenge rank are summarized in Table 4.3. We also report the average Surface Dice SCore at 1mm (SDSC) and the median Hausdorff Distance at 95% (HD95) as defined in (Nikolov et al., 2021). Our baseline method, developed in (Andrearczyk, Oreiller, Vallières, Castelli, et al., 2020) and provided to participants as an example on our GitHub repository, obtains an average

Table 4.3: Summary of the challenge results as of April 2021. The average DSC, precision, recall, SDSC and median HD95 are reported for the baseline algorithms and every team (the best result of each team). The unit of the HD95 is [mm]. The participant names are reported when no team name was provided. The ranking is only provided for teams that presented their method in a paper submission. The post-challenge results are denoted by an asterisk *. Bold values represent the best scores for each metric, excluding post-challenge results since we do not have any information about their method.

Team	DSC	HD95	Precision	Recall	SDSC	Rank
paar*	0.7624	3.27	0.8304	0.7490	0.6167	-
andrei.iantsen (Iantsen, Visvikis, et al., 2021)	**0.7591**	**3.27**	0.8333	0.7400	**0.6010**	1
junma (J. Ma & Yang, 2021)	0.7525	**3.27**	0.8384	0.7174	0.6003	2
Fuller_MDA_Lab*	0.7523	3.27	0.7838	0.7685	0.6168	-
supratik_bose*	0.7440	3.27	0.8350	0.7085	0.5822	-
badger*	0.7377	3.27	0.8143	0.7160	0.5800	-
badger (J. Xie & Peng, 2021)	0.7355	**3.27**	0.8326	0.7024	0.5735	3
deepX (Yuan, 2021)	0.7318	3.54	0.7851	0.7319	0.5528	4
flash*	0.7280	3.54	0.8020	0.7083	0.5650	-
AIView_sjtu (H. Chen et al., 2021)	0.7241	3.33	**0.8479**	0.6701	0.5598	5
DCPT	0.7049	4.10	0.7651	0.7047	0.5562	-
xuefeng (Ghimire et al., 2021)	0.6911	5.06	0.7525	0.6928	0.5011	6
ucl_charp	0.6765	5.42	0.7231	0.7257	0.5194	-
QuritLab (Yousefirizi & Rahmim, 2021)	0.6677	5.64	0.7289	0.7164	0.5086	7
Unipa	0.6674	4.10	0.7143	0.7039	0.4902	-
Baseline 3D PET/CT	0.6610	21.88	0.5909	**0.8534**	0.4502	-
Baseline 2D PET/CT	0.6588	26.81	0.6242	0.7629	0.4796	-
HFHSegTeam (Zhu et al., 2021)	0.6441	14.27	0.6938	0.6670	0.4922	8
UESTC_501	0.6382	5.16	0.6455	0.6874	0.4339	-
Fuller_MDA_Lab (M. Naser et al., 2021)	0.6373	5.06	0.7546	0.6283	0.4730	9
Yone*	0.6341	5.92	0.7690	0.6640	0.4513	-
Baseline 3D PET	0.6306	24.95	0.5768	0.8214	0.4399	-
Baseline 2D PET	0.6284	27.62	0.6470	0.6666	0.4231	-
Maastro-Deep-L. (Rao et al., 2021)	0.5874	29.56	0.6560	0.6142	0.4118	10
Yone	0.5737	21.46	0.6606	0.5590	0.4216	-
SC_109	0.5633	5.64	0.7652	0.5022	0.3542	-
Roque	0.5606	14.94	0.5850	0.6843	0.3601	-
Baseline 2D CT	0.3071	27.54	0.3477	0.3574	0.1847	-
Baseline 3D CT	0.2729	32.02	0.2154	0.5874	0.1218	-

DSC of 0.6588 and 0.6610 with the 2D and 3D implementations, respectively. Results on individual modalities are also reported for comparison.

The results from the participants (excluding post-challenge submissions) range from an average DSC of 0.5606 to 0.7591. Iantsen, Visvikis, et al., 2021 (participant andrei.iantsen) obtained the best overall results with an average DSC of 0.7591, an average precision of 0.8332 and an average recall of 0.7400. This result (DSC) is not significantly higher than the second-best participant (J. Ma & Yang, 2021) (p-value of 0.3517 with a one-tailed Wilcoxon test). The statistical comparison of the score of each team is done in Figure C.1 with the one-tailed Wilcoxon test and corrected for multiple hypotheses testing. Across all participants, the average precision ranges from 0.5850 to 0.8479. The recall ranges from 0.5022 to 0.8534, with the latter surprisingly obtained by the 3D PET/CT baseline


(a) CHUV017, DSC=0.8159

(b) CHUV023, DSC=0.6832

(c) CHUV001, DSC=0.1118

(d) CHUV019, DSC=0.0000

Figure 4.2: Examples of results of the winning algorithm (`andrei.iantsen` (Iantsen, Visvikis, et al., 2021)). The automatic segmentation results (green) and ground truth annotations (red) are displayed on 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the entire image (see Eq. 4.1). (a), (b) Excellent segmentation results, detecting the GTVt of the primary oropharyngeal tumor localized at the base of the tongue and discarding the laterocervical lymph nodes despite high FDG uptake on PET. (c) Incorrect segmentation of the top volume at the level of the soft palate; (d) Incorrect segmentation of the smaller volume below the level of the hyoid bone.

(although with low precision, reflecting a trend to over-segment as compared to other algorithms). The median HD95 ranges from 3.27 to 32.02 [mm]. We chose to report the median since a value of $+\infty$ is attributed when the prediction is null. 3.27 [mm] is a highly observed value for HD95, which is probably due to the coarse axial resolution of the CT on the test set as we computed the performance in the original CT resolution (see C.1).

Note that two participants decided to withdraw their submissions due to very low scores. We allowed them to do so since their low scores were due to incorrect post-processing (*e.g.* setting incorrect pixel spacing or image origin), which was not representative of the performance of their algorithms. The distributions of DSCs across patients and across participants are reported in Figures. 4.3 and 4.4 respectively. Examples of segmentation results (TPs on top row, and FPs on bottom row) are shown in Fig. 4.2.

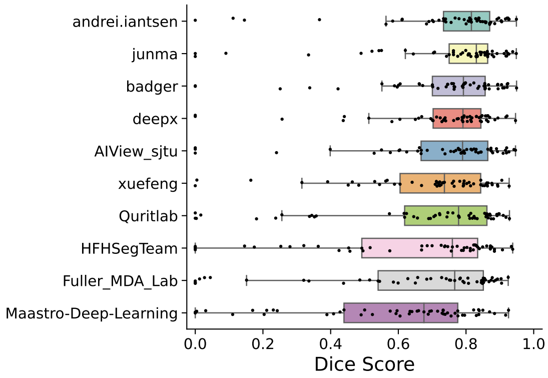


Figure 4.3: Box plots of the distribution of the 53 test DSCs for each participant, ordered by decreasing rank.

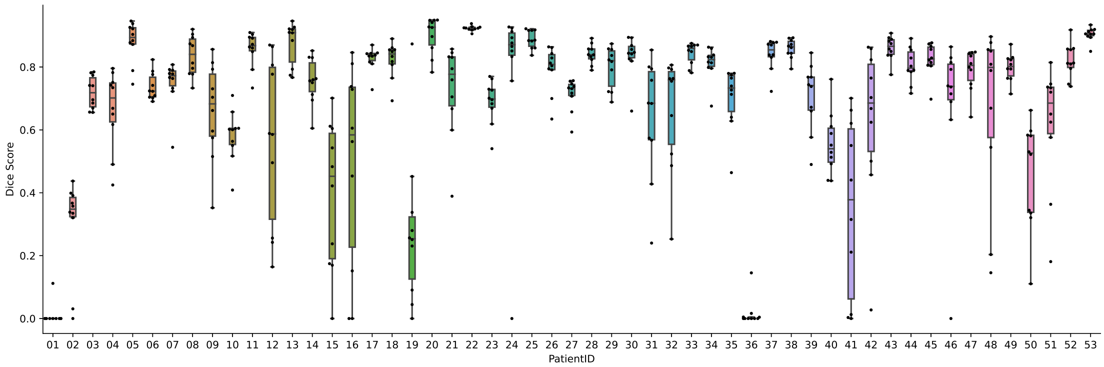


Figure 4.4: Box plots of the distribution of DSCs across the 10 participants for each of the 53 patients in the test set.

4.4.4 Inter-Observer Agreement

We realized that it was crucial to also define the baseline for human observers performing the GTVt delineation task (*i.e.* segmentation), as well as their agreement. Three observers, *i.e.* two experts in radiation oncology and one nuclear physician, annotated the same 21 cases drawn randomly from the training and test sets and coming from all five centers. These 21 cases were chosen to represent approximately 10 % of the dataset. It is worth noting that annotating the entire dataset four times was too costly. They were asked to delineate as close as possible the true tumoral volume as the aim is for radiomics studies. Together with the official challenge delineations, it amounts to four observers. All unique pairs of observers were considered, resulting in six pairs of comparisons. We computed the average DSC of all the pairs, *i.e.* all possible pairs of the four observers, which resulted in an average DSC of 0.6110. It is worth noting that for a faithful delineation of the tumor, a contrast-enhanced CT or an MRI image is required. Furthermore, there are no clinical guidelines for the task of segmenting GTVt on PET/CT fusion. Moreover, the clinical information (*e.g.* physical examination) brings essential information to decide whether an abnormal structure is malignant. In this agreement, the observers were asked to perform this task with the PET/CT images only. Similar agreements were reported in the literature. Gudi et al., 2017 reported the agreement of three observers with an average DSC of 0.57 using only the CT images for annotation and 0.69 using both PET and CT.

4.4.5 Ensemble of Participants

In this section, we evaluate the possibility to ensemble the different participants' results into a "super-algorithm". Such analyses often revealed superior performances to all submitted runs (Menze et al., 2014), leveraging the diversity of the different methods (Hastie et al., 2009). We ensemble the (binary) predictions of all participants (with paper submissions, *i.e.* 10 participants) using the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm (Warfield et al., 2004). This ensemble of predictions obtains an average DSC of 0.7574, a precision of 0.7301, and a recall of 0.8439. This result is better than the average performance of all participants (0.6931) and is slightly, but not significantly, outperformed by the best score of 0.7591 (p-value= 0.9230). A simpler ensembling method is computed by taking the average of the 10 teams for each patient, and then, thresholding to 0.5 to obtain a binary prediction. This average prediction scores a DSC of 0.7426 which is not as good as the STAPLE ensembling (p-value= 0.044). Note that several participants already reported results obtained as an ensemble of multiple independent network predictions. (see Table 4.2).

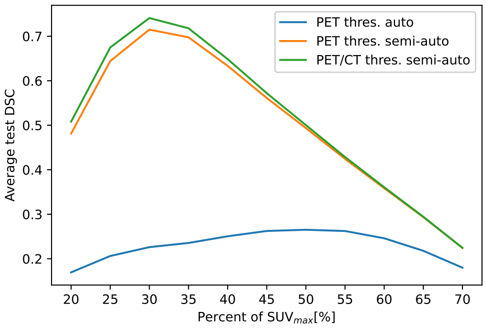


Figure 4.5: Segmentation performance of PET thresholding-based method at different percentages of maximum SUV. Three results are reported: the automatic PET threshold, the semi-automatic PET threshold (indicating the location of the ground truth GTVt), and the semi-automatic PET and CT (for removing the air) threshold. Best result at 0.3: DSC 0.7491, 0.8179 0.7672

4.4.6 PET Thresholding

PET thresholding is *de facto* the most widely used method for lesion segmentation, at least in clinical routine, often via an initial manual delineation of the field of interest. As a comparison to the results obtained by the participants using deep learning automatic segmentation algorithms, we evaluate simple PET thresholding methods (automatic and semi-automatic). For the fully automatic threshold method, we simply threshold the PET image at a given percentage of the maximum SUV value within the bounding box.

For the semi-automatic threshold method, we mimic a manual indication of the GTVt followed by a threshold of the PET values. To this end, we threshold the PET image, compute the 26-connected components and retain the component that overlaps with the ground truth GTVt (or multiple components if more than one overlap with the ground truth GTVt). In Fig. 4.5, we report the results of both methods on the test set when varying the percentage of the maximum SUV used for thresholding. Finally, we also evaluate the same semi-automatic thresholding method with an additional threshold on the CT images (at -150 HU) to remove the air from the predictions. The best results, with an average DSC of 0.7409, are obtained with this semi-automatic PET/CT threshold at 30% of the maximum SUV value, which is aligned with previous findings, including in the context of the identification of predictive biomarkers (Castelli et al., 2017).

4.4.7 Tumor Size and Segmentation Performance

In this section, we evaluate how the algorithms perform for different tumor sizes. To this end, we explore the correlation of tumor size with the performance of the algorithms. The tumor size is calculated as the voxel count inside the ground truth GTVt multiplied

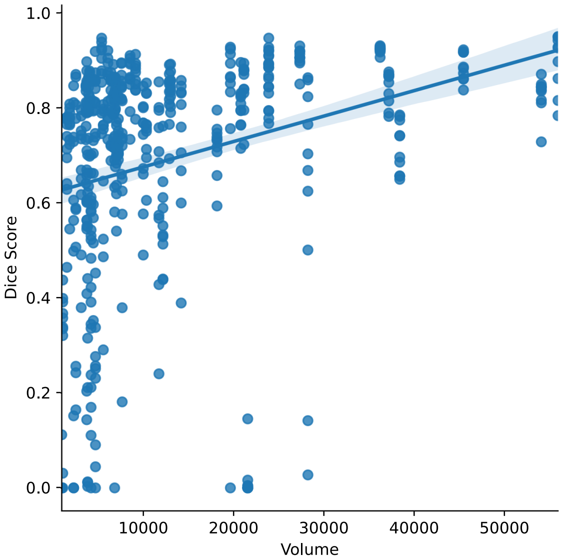


Figure 4.6: Scatter plot of DSC vs. tumor volume (voxel count in the VOI) for 10 participants. The corresponding Spearman correlation is 0.43.

by the voxel volume. The Spearman correlation across all ten participants and all tumors is 0.4301 (p-value$<$ 0.001). In Fig. 4.6, we illustrate this correlation with a scatter plot of the DSC as a function of tumor size. Fig. 4.7 relates the performance for each of the 10 algorithms for four tumor size groups. This figure was generated by grouping the 53 test cases in 4 bins (*i.e.* intervals) of 13, 13, 13, and 14 cases, respectively. The average DSC was then computed for each team in each bin.

4.4.8 Analysis of False Positives

In this section, we want to evaluate, for a given algorithm, whether FPs are generally occurring in the surroundings of the ground truth GTVt, or biased towards other regions with high FDG uptakes such as the lymph nodes or other zones with inflammation. To this end, we compute the shortest Euclidean distance of each FP voxel to the ground truth GTVt. We then aggregate these distances for all test cases and report these values into a histogram in Fig. 4.8. Similarly, we compute the distance of each FP voxel to the ground truth GTVn (lymph nodes) and report the histogram on the same figure. We compute this analysis for the best participant (`andrei.iantsen`), as well as for the baseline (3D PET/CT U-Net) since it was the approach with the largest recall but low precision. Note that we only compute the histogram of the FP voxels to avoid squashing the counts of the non-zero bins due to the large number of TPs with a distance to the GTVt of zero (first bin).

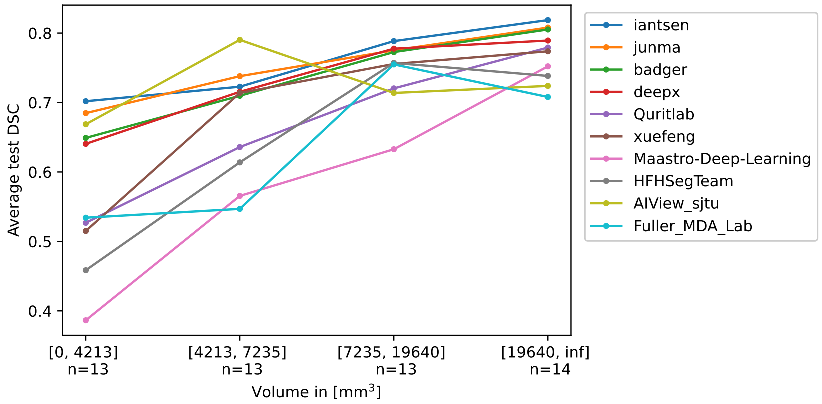
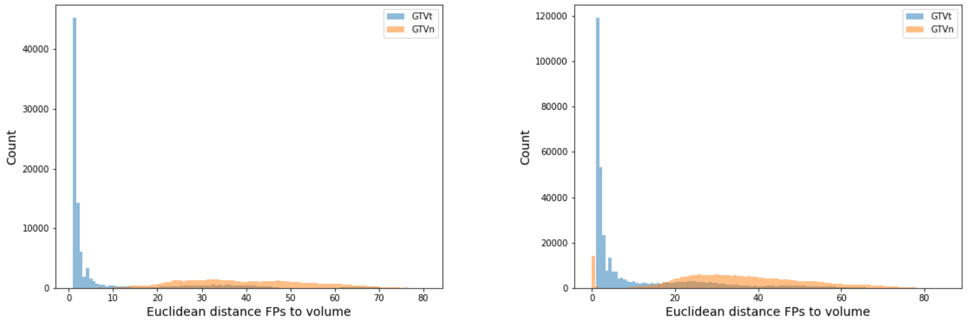


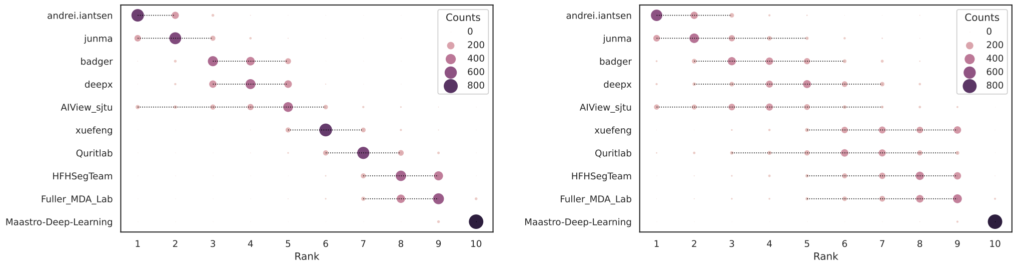
Figure 4.7: Average DSC of each team's algorithm in function of the volume of the tumors. This figure was generated by distributing the 53 test volumes in 4 bins of n =13, 13, 13, and 14 each and then computing the average DSC for each bin.



(a) `andrei.iantsen`

(b) baseline 3D PET/CT

Figure 4.8: Histogram of the Euclidean distance of the FP voxels to the closest ground truth GTVt voxel and GTVn voxel. We evaluate here the prediction of the first ranked participant (`andrei.iantsen`) (a) and our baseline 3D PET/CT (b). For comparison, the False Discovery Rate (FDR), i.e. FP/(FP+TP) is 0.15, with 544,343 TPs in (a) and FDR = 0.37 with 621,413 TPs in (b).



(a) Rank based on average DSC (b) (Alternative) rank based on average rank DSC

Figure 4.9: Ranking robustness against changes in test data. The robustness is assessed by ranking 1000 bootstraps of the test set. The size of the circles is proportional to the number of times a team obtained the corresponding rank for each bootstrap. The dashed lines represent the confidence intervals at 95 % computed from the bootstrap analysis. The current ranking, *i.e.* the one used in this challenge, is obtained by averaging the DSCs across all test cases. The alternative ranking is computed by averaging the rankings of each team across the test cases.

4.4.9 Ranking Robustness

Ranking robustness against changes in the test set is assessed by evaluating the variation of the ranking on 1000 bootstrap repetitions of the test set. We also compared the current ranking against an alternate ranking defined as follows. This alternative ranking was computed based on the average ranking across all cases. If multiple teams obtain the same rank for one case, the average rank is attributed to these teams. For instance, if three participants score 0 on a given case, the average rank of $\frac{8+9+10}{10} = 9$ is attributed to all of them for this case.

Fig. 4.9 depicts the results of the bootstrap analysis for the two rankings. We also computed the Kendall rank correlation coefficient between the ranking of each bootstrap and the ranking on the whole test set. We obtained 0.8772 (0.7333 - 1.0000) and 0.7335 (0.4658 - 0.9111) for the current ranking and alternate ranking, respectively. The numbers in parenthesis are the confidence intervals at 95 % computed with the bootstrap eanalysis. The methodology used in this section to report ranking robustness is inspired by the challengeR toolkit (Wiesenfarth et al., 2021).

4.5 Discussion

This section interprets and discusses the results reported in Section 4.4. We first discuss and report the overall challenge participation and main lessons learned. Second, the segmentation performance achieved by all participating methods is interpreted. Finally, we report the current limitations and sources of errors of this challenge.

4.5.1 Participation and Main Lessons Learned

This challenge allowed us to compare state-of-the-art algorithms developed by 18 teams across the world on the task of primary H&N tumor segmentation in PET/CT images. Excellent results were obtained with the first ranked team reaching 0.7591 average DSC, 0.8332 precision, and 0.7400 recall. In Table 4.2, we attempted to group the results based on important elements of the algorithms. In particular, we identified several elements important for addressing the task. All participants used U-Net based architectures, mostly 3D. Preprocessing, normalization, data augmentation, and ensembling seem to play an important role in the final results. Most of these trends (see also algorithms description in Section 4.4.2) and results can be found in other medical imaging segmentation challenges (J. Ma, 2021; Menze et al., 2014). An interesting comparison of several challenges (including HECKTOR 2020) and algorithms focusing on automatic segmentation in medical images can be found in (J. Ma, 2021).

We note, however, that it is a difficult task to characterize algorithms with only a few descriptions and to assign good performance to specific parts. The methods are highly complex with high degrees of freedom and many hyper-parameters that can all have a strong influence on segmentation performance. Simple modifications such as the number of training iterations or the learning rate can have a large impact on the results and cannot be exhaustively listed and compared. For this analysis, we asked the participants to specifically report a set of characteristics of their algorithms to be able to compare them in Table 4.2. More information will be asked in the future editions of HECKTOR to enhance comparison.

The ranking used in this edition was based on the average DSC. The results of Section 4.4.9 show that this approach is more robust to changes in the test set. These findings are corroborated by Maier-Hein et al., 2018 where they showed that ranking based on averaged metrics are more consistent for changes in test data.

4.5.2 Overall Segmentation Performance

As shown in Fig. 4.4, some cases were incorrectly segmented by most or all participants, *e.g.* CHUV01 and CHUV36. On the contrary, some cases were correctly segmented by most participants (*e.g.* CHUV22 and CHUV53), and others showed a large variability across participants' algorithms (*e.g.* CHUV16 and CHUV41). These differences, as confirmed by further evaluations in Sections 4.4.8, 4.4.7, originate from the tumor size, the SUVs within the GTVt, and the presence of lymph nodes or other regions with high SUVs. Some examples are illustrated in Fig. 4.2.

The participants' algorithms obtained better results than the inter-observer agreement. This comparison, however, should be put into perspective. First, the cases used in the agreement were different from the test set. Second, one annotator, the one who annotated the entire dataset for the challenge, had extra information since he corrected the radiotherapy annotations whereas the others were asked to draw the segmentation from scratch without any further information than the raw PET/CT data. Finally, some annotators delineated closer to radiotherapy requirements, *i.e.* with large annotations, resulting in higher disagreement. To alleviate this issue, we are currently developing clear guidelines for the next iteration of the challenge.

The results can also be compared with a simple PET thresholding method (see Section 4.4.6), often used in radiomics studies (Castelli et al., 2017; Erdi et al., 1997). The latter obtained an average DSC of 0.7409 when used in a semi-automatic manner. This result is significantly lower than the performance of the best participants (0.7591, p-value of 0.0237) and must be considered with precaution since the segmentation was highly guided toward the true tumor location and the threshold was optimized on the test set. With a fully automatic threshold of the PET image in the oropharynx region, we only obtain 0.2652 due to various regions, including lymph nodes, with high SUVs. The best semi-automatic threshold method was obtained with a threshold around 30% of the maximum SUV, as frequently used to measure the metabolic response characteristic of the tumor, *e.g.* 36-44% for best approximation of tumor volume (Erdi et al., 1997), 40 to 68% of SUV max for best radiomics results in DFS prediction (Castelli et al., 2017; Creff et al., 2020). Overall, this suggests that the segmentation algorithms can leverage the wealth of both PET and CT images (*i.e.* metabolic and anatomical/structural tumor properties) to provide more advanced segmentation rules when compared to simple PET thresholding. This is also corroborated by the consistent superiority of algorithms using both PET and CT imaging modalities when compared to using PET only.

The ensemble of participants' methods (see Section 4.4.5) reached a good consensus with an average DSC of 0.7574 and a rather high recall (0.8438) and low precision (0.7301) as compared to other results in the same range. While this is not better than the first rank result, it would likely achieve an excellent generalization to other data.

4.5.3 Detailed Performance Analysis

The analysis of tumor sizes in Section 4.4.7 (Figs. 4.6 and 4.7) showed that they are correlated with the segmentation performance. These results seem to show that the small tumor sizes are more difficult to segment than the large ones. More precisely, smaller tumors are less consistently well segmented, resulting in a large variation of performance. This is not surprising since small lesions suffer from a higher partial volume effect which increases the relative difficulty to define the boundary of the tumor (Foster et al., 2014). Moreover, the volumetric (or 3D) DSC is largely dependent on the volume sizes. A contour deviation of ±1mm around the true tumor boundary, for instance, will affect DSC values more for small tumors than the large ones, resulting in a negligible chance for the latter.

In Fig. 4.8 (Section 4.4.8), we analyzed the spatial arrangement of FPs segmented voxels. We conducted this experiment for the first ranked results and our baseline. In both cases, the majority of FP voxels are located in the surrounding of the GTVt, as shown in Fig. 4.8. As illustrated in the same figure, the FPs of the best results are not located near the lymph nodes, whereas a lot of FPs of the baseline are located in the lymph nodes and their surroundings. This suggests that, unlike the baseline, the best algorithm relies on true tumoral patterns and not only on FDG uptake.

4.5.4 Limitations and Sources of Errors

The main limitation of the current challenge is the lack of more precise GTVt ground truth. The annotations were made on the PET/CT fusion without using other modalities such as contrast-enhanced CT or MR which allow delineating the tumor more faithfully. This limitation is illustrated by the results of the inter-observer agreement mentioned in Section 4.4.4, where the average DSC of 0.6110 highlighted the difficulty of the task. A source of error, therefore, originates from the degree of subjectivity and the lack of guidelines in the annotation and correction of the expert.

Another limitation of this challenge is the lack of test data with exact ground truth. To obtain such data, phantom and simulation can be used. This enables the evaluation of performances of models on data where the exact ground truth is known. Hatt et al., 2017 claim that for a good benchmark in PET segmentation, one must include simulated and phantom test images in addition to clinical test data.

In this challenge, we provided the participants with a bounding box to decrease the difficulty of the task. This can be seen as a limitation since the resulting methods are not fully automatic, but these bounding boxes cover a large portion of the original image and are easy to detect automatically (Andrearczyk, Oreiller, & Depeursinge, 2020).

4.6 Conclusions

This paper presents the HECKTOR 2020 challenge on the segmentation of the primary tumor of oropharyngeal H&N cancer in FDG PET/CT. Detailed information was reported on the dataset, participation, and segmentation performance. Good participation with 18 teams and 10 participants' publications allowed us to compare state-of-the-art segmentation methods on this challenging task. The results are very satisfactory with the winning team achieving an average DSC of 0.7591, which is superior to the inter-observer agreement (average DSC 0.6110). These results were obtained with a strict testing scheme as the test cases were all from an unseen center. It is reasonable to expect better results if the proposed methods are fine-tuned on few examples from this center. All participants used U-Net based deep learning models, most of them with a 3D architecture and standard pre-processing techniques. We could identify several key elements that seem to have led to good results, including normalization, data augmentation, and ensembling of multiple models.

Preliminary experiments show that fully automatic radiomics methods are on pair or surpass radiomics models based on feature extraction from manual annotations (Andrearczyk, Fontaine, et al., 2021; Fontaine et al., 2021). These preliminary results are very encouraging and demonstrate that we are one step closer to analyzing very large-scale cohorts for radiomics validation.

While focusing on H&N cancer in HECKTOR, we believe that many of the methods developed and lessons learned will generalize to the automatic segmentation of other types of cancer imaged in PET/CT images (*e.g.* lung, melanoma).

In future editions, we aim to increase the size of the dataset and propose other clinically relevant tasks such as the segmentation of lymph nodes and the prediction of patient outcome (*e.g.* disease-free survival).

Acknowledgments

We thank all the teams for their valuable work. This challenge and the winner prize were kindly sponsored by Siemens Healthineers Switzerland. This work was also partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069) and the Swiss Personalized Health Network (SPHN, via the IMAGINE and QA4IQI projects).

Author contributions

Valentin Oreiller and Vincent Andrearczyk:
Design of the task and of the challenge, organization of the challenge, development of baseline algorithms and post-challenge analyses, writing of the paper.

Mario Jreige: Design of the task and of the challenge, organization of the challenge, quality control/annotations, annotations for inter-annotator agreement.

Martin Vallières:
Design of the task and of the challenge, provided the initial data and annotations for the training set (Vallieres et al., 2017), revision of the paper.

Joel Castelli, Hesham Elhalawani and Sarah Boughdad:
Design of the task and of the challenge, annotations for inter-annotator agreement.

Simeng Zhu, Juanying Xie, Ying Peng, Andrei Iantsen, Mathieu Hatt, Yading Yuan, Jun Ma, Xiaoping Yang, Chinmay Rao, Suraj Pai, Kanchan Ghimire, Xue Feng, Mohamed A.

Naser, Clifton D. Fuller, Fereshteh Yousefirizi, Arman Rahmim, Huai Chen and Lisheng Wang:
Participation to the challenge, publication of their method, revision of the paper.

John O. Prior:
Design of the task and of the challenge, organization of the challenge, revision of the paper.

Adrien Depeursinge:
Design of the task and of the challenge, organization of the challenge, development of baseline algorithms and post-challenge analyses, writing of the paper.

5 Overview of HECKTOR 2021

The paper presented in this chapter is:

- Vincent Andrearczyk*, Valentin Oreiller*, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt†, and Adrien Depeursinge†. "Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images." In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pp. 1-37. Springer, Cham, 2021.

* both authors contributed equally.
† both authors contributed equally.

Abstract

This paper presents an overview of the second edition of the HEad and neCK TumOR (HECKTOR) challenge, organized as a satellite event of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021. The challenge is composed of three tasks related to the automatic analysis of PET/CT images for patients with Head and Neck cancer (H&N), focusing on the oropharynx region. *Task 1* is the automatic segmentation of H&N primary Gross Tumor Volume (GTVt) in FDG-PET/CT images. *Task 2* is the automatic prediction of Progression Free Survival (PFS) from the same FDG-PET/CT. Finally, *Task 3* is the same as Task 2 with ground truth GTVt annotations provided to the participants. The data were collected from six centers for a total of 325 images, split into 224 training and 101 testing cases. The interest in the challenge was highlighted by the important participation with 103 registered teams and 448 result submissions. The best methods obtained a Dice Similarity Coefficient (DSC) of 0.7591 in the first task, and a Concordance index (C-index) of 0.7196 and 0.6978 in Tasks 2 and 3, respectively. In all tasks, simplicity of the approach was found to be key

to ensure generalization performance. The comparison of the PFS prediction performance in Tasks 2 and 3 suggests that providing the GTVt contour was not crucial to achieve best results, which indicates that fully automatic methods can be used. This potentially obviates the need for GTVt contouring, opening avenues for reproducible and large scale radiomics studies including thousands potential subjects.

5.1 Introduction: Research Context

The prediction of disease characteristics and outcomes using quantitative image biomarkers from medical images (i.e. radiomics) has shown tremendous potential to optimize and personalize patient care, particularly in the context of Head and Neck (H&N) tumors (Vallieres et al., 2017). FluoroDeoxyGlucose (FDG)-Positron Emission Tomography (PET) and Computed Tomography (CT) imaging are the modalities of choice for the initial staging and follow-up of H&N cancer, as well as for radiotherapy planning purposes. Yet, both gross tumor volume (GTV) delineations in radiotherapy planning and radiomics analyses aiming at predicting outcome rely on an expensive and error-prone manual or semi-automatic annotation process of Volumes of Interest (VOI) in three dimensions. The fully automatic segmentation of H&N tumors from FDG-PET/CT images could therefore enable faster and more reproducible GTV definition as well as the validation of radiomics models on very large cohorts. Besides, fully automatic segmentation algorithms could also facilitate the application of validated models to patients' images in routine clinical workflows. By focusing on metabolic and morphological tissue properties respectively, PET and CT images provide complementary and synergistic information for cancerous lesion segmentation and patient outcome prediction. The HEad and neCK TumOR segmentation and outcome prediction from PET/CT images (HECKTOR)[I] challenge aimed at identifying the best methods to leverage the rich bi-modal information in the context of H&N primary tumor segmentation and outcome prediction. The analysis of the results provides precious information on the adequacy of the image analysis methods for the different tasks and the feasibility of large-scale and reproducible radiomics studies.

The potential of PET information for automatically segmenting tumors has been long exploited in the literature. For an in-depth review of automatic segmentation of PET images in the pre-deep learning era, see (Foster et al., 2014; Hatt et al., 2017) covering methods such as fixed or adaptive thresholding, active contours, statistical learning, and mixture models. The need for a standardized evaluation of PET automatic segmentation methods and a comparison study between all the current algorithms was highlighted in (Hatt et al., 2017). The first challenge on tumor segmentation in PET images was proposed at MICCAI 2016[II] by Hatt, Laurent, Ouahabi, Fayad, Tan, Li, Lu, Jaouen, Tauber, Czakon, et al., 2018b, implementing evaluation recommendations published previously by the AAPM (American Association of Physicists in Medicine) Task group

[I] https://www.aicrowd.com/challenges/miccai-2021-hecktor, as of October 2021.

[II] https://portal.fli-iam.irisa.fr/petseg-challenge/overview#_ftn1, as of October 2020.

211 (Hatt et al., 2017). Multi-modal analyses of PET and CT images have also recently been proposed for different tasks, including lung cancer segmentation in (A. Kumar et al., 2019; Li et al., 2019; X. Zhao et al., 2018; Zhong et al., 2018) and bone lesion detection in (Xu et al., 2018). In (Andrearczyk, Oreiller, Vallières, Castelli, et al., 2020), we developed a baseline Convolutional Neural Network (CNN) approach based on a leave-one-center-out cross-validation on the training data of the HECKTOR challenge. Promising results were obtained with limitations that motivated additional data curation, data cleaning and the creation of the first HECKTOR challenge in 2020 (Andrearczyk, Oreiller, Jreige, et al., 2020; Oreiller et al., 2021). This first edition compared segmentation architectures as well as the complementarity of the two modalities for the segmentation of GTVt in H&N.

In this second edition of the challenge, we propose a larger dataset, including a new center to better evaluate the generalization of the algorithms, as well as new tasks of prediction of Progression-Free Survival (PFS). Preliminary studies of automatic PFS prediction were performed with standard radiomics (Fontaine et al., 2021) and deep learning models (Andrearczyk, Fontaine, et al., 2021) prior to challenge design. The proposed dataset comprises data from six centers. Five centers are used for the training data and two for testing (data from the sixth center are split between training and testing sets). The task is challenging due to, among others, the variation in image acquisition and quality across centers (the test set contains data from a domain not represented in the training set) and the presence of lymph nodes with high metabolic responses in the PET images.

The critical consequences of the lack of quality control in challenge designs were shown in (Maier-Hein et al., 2018) including reproducibility and interpretation of the results often hampered by the lack of provided relevant information and the use of non-robust ranking of algorithms. Solutions were proposed in the form of the Biomedical Image Analysis challengeS (BIAS) (Maier-Hein et al., 2020) guidelines for reporting the results. This paper presents an overview of the challenge following these guidelines.

Individual participants' papers reporting their methods and results were submitted to the challenge organizers. Reviews were organized by the organizers and the papers of the participants are published in the LNCS challenges proceedings (An et al., 2022; Bourigault et al., 2022; Cho et al., 2022; De Biase et al., 2022; Fatan et al., 2022; Ghimire et al., 2022; Huynh et al., 2022; Juanco-Müller et al., 2022; Lang et al., 2022; J. Lee et al., 2022; T. Liu et al., 2022; Lu et al., 2022; B. Ma et al., 2022; Martinez-Larraz et al., 2022; Meng et al., 2022; Murugesan et al., 2022; M. A. Naser, Wahid, Mohamed, et al., 2022; M. A. Naser, Wahid, van Dijk, et al., 2022; Paeenafrakati et al., 2022; Qayyum et al., 2022; Ren et al., 2022; Saeed et al., 2022; Starke et al., 2022; Wahid et al., 2022; G. Wang et al., 2022; J. Wang et al., 2022; J. Xie & Peng, 2022; Yousefirizi et al., 2022; Yuan et al., 2022). When participating in multiple tasks, participants could submit one or multiple papers.

The manuscript is organized as follows. The challenge dataset is described in Section 5.2. The tasks descriptions, including challenge design, algorithms summaries and results, are split into two sections. The segmentation task (Task 1) is presented in Section 5.3, and the outcome prediction tasks (Tasks 2 and 3) are described in Section 5.4 . Section 5.5 discusses the results and findings and Section 5.6 concludes the paper.

5.2 Dataset

5.2.1 Mission of the Challenge

Biomedical application

The participating algorithms target the following fields of application: diagnosis, prognosis and research. The participating teams' algorithms were designed for either or both image segmentation (i.e., classifying voxels as either primary tumor or background) and PFS prediction (i.e., ranking patients according to a predicted risk of progression).

Cohorts

As suggested in (Maier-Hein et al., 2020), we refer to the patients from whom the image data were acquired as the challenge cohort. The target cohort[III] comprises patients received for initial staging of H&N cancer. The clinical goals are two-fold; the automatically segmented regions can be used as a basis for (i) treatment planning in radiotherapy, (ii) further radiomics studies to predict clinical outcomes such as overall patient survival, disease-free survival, response to therapy or tumor aggressiveness. Note that the PFS outcome prediction task does not necessarily have to rely on the output of the segmentation task. In the former case (i), the regions will need to be further refined or extended for optimal dose delivery and control. The challenge cohort[IV] includes patients with histologically proven H&N cancer who underwent radiotherapy treatment planning. The data were acquired from six centers (four for the training, one for the testing, and one for both) with variations in the scanner manufacturers and acquisition protocols. The data contain PET and CT imaging modalities as well as clinical information including age, sex, acquisition center, TNM staging, HPV status and alcohol. A detailed description of the annotations is provided in Section 5.2.2.

Target entity

The data origin, i.e. the region from which the image data were acquired, varied from the head region only to the whole body. While we provided the data as acquired, we

[III] The target cohort refers to the subjects from whom the data would be acquired in the final biomedical application. It is mentioned for additional information as suggested in BIAS, although all data provided for the challenge are part of the challenge cohort.

[IV] The challenge cohort refers to the subjects from whom the challenge data were acquired.

Table 5.1: List of the hospital centers in Canada (CA), Switzerland (CH) and France (FR) and number of cases, with a total of 224 training and 101 test cases.

Center	Split	# cases
HGJ: Hôpital Général Juif, Montréal, CA	Train	55
CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke, CA	Train	72
HMR: Hôpital Maisonneuve-Rosemont, Montréal, CA	Train	18
CHUM: Centre Hospitalier de l'Université de Montréal, Montréal, CA	Train	56
CHUP: Centre Hospitalier Universitaire Poitiers, FR	Train	23
Total	Train	224
CHUV: Centre Hospitalier Universitaire Vaudois, CH	Test	53
CHUP: Centre Hospitalier Universitaire Poitiers, FR	Test	48
Total	Test	101

Table 5.2: List of scanners used in the various centers.

Center	Device
HGJ	hybrid PET/CT scanner (Discovery ST, GE Healthcare)
CHUS	hybrid PET/CT scanner (Gemini GXL 16, Philips)
HMR	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUM	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUV	hybrid PET/CT scanner (Discovery D690 ToF, GE Healthcare)
CHUP	hybrid PET/CT scanner (Biograph mCT 40 ToF, Siemens)

limited the analysis to the oropharynx region and provided a semi-automatically detected bounding-box locating the oropharynx region (Andrearczyk, Oreiller, & Depeursinge, 2020), as illustrated in Fig. 5.1. The participants could use the entire images if they wished, but the predictions were evaluated only within these bounding-boxes.

5.2.2 Challenge Dataset

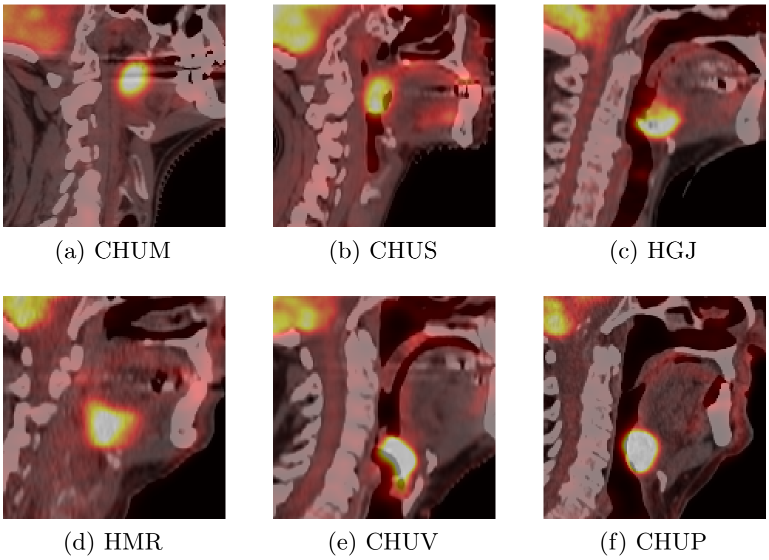
Data source

The data were acquired from six centers as detailed in Table 5.1. It consists of PET/CT images of patients with H&N cancer located in the oropharynx region. The devices and imaging protocols used to acquire the data are described in Table 5.2. Additional information about the image acquisition is provided in Appendix D.2.

Training and test case characteristics

The training data comprise 224 cases from five centers (HGJ, HMR[V], CHUM, CHUS and CHUP). Originally, the dataset in (Vallieres et al., 2017) contained 298 cases, among which we selected the cases with oropharynx cancer. The test data contain 101 cases from a fifth center CHUV (n=53) and CHUP (n=48). Examples of PET/CT images

[V] For simplicity, these centers were renamed CHGJ and CHMR during the challenge.


(a) CHUM (b) CHUS (c) HGJ

(d) HMR (e) CHUV (f) CHUP

Figure 5.1: Case examples of 2D sagittal slices of fused PET/CT images from each of the six centers. The CT (grayscale) window in Hounsfield unit is $[-140, 260]$ and the PET window in SUV is $[0, 12]$, represented in a "hot" colormap.

of each center are shown in Fig. 5.1. Each case includes a CT image, a PET image and a GTVt mask (for the training cases) in the Neuroimaging Informatics Technology Initiative (NIfTI) format, as well as patient information (e.g. age, sex) and center. A bounding-box of size 144 × 144 × 144 mm^3 locating the oropharynx region was also provided. Details of the semi-automatic region detection can be found in (Andrearczyk, Oreiller, & Depeursinge, 2020).

Finally, to provide a fair comparison, participants who wanted to use additional external data for training were asked to also report results using only the HECKTOR data and discuss differences in the results. However, no participant used external data in this edition.

Annotation characteristics

For the HGJ, CHUS, HMR, and CHUM centers, initial annotations, i.e. 3D contours of the GTVt, were made by expert radiation oncologists and were later re-annotated as described below. Details of the initial annotations of the training set can be found in (Vallieres et al., 2017). In particular, 40% (80 cases) of the training radiotherapy contours were directly drawn on the CT of the PET/CT scan and thereafter used for treatment planning. The remaining 60% of the training radiotherapy contours were drawn on a different CT scan dedicated to treatment planning and were then registered to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM (MIM software Inc., Cleveland, OH). The initial contours of the test set were all directly drawn on the CT of the PET/CT scan.

The original contours for the CHUV center were drawn by an expert radiation oncologist for radiomics purpose (Castelli et al., 2019). The expert contoured the tumors on the PET/CT scan. The delineation from the CHUP center were obtained semi-automatically with a Fuzzy Locally Adaptive Bayesian (FLAB) segmentation (Hatt et al., 2009) and corrected by an expert radiation oncologist based on the corresponding CT information. These contours were obtained on the PET images only.

Given the heterogeneous nature of the original contours, a re-annotation of the VOIs was performed. During the first edition of HECKTOR (HGJ, CHUS, HMR, CHUM, and CHUV), the re-annotation was supervised by an expert who is both radiologist and nuclear medicine physician. Two non-experts (organizers of the challenge) made an initial cleaning in order to facilitate the expert's work. The expert either validated or edited the VOIs. The Siemens Syngo.Via RT Image Suite was used to edit the contours in 3D with fused PET/CT images. Most of the contours were re-drawn completely, and the original segmentations were used to localize the tumor and discriminate between malignant versus benign high metabolic regions.

For the data added to the current HECKTOR edition (CHUP), the re-annotation was performed by three experts: one nuclear medicine physician, one radiation oncologist and one who is both radiologist and nuclear medicine physician. The 71 cases were divided between the three experts and each annotation was then checked by all three of them. This re-annotation was performed in a centralized fashion with the MIM software, and the verification of the contours was made possible by the MIM Cloud platform [VI]. Guidelines for re-annotating the images were developed by our experts and are stated in the following.

Oropharyngeal lesions are contoured on PET/CT using information from PET and unenhanced CT acquisitions. The contouring includes the entire edges of the morphologic anomaly as depicted on unenhanced CT (mainly visualized as a mass effect) and the corresponding hypermetabolic volume, using PET acquisition, unenhanced CT and PET/CT fusion visualizations based on automatic co-registration. The contouring excludes the hypermetablic activity projecting outside the physical limits of the lesion (for example in the lumen of the airway or on the bony structures with no morphologic evidence of local invasion). The standardized nomenclature per AAPM TG-263 is “GTVp". For more specific situations, clinical nodal category was verified to ensure the exclusion of nearby FDG-avid and/or enlarged lymph nodes (e.g. submandibular, high level II, and retropharyngeal). In case of tonsillar fossa or base of tongue fullness/enlargement without corresponding FDG avidity, the clinical datasheet was reviewed to exclude patients with pre-radiation tonsillectomy or extensive biopsy.

[VI] https://mim-cloud.appspot.com/ as of December 2021.

Data preprocessing methods

No preprocessing was performed on the images to reflect the diversity of clinical data and to leave full flexibility to the participants. However, we provided various pieces of code to load, crop and resample the data, as well as to evaluate the results on our GitHub repository[VII]. This code was provided as a suggestion to help the participants and to maximize transparency, but the participants were free to use other methods.

Sources of errors

In (Oreiller et al., 2021), we reported an inter-observer (four observers) agreement of 0.6110 on a subset of the HECKTOR 2020 data containing 21 randomly drawn cases. Similar agreements were reported in the literature (Gudi et al., 2017) with an average DSC agreement of three observers of 0.57 using only the CT images for annotation and 0.69 using both PET and CT. A source of error therefore originates from the degree of subjectivity in the annotation and correction of the expert. Another source of error is the difference in the re-annotation between the centers used in HECKTOR 2020 and the one added in HECKTOR 2021. In HECKTOR 2020, the re-annotation was checked by only one expert while for HECKTOR 2021 three experts participated in the re-annotation. Moreover, the softwares used were different.

Another source of error comes from the lack of CT images with a contrast agent for a more accurate delineation of the primary tumor.

Institutional review boards

Institutional Review Boards (IRB) of all participating institutions permitted the use of images and clinical data, either fully anonymized or coded, from all cases for research purposes only. Retrospective analyses were performed following the relevant guidelines and regulations as approved by the respective institutional ethical committees with protocol numbers: MM-JGH-CR15-50 (HGJ, CHUS, HMR, CHUM) and CER-VD 2018-01513 (CHUV). In the case of CHUP, ethical review and approval were waived because data were already collected for routine patient management before analysis, in which patients provided informed consent. No additional data was specifically collected for the present challenge.

5.3 Task 1: Segmentation

5.3.1 Methods: Reporting of Challenge Design

A summary of the information on the challenge organization is provided in Appendix D.1, following the BIAS recommendations.

[VII] github.com/voreille/hecktor, as of December 2021.

Assessment aim

The assessment aim for the segmentation task is the following; evaluate the feasibility of fully automatic GTVt segmentation for H&N cancers in the oropharyngeal region via the identification of the most accurate segmentation algorithm. The performance of the latter is identified by computing the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) at 95^{th} percentile (HD95) between prediction and manual expert annotations.

DSC measures volumetric overlap between segmentation results and annotations. It is a good measure of segmentation for imbalanced segmentation problems, i.e. the region to segment is small as compared to the image size. DSC is commonly used in the evaluation and ranking of segmentation algorithms and particularly tumor segmentation tasks (Gudi et al., 2017; Moe et al., 2019).

In 3D, the HD computes the maximal distance between the surfaces of two segmentations. It provides an insight on how close the boundaries of the prediction and the ground truth are. The HD95 measure the 95^{th} quantile of the distribution of surface distances instead of the maximum. This metric is more robust towards outlier segmented pixels than the HD and thus is often used to evaluate automatic algorithms (Kuijf et al., 2019; Z. Liu et al., 2020).

Assessment Method Participants were given access to the test cases without the ground truth annotations and were asked to submit the results of their algorithms on the test cases on the AIcrowd platform.

Results were ranked using the DSC and HD95, both computed on images cropped using the provided bounding-boxes (see Section 5.2.2) in the original CT resolution. If the submitted results were in a resolution different from the CT resolution, we applied nearest-neighbor interpolation before evaluation.

The two metrics are defined for set A (ground truth volumes) and set B (predicted volumes) as follow:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{5.1}$$

where $|\cdot|$ is the set cardinality and

$$\mathrm{HD95}(A,B) = P_{95}\left\{\sup_{a\in A}\inf_{b\in B} \mathrm{d}(a,b), \sup_{b\in B}\inf_{a\in A} \mathrm{d}(a,b)\right\}, \tag{5.2}$$

where $\mathrm{d}(a,b)$ is the Euclidean distance between points a and b, sup and inf are the supremum and infimum, respectively. P_{95} is the 95^{th} percentile.

The ranking was computed from the average DSC and median HD95 across all cases. Since the HD95 is unbounded, i.e. it is infinity when there is no prediction, we choose the median instead of the mean for aggregation. The two metrics are ranked separately and the final rank is obtained by Borda counting. This ranking method was used first to determine the best submission of each participating team (ranking the 1 to 5 submissions), then to obtain the final ranking (across all participants). Each participating team had the opportunity to submit up to five (valid) runs. The final ranking is reported in Section 5.3.2 and discussed in Section 5.5.

Missing values (i.e. missing predictions on one or multiple patients) did not occur in the submitted results but would have been treated as DSC of zero and a HD95 of infinity. In the case of tied rank (which was very unlikely due to the computation of the results average of 53 DSCs), we considered precision as the second ranking metric. The evaluation implementation can be found on our GitHub repository[VIII] and was made available to the participants to maximize transparency.

5.3.2 Results: Reporting of Segmentation Task Outcome

Participation

As of September 14 2021 (submission deadline), the number of registered teams was 44 for Task 1, 30 for Task 2 and 8 for Task 3. A team is made of at least one participant and not all participants that signed the End User Agreement (EUA) registered a team. Each team could submit up to five valid submissions. By the submission deadline, we had received 448 results submissions, including valid and invalid ones (i.e. not graded due to format errors). This participation was much higher than last year's challenge with 83 submissions and highlights the growing interest in the challenge.

In this section, we present the algorithms and results of participants who submitted a paper (An et al., 2022; Bourigault et al., 2022; Cho et al., 2022; De Biase et al., 2022; Fatan et al., 2022; Ghimire et al., 2022; Juanco-Müller et al., 2022; Lang et al., 2022; J. Lee et al., 2022; T. Liu et al., 2022; Lu et al., 2022; Martinez-Larraz et al., 2022; Meng et al., 2022; Murugesan et al., 2022; M. A. Naser, Wahid, van Dijk, et al., 2022; Paeenafrakati et al., 2022; Qayyum et al., 2022; Ren et al., 2022; G. Wang et al., 2022; J. Wang et al., 2022; J. Xie & Peng, 2022; Yousefirizi et al., 2022; Yuan et al., 2022). An exhaustive list of the results can be seen on the leaderboard[IX].

[VIII] github.com/voreille/hecktor, as of December 2021.

[IX] https://www.aicrowd.com/challenges/miccai-2021-hecktor/leaderboards?challenge_leaderboard_extra_id=667&challenge_round_id=879

Segmentation: summary of participants' methods

This section summarizes the approaches proposed by all teams for the automatic segmentation of the primary tumor (Task 1). Table 5.3 provides a synthetic comparison of the methodological choices and design. All methods are further detailed in dedicated paragraphs. The paragraphs are ordered according to the official ranking, starting with the winners of Task 1.

			Preprocess.				Data augmentation					Model archit.				Loss				Training/ evaluation				
Team	Dice	HD95	iso-resampling	CT clipping	Min-max norm.	Standardization	Rotation	Scaling	Flipping	Noise addition	Other	U-Net	Attention	Res. connection	SE norm. (Iantsen, Visvikis, et al., 2021)	Dice	Cross-entropy	Focal (Lin et al., 2017)	Else	Optimizer	nnU-Net (Isensee et al., 2021)	LR decay	Cross-validation	Ensembling
Pengy (J. Xie & Peng, 2022)	0.7785	3.0882	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓		✓		SGD	✓	✓	✓	5
SJTU EIEE 2-426Lab^{X} (An et al., 2022)	0.7733	3.0882	✓	✓	✓		✓		✓			✓		✓	✓	✓	✓	✓	✓	Adam	✓	✓	✓	9
HiLab (Lu et al., 2022)	0.7735	3.0882	✓	✓		✓						✓	✓		✓	✓		✓		Adam		✓	✓	14
BCIOQurit (Yousefirizi & Rahmim, 2021)	0.7709	3.0882	✓	✓		✓	✓	✓	✓			✓		✓				✓	✓	Adam	✓	✓	✓	10
Aarhus Oslo (Ren et al., 2022)	0.7790	3.1549	✓				✓	✓				✓				✓	✓			Adam	✓	✓	✓	3
Fuller MDA (M. A. Naser, Wahid, van Dijk, et al., 2022)	0.7702	3.1432	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓				Adam		✓	✓	10
UMCG (De Biase et al., 2022)	0.7621	3.1432	✓			✓						✓	✓					✓	✓	Adam		✓	✓	5
Siat (G. Wang et al., 2022)	0.7681	3.1549	✓	✓		✓	✓		✓			✓	✓	✓		✓		✓		na		na	✓	5
Heck Uihak (Cho et al., 2022)	0.7656	3.1549	✓	✓	✓	✓	✓		✓			✓	✓		✓	✓		✓		Adam		✓	✓	5
BMIT USYD (Meng et al., 2022)	0.7453	3.1549	✓	✓	✓	✓	✓		✓			✓				✓				Adam		✓	✓	10
DeepX (Yuan et al., 2022)	0.7602	3.2700	✓	✓		✓			✓		✓	✓	✓	✓	✓	✓				Adam		✓	✓	15
Emmanuelle Bourigault (Bourigault et al., 2022)	0.7595	3.2700	✓	✓		✓	✓		✓			✓	✓	✓	✓	✓		✓		Adam		✓	✓	5
C235 (T. Liu et al., 2022)	0.7565	3.2700	✓	✓	✓	✓						✓	✓	✓		✓		✓		Adam		✓		5
Abdul Qayyum(Qayyum et al., 2022)	0.7487	3.2700			✓				✓		✓	✓		✓		✓	✓			Adam			✓	
RedNeucon (Martinez-Larraz et al., 2022)	0.7400	3.2700	na	na	na	na	✓	✓	✓			✓	✓			✓	✓		✓	Adam		✓		25
DMLang (Lang et al., 2022)	0.7046	4.0265	✓	✓	✓		✓		✓	✓	✓	✓				✓				Adam				
Xuefeng (Ghimire et al., 2022)	0.6851	4.1932	✓	✓		✓	✓	✓	✓	✓		✓				✓	✓			SGD		✓	✓	
Qurit Tecvico (Paeenafrakati et al., 2022)	0.6771	5.4208			✓							✓				✓				Adam				
Vokyj (Juanco-Müller et al., 2022)	0.6331	6.1267	✓	✓		✓	✓									✓		✓		Adam		✓		
TECVICO Corp Family (Fatan et al., 2022)	0.6357	6.3718	na	na	na	na						✓		✓		✓				Adam				2
BAMF health (Murugesan et al., 2022)	0.7795	3.0571	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓		✓	✓			SGD	✓		✓	10
Wangjiao (J. Wang et al., 2022)	0.7628	3.2700	✓	✓		✓						✓	✓	✓	✓	✓		✓		Adam		✓	✓	6

Table 5.3: Synthetic comparison of segmentation methods and results. More details are available in Section 5.3.2. The number of used models is reported in the last column when ensembling was used. "na" stands for "not available".

J. Xie and Peng, 2022 (team “Pengy”) used a well-tuned patch-based 3D nnU-Net (Isensee et al., 2021) with standard pre-processing and training scheme, where the learning rate is adjusted dynamically using polyLR (L. Chen et al., 2016). The Squeeze and Excitation (SE) normalization (Iantsen, Visvikis, et al., 2021) was also one of the main ingredient of their approach. The approach is straighforward yet efficient as they ranked first for Task 1. Five models are trained in a five-fold cross-validation with random data augmentation including rotation, scaling, mirroring, Gaussian noise and Gamma correction. The five test predictions are ensembled via probability averaging for the final results.

An et al., 2022 (team “SJTU EIEE 2-426Lab”) proposed a framework which is based on the subsequent application of three different U-Nets. The first U-Net is used to coarsely segment the tumor and then select a bounding-box. Then, the second network performs a finer segmentation on the smaller bounding box. Finally, the last network takes as input the concatenation of PET, CT and the previous segmentation to refine the predictions. They trained the three networks with different objectives. The first one was trained to optimize the recall rate, and the two subsequent ones were trained to optimize the Dice score. All objectives were implemented with the F-loss which includes a hyper-parameter allowing to balance between recall and Dice. The final prediction was obtained through majority voting on three different predictions: an ensemble of five nnU-Nets (Isensee et al., 2021) (trained on five different folds), an ensemble of three U-Nets with SE normalization (Iantsen, Visvikis, et al., 2021), and the predictions made by the proposed model.

Lu et al., 2022 (team “HiLab”) employed an ensemble of various 3D U-Nets, including the eight models used in (Iantsen, Visvikis, et al., 2021), winner of HECKTOR 2020, five models trained with leave-one-center-out, and one model combining a priori and a posteriori attention. In this last model, the normalized PET image was used as a priori attention map for segmentation on the CT image. Mix-up was also used, mixing PET and CT in the training set to construct a new domain to account for the domain shift in the test set. All 14 predictions were averaged and thresholded to 0.5 for the final ensembled prediction.

Yousefirizi et al., 2022 (team “BCIOqurit”) used a 3D nnU-Net with SE normalization (Iantsen, Visvikis, et al., 2021) trained on a leave-one-center-out with a combination of a "unified" focal and Mumford-Shah (Kim & Ye, 2019) losses taking the advantage of distribution, region, and boundary-based loss functions.

Ren et al., 2022 (team “Aarhus Oslo”) proposed a 3D nnU-Net with various PET normalization methods, namely PET-clip and PET-sin. The former clips the Standardized Uptake Values (SUV) range in [0,5] and the latter transforms monotonic spatial SUV increase into onion rings via a sine transform of SUV. Loss functions were also combined and compared (Dice, Cross-Entropy, Focal and TopK). No strong global trend was observed on the influence of the normalization or loss.

In (M. A. Naser, Wahid, van Dijk, et al., 2022), Naser et al. (team “Fuller MDA”) used an ensemble of 3D residual U-Nets trained on a 10-fold CV resulting in 10 different models. The ensemble was performed either by STAPLE or majority voting on the binarized predictions. Models with different numbers of channels were also compared. The best combination was the one with fewer feature maps and ensembled with majority voting.

De Biase et al., 2022 (team “UMCG”) compared two methods: (i) Co-learning Multi-Modal PET/CT adapted from (Xue et al., 2021), which takes as input PET and CT as two separate images, outputs two masks that are averaged and (ii) Skip-scSE Multi-Scale Attention, which concatenates PET and CT in the channel dimension. The Skip-scSE models clearly outperformed the other. Ensembling (i) and (ii) provided worse results.

G. Wang et al., 2022 (team “Siat”) used an ensemble of 3D U-Nets with multi-channel attention mechanisms. For each channel in the input data, this attention module outputs a weighted combination of filter outputs from three receptive fields over the input. A comparison with a standard 3D Vnet without attention showed the benefit of the latter.

Cho et al., 2022 (team “Heck Uihak”) used a backbone 3D U-Net that takes as input PET/CT images and outputs the predictions. This backbone U-Net is coupled with an attention module. The attention module was designed around a U-Net architecture and takes as input the PET images and produces attention maps. These attention maps are then multiplied with the skip connections of the backbone U-Net. The whole pipeline was trained with a sum of a Dice loss and a focal loss.

Meng et al., 2022 (team “BMIT USYD”) used multi-task learning scheme to address Tasks 1 and 2. A modified 3D U-Net was used for segmentation. Its output is a voxel-wise tumor probability that is fed together with PET/CT to a 3D denseNet. Ensembling was used to produce the final output.

Yuan et al., 2022 (team “DeepX”) proposed a 3D U-Net with scale attention which is referred to as Scale Attention Network (SA-Net). The skip connections were replaced by an attention block and the concatenation was replaced by a summation. The attention block takes as input the skip connections at all the scales and output an attention map which is added to the feature maps of the decoder. The attention blocks include a SE block. The encoder and decoder include ResNet-like blocks containing a SE block. An ensemble of 15 models was used for the final prediction (5 from the 5-fold CV with input size $144 \times 144 \times 144$ at 1mm^3, 5 from the 5-fold CV with input size $128 \times 128 \times 128$ at $1.25 \times 1.25 \times 1.25$mm^3, and 5 from a leave-one-center-out CV with input size $144 \times 144 \times 144$ at 1mm^3).

Bourigault et al., 2022 (team “Emmanuelle Bourigault”) proposed a full scale 3D U-Net architecture with attention, residual connections and SE norm. Conditional random fields was applied as post-processing.

T. Liu et al., 2022 (team “C235”) proposed a model based on 3D U-Net supplemented with a simple attention module referred to as SimAM. Different from channel-wise and spatial-wise attention mechanisms, SimAM generates the corresponding weight for each pixel in each channel and spatial position. They compared their model to last year’s winning algorithm based on SE Norm and report a small but consistent increase in segmentation performance when using the proposed SimAM attention module, which also resulted in models with about 20 times less parameters.

Qayyum et al., 2022 (team “Abdul Qayyum”) proposed to use a 3D U-Net with 3D inception as well as squeeze and excitation modules with residual connections. They extended the 2D inception module into 3D with extra 3D depth-wise layers for semantic segmentation. The comparison with and without the inception module showed a systematic improvement associated with the latter.

Martinez-Larraz et al., 2022 (team “RedNeucon”) ensembled a total of 25 models: a combination of 2D (trained on axial, sagittal and coronal planes) and 3D U-Nets, all trained on cross-validation and on the full dataset.

Lang et al., 2022 (team “DMLang”) used a network based on a 3D U-Net. The main modification is that the skip connections were linked directly after the downsampling. They also optimized the kernel size and the strides of the convolutions.

Ghimire et al., 2022 (team “Xuefeng”) developed a patch-based 3D U-Net with overlapping sliding window at test time. Deep supervision technique was applied to the network, where the computation of loss occurs at each decoding block. Various patch sizes, modality combination and convolution types were compared. Results suggest that larger patch size, bi-modal inputs, and conventional convolution (i.e. not dilated) was better.

Paeenafrakati et al., 2022 (team “Qurit Tecvico”) proposed to use 3D U-Net or 3D U-NeTr (U-Net with transformers) to segment the GTVt. The network’s input consists of a one-channel image. This image was obtained by image-level fusion techniques to combine information of both PET and CT images. They assessed ten different image fusion methods. To select the best combination of architecture and fusion method, they used a validation set of 23 images. The best combination was a U-Net architecture with the Laplacian pyramid method for fusion. This model obtained a DSC of, respectively, 0.81 and 0.68 on the validation and test set.

Juanco-Müller et al., 2022 (team “Vokyj”) proposed a model trained on supervoxels (obtained with Simple Linear Iterative Clustering, SLIC), motivated by the efficiency of the latter. The model is composed of an MLP encoder and graph CNN decoder. The models were trained on extracted patches of size 72x72x72.

Fatan et al., 2022 (team “TECVICO Corp Family”) employed a 3D U-Net with autoencoder regularization (Myronenko, 2018) trained on various fusions of PET and CT images. The best results were obtained with a Laplacian pyramid-sparse representation mixture.

J. Lee et al., 2022 (team “Neurophet”) used a dual path encoder (PET, CT) whose paths are coupled by a shared-weight cross-information module in each layer of the encoding path of the 3D U-Net architecture. The cross-attention module performs global average pooling over the feature channels resulting from convolutional blocks in both paths and feeds the resulting pooled features into a weight-shared fully connected layer. Its output, two (transformed) feature vectors are added elementwise and activated using a sigmoid function. The final output of each layer in the encoding part is obtained by multiplication of the features in each of the two paths with these cross-attention weights. The study used the generalized dice loss as training metric. Five separate models were built, using data from four centers for training and data from the 5th center for evaluation (average DSC 0.6808). Predictions on the test set (DSC 0.7367) were obtained by majority voting across the segmentation results of all 5 models.

Murugesan et al., 2022 (team “BAMF Health”) proposed to ensemble the predictions of 3D nnU-Nets (with and without residual connections) using adaptive ensembling to eliminate false positives. A selective ensemble of 8 test-time augmentations and 10 folds (5 U-Nets and 5 residual U-Nets) was used for the final segmentation output.

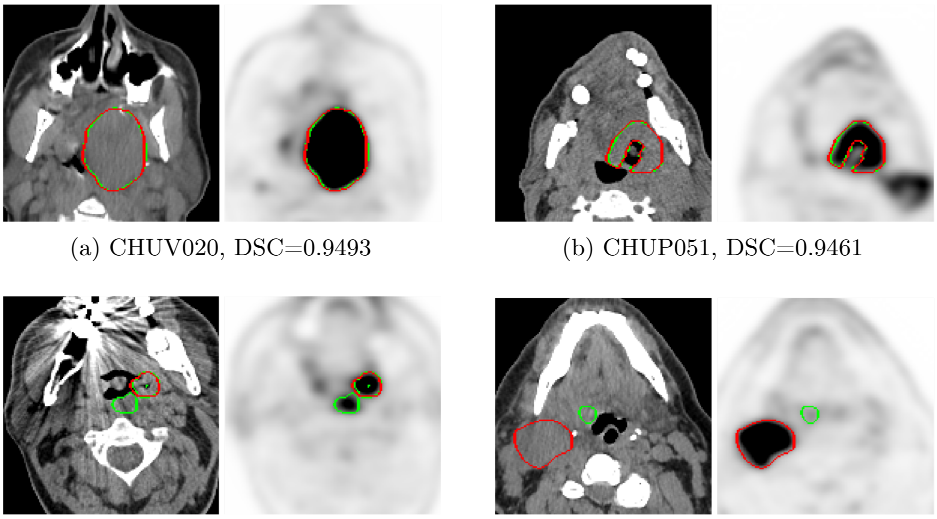
J. Wang et al., 2022 (team “Wangjiao”) used a combination of convolutional and transformer blocks in a U-Net model with attention (global context and channel) in the decoder. The model was trained with squeeze and excitation, and a Dice and Focal loss.

In Table 5.3, we summarize some of the main components of the participants’ algorithms, including model architecture, preprocessing, training scheme and postprocessing.

Results

The results, including average DSC and HD95 are summarized in Table 5.3 with an algorithm summary. The two results at the bottom of the table without a rank were made ineligible to the ranking due to an excessive number of submissions on the HECKTOR 2020 dataset (on the online leaderboard) resulting in an overfit of the 2020 test set which represents half of the 2021 test set.

The results from the participants range from an average DSC of 0.6331 to 0.7785 and the median HD95 from 6.3718 to 3.0882. J. Xie and Peng, 2022 (team “Pengy”) obtained the best overall results with an average DSC of 0.7785 and a median HD95 of 3.0882. Examples of segmentation results (true positives on top row, and false positives on bottom row) are shown in Fig. 5.2.



(a) CHUV020, DSC=0.9493

(b) CHUP051, DSC=0.9461

(c) CHUP063, DSC=0.3884

(d) CHUV036, DSC=0.0000

Figure 5.2: Examples of results of the winning team (Pengy (J. Xie & Peng, 2022)). The automatic segmentation results (green) and ground truth annotations (red) are displayed on an overlay of 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the whole image, see Eq. (5.1).

5.4 Tasks 2 and 3: Outcome Prediction

In order to expand the scope of the challenge compared to the previous installment (2020) that focused on a single task dedicated to the automatic segmentation of GTVt (i.e., same as the updated Task 1 in the 2021 edition), it was decided to add a task with the aim of predicting outcome, i.e. Progression-Free Survival (PFS).

5.4.1 Methods: Reporting of Challenge Design

It was chosen to carry out this task on the same patients dataset used for Task 1, exploiting both the available clinical information and the multimodal FDG-PET/CT images. The available clinical factors included center, age, gender, TNM 7/8th edition staging and clinical stage, tobacco and alcohol consumption, performance status, HPV status, treatment (radiotherapy only or chemoradiotherapy). The information regarding tobacco and alcohol consumption, performance and HPV status was available only for some patients. For five patients from the training set, the weight was unknown and was set at 75kg to compute SUV values. Of note, this outcome prediction task was subdivided into two different tasks that participants could choose to tackle separately: Task 3 provided the same data as Task 2, with the exception of providing, in addition, the reference expert contours (i.e., ground-truth of the GTVt). In order to avoid providing the reference contours to participants that could also participate in Task 1, we relied on a

Docker-based submission procedure: participants had to encapsulate their algorithm in a Docker and submit it on the challenge platform. The organizers then ran the Dockers on the test data locally, in order to compute the performance. In such a way, the participants never had direct access to the reference contours of the test set, although they could incorporate them in their algorithms the way they saw fit.

Assessment aim

The chosen clinical endpoint to predict was PFS. Progression was defined based on Response Evaluation Criteria In Solid Tumors (RECIST) criteria, i.e., either a size increase of known lesions (i.e., change of T and or N), or appearance of new lesions (i.e., change of N and/or M). Disease-specific death was also considered a progression event for patients previously considered stable. In the training set, participants were provided with the survival endpoint to predict, censoring and time-to-event between PET/CT scan and event (in days).

Assessment Method

For Task 2, challengers had to submit a CSV file containing the patient IDs with the outputs of the model as a predicted risk score anti-concordant with the PFS in days. For Task 3, the challengers had to submit a Docker encapsulating their method which was run by the organizers on the test set, producing the CSV file for evaluation. Thus for both tasks, the performance of the output predicted scores were evaluated using the Concordance index (C-index) (Harrell et al., 1982) on the test data. The C-index quantifies the model's ability to provide an accurate ranking of the survival times based on the computed individual risk scores, generalizing the Area Under the ROC Curve (AUC). It can account for censored data and represents the global assessment of the model discrimination power. Therefore the final ranking was based on the best C-index value obtained on the test set, out of the maximum of 5 submissions per team. The C-index computation is based on the implementation found in the Lifelines library (Davidson-Pilon, 2019) and adapted to handle missing values that are counted as non-concordant.

5.4.2 Results: Reporting of Challenge Outcome

Participation

Thirty different teams submitted a total of 149 valid submissions to Task 2. Eighteen corresponding papers were submitted, which made the submissions eligible for final ranking and prize. Probably because of the added complexity of Task 3 requiring encapsulating the method in a Docker, only 8 teams submitted a total of 27 valid submissions. All these 8 teams also participated in Task 2, with 7 corresponding papers.

Outcome prediction: summary of participants' methods

The following describes the approach of each team participating in Task 2 (and 3 for some), in the order of the Task 2 ranking. Table 5.4 provides a synthetic comparison of the methodological choices and designs for these tasks.

		Pre-processing						Segment.			Image features				Modeling and training approach													Masks		
Team	**C-index Task 2**	Iso-resampling	CT clipping	Min-max norm.	Standardization	PET/CT fusion	Further cropping	Relies on Task 1	Additional segm.	No segmentation	Deep features	Large radiomics set	Volume, shape	IBSI compliant	Ensembling	Deep model	Algo. RF, SVM...	Feature selection	PET as input	CT as input	PET/CT fusion	Use clinical var.	Imputed missing	Cross-val.	Augmentation	**C-index Task 3**	GT masks	Task 1 masks	PET thresh. masks
BioMedIA (Saeed et al., 2022)	0.7196			✓		✓	✓			✓	✓					✓					✓	✓				na			
Fuller MDA (M. A. Naser, Wahid, Mohamed, et al., 2022)	0.6938		✓	✓	✓					✓	✓				✓	✓			✓	✓		✓			✓	0.6978	✓	✓	
Qurit Tecvico (Paeenafrakati et al., 2022)	0.6828			✓				✓				✓		✓	✓		✓	✓			✓	✓		✓		na			
BMIT_USYD (Meng et al., 2022)	0.6710		✓	✓	✓		✓	✓			✓				✓	✓			✓	✓		✓			✓	na			
DMLang (Lang et al., 2022)	0.6681		✓	✓				✓			✓					✓			✓	✓		✓			✓	na			
TECVICO_C. (Fatan et al., 2022)	0.6608					✓	✓	✓				✓		✓			✓	✓			✓	✓		✓		na			
BAMF Health (Murugesan et al., 2022)	0.6602		✓	✓	✓			✓	✓			✓		✓			✓	✓	✓	✓		✓	✓	✓		0.6602			✓
ia-h-ai (Starke et al., 2022)	0.6592											✓		✓			✓	✓	✓	✓		✓		✓		0.6592			✓
Neurophet (J. Lee et al., 2022)	0.6495							✓					✓				✓					✓		✓		na			
UMCG (B. Ma et al., 2022)	0.6445		✓	✓	✓	✓				✓	✓				✓	✓			✓	✓	✓	✓		✓	✓	0.6373	✓	✓	
Aarhus Oslo (Huynh et al., 2022)	0.6391																					✓		✓		na			
RedNeucon (Martinez-Larraz et al., 2022)	0.6280							✓	✓			✓					✓	✓				✓		✓		na			
Emmanuelle B. (Bourigault et al., 2022)	0.6223	✓	✓		✓			✓			✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓		na			
BCIOQurit (Yousefirizi et al., 2022)	0.6116	✓	✓		✓			✓			✓	✓		✓			✓	✓	✓	✓		✓		✓	✓	0.4903	✓		
Vokyj (Juanco-Müller et al., 2022)	0.5937	✓	✓		✓			✓					✓				✓					✓			✓	na			
Xuefeng (Ghimire et al., 2022)	0.5510	✓	✓		✓			✓					✓				✓									0.5089	✓		
DeepX (Yuan et al., 2022)	0.5290	✓	✓		✓			✓				✓		✓			✓		✓	✓		✓				na			

Table 5.4: Synthetic comparison of outcome prediction methods. More details are available in Section 5.4.2. All participants of task 3 also participated in task 2.

Saeed et al., 2022 (team "BiomedIA") first experimented with the clinical variables and determined that better prediction was achieved using only variables with complete values, compared to using all variables with imputing missing values. They elected to first implement a fusion of PET and CT images by averaging them into a new single PET/CT image that would be further cropped (2 different sizes of 50x50x50 and 80x80x80 were tested) to form the main input to their solution based on a 3D CNN (Deep-CR) trained to extract features which were then fed into Multi-Task Logistic Regression (MTLR, a sequence of logistic regression models created at various timelines to evaluate the probability of the event happening) improved by integrating neural networks to achieve nonlinearity, along with the clinical variables. Two different models were compared as the input to MTLR: either a CNN with 3 paths (for PET, CT and fused PET/CT) or only one using only fused PET/CT. The batch size, learning rate, and dropout were experimentally set to 16, 0.016, and 0.2 respectively for the training. The model was trained for 100 epochs using Adam optimizer. No cross-validation or data augmentation was used. Of note, the results of CNN and MTLR (i.e., exploiting both images and clinical variables) were averaged with the prediction of a Cox model using only clinical variables to obtain the best result. This team won Task 2 with a C-index of 0.72 but did not participate in Task 3.

M. A. Naser, Wahid, Mohamed, et al., 2022 (team "Fuller MDA") also adopted an approach based on deep learning. Clinical variables without missing values were transformed into an image matrix in order to be fed along with PET and CT images (rescaled and z-score normalized) as separate channels to a DenseNet121 CNN. Adopting a 10-fold cross-validation scheme, the model was trained either only with 2 channels (PET and CT) or 3 (adding the clinical), with data augmentation, for 800 iterations with a decreasing learning rate, the Adam optimizer and a negative log-likelihood loss. Of note, the PFS was discretized into 20 discrete intervals for the output of the network. Two different approaches of ensembling the various models obtained over the 10 folds (consensus or average) were implemented. The best result (0.694, rank 2) was obtained with the Image+Clinical consensus model. The team also participated in Task 3 where they used ground-truth masks as an additional input channel to the same network (Wahid et al., 2022), achieving the first rank with a C-index of 0.70.

Paeenafrakati et al., 2022 (team "Qurit Tecvico") implemented a classical radiomics approach, where a large set of IBSI-compliant features were extracted with the SERA package (Ashrafinia, 2019) from the delineated tumor (based on the output of their solution for Task 1) in PET, CT as well as a fusion of PET/CT (of note, 10 different fusion techniques were explored). The features were then selected through 13 different dimensionality reduction techniques and 15 different selection methods and combined along with clinical variables, into several models with 5-fold cross-validation (the entire training set was used for each approach) through the use of 8 different survival prediction algorithms. The best performance (0.68) in the test set was obtained with an ensemble voting of these various algorithms, obtaining third rank in Task 2 (the team did not participate in Task 3).

Meng et al., 2022 (team “BMIT USYD”) proposed a single unified framework to achieve both segmentation (Task 1) and outcome prediction (Task 2, no participation in Task 3). They first selected a few relevant clinical variables to take into account by performing a univariate/multivariate analysis, retaining only HPV status, performance status, and M stage. Their proposed model is composed of two main components: a U-Net based network for segmentation and a DenseNet based cascaded survival network. Both extract deep features that are fed into fully connected layers for outcome prediction and are trained in an end-to-end manner to minimize the combined loss of segmentation and survival prediction losses, with Adam optimizer, a batch size of 8 for 10000 iterations, with a decreasing learning rate. Clinical factors were concatenated in the non-activated fully connected layer. Of note, both the segmentation output and the cropped, normalized PET and CT images are fed to the DenseNet cascaded survival network. Data augmentation (random translations, rotations and flipping) was applied. Ten different models were trained, the first 5 through leave-one-center-out cross-validation and the next five with 5-fold cross-validation. The ensemble of these achieved a C-index of 0.671 in the test set.

Lang et al., 2022 (team “DMlang”) relied on the segmentation output of Task 1 (or on the reference contours in training) to generate cropped bounding-boxes as inputs to their approach for predicting outcome, which relied on extracting deep features from PET and CT images thanks to a pre-trained C3D network designed to classify video clips. In order to feed PET and CT images to this C3D model, each 3 consecutive slices were fed to the color channels. The obtained PET and CT features were then concatenated and fed to a dense layer, which was then concatenated with clinical variables. Each output neuron represented the conditional probability of surviving a discrete time interval (the best model involved layers of size 512 and 256 and an output size of 15 corresponding to time intervals covering a maximum of 10 years of survival with the first 5 years split into intervals of half a year and all subsequent intervals with a width of one year). The same data augmentation as for the segmentation task was used. For training this network, a batch size of 16 was applied and 75 epochs were used with the Adam optimizer to minimize the negative log-likelihood. For model optimization, hyper-parameters were tuned manually. Of note, the team did not rely on ensemble of models nor on cross-validation, but generated a single stratified split of the training data. The trained model achieved a C-index of 0.668. The team did not participate in Task 3.

Fatan et al., 2022 (team “TECVICO Corp Family”) used a similar PET/CT fusion approach (5 different techniques) and cropping as the team "Qurit_Tecvivo", extracted 215 IBSI-compliant radiomics features with the same package (SERA), that were fed into a number of feature selection techniques (7) and classifiers (5). They did not perform an ensemble of these but selected the best model in cross-validation during training. The best combination (LP-SR fusion and the classifier GlmBoost) obtained 0.66 in the test set. They did not participate in Task 3.

Murugesan et al., 2022 (team “BAMF Health”) participated in both Tasks 2 and 3. Interestingly, their best results were obtained using the tumor masks by their segmentation method of Task 1, instead of the reference contours. Their solution was based on standard IBSI-compliant radiomics features extracted with Pyradiomics from PET and CT images after z-score normalization of intensities. In addition, in-house features calculating the number of uptakes and their volumes in each PET/CT were calculated through thresholding of PET SUVs. All clinical variables were exploited, missing values were imputed using the mean value of provided variables. Before further exploitation of the radiomics features, they were standardized using their mean and standard deviation. Then principal component analysis was applied to the features, capturing 95 of information. Variable importance combined with fast unified random forests for survival, regression, and classification was used for modeling through repeated random sub-sampling validation over 100 multiple random splits, in order to look for an optimal combination of features and to optimize hyper-parameters. The best result in the test set was obtained with a model relying on PCA components, with a 0.66 C-index (for both Tasks 2 and 3).

Starke et al., 2022 (team “ia-h-ai”) built a strategy based on standard radiomics modeling, addressing both Tasks 2 and 3. They first strategically split the training data into 3 folds, ensuring that for each split, one of the centers is present in the validation set but not the training. Clinical factors were all considered, by imputing missing values through k-nearest neighbor (k=20). They used either the provided reference volumes or alternative ones obtained through thresholding the PET intensities with SUV $>$ 2.5. 172 IBSI-compliant handcrafted features were then extracted from both PET and CT images volumes of interest using Pyradiomics. They established some baseline models through Cox proportional hazards models exploiting only the clinical variables, then moved to more advanced modeling relying on random survival forest, still using only clinical variables. In order to add image features to the models, they performed feature selection through three different processes: stability (L1-regularized Cox regression applied to multiple bootstrapped datasets for a range of regularization strength parameters), permutation-based feature importance and finally sequential feature selection. This allowed them to retain only a small number of features for the actual modeling step, where they compared different approaches using random forest survival (300 trees): fully automated feature selection and combination or different ways of manually selecting features, including a priori selection based on literature. They consistently obtained better performance on the test set by relying on the alternative volumes of interest (thresholded at SUV $>$ 2.5, leading to volumes larger than the reference ground-truth contours), and models with hand-picked features, contrary to fully automatic selection that demonstrated overfitting.

J. Lee et al., 2022 (team “Neurophet”) exploited only clinical variables (missing values were coded as 0 or -1 depending on the variable) and segmented volumes from Task 1 (i.e. only 1 feature, the tumor volume) to train a random forest survival model through 5-fold randomized cross-validation with 100 iterations. Of note, the center ID was added as a clinical factor. The proposed model achieved a C-index of 0.65 on the test set, with a higher performance than the same model without tumor volume (0.64). The team did not participate in Task 3.

B. Ma et al., 2022 (team "UMCG") proposed a pipeline based on deep learning as well, consisting of three parts: 1) the pyramid autoencoder of a 3D Resnet extracting image features from both CT and PET, 2) a feed-forward feature selection to remove the redundant image and clinical features, and 3) a DeepSurv (a Cox deep network) for survival prediction. Clinical variables were used but missing values were not imputed, rather described as an additional class (i.e., unknown). PET and CT images were pre-processed and a new PET/CT image obtained by summation of PET and CT was used as a third input to the autoencoder. The segmentation masks were not used for Task 2, but were used for Task 3 in order to extract the tumor region in two different ways, both being used as inputs to the network. This pipeline was trained on using different splits of the training set (leave-one-center out and random selection of 179 patients for training and 45 for validations), resulting in 6-fold cross-validation. The Autoencoders were trained using the Adam optimizer with the initial learning rate 0.001 and data augmentation for 80 epochs. The official DeepSurv was trained for 5000 steps with the default settings. A total of 30 DeepSurv models were trained in each fold and the 3 models with the highest validation set C-index were selected. In total 18 models were obtained and their predicted risk scores are averaged to obtain the final result: 0.6445 and 0.6373 C-index in the test set for Task 2 and 3 respectively.

Huynh et al., 2022 (team "Aarhus Oslo") team compared a conventional radiomics approach (although without tumor delineation, i.e., features were extracted from the whole bounding-box) and a deep learning approach in Task 2 only. Both used the provided bounding-box of PET and CT images as inputs, and in the case of the deep learning approach, an additional pre-processing step was applied to PET images in order to reduce the variability of images due to various centers based on a sin transform. For the standard radiomics approach, only clinical variables without missing values were exploited, whereas they were not used in the case of the deep learning approach. In the standard radiomics modeling, over 100 IBSI-compliant features were calculated but only a handful were manually selected based on literature and further used: one from CT and 4 from PET. These features (and clinical variables) were then fed to 2 ensemble models: random forest and gradient boosting. Hyper-parameters (number of trees, maximum depth for each tree, and learning rate, loss function tuning) were tuned using grid-search, and models were trained and evaluated using 5-fold cross-validation. In the case of deep learning, only CT and PET-sin images were used as input of a CNNs built with the encoder part of the SE Norm U-Net model (Iantsen, Visvikis, et al., 2021) with three fully connected layers (4096, 512, and 1 units) added to the top. Five-fold cross-validation was also used. Each model was trained for 150 epochs using the Adam optimizer with a batch size of 4. The initial learning rate was set to 3e-6 and the loss was defined as a fusion of the Canberra distance loss and Huber loss ($\delta = 1$). Based on the results of cross-validation in training, the four following models were evaluated on the test set: Gradient boosting trained on either clinical factors (either all or only uncensored data) or both clinical factors and selected radiomics features and ensemble based on mean predicted values of five-fold deep learning models trained on FDG-PET/CT. All models had near-random performance in the test set, except the clinical-only model built with gradient boosting (0.66).

Martinez-Larraz et al., 2022 (team “RedNeucon”) implemented a conventional radiomics approach based on the extraction of handcrafted features from PET and CT with a Matlab toolbox, from the reference contour volumes and the segmentation output of Task 1, as well as an additional volume of interest generated by determining a two pixel inward and outward the contours to get a tumor “boundary region”. Only clinical variables without missing values were used. Features were then selected after ranking according to 2 methods, ranking for classification using a Fisher F-Test and an algorithm based on K-nearest neighbors. When two features showed a correlation above 0.5, the best one was kept. Three different modeling algorithms were compared in 5-fold cross-validation: Gaussian Process Regression (GPR), an Ensembled Bagged of trees and a Support Vector Machine. The best result on the test set (0.628) was obtained with the GPR with 35 features.

Bourigault et al., 2022 (team “Emmanuelle Bourigault”) proposed a Cox proportional hazard regression model using a combination of clinical, radiomic, and deep learning features from PET/CT images. All clinical variables were exploited, after imputing missing values using a function of available ones. IBSI-compliant handcrafted radiomics features including wavelet-filtered ones were calculated using Pyradiomics and were combined with deep features from the 3D U-Net used in the segmentation Task 1, in addition to clinical variables. Spearman rank correlation above 0.8 was used to eliminate intercorrelated features. Feature selection was performed using Lasso regression with 5-fold cross-validation, reducing the set of 270 variables to 70 (7 clinical, 14 radiomics and 49 deep). Three different models were implemented for modeling: Cox proportional hazard regression model, random survival forest and Deepsurv (a Cox proportional hazards deep neural network). All three models were trained with different combinations of the selected clinical, radiomics (PET, CT or PET/CT) and deep features. The best performance in validation was obtained with the Cox model using clinical + CT radiomics + deep learning features, although in the test set its final performance was 0.62.

Yousefirizi et al., 2022 (team “BCIOqurit”) proposed training a proportional hazard Cox model with a multilayer perceptron neural net backbone to predict the score for each patient. This Cox model was trained on a number of PET and CT radiomics features extracted from the segmented lesions, patient demographics, and encoder features provided from the penultimate layer of a multi-input 2D PET/CT convolutional neural network tasked with predicting time-to-event for each lesion. A grid search over several feature selection and classifiers methods identified 192 unique combinations of radiomics features that were used to train the overall Cox model with the Adam optimizer, a learning rate of 0.0024, a batch size of 32, and an early stopping method that monitored the validation loss. A 10-fold cross-validation scheme was used and an ensemble model of these achieved a C-index score of 0.612 in the test set.

Juanco-Müller et al., 2022 (team "Vokyj") proposed to fit a Weibull accelerated failure time model with clinical factors and the shape descriptors of the segmented tumor (output of Task 1). M-stage and two shape features (Euler number and Surface Area) were the most predictive of PFS, the model achieving a performance of 0.59 in the test set. The team did not participate in Task 3.

Ghimire et al., 2022 (team "Xuefeng") implemented a straightforward approach that consisted in calculating the tumor volume and tumor surface area of the Task 1 segmentation outputs, as well as the classification output from the segmentation network trained to classify the input images into 6 different classes of PFS (which was first discretized into 6 bins). These imaging features were then combined with all available clinical factors, for which missing values were imputed with the median value for numerical variables and mode value for categorical ones. All features were then normalized to zero mean and 1 standard deviation for a linear model to be fitted to the training data. The model was applied to both Tasks 2 and 3, using the reference contours instead of the Task 1 segmentation results, leading to C-index values of 0.43 and 0.51 respectively.

Yuan et al., 2022 (team "DeepX") implemented a standard radiomics approach, extracting more than 200 IBSI-compliant handcrafted features with Pyradiomics, from both PET and CT images using the segmentation output of Task 1, which were then manually ranked and selected according to their concordance index. Regarding clinical variables, only age was used. The 7 selected features were evaluated independently or combined through averaging concordance ranking, obtaining their best C-index of 0.53 in the test set.

5.5 Discussion: Putting the Results into Context

Outcomes and findings of participating methods are summarized in Section 5.5.1 for all three tasks. In general, we observed that simplicity was beneficial for generalization and that sophisticated methods tend to overfit the training/validation. Despite the diversity in terms of centers and image acquisition, no specific feature or image harmonization method was employed, which could be one avenue for improving generalization abilities of the methods in all tasks (Atul Mali et al., 2021).

The combined scope of the three proposed tasks also allowed the emergence of very interesting findings concerning the relationship of the GTVt contouring task and PFS prediction. In a nutshell, ground truth ROIs were not providing top results, even though they were re-annotated in a centralized fashion to be dedicated for radiomics (Fontaine et al., 2022a). Simple PET thresholded and bounding-boxes for deep learning outperformed the use of ground truth ROI. This suggests that algorithms looking elsewhere than the GTVt is beneficial (e.g. tumoral environment, nodal metastases). Fully automatic algorithms are expected to provide optimal results, which was already highlighted by several papers in the context of the HECKTOR challenge (Andrearczyk, Fontaine, et al., 2021; Fontaine et al., 2021; Murugesan et al., 2022; M. A. Naser, Wahid, Mohamed, et al., 2022; Starke et al., 2022). This potentially obviates the need for GTVt contouring, opening avenues for reproducible and large scale radiomics studies including thousands potential subjects.

5.5.1 Outcomes and Findings

A major benefit of this challenge is to compare various algorithms developed by teams from all around the world on the same dataset and task, with held-out test data.

We distinguish here between the technical and biomedical impact. The main technical impact of the challenge is the comparison of state-of-the-art algorithms on the provided data. We identified key elements for addressing the task: 3D U-Net, preprocessing, normalization, data augmentation and ensembling, as summarized in Tables 5.3 and 5.4. The main biomedical impact of the results is the opportunity to generate large cohorts with automatic tumor segmentation for comprehensive radiomics studies, as well as to define and further push state of the art performance.

Task 1: Automatic segmentation of the GTVt
The best methods obtain excellent results with DSCs above 0.75, better than inter-observer variability (DSC 0.61) performed on a subset of our data and similar variability reported in the literature (DSCs of 0.57 and 0.69 on CT and PET/CT respectively) (Gudi et al., 2017). Note that without injected contrast CT, delineating the exact contour of the tumor is very difficult. Thus, the inter-observer DSC could be low only due to disagreements at the border of the tumor, without taking into account the error rate due to the segmentation of non-malignant structures (if any). For that reason, defining the task as solved solely based on the DSC is not sufficient. In the context of this challenge, we can therefore define the task as solved if the algorithms follow these three criteria:

1. Higher or similar DSC than inter-observers agreement.
2. Detect all the primary tumors in the oropharynx region (i.e. segmentation not evaluated at the pixel level, rather at the occurrence level).
3. Similarly, detect only the primary tumors in the oropharynx region (discarding lymph nodes and other potentially false positives).

According to these criteria, the task is partially solved. The first criterion, evaluating the segmentation at the pixel level, is fulfilled. At the occurrence level (criteria 2 and 3), however, even the algorithms with the highest DSC output FP and FN regions. Besides, there is still a lot of work to do on highly related tasks, including the segmentation of lymph nodes, the development of super-annotator ground truth as well as the agreement of multiple annotators, and, finally, the prediction of patient outcome following the tumor segmentation.

Similarly to last year's challenge, we identified the same key elements that cause the algorithms to fail in poorly segmented cases. These elements are as follows; low FDG uptake on PET, primary tumor that looks like a lymph node, abnormal uptake in the tongue and tumor present at the border of the oropharynx region. Some examples are illustrated in Fig. 5.1.

Tasks 2 and 3: Predicting PFS

The challengers relied on a variety of approaches and tackled the task quite differently (Table 5.4). A few teams relied on deep learning exclusively, whereas others exploited more classical radiomics pipelines. Some teams also implemented various combinations of both. PET and CT images were also exploited in several different ways. Either as separate inputs or through various fusion techniques, for either deep learning or classical radiomics analysis. Interestingly, despite the recent rise of interest in the development of methods dedicated to the harmonization of multicentric data, either in the image domain through image processing or deep learning based image synthesis (Choe et al., 2019) or in the features domain through batch-harmonization techniques such as ComBat (Da-ano et al., 2020), none of the teams implemented specific multicentric harmonization techniques, beyond usual approaches to take into account the diversity of the images in the training and testing sets by relying on, for example, leave-one-center-out cross-validation and image intensities rescaling or z-score normalization. The use of clinical variables was also the opportunity for challengers to deploy different approaches. Among the methods using deep networks, some encoded the clinical information into images to feed them as input to the deep networks, whereas others integrated them as vectors concatenated in other layers. Some teams elected to rely only on clinical factors without missing values, whereas others implemented some way of imputing missing values in order to exploit all available variables. In addition, some teams pre-selected only a subset of the clinical variables with prior knowledge. Interestingly, some challengers obtained their best performance by building models relying only on clinical variables. Finally, most teams who participated in Task 1 relied on their segmentation output in Tasks 2 and 3, however, a few explored additional or alternative volumes of interest. Interestingly for Task 3, some challengers obtained better results using alternative segmentation or Task 1 outputs instead of the provided reference contours.

Predicting PFS was the objective of both Tasks 2 and 3
The only difference was that the GTVt ROI was provided for Task 3, but not for Task 2. One surprising trend showed that the predictive performance was found to be slightly higher when the GTVt ROI was not used (Task 2), which could be the result of the following. First, fewer teams participated in Task 3, which can be partially explained by the requirement to submit a Docker container instead of direct prediction of hazard scores. Second, for deep learning-based radiomics, using input ROIs is less straightforward than handcrafted radiomics, which makes input contour less relevant. Nevertheless, Starke et al., 2022 used a classical radiomics pipeline and observed that ROIs based on a simple PET-based thresholding approach systematically outperformed a model based on features extracted from the provided GTVt. This suggests that prognostically relevant information is contained not only in the primary tumor area, but also in other (metabolically active) parts such as the lymph nodes. Similar results have been obtained recently in different tumor localizations. For instance, it was shown in uterine cancer that radiomics features extracted from the entire uterus organ in MRI rather than the tumor only led to more accurate models (H. Xie et al., 2019). In cervical cancer patients, specific SUV thresholds in PET images led to more accurate metrics (Leseur et al., 2016), even though this threshold might not be the more accurate to delineate the metabolic uptake tumor volume. Finally, a study in non-small cell lung cancer recently showed that radiomics features extracted from a large volume of interest containing the primary tumor and the surrounding healthy tissues in PET/CT images could be used to train models as accurate as those trained on features extracted from the delineate tumor, provided a consensus of several machine learning algorithms is used for the prediction (Sepehri et al., 2021).

5.5.2 Limitations of the Challenge

The dataset provided in this challenge suffers from several limitations. First, the contours were mainly drawn based on the PET/CT fusion which is not sufficient to clearly delineate the tumor. Other methods such as MRI with gadolinium or contrast CT are the gold standard to obtain the true contours for radiation oncology. Since the target clinical application is radiomics, however, the precision of the contours is not as important as for radiotherapy planning.

Another limitation comes from the definition of the task, given that only one segmentation was drawn on the fusion of PET and CT. For radiomics analysis, it could be beneficial to consider one segmentation per modality since the PET signal is often not contained in the fusion-based segmentation due to the poor spatial resolution of this modality.

5.6 Conclusions

This paper presented a general overview of the HECKTOR challenge including the data, the participation, main results and discussions. The proposed tasks were the segmentation of the primary tumor in oropharyngeal cancer as well as the PFS prediction. The participation was high, with 20, 17, and 6 eligible teams for tasks 1, 2, and 3, respectively. The participation doubled compared to the previous edition, which shows the growing interest in automatic lesion segmentation for H&N cancer.

The task proposed this year was to segment the primary tumor in PET/CT images. This task is not as simple as thresholding the PET image since we target only the primary tumor and the region covered by high PET activation is often too large, going beyond the limits of the tumor tissues. Deep learning methods based on U-Net models were mostly used in the challenge. Interesting ideas were implemented to combine PET and CT complementary information. Model ensembling, as well as data preprocessing and augmentation, seem to have played an important role in achieving top-ranking results.

Acknowledgments

The organizers thank all the teams for their participation and valuable work. This challenge and the winner prizes were sponsored by Siemens Healthineers Switzerland, Bioemtech Greece and Aquilab France (500€ each, for Task 1, 2 and 3 respectively). The software used to centralise the quality control of the GTVt regions was MIM (MIM software Inc., Cleveland,OH), which kindly supported the challenge via free licences. This work was also partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069) and the Swiss Personalized Health Network (SPHN, via the IMAGINE and QA4IQI projects).

6 Discussion

This Chapter is organized as follows. In Section 6.1 we discussed the results of the LRI CNNs presented in Chapter 2 and Chapter 3. Section 6.2 resumes the results of the HECKTOR challenge presented in Chapter 4 and Chapter 5.

6.1 LRI CNNs

In Chapter 2, we implemented two different kinds of 3D LRI CNNs: a bispectral CNN (SSB-CNN) and a spectral CNN (SSE-CNN). We showed the efficiency in terms of accuracy, numbers of parameters, and data efficiency of the bispectral over the standard CNN (Z3-CNN). The experiments were performed on two datasets, one artificial and a subset of the NLST trial.

The artificial dataset was specifically designed to give an advantage to LRI methods, which is what we observed. We obtained similar results in (Andrearczyk, Fageot, et al., 2019) with other LRI architectures. On this artificial dataset, every LRI CNNs outperformed the standard CNN.

This trend was less evident in the NLST dataset. This dataset comprises a training set of 392 CT images of contoured lung nodules. The task is to discriminate between malignant and benign nodules. We tested all the methods on a held-out test set of 93 nodules. The bispectral CNN showed its superiority in this biomedical dataset compared to the spectral CNN. This is probably thanks to the more complete representation of the bispectrum over the spectrum as explained in Section 2.3.5 and illustrated in the toy experiments of Section 2.4.1. The bispectral CNN was also better than the standard CNN, indicating that LRI may be suitable for this data type.

In (Andrearczyk, Fageot, et al., 2020), we comprehensively analyzed different 3D LRI approaches by comparing a standard CNN, a steerable LRI CNN, a G-LRI CNN, and a spectral LRI CNN. The steerable LRI CNN was implementing to the $\mathcal{G}^{\text{steer}}$ introduced in

Section 1.1.9, but in 3D, thus, implemented thanks to the SHs. For this network, multiple discretizations of $SO(3)$ were evaluated, and the best performing one was with a number of $M = 72$ orientations homogeneously distributed on $SO(3)$. The G-LRI CNN is an LRI CNN inspired by the G-CNN of (T. Cohen & Welling, 2016b) and consists of one convolution layer with the same kernels rotated at right-angle rotations summing up to a total number of orientations $M = 24$. Since both networks, the steerable LRI and G-LRI CNN, output response maps at different orientations, an orientation max-pooling was performed at the end of the LRI layer.

In (Andrearczyk, Fageot, et al., 2020) we used a similar architecture than the one of Chapter 2 with the difference that the MLP (see Figure 1.10) contained one additional fully connected layer of 128 hidden units. For the work presented in Chapter 2, we chose a simpler architecture since we wanted to show the effect of the number of convolutional parameters. The fully connected layer with 128 hidden units increased the number of total learnable parameters, potentially hiding performance differences between the various LRI CNN designs.

The best-performing model on the NLST dataset was the G-CNN, achieving 0.877 ± 0.022 accuracy. The bispectral CNN was not evaluated in this study. However, with the same architecture, it achieved an accuracy of 0.878 ± 0.02, slightly outperforming the G-LRI CNN. These results are summarized in Table 6.1.

Table 6.1: Evaluation of the different LRI CNNs and comparison with a baseline radiomics model on the NLST dataset. The results indicates the mean Accuracy$\pm\sigma$ where sigma is the standard deviation obtained with 10 different initializations of the networks. The results for the baseline radiomics model are expressed as the accuracy obtained on the test set and the numbers in parenthesis indicates the confidence interval at 95% evaluated with a bootstrap analysis.

Model	Filters	SH degree	Orientations	Accuracy$_{\pm\sigma}$	Parameters
Standard CNN	2	-	-	0.808$_{\pm 0.02}$	1466
Standard CNN	144	-	-	0.833$_{\pm 0.02}$	105410
S-LRI	4	2	72	0.842$_{\pm 0.03}$	1034
SSE-LRI	4	1	-	0.824$_{\pm 0.02}$	2030
G-LRI	4	-	24	0.877$_{\pm 0.02}$	3818
SSB-LRI	4	3	-	0.878$_{\pm 0.02}$	5158
Radiomics	-	-	-	**0.935** (0.908-0.950)	-

We also evaluated a baseline radiomics model with standard radiomics features extracted from Pyradiomics (Van Griethuysen et al., 2017) and a standard machine learning pipeline. This baseline achieved the best results on the NLST data with an accuracy of 0.935 (0.908 - 0.950). The number in parenthesis indicates the interval at 95% approximated by bootstrap analysis. The results between the CNN and the baseline radiomics model are expressed differently since the standard deviation for the network is evaluated for ten different initializations. The radiomics baseline cannot be evaluated in this way. Thus, we used an approximation of the confidence interval via the bootstrap method to show the potential performance variation of this model.

The results from this radiomics baseline are the reason why we did not publish the manuscript of Chapter 2. The results of the radiomics model outperformed the proposed LRI CNNs, probably because the architecture tested was too shallow. With only one convolutional layer, CNN may not be able to model complex non-linear relationships that radiomics features can capture. For this reason, we then implemented the bispectral LRI layer in a deeper model.

We did it in 2D to simplify the problem and to be able to test this model on data containing directional patterns, which is the subject of Chapter 3. This work included the bispectral layer into a 2D U-Net architecture to segment nuclei instances in histopathological images obtained from the MoNuSeg2018 challenge (N. Kumar et al., 2019). The similarity between the CHs and the SHs made the implementation very close to the 3D bispectral layer.

The accuracy (F-score) results were not better than the U-Net with standard convolutions. However, the output predictions of the LRI U-Net were more robust towards input rotations. We showed in Section 3.3 that 90° rotations of the input had a substantial impact on the prediction of the U-Net with standard convolutions, which had a difference between predictions of around 8% RMSE. In contrast, the bispectral U-Net had an RMSE of around 2.5e-5%. These results were obtained even though both networks were trained with data augmentation, including rotations. While it requires further investigation, this suggests that standard CNN can lead to unstable predictions.

The network architecture and the dataset used to test the LRI U-Net were highly inspired by (Lafarge et al., 2019), where they used a G-CNN to perform the same task. The results they reported are slightly better. Their top performing network achieved an F-score of 0.771 ± 0.06 whereas our top performing bispectral U-Net obtained an F-score of 0.716 ± 0.03. Nonetheless, our results are not directly comparable to theirs since we did not use the same train/test split data. Furthermore, we did ten repetitions with ten different splits, and they did three repetitions with only one split. Varying splits usually lead to higher variance in performance. Thus we cannot directly compare the results of these two works. Moreover, the instance segmentation of the cell nuclei requires one additional post-processing step, which is essential in the final results. Differences in the implementation of this post-processing step may impact the difference in results observed.

To conclude, further experiments where bispectral U-Nets are tested against other G-CNN architecture must be done to determine if LRI is adapted for this task. The G-CNNs-based methods may have an advantage since these methods do not discard the local orientation of the filters and propagate the rotation equivariance throughout the network.

In future work, we will investigate the 3D bispectral layer developed in Chapter 2 with an architecture similar to 3 on 3D segmentation tasks and compare it with other 3D CNNs baseline, including G-convolution-based networks.

From the implementation of the 2D U-Net, we learned valuable lessons that can be applied to the design of the 3D bispectral layer. For instance, we had to use a custom pixel-wise non-linearity directly after the output of the bispectral layer, which was $\sigma(x) = \log(1 + |x|)\text{sgn}(x)$. We introduced this non-linearity to make the layer behaves closer to a linear one. The problem we faced is that when multiple convolutional layers are stacked to form complex CNNs, the initialization of the weights must be made in order to maintain a similar variance between the input layer and output layer to avoid exploding or vanishing gradients (Glorot & Bengio, 2010; He et al., 2015). The bispectrum multiplies three feature maps together (Equation 1.42). Thus the variance between the input and output was very unstable. Using the non-linearity mentioned above helped reduce this variance and initialize the networks.

Another direction for future work could be to evaluate bispectral features for radiomics. The implementation of the LRI operators explained in Section 1.1.9 could be applied within a standard radiomics framework. Some experiments must be made to choose the best radial profiles, and we could evaluate these features against popular radiomics features on the HECKTOR dataset. If the results are conclusive, these features could be quickly implemented in the QuantImage platform to provide clinicians with more robust LRI features.

6.2 HECKTOR Challenge

In the first edition of the challenge, 64 teams were registered, and ten eligible teams submitted a paper. The winner of HECKTOR 2020 obtained a DSC of 0.7591 on the testing set with the method based on a U-Net architecture (Iantsen, Ferreira, et al., 2021).

Experiment on semi-automatic thresholding algorithms, Chapter 4 Section 4.4.6, as well as baseline CNNs, trained only on the PET images, showed that this modality contains most of the signal. This makes sense since the contours were drawn on PET/CT fusion, and the PET signal was used as a surrogate contrast agent for regions with low contrast on the CT. The semi-automatic method, as well as the baseline CNN which used both modalities during the training and the prediction, had a slightly better performance than methods using only the PET modality, showing that the complementary information of the CT is essential to obtain the top-performing segmentation method on this dataset.

We also organized an inter-observer agreement on a subset of the HECKTOR 2020 data, which had an average DSC of 0.6110, similar to previously published agreements(Gudi et al., 2017). In this study, they reported the agreement of three observers with an average DSC of 0.57 using only the CT images for annotation and 0.69 using both PET and CT.

This low agreement illustrates the difficulty of the task for a human observer.

In HECKTOR 2021, 44 teams registered for the primary tumor segmentation task, among which 23 submitted a paper The outcome prediction task had 30 registered teams for 18 submitted papers. The participation in this second edition was higher than in the previous one, showing the community's growing interest in our challenge.

The winner of the segmentation task achieved a DSC of 0.78 with an architecture (J. Xie & Peng, 2022) similar to the winner of HECKTOR 2020. The increase in performance was marginal compared to the previous iteration of the challenge. It can be explained by introducing a new center where tumors were, on average bigger. The median volume of the primary tumor was 7017 and 10879 mm^3 for the testing set of HECKTOR 2020 and the added center, respectively. As a result, the winner of HECKTOR 2021 achieved a DSC of 0.7659 on the test set of HECKTOR 2020, which is slightly better than the winner of HECKTOR 2020 (DSC=0.759). The winner obtained a DSC of 0.7923 on the center added to the test of HECKTOR 2021. We showed in Section 4.4.7 that the tumor size significantly impacts the DSC, which would explain why the results of HECKTOR 2021 are slightly better than the first edition. Furthermore, part of the new center, 25 images, among a total of 75 images, were added to the training of HECKTOR 2021. Hence, the networks had already seen this center during the training phase, which may be another reason they performed better in this new center. Furthermore, we observed that the sixth best performing teams obtained a DSC above 0.77, illustrating the closeness of the different methods in terms of performance.

In conclusion, there was no significant improvement in primary tumor segmentation between HECKTOR 2020 and HECKTOR 2021. Given the low inter-observer agreement, it is likely that finding methods that outperform the current approaches will overfit the data and will not improve the actual segmentation performance. It is important to note that only one annotator contoured the data of HECKTOR 2020; thus, overfitting these data is more likely. In the subsequent editions of the challenge multiple annotators were involved, thus mitigating the risk of overfitting one annotator.

Since it seems that a plateau was reached in the segmentation performance, we decided to complexify the task and stop providing the fixed-size bounding boxes in HECKTOR 2022. Currently, the participants have to automatically preprocess the image first to identify the oropharyngeal region and then apply their method to segment the primary tumor and the lymph nodes.

For the PFS prediction, some interesting insights were highlighted. First, this task was proposed as two separate tasks: Tasks 2 and 3. In Task 2, during the training phase, the team had access to the images, contours, and clinical data. During the evaluation phase, they only had access to the image and clinical data. The task was to learn from the entire image or to segment the image with an algorithm of Task 1 followed by a classical radiomics pipeline to predict PFS. Task 3 was designed to allow participants access to the contoured test data. We used a Docker-based[I] submission system to avoid providing the ground truth for Task 1. Thus, the participants could train their method with the contoured data and submit their models to our team, who would run them on the test images with the ground-truth annotations.

Interestingly, fully automatic methods performed slightly better (Task 2) than methods that had access to the contours of the primary tumor (Task 3). This observation must be taken with caution since the number of participants for both tasks was different: 17 teams participated in Task 2 against 6 in Task 3.

Moreover, one team (Starke et al., 2022), who used a standard radiomics pipeline, observed systematically better results with ROIs based on simple PET thresholding then with the provided ground-truth ROIs. This may indicate that important information is contained outside the primary tumor segmentation, maybe in the lymph nodes or in the border of the primary tumor. One limitation of this hypothesis is that we provided only the contours for the primary tumor. Since in HECKTOR 2022, we also introduced the GTVn segmentation, we will see if this trend holds. If it is the case, then it would suggest that we may have extended the ROIs for future radiomics analysis, as it was also highlighted in different cancer locations (Leseur et al., 2016; Sepehri et al., 2021; H. Xie et al., 2019).

The fact that fully automatic approaches (Task 2) seemed to perform better than standard radiomics has exciting implications since it means that we are a step closer to testing these models on large-scale datasets without the need to annotate them.

For HECKTOR 2022, we organized a more comprehensive inter-observer agreement to test the variation of delineation on GTVt and GTVn. We hope to observe a better agreement than the one previously conducted since the radiologists have agreed on a set of guidelines for the segmentation of the GTVt and GTVn.

As future work and since we will soon have the results of HECKTOR 2022, we will test if simple ROIs based on PET-thresholding yield similar results to HECKTOR 2021. Although, this time, we will be able to evaluate if the gain in performance comes from the GTVn or elsewhere. Furthermore, an excellent post-challenge analysis would be to evaluate the best performing fully automatic method with methods from interpretability in order to determine which part of the image is important for outcome prediction.

[I] https://www.docker.com/ as of August 2022.

A 3D Bispectral LRI CNN

A.1 Clebsch-Gordan Matrices

Let us fix $n_1, n_2 \geq 0$. The Clebsch-Gordan matrix C_{n_1,n_2} is characterized by the fact that it block-diagonalizes the Kronecker product of two Wigner-D matrices as

$$\mathrm{D}_{n_1}(\mathrm{R}) \otimes \mathrm{D}_{n_2}(\mathrm{R}) = \mathrm{C}_{n_1,n_2} \left[\bigoplus_{i=|n_1-n_2|}^{n_1+n_2} \mathrm{D}_i(\mathrm{R}) \right] \mathrm{C}^{\dagger}_{n_1,n_2} \tag{A.1}$$

for any matrix rotation $\mathrm{R} \in SO(3)$. This means in particular that C_{n_1,n_2} has $\sum_{n=|n_1-n_2|}^{n_1+n_2}(2n+1)$ rows and $(2n_1+1)(2n_2+1)$ columns. These two numbers are actually equal, hence $\mathrm{C}_{n_1,n_2} \in \mathbb{R}^{(2n_1+1)(2n_2+1)\times(2n_1+1)(2n_2+1)}$, but the relation (A.1) also reveals the structure of the matrix, whose coefficients are indexed as $\mathrm{C}_{n_1,n_2}[(n,m),(m_1,m_2)]$, with $n \in \{|n_1-n_2|, \dots, (n_1+n_2)\}$, $m_1 \in \{-n_1, \dots, n_1\}$, and $m_2 \in \{-n_2, \dots, n_2\}$. In the literature, the Clebsch-Gordan coefficients are often written with bracket notations, that reveal some of their symmetries (Alex et al., 2011). Moreover, the Clebsch-Gordan matrix has many 0 entries. We indeed have that

$$\begin{aligned}\mathrm{C}_{n_1,n_2}[(n,m),(m_1,m_2)] &= 0 \text{ if } m \neq m_1+m_2 \\ &= \langle n_1 m_1 n_2 m_2 | n(m_1+m_2)\rangle,\end{aligned}$$

where $\langle | \rangle$ is the bracket notation used for instance in (Chaichian & Hagedorn, 1998, Chapter 5.3.1).

A.2 Proof of Proposition 3

The equivariance to translations is simpler and similar to the equivariance to rotations, therefore we skip it (it simply uses that $(I(\cdot - \boldsymbol{x}_0) * \kappa_n^m)(\boldsymbol{x}) = (I * \kappa_n^m)(\boldsymbol{x} - \boldsymbol{x}_0)$). Let $\boldsymbol{\mathcal{F}}_n(\boldsymbol{x})$ and $\boldsymbol{\mathcal{F}}'_n(\boldsymbol{x})$ be the Fourier feature maps of I and $I(\mathrm{R}_0 \cdot)$ respectively, with $\mathrm{R}_0 \in SO(3)$.

According to (2.4) applied to $\mathrm{R} = \mathrm{R}_0^{-1}$, we have that

$$\kappa_n^m(\mathrm{R}_0^{-1}\cdot) = \sum_{m'=-n}^{n} \mathrm{D}_n(\mathrm{R}_0^{-1})_{m,m'}\kappa_n^m. \tag{A.2}$$

Moreover, we have that $(I(\mathrm{R}_0\cdot) * \kappa_n^m)(\boldsymbol{x}) = (I * \kappa_n^m(\mathrm{R}_0^{-1}\cdot))(\mathrm{R}_0\boldsymbol{x})$. Together with (A.2), this implies that

$$\boldsymbol{\mathcal{F}'}_n(\boldsymbol{x}) = ((I(\mathrm{R}_0\cdot) * \kappa_n^m)(\boldsymbol{x}))_m = \boldsymbol{\mathcal{F}}(\mathrm{R}_0\boldsymbol{x})\mathrm{D}_n(\mathrm{R}_0^{-1}\boldsymbol{x}). \tag{A.3}$$

This implies that

$$\begin{aligned}
\mathcal{G}^{\mathrm{SSB}}_{n,n',\ell}\{I(\mathrm{R}_0\cdot)\}(\boldsymbol{x}) &= \mathcal{B}\{\boldsymbol{\mathcal{F}'}_n(\boldsymbol{x}), \boldsymbol{\mathcal{F}'}_{n'}(\boldsymbol{x}), \boldsymbol{\mathcal{F}'}_\ell(\boldsymbol{x})\}\\
&= \mathcal{B}\{\boldsymbol{\mathcal{F}}_n(\mathrm{R}_0\boldsymbol{x})\mathrm{D}_n(\mathrm{R}_0^{-1}), \ldots\\
&\qquad \boldsymbol{\mathcal{F}}_{n'}(\mathrm{R}_0\boldsymbol{x})\mathrm{D}_{n'}(\mathrm{R}_0^{-1}), \boldsymbol{\mathcal{F}}_\ell(\mathrm{R}_0\boldsymbol{x})\mathrm{D}_\ell(\mathrm{R}_0^{-1})\}\\
&= \mathcal{B}\{\boldsymbol{\mathcal{F}}_n(\mathrm{R}_0\boldsymbol{x}), \boldsymbol{\mathcal{F}}_{n'}(\mathrm{R}_0\boldsymbol{x}), \boldsymbol{\mathcal{F}}_\ell(\mathrm{R}_0\boldsymbol{x})\}\\
&= \mathcal{G}^{\mathrm{SSB}}_{n,n',\ell}\{I\}(\mathrm{R}_0\boldsymbol{x}),
\end{aligned}$$

where we used the invariance of the bispectrum for the third equality. This demonstrates the equivariance of the operator $\mathcal{G}^{\mathrm{SSB}}_{n,n',\ell}$ with respect to rotations. Finally, the locality simply follows from the fact that the convolution $I * \kappa_n^m(\boldsymbol{x})$ depends on the values of $I(\boldsymbol{x} - \boldsymbol{y})$ with $\boldsymbol{y}$ in the support of κ_n^m, which is bounded as soon as h_n is compactly supported, what we assumed.

B 2D Bispectral LRI U-Net

B.1 Proof of Theorem 1

For any $n \in \mathbb{Z}$, the operator $I \mapsto \mathcal{C}_n\{I\} = I * \kappa_n$ is equivariant to translations (as a convolution) and local (due to the fact that h, and therefore $\kappa_n(\rho, \theta) = h(\rho)\mathrm{e}^{\mathrm{j}n\theta}$, have a finite support). Then, the operator $\mathcal{G}_{n,n'}$ inherits these properties, as is clear from its definition in Eq. (3.4).

We moreover observe that, for any rotation matrix R_{θ_0} associated with the angle θ_0, we have that

$$I(R_{\theta_0}\cdot) * \kappa_n = (I * \kappa_n(R_{\theta_0}^{-1}\cdot))(R_{\theta_0}\cdot) = e^{\mathrm{j}n\theta_0}(I * \kappa_n)(R_{\theta_0}\cdot), \tag{B.1}$$

where the first inequality comes from the relation $f(R_{\theta_0}) * g = (f * g(R_{\theta_0}^{-1}\cdot))(R_{\theta_0}\cdot)$ valid for any $f, g \in L_2(\mathbb{R}^2)$ and the second uses that $\kappa_n(R_{\theta_0}\cdot) = e^{\mathrm{j}n\theta_0}\kappa_n$. This implies that, for any image I and any $\boldsymbol{x} \in \mathbb{R}^2$,

$$\mathcal{C}_n\{I(\mathrm{R}_{\theta_0}\cdot)\} = I * (\kappa_n(R_{\theta_0}\cdot))\,(\boldsymbol{x}) = e^{\mathrm{j}n\theta_0}\mathcal{C}_n\{I\}(\mathrm{R}_{\theta_0}\boldsymbol{x}). \tag{B.2}$$

We have therefore that

$$\begin{aligned}\mathcal{G}_{n,n'}\{I(R_{\theta_0}\cdot)\}(\boldsymbol{x}) &= e^{\mathrm{j}n\theta_0}e^{\mathrm{j}n'\theta_0}\overline{e^{\mathrm{j}(n+n')\theta_0}}\mathcal{C}_n\{I\}(\mathrm{R}_{\theta_0}\boldsymbol{x})\mathcal{C}_{n'}\{I\}(\mathrm{R}_{\theta_0}\boldsymbol{x})\mathcal{C}_{n+n'}\{I\}(\mathrm{R}_{\theta_0}\boldsymbol{x})\\ &= \mathcal{G}_{n,n'}\{I\}(R_{\theta_0}\boldsymbol{x})\end{aligned}$$

and the operator $\mathcal{G}_{n,n'}$ is globally rotation equivariant. Being local and equivariant to shifts and rotations, $\mathcal{G}_{n,n'}$ is LRI.

B.2 Additional Results

In this Appendix, we report additional results. First, we computed additional metrics to investigate the performance difference between the bispectral and the standard U-Nets. The precision and recall for the bispectral U-Net (N=7) were 0.7004 ± 0.0617 and 0.7500 ± 0.0350, respectively. For the standard U-Net, we obtained 0.7156 ± 0.0505 and 0.7686 ± 0.0400. We also trained a standard U-Net without masking the kernels, obtaining an F-score of 0.7318 ± 0.0220.

Second, we compare the computational time between the two approaches. The average forward time on 80 1000x1000 images for the bispectral U-Net with $N = 0, 2, 4$, and 7 were 0.14, 0.42, 1.05, and 2.17 seconds respectively. The average forward time for the standard U-Net was 0.07 seconds.

Finally, we report some examples of predictions made by the bispectral and standard U-Net. Fig. B.1 illustrates the predictions on a patch where the bispectral U-Net outperforms the standard U-Net. Fig. B.2 shows a patch where the bispectral U-Net over segments and thus performs worse than the spectral U-Net.

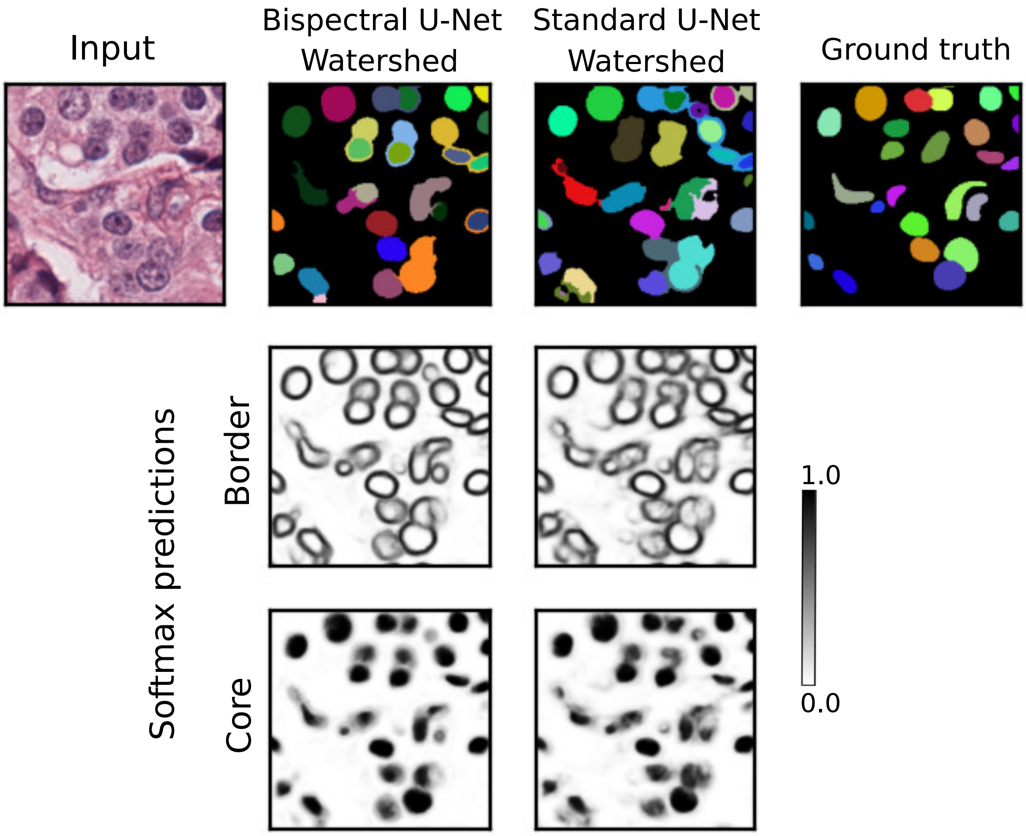


Figure B.1: Illustration of predictions where the bispectral U-Net outperforms the standard U-Net. The F-score on this patch for the bispectral and standard U-Net are, respectively, 0.8929 and 0.7931. The colors in the top row images are used to highlight the different nucleus instances obtained after the application of the watershed algorithm.

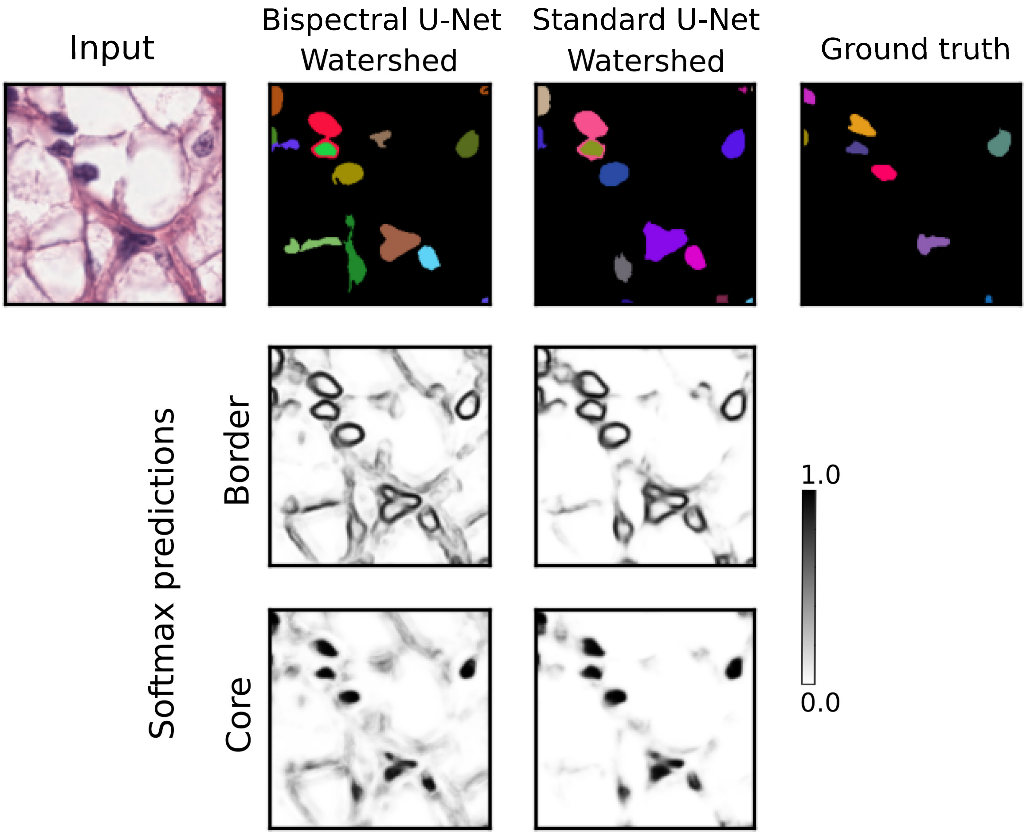


Figure B.2: Illustration of predictions where the bispectral U-Net over segments. The F-score on this patch for the bispectral and standard U-Net are, respectively, 0.6667 and 0.7619. The colors in the top row images are used to highlight the different nucleus instances obtained after the application of the watershed algorithm.

The quantitative results seem to indicate that the bispectral U-Net always performed a little worse than the standard U-Net. However, as highlighted in Fig. B.1 and B.2, it is difficult to assess whether these differences come from the post-processing step.

B.3 Comparison with Spectral U-Net

This appendix describes the results with a similar architecture to the bispectral U-Net. However, the invariant used here is the spectrum yielding a spectral U-Net, which is very similar to the design proposed in (Eickenberg et al., 2017) or in (Andrearczyk, Fageot, et al., 2020). The design of the LRI layer is almost the same and the associated LRI operator is defined as:

$$\mathcal{G}_n\{I\}(\boldsymbol{x}) = \mathcal{C}_n\{I\}(\boldsymbol{x})\overline{\mathcal{C}_n\{I\}(\boldsymbol{x})}. \tag{B.3}$$

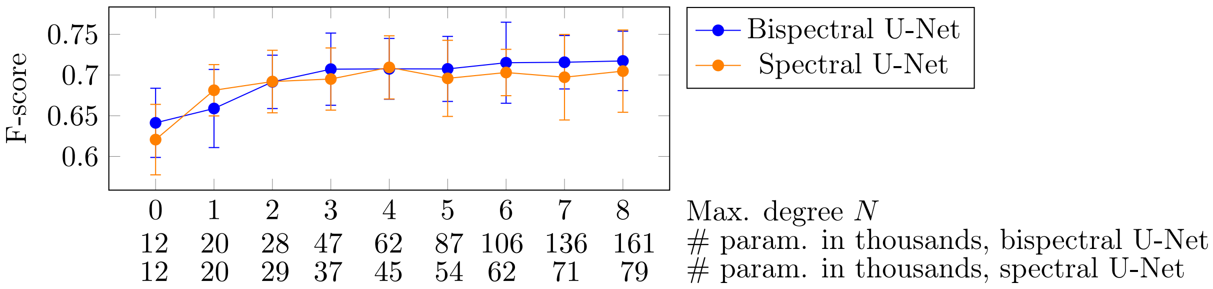


Figure B.3: Performance evaluation of the different networks. The average F-scores across 10 repetitions of the proposed bispectral (blue) and spectral (orange) U-Net evaluated at different maximum degrees N are reported.

This invariant is equivalent to taking the modulus, *i.e.* spectrum, of the Fourier coefficients $\mathcal{C}_n$. The results are reported in Fig. B.3.

The results suggest slightly better performance for the bispectral U-Net, but the difference remains too marginal to conclude.

C Overview and Post-Analyses of HECKTOR 2020

C.1 Participants' Algorithms Summary

In (Iantsen, Visvikis, et al., 2021), Iantsen et al. proposed a model based on a U-Net architecture with residual layers and supplemented with 'Squeeze and Excitation' (SE) normalization, previously developed by the same authors for brain tumor segmentation. An unweighted sum of soft Dice loss and Focal Loss was used for training. The test results were obtained as an ensemble of eight models trained and validated on different splits of the training set. No data augmentation was performed.

In (J. Ma & Yang, 2021), Ma and Yang used a combination of U-Nets and hybrid active contours. First, 3D U-Nets are trained to segment the tumor (with a cross-validation on the training set). Then, the segmentation uncertainty is estimated by model ensembles on the test set to select the cases with high uncertainties. Finally, the authors used a hybrid active contour model to refine the high uncertainty cases. The U-Nets were trained with an unweighted combination of Dice loss and top-K loss. No data augmentation was used.

In (Zhu et al., 2021), Zhu et al. used a two steps approach. First, a classification network (based on ResNet) selects the axial slices which may contain the tumor. These slices are then segmented using a 2D U-Net to generate the binary output masks. Data augmentation was applied by shifting the crop around the provided bounding boxes and the U-Net was trained with a soft Dice loss. The preprocessing includes clipping the CT and the PET, standardizing the HU within the cropped volume and scaling the range of the PET to correspond to the CT range by dividing it by a factor of 10.

In (Yuan, 2021), Yuan proposed to integrate information across different scales by using a dynamic Scale Attention Network (SA-Net), based on a U-Net architecture. Their network incorporates low-level details with high-level semantics from feature maps at different scales. The network was trained with standard data augmentation and with a Jaccard distance loss, previously developed by the authors. The results on the test set were obtained as an ensemble of ten models.

In (H. Chen et al., 2021), Chen et al. proposed a three-step framework with iterative refinement of the results. In this approach, multiple 3D U-Nets are trained one-by-one using a Dice loss without data augmentation. The predictions and features of previous models are captured as additional information for the next one to further refine the segmentation.

In (Ghimire et al., 2021), Ghimire et al. developed a patch-based approach to tackle the memory issue associated with 3D images and networks. They used an ensemble of conventional convolutions (with small receptive fields capturing fine details) and dilated convolutions (with a larger receptive field of capturing global information). They trained their model with a weighted cross-entropy and dice loss and random left-right flips of the patches were applied for data augmentation. Finally, an ensemble of the best two models selected during cross-validation was used for predicting the segmentation of the test data.

In (Yousefirizi & Rahmim, 2021), Yousefirizi and Rahmim proposed a deep 3D model based on SegAN, a generative adversarial network (GAN) for medical image segmentation. An improved polyphase V-net (to help preserve boundary details) is used for the generator and the discriminator network has a similar structure to the encoder part of the former. The networks were trained using a combination of Mumford-Shah (MS) and multi-scale Mean Absolute Error (MAE) losses, without data augmentation.

In (J. Xie & Peng, 2021), Xie and Peng proposed a 3D scSE nnU-Net model, improving upon the 3D nnU-Net by integrating the spatial and channel 'Squeeze and Excitation' (scSE) blocks. They trained the model with a weighted combination of Dice and cross-entropy losses, together with standard data augmentation techniques (rotation, scaling etc.). To preprocess the CT images an automated level-window-like clipping of intensity values is performed based on the 0.5 and 99.5th percentile of these values. The intensity values of the PET are standardized by subtracting the mean and then, by dividing by the standard deviation of the image.

In (M. Naser et al., 2021), Naser et al. used a variant of 2D and 3D U-Net (we report the best result, with the 3D model). The models were trained with a combination of Dice and cross-entropy losses with standard data augmentation.

In (Rao et al., 2021), Rao et al. proposed an ensemble of two methods, namely a 3D U-Net and another 2D U-Net variant with 3D context. A top-k loss was used to train the models without data augmentation.

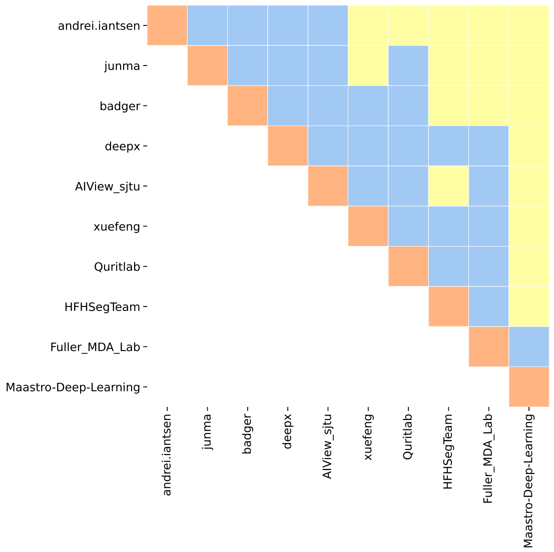


Figure C.1: The significance matrix represents significant tests for the one-sided Wilcoxon signed-rank test at a 5% significance level, adjusted for multiple comparisons with the Holm-Bonferroni method for 45 hypotheses. For each pair, the alternative hypothesis is that the best team has a greater score. For instance, for the `andrei.iantsen-junma` pair the alternative is that `andrei.iantsen` has a better DSC than `junma`. The yellow color indicates that the team on the line of the matrix has significantly better DSC than the team on the column. Blue color means no significant difference. Orange color is used as a visual guide to show pairs of identical teams.

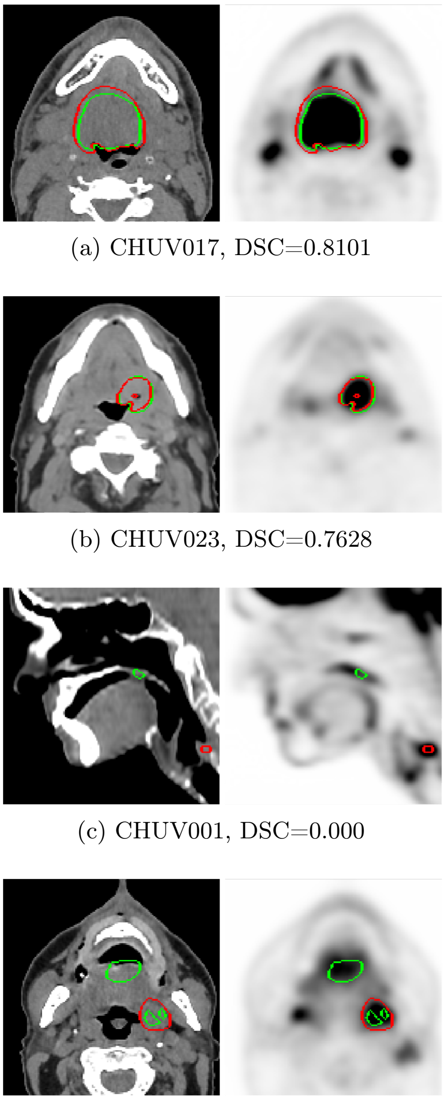
C.2 HECKTOR 2020: Additional plots

C.3 HECKTOR 2020: Centers statistics

In Table C.1, we report the differences between the five centers in terms of image properties such as devices, pixel spacing and slice spacing. We also disclose the distribution of GTVt volumes in Fig. C.6 and Table C.2.

Table C.1: Statistics of the different centers. GTVt volumes are computed after iso-resampling at $1 \times 1 \times 1$ mm^3. The GTVt volumes are reported in cm^3 as average plus the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile in parenthesis. All devices are hybrid PET/CT.

Center	Pixel spacing CT	Slice spacing CT	Pixel spacing PT	Slice spacing PT	Device
HGJ	0.98 (0.98 - 0.98)	3.27 (3.27 - 3.27)	3.52 (3.52 - 4.69)	3.27 (3.27 - 3.27)	Discovery ST, GE Healthcare
CHUS	1.17(0.68- 1.17)	3.00 (2.00 - 5.00)	4.00 (4.00 - 4.00)	4.00 (4.00 - 4.00)	GeminiGXL 16, Philips
HMR	0.98 (0.98 1.37)	3.27 (3.27 - 3.27)	3.52 (3.52 - 5.47)	3.27 (3.27 - 3.27)	Discovery STE, GE Healthcare
CHUM	0.98 (0.98 - 1.37)	1.50 (1.50 - 3.27)	4.00 (3.52 - 5.47)	4.00 (3.27 - 4.06)	Discovery STE, GE Healthcare
CHUV	1.37 (0.98 - 1.37)	3.27 (1.00 - 4.25)	2.73 (2.73 - 3.91)	3.27 (3.27 - 4.25)	Discovery D690 TOF, GE Healthcare



(a) CHUV017, DSC=0.8101

(b) CHUV023, DSC=0.7628

(c) CHUV001, DSC=0.000

(d) CHUV019, DSC=0.0905

Figure C.2: Examples of results of the second algorithm (`junma` (J. Ma & Yang, 2021)). The automatic segmentation results (green) and ground truth annotations (red) are displayed on 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the entire image (see Eq. 4.1). (a), (b) Excellent segmentation results, detecting the GTVt of the primary oropharyngeal tumor localized at the base of the tongue and discarding the laterocervical lymph nodes despite high FDG uptake on PET. (c) Incorrect segmentation of the top volume at the level of the soft palate; (d) Incorrect segmentation of the smaller volume below the level of the hyoid bone.

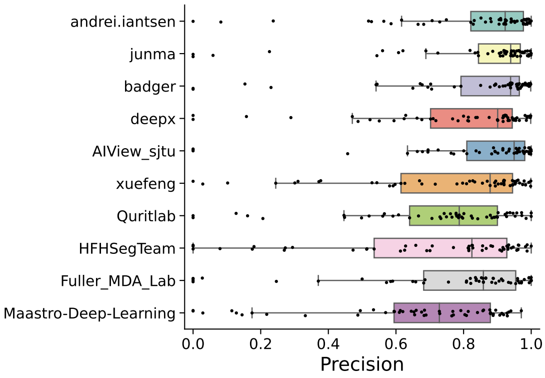


Figure C.3: Box plots of the distribution of the precision on the 53 test cases for each participant, ordered by decreasing rank.

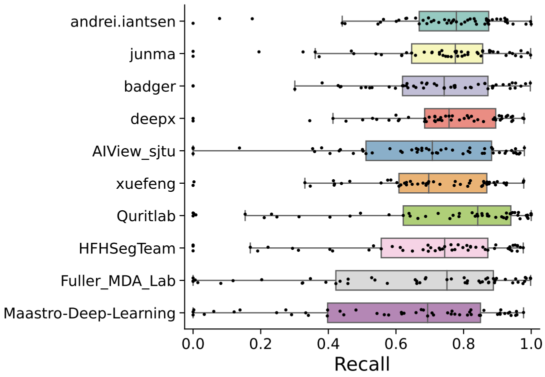


Figure C.4: Box plots of the distribution of the recall on the 53 test cases for each participant, ordered by decreasing rank.

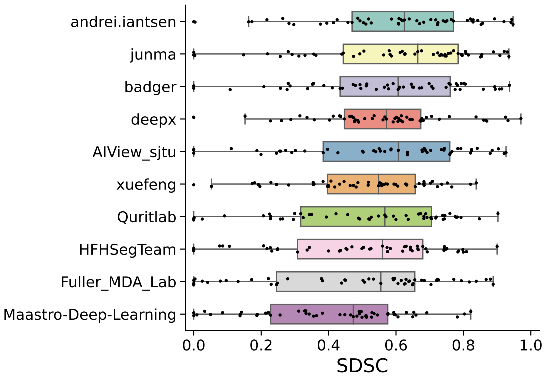


Figure C.5: Box plots of the distribution of the 53 test SDSCs for each participant, ordered by decreasing rank.

Table C.2: Average GTVt volume for the five center used in this challenge. The numbers in parenthesis represent the 5th and 95th respectively.

Center	GTVt volume
HGJ	14.913 (2.263 - 38.879)
CHUS	14.209 (1.837 - 42.967)
HMR	23.622 (2.412 - 88.785)
CHUM	9.866 (1.358 - 24.884)
CHUV	13.317 (1.725 - 41.212)

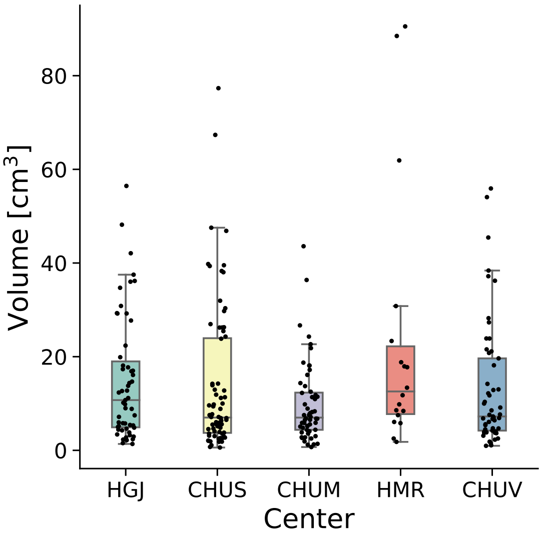


Figure C.6: Box plots of the distribution of the GTVt volumes per center.

D Overview of HECKTOR 2021

D.1 Challenge Information

In this appendix, we list important information about the challenge as suggested in the BIAS guidelines (Maier-Hein et al., 2020).

Challenge name

HEad and neCK TumOR segmentation and outcome prediction challenge (HECKTOR) 2021

Organizing team

(Authors of this paper) Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt and Adrien Depeursinge

Life cycle type

A fixed submission deadline was set for the challenge results.

Challenge venue and platform

The challenge is associated with MICCAI 2021. Information on the challenge is available on the website, together with the link to download the data, the submission platform and the leaderboard[I].

Participation policies

(a) Task 1: Algorithms producing fully-automatic segmentation of the test cases were

[I] www.aicrowd.com/challenges/hecktor

allowed. Task 2 and 3: Algorithms producing fully-automatic PFS risk score prediction of the test cases were allowed.
(b) The data used to train algorithms was not restricted. If using external data (private or public), participants were asked to also report results using only the HECKTOR data.
(c) Members of the organizers' institutes could participate in the challenge but were not eligible for awards.
(d) Task 1: The award was 500 euros, sponsored by Siemens Healthineers Switzerland. Task 2: The award was 500 euros, sponsored by Aquilab. Task 3: The award was 500 euros, sponsored by Bioemtech.
(e) Policy for results announcement: The results were made available on the AIcrowd leaderboard and the best three results of each task were announced publicly. Once participants submitted their results on the test set to the challenge organizers via the challenge website, they were considered fully vested in the challenge, so that their performance results (without identifying the participant unless permission was granted) became part of any presentations, publications, or subsequent analyses derived from the challenge at the discretion of the organizers.
(f) Publication policy: This overview paper was written by the organizing team's members. The participating teams were encouraged to submit a paper describing their method. The participants can publish their results separately elsewhere when citing the overview paper, and (if so) no embargo will be applied.

Submission method
Submission instructions are available on the website[II] and are reported in the following.
Task 1: Results should be provided as a single binary mask per patient (1 in the predicted GTVt) in .nii.gz format. The resolution of this mask should be the same as the original CT resolution and the volume cropped using the provided bounding-boxes. The participants should pay attention to saving NIfTI volumes with the correct pixel spacing and origin with respect to the original reference frame. The NIfTI files should be named [PatientID].nii.gz, matching the patient names, e.g. CHUV001.nii.gz and placed in a folder. This folder should be zipped before submission. If results are submitted without cropping and/or resampling, we will employ nearest neighbor interpolation given that the coordinate system is provided.
Task 2: Results should be submitted as a CSV file containing the patient ID as "PatientID" and the output of the model (continuous) as "Prediction". An individual output should be anti-concordant with the PFS in days (i.e., the model should output a predicted risk score).
Task 3: For this task, the developed methods will be evaluated on the testing set by the organizers by running them within a docker provided by the challengers. Practically, your method should process one patient at a time. It should take 3 nifty files as inputs (file 1: the PET image, file 2: the CT image, file 3: the provided ground-trugh segmentation

[II] https://www.aicrowd.com/challenges/miccai-2021-hecktor#results-submission-format

mask, all 3 files have the same dimensions, the ground-truth mask contains only 2 values: 0 for the background, 1 for the tumor), and should output the predicted risk score produced by your model.
Participants were allowed five valid submissions per task. The best result was reported for each task/team. For a team submitting multiple runs to task one, the best result was determined as the highest ranking result within these runs (see ranking description in Section 5.3.1).

Challenge schedule
The schedule of the challenge, including modifications, is reported in the following.

- the release date of the training cases: ~~June 01~~ June 04 2021
- the release date of the test cases: ~~Aug. 01~~ Aug. 06 2021
- the submission date(s): opens Sept. 01 2021 closes Sept. 10 Sept. 14 2021 (23:59 UTC-10)
- paper abstract submission deadline: Sept. 15 2021 (23:59 UTC-10)
- full paper submission deadline: Sept. 17 2021 (23:59 UTC-10)
- the release date of the ranking: ~~Sept. 17 2021~~ Sept. 27 2021
- associated workshop days: Sept. 27 2021

Ethics approval
Montreal: CHUM, CHUS, HGJ, HMR data (training): The ethics approval was granted by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).
Lausanne: CHUV data (testing): The ethics approval was obtained from the Commission cantonale (VD) d'éthique de la recherche sur l'être humain (CER-VD) with protocol number: 2018-01513.
Poitiers: CHUP data (partly training and testing): The fully anonymized data originates from patients who consent to the use of their data for research purposes.

Data usage agreement
The participants had to fill out and sign an end-user-agreement in order to be granted access to the data. The form can be found under the Resources tab of the HECKTOR website.

Code availability

The evaluation software was made available on our github page[III]. The participating teams decided whether they wanted to disclose their code (they were encouraged to do so).

Conflict of interest

No conflict of interest applies. Fundings are specified in the acknowledgments. Only the organizers had access to the test cases' ground truth contours.

Author contributions

Vincent Andrearczyk:
Design of the tasks and of the challenge, writing of the proposal, development of baseline algorithms, development of the AIcrowd website, writing of the overview paper, organization of the challenge event, organization of the submission and reviewing process of the participants' papers.

Valentin Oreiller:
Design of the tasks and of the challenge, writing of the proposal, development of the AIcrowd website, development of the evaluation code, writing of the overview paper, organization of the challenge event, organization of the submission and reviewing process of the papers.

Sarah Boughdad:
Design of the tasks and of the challenge, annotations.

Catherine Cheze Le Rest:
Design of the tasks and of the challenge, annotations.

Hesham Elhalawani:
Design of the tasks and of the challenge, annotations.

Mario Jreige:
Design of the tasks and of the challenge, quality control/annotations, annotations, revision of the paper and accepted the last version of the submitted paper.

John O. Prior:
Design of the tasks and of the challenge, revision of the paper and accepted the last version of the submitted paper.

Martin Vallières:

[III] github.com/voreille/hecktor

Design of the tasks and of the challenge, provided the initial data and annotations for the training set (Vallieres et al., 2017), revision of the paper and accepted the last version of the submitted paper.

Dimitris Visvikis:
Design of the task and challenge.

Mathieu Hatt:
Design of the tasks and of the challenge, writing of the proposal, writing of the overview paper, organization of the challenge event.

Adrien Depeursinge:
Design of the tasks and of the challenge, writing of the proposal, writing of the overview paper, organization of the challenge event.

D.2 Image Acquisition Details

HGJ: For the PET portion of the FDG-PET/CT scan, a median of 584 MBq (range: 368-715) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 180-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a span (axial mash) of 5. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 × 3.52 mm 2 (range: 3.52-4.69). For the CT portion of the FDG-PET/CT scan, an energy of 140 kVp with an exposure of 12 mAs was used. The CT slice thickness resolution was 3.75 mm and the median in-plane resolution was 0.98 × 0.98 mm 2 for all patients.

CHUS: For the PET portion of the FDG-PET/CT scan, a median of 325 MBq (range: 165-517) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 150 s (range: 120-151) per bed position. Attenuation corrected images were reconstructed using a LOR-RAMLA iterative algorithm. The FDG-PET slice thickness resolution was 4 mm and the median in-plane resolution was 4×4 mm 2 for all patients. For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 12-140) with a median exposure of 210 mAs (range: 43-250) was used. The median CT slice thickness resolution was 3 mm (range: 2-5) and the median in-plane resolution was 1.17 × 1.17 mm 2 (range: 0.68-1.17).

HMR: For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227-859) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 360 s (range: 120-360) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3-5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 × 3.52 mm 2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120-140) with a median exposure of 11 mAs (range: 5-16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98 × 0.98 mm 2 (range: 0.98-1.37).

CHUM: For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199-3182) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 120-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 3 (range: 3-5). The median FDG-PET slice thickness resolution was 4 mm (range: 3.27-4) and the median in-plane resolution was 4 × 4 mm 2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 120-140) with a median exposure of 350 mAs (range: 5-350) was used. The median CT slice thickness resolution was 1.5 mm (range: 1.5-3.75) and the median in-plane resolution was 0.98 × 0.98 mm 2 (range: 0.98-1.37).

CHUV: The patients fasted at least 4h before the injection of 4 Mbq/kg of(18F)-FDG (Flucis). Blood glucose levels were checked before the injection of (18F)-FDG. If not contra-indicated, intravenous contrast agents were administered before CT scanning. After a 60-min uptake period of rest, patients were imaged with the PET/CT imaging system. First, a CT (120 kV, 80 mA, 0.8-s rotation time, slice thickness 3.75 mm) was performed from the base of the skull to the mid-thigh. PET scanning was performed immediately after acquisition of the CT. Images were acquired from the base of the skull to the mid-thigh (3 min/bed position). PET images were reconstructed by using an ordered-subset expectation maximization iterative reconstruction (OSEM) (two iterations, 28 subsets) and an iterative fully 3D (DiscoveryST). CT data were used for attenuation calculation.

CHUP: PET/CT acquisition began after 6 hours of fasting and 60±5 min after injection of 3 MBq/kg of 18F-FDG (421±98 MBq, range 220-695 MBq). Non-contrast-enhanced, non-respiratory gated (free breathing) CT images were acquired for attenuation correction (120 kVp, Care Dose® current modulation system) with an in-plane resolution of 0.853×0.853 mm2 and a 5 mm slice thickness. PET data were acquired using 2.5 minutes per bed position routine protocol and images were reconstructed using a CT-based attenuation correction and the OSEM-TrueX-TOF algorithm (with time-of-flight and spatial resolution modeling, 3 iterations and 21 subsets, 5 mm 3D Gaussian post-filtering, voxel size 4×4×4 mm3).

Bibliography

Abler, D., Schaer, R., Oreiller, V., Verma, H., Reichenbach, J., Aidonopoulos, O., Evéquoz, F., Jreige, M., Prior, J. O., & Depeursinge, A. (in preparation). QuantImage v2: an integrated clinician-in-the-loop cloud platform for radiomics and machine learning research. *Radiology.*

Ablowitz, M. J., Kaup, D. J., Newell, A. C., & Segur, H. (1974). The inverse scattering transform-fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, *53*(4), 249–315.

Ahonen, T., Matas, J., He, C., & Pietikäinen, M. (2009). Rotation invariant image description with local binary pattern histogram Fourier features. *Scandinavian Conference on Image Analysis*, 61–70.

Alex, A., Kalus, M., Huckleberry, A., & von Delft, J. (2011). A numerical algorithm for the explicit calculation of $SU(N)$ and $SL(N, C)$ Clebsch–Gordan coefficients. *Journal of Mathematical Physics*, *52*(2), 023507.

Al-Ibraheem, A., Buck, A., Krause, B. J., Scheidhauer, K., & Schwaiger, M. (2009). Clinical applications of fdg pet and pet/ct in head and neck cancer. *Journal of oncology*, *2009.*

An, C., Chen, H., & Wang, L. (2022). A coarse-to-fine framework for head and neck tumor segmentation in CT and PET images. *Lecture Notes in Computer Science (LNCS) Challenges.*

Andrearczyk, V., & Depeursinge, A. (2018). Rotational 3D texture classification using group equivariant CNNs. *arXiv:1810.06889.*

Andrearczyk, V., & Whelan, P. (2016). Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, *84*, 63–69.

Andrearczyk, V., Fageot, J., Oreiller, V., Montet, X., & Depeursinge, A. (2019). Exploring local rotation invariance in 3D CNNs with steerable filters. *International Conference on Medical Imaging with Deep Learning.*

Andrearczyk, V., Fageot, J., Oreiller, V., Montet, X., & Depeursinge, A. (2020). Local Rotation Invariance in 3D CNNs. *Medical Image Analysis.*

Andrearczyk, V., Fontaine, P., Oreiller, V., Castelli, J., Jreige, M., O. Prior, J., & Depeursinge, A. (2021). Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. *PRedictive Intelligence in MEdicine (PRIME at MICCAI).*

Andrearczyk, V., Oreiller, V., Boughdad, S., Cheze-Le-Rest, C., Elhalawani, H., Jreige, M., Prior, J. O., Vallières, M., Visvikis, D., Hatt, M., & Depeursinge, A. (2022). Overview of the hecktor challenge at miccai 2021: automatic head and neck tumor segmentation and outcome prediction in pet/ct images. *Head and Neck Tumor Segmentation and Outcome Prediction*, 1–37. https://doi.org/10.1007/978-3-030-98253-9_1

Andrearczyk, V., Oreiller, V., & Depeursinge, A. (2020). Oropharynx detection in PET-CT for tumor segmentation. *Irish Machine Vision and Image Processing.*

Andrearczyk, V., Oreiller, V., & Depeursinge, A. (Eds.). (2021). *Head and neck tumor segmentation* (Vol. 12603). Springer International Publishing. https://link.springer.com/book/10.1007%2F978-3-030-67194-5

Andrearczyk, V., Oreiller, V., Fageot, J., Montet, X., & Depeursinge, A. (2019a). Solid spherical energy (SSE) CNNs for efficient 3D medical image analysis. *Irish Machine Vision and Image Processing Conference.*

Andrearczyk, V., Oreiller, V., Fageot, J., Montet, X., & Depeursinge, A. (2019b). Solid Spherical Energy (SSE) CNNs for Efficient 3D Medical Image Analysis. *Irish Machine Vision and Image Processing Conference*, 37–44.

Andrearczyk, V., Oreiller, V., Jreige, M., Vallières, M., Castelli, J., Elhalawani, H., Boughdad, S., Prior, J. O., & Depeursinge, A. (2020). Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, 1–21.

Andrearczyk, V., Oreiller, V., Vallières, M., Castelli, J., Elhalawani, H., Jreige, M., Boughdad, S., Prior, J. O., & Depeursinge, A. (2020). Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. *International Conference on Medical Imaging with Deep Learning (MIDL).*

Andrearczyk, V., Oreiller, V., Vallières, M., Jreige, M., Prior, J. O., & Depeursinge, A. (2021). Overview of the HECKTOR challenge at MICCAI 2020: Automatic head and neck tumor segmentation in PET/CT. *Lecture Notes in Computer Science (LNCS) Challenges.*

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. (2022). The medical segmentation decathlon. *Nature Communications*, *13*(1), 1–13.

Aristophanous, M., Penney, B. C., Martel, M. K., & Pelizzari, C. A. (2007). A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography. *Medical physics*, *34*(11), 4223–4235.

Artin, M. (2018). *Algebra.* Prentice Hall.

Ashrafinia, S. (2019). *Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics* (Doctoral dissertation). The Johns Hopkins University.

Atul Mali, S., Ibrahim, A., Woodruff, H. C., Andrearczyk, V., Müller, H., Primakov, S., Salahuddin, Z., Chatterjee, A., & Lambin, P. (2021). Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *Journal of Personalized Medicine*, *11*(9).

Banerjee, J., Moelker, A., Niessen, W. J., & Walsum, T. v. (2012). 3d lbp-based rotationally invariant region description. *Asian Conference on Computer Vision*, 26–37.

Bartelt, H., Lohmann, A. W., & Wirnitzer, B. (1984). Phase and amplitude recovery from bispectra. *Applied optics*, *23*(18), 3121–3129.

Bekkers, E. J., Lafarge, M. W., Veta, M., Eppenhof, K. A., Pluim, J. P., & Duits, R. (2018). Roto-translation covariant convolutional networks for medical image analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 440–448.

Bekkers, E. J. (2019). B-Spline CNNs on Lie groups. *arXiv preprint arXiv:1909.12057*.

Blanc-Durand, P., Van Der Gucht, A., Schaefer, N., Itti, E., & Prior, J. O. (2018). Automatic lesion detection and segmentation of 18F-FET PET in gliomas: A full 3D U-Net convolutional neural network study. *PloS One*, *13*(4), e0195798.

Bogowicz, M., Riesterer, O., Stark, L. S., Studer, G., Unkelbach, J., Guckenberger, M., & Tanadini-Lang, S. (2017). Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta oncologica*, *56*(11), 1531–1536.

Bonner, J. A., Harari, P. M., Giralt, J., Cohen, R. B., Jones, C. U., Sur, R. K., Raben, D., Baselga, J., Spencer, S. A., Zhu, J., et al. (2010). Radiotherapy plus Cetuximab for locoregionally advanced Head and Neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between Cetuximab-induced rash and survival. *The Lancet Oncology*, *11*(1), 21–28.

Bourigault, E., McGowan, D. R., Mehranian, A., & Papiez, B. W. (2022). Multimodal PET/CT tumour segmentation and prediction of progression-free survival using a full-scale UNet with attention. *Lecture Notes in Computer Science (LNCS) Challenges*.

Braakhuis, B., Brakenhoff, R., & Leemans, C. R. (2012). Treatment choice for locally advanced head and neck cancers on the basis of risk factors: biological risk factors. *Annals of Oncology*, *23*, x173–x177.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: fast and flexible image augmentations. *Information*, *11*(2). https://doi.org/10.3390/info11020125

Castelli, J., Depeursinge, A., De Bari, B., Devillers, A., De Crevoisier, R., Bourhis, J., & Prior, J. O. (2017). Metabolic tumor volume and total lesion glycolysis in oropharyngeal cancer treated with definitive radiotherapy: which threshold is the best predictor of local control? *Clinical nuclear medicine*, *42*(6), e281–e285.

Castelli, J., Depeursinge, A., Devillers, A., Campillo-Gimenez, B., Dicente, Y., Prior, J., Chajon, E., Jegoux, F., Sire, C., Acosta, O., et al. (2019). PET-based prognostic survival model after radiotherapy for head and neck cancer. *European journal of nuclear medicine and molecular imaging*, *46*(3), 638–649.

Chaichian, M., & Hagedorn, R. (1998). *Symmetries in quantum mechanics: from angular momentum to supersymmetry* (1st ed.). CRC Press.

Chajon, E., Lafond, C., Louvel, G., Castelli, J., Williaume, D., Henry, O., Jégoux, F., Vauléon, E., Manens, J.-P., Le Prisé, E., et al. (2013). Salivary gland-sparing other than parotid-sparing in definitive Head-and-Neck intensity-modulated radiotherapy does not seem to jeopardize local control. *Radiation Oncology*, *8*(1), 132.

Chen, H., Chen, H., & Wang, L. (2021). Iteratively refine the segmentation of head and neck tumor in FDG-PET and CT images. *Lecture Notes in Computer Science (LNCS) Challenges*.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *CoRR*, *abs/1606.00915*.

Chen, M.-K., Chen, T. H.-H., Liu, J.-P., Chang, C.-C., & Chie, W.-C. (2004). Better prediction of prognosis for patients with nasopharyngeal carcinoma using primary tumor volume. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *100*(10), 2160–2166.

Chenouard, N., & Unser, M. (2012a). 3D steerable wavelets in practice. *IEEE Transactions on Image Processing*, *21*(11), 4522–4533.

Chenouard, N., & Unser, M. (2012b). 3D Steerable Wavelets in Practice. *IEEE Transactions on Image Processing*, *21*(11), 4522–4533.

Cho, M., Choi, Y., Hwang, D., Yie, S. Y., Kim, H., & Lee, J. S. (2022). Multimodal spatial attention network for automatic head and neck tumor segmentation in FDG-PET and CT images. *Lecture Notes in Computer Science (LNCS) Challenges*.

Choe, J., Lee, S., Do, K.-H., Lee, G., Lee, J.-G., Lee, S., & Seo, J. (2019). Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodule. *Radiology*, *292*(2), 365–373.

Chow, L. Q. (2020). Head and neck cancer. *New England Journal of Medicine*, *382*(1), 60–72.

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, *26*(6), 1045–1057.

Cohen, T., Weiler, M., Kicanaoglu, B., & Welling, M. (2019). Gauge equivariant convolutional networks and the icosahedral cnn. *International conference on Machine learning*, 1321–1330.

Cohen, T., & Welling, M. (2016a). Group equivariant convolutional networks. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (pp. 2990–2999). PMLR.

Cohen, T., & Welling, M. (2016b). Group equivariant convolutional networks. *International conference on machine learning*, 2990–2999.

Cohen, T. S., Geiger, M., & Weiler, M. (2019). A general theory of equivariant CNNs on homogeneous spaces. *Advances in Neural Information Processing Systems*, 9142–9153.

Cohen, T., Geiger, M., Köhler, J., & Welling, M. (2018). Spherical CNNs. *arXiv preprint arXiv:1801.10130.*

Cohen, T., & Welling, M. (2016c). Steerable CNNs. *arXiv:1612.08498.*

Coxeter, H. S. M. (1961). *Introduction to geometry.* New York, London.

Creff, G., Devillers, A., Depeursinge, A., Palard-Novello, X., Acosta, O., Jegoux, F., & Castelli, J. (2020). Evaluation of the prognostic value of FDG PET/CT parameters for patients with surgically treated head and neck cancer: A systematic review. *JAMA Otolaryngology–Head & Neck Surgery*, *146*(5), 471–479.

Da-ano, R., Masson, I., Lucia, F., Dorée, M., Robin, P., Alfieri, J., Rousseau, C., Mervoyer, A., Reinhold, C., Castelli, J., De Crevoisier, R., Ramée, J., Pradier, O., Schick, U., Visvikis, D., & Hatt, M. (2020). Performance comparison of modified ComBat for harmonization of radiomic features for multicentric studies. *Scientific Reports*, *10*(1), 102488.

Davidson-Pilon, C. (2019). Lifelines: survival analysis in Python. *Journal of Open Source Software*, *4*(40), 1317.

De Biase, A., Tang, W., Sourlos, N., Ma, B., Guo, J., Sijtsema, N. M., & van Ooijen, P. (2022). Skip-SCSE multi-scale attention and co-learning method for oropharyngeal tumor segmentation on multi-modal PET-CT images. *Lecture Notes in Computer Science (LNCS) Challenges.*

del Toro, O. A. J., Goksel, O., Menze, B., Müller, H., Langs, G., Weber, M.-A., Eggel, I., Gruenberg, K., Holzer, M., et al. (n.d.). VISCERAL–VISual Concept Extraction challenge in RAdioLogy: ISBI 2014 challenge organization.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Depeursinge, A., Fageot, J., Andrearczyk, V., Ward, J. P., & Unser, M. (2018). Rotation invariance and directional sensitivity: Spherical harmonics versus radiomics features. *International Workshop on Machine Learning in Medical Imaging*, 107–115.

Depeursinge, A., Fageot, J., & Al-Kadi, O. S. (2017a). Fundamentals of texture processing for biomedical image analysis: a general definition and problem formulation. *Biomedical texture analysis: fundamentals, applications and tools* (pp. 1–27). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-812133-7.00001-6

Depeursinge, A., Fageot, J., & Al-Kadi, O. S. (2017b). Fundamentals of texture processing for biomedical image analysis: a general definition and problem formulation. *Biomedical texture analysis* (pp. 1–27). Elsevier.

Depeursinge, A., Pad, P., Chin, A. S., Leung, A. N., Rubin, D. L., Müller, H., & Unser, M. (2015). Optimized steerable wavelets for texture analysis of lung tissue in 3-d ct: classification of usual interstitial pneumonia. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 403–406.

Depeursinge, A., Püspöki, Z., Ward, J. P., & Unser, M. (2017a). Steerable Wavelet Machines (SWM): Learning Moving Frames for Texture Classification. *IEEE Transactions on Image Processing*, *26*(4), 1626–1636.

Depeursinge, A., Püspöki, Z., Ward, J. P., & Unser, M. (2017b). Steerable wavelet machines (SWM): learning moving frames for texture classification. *IEEE Transactions on Image Processing*, *26*(4), 1626–1636.

Dicente Cid, Y., Müller, H., Platon, A., Poletti, P., & Depeursinge, A. (2017a). 3D solid texture classification using locally-oriented wavelet transforms. *IEEE Transactions on Image Processing*, *26*, 1899–1910.

Dicente Cid, Y., Müller, H., Platon, A., Poletti, P.-A., & Depeursinge, A. (2017b). 3-d solid texture classification using locally-oriented wavelet transforms. *IEEE Transactions on Image Processing*, *26*(4), 1899–1910. https://doi.org/10.1109/TIP.2017.2665041

Driscoll, J. R., & Healy, D. M. (1994). Computing fourier transforms and convolutions on the 2-sphere. *Advances in applied mathematics*, *15*(2), 202–250.

Dumont, B., Maggio, S., & Montalvo, P. (2018). Robustness of rotation-equivariant networks to adversarial perturbations. *arXiv preprint arXiv:1802.06627*.

Eickenberg, M., Exarchakis, G., Hirn, M., & Mallat, S. (2017). Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3D electronic densities. *Advances in Neural Information Processing Systems*, 6540–6549.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, *45*(2), 228–247.

Erdi, Y. E., Mawlawi, O., Larson, S. M., Imbriaco, M., Yeung, H., Finn, R., & Humm, J. L. (1997). Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *80*(S12), 2505–2509.

Fageot, J., Uhlmann, V., Püspöki, Z., Beck, B., Unser, M., & Depeursinge, A. (2018). Principled design and implementation of steerable detectors. *arXiv preprint arXiv:1811.00863*.

Fageot, J., Uhlmann, V., Püspöki, Z., Beck, B., Unser, M., & Depeursinge, A. (2021). Principled design and implementation of steerable detectors. *IEEE Transactions on Image Processing*, *30*, 4465–4478.

Falcão, A. X., Stolfi, J., & de Alencar Lotufo, R. (2004). The image foresting transform: theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 19–29.

Fatan, M., Hosseinzadeh, M., Askari, D., Sheykhi, H., Rezaeijo, S. M., & Salmanpoor, M. R. (2022). Fusion-based head and neck tumor segmentation and survival prediction using robust deep learning techniques and advanced hybrid machine learning systems. *Lecture Notes in Computer Science (LNCS) Challenges.*

Flusser, J., Zitova, B., & Suk, T. (2009). *Moments and moment invariants in pattern recognition.* John Wiley & Sons.

Fontaine, P., Andrearczyk, V., Oreiller, V., Abler, D., Castelli, J., Acosta, O., De Crevoisier, R., Vallières, M., Jreige, M., Prior, J. O., & Depeursinge, A. (2022a). Cleaning radiotherapy contours for radiomics studies, is it worth it? a head and neck cancer study. *Clinical and Translational Radiation Oncology.*

Fontaine, P., Andrearczyk, V., Oreiller, V., Abler, D., Castelli, J., Acosta, O., De Crevoisier, R., Vallières, M., Jreige, M., Prior, J. O., & Depeursinge, A. (2022b). Cleaning radiotherapy contours for radiomics studies, is it worth it? a head and neck cancer study. *Clinical and Translational Radiation Oncology*, *33*, 153–158.

Fontaine, P., Andrearczyk, V., Oreiller, V., Castelli, J., Jreige, M., O. Prior, J., & Depeursinge, A. (2021). Fully automatic head and neck cancer prognosis prediction in PET/CT. *Multimodal Learning and Fusion Across Scales for Clinical Decision Support (ML-CDS at MICCAI).*

Foster, B., Bagci, U., Mansoor, A., Xu, Z., & Mollura, D. J. (2014). A review on segmentation of positron emission tomography images. *Computers in biology and medicine*, *50*, 76–96.

Freeman, W., & Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9), 891–906.

Fu, X., Bi, L., Kumar, A., Fulham, M., & Kim, J. (2021). Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics.*

Gallier, J. (2009). Notes on spherical harmonics and linear representations of Lie groups [Accessed: 2020-04-13].

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer graphics and image processing*, *4*(2), 172–179.

Ghimire, K., Chen, Q., & Feng, X. (2021). Patch-based 3D UNet for head and neck tumor segmentation with an ensemble of conventional and dilated convolutions. *Lecture Notes in Computer Science (LNCS) Challenges.*

Ghimire, K., Chen, Q., & Feng, X. (2022). Head and neck tumor segmentation with deeply-supervised 3D UNet and progression-free survival prediction with linear model. *Lecture Notes in Computer Science (LNCS) Challenges.*

Giannakis, G. B. (1989). Signal reconstruction from multiple correlations: frequency-and time-domain approaches. *JOSA A*, *6*(5), 682–697.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al. (2018). NiftyNet: A deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, *158*, 113–122.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016a). Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, *278*(2), 563–577.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016b). Radiomics: images are more than pictures, they are data. *Radiology*, *278*(2), 563.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.

Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159.

Groendahl, A. R., Knudtsen, I. S., Huynh, B. N., Mulstad, M., Moe, Y. M., Knuth, F., Tomic, O., Indahl, U. G., Torheim, T., Dale, E., et al. (2021). A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Physics in Medicine & Biology*, *66*(6), 065012.

Gudi, S., Ghosh-Laskar, S., Agarwal, J. P., Chaudhari, S., Rangarajan, V., Paul, S. N., Upreti, R., Murthy, V., Budrukkar, A., & Gupta, T. (2017). Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site. *Journal of medical imaging and radiation sciences*, *48*(2), 184–192.

Guiot, J., Vaidyanathan, A., Deprez, L., Zerka, F., Danthine, D., Frix, A.-N., Lambin, P., Bottari, F., Tsoutzidis, N., Miraglio, B., et al. (2022). A review in radiomics: making personalized medicine a reality via routine imaging. *Medicinal Research Reviews*, *42*(1), 426–440.

Guo, Z., Guo, N., Gong, K., Li, Q., et al. (2019). Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Physics in Medicine & Biology*, *64*(20), 205015.

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610–621.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York.

Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S., Li, L., Lu, W., Jaouen, V., Tauber, C., Czakon, J., et al. (2018a). The first miccai challenge on pet tumor segmentation. *Medical image analysis*, *44*, 177–195.

Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S., Li, L., Lu, W., Jaouen, V., Tauber, C., Czakon, J., Drapejkowski, F., Dyrka, W., Camarasu-Pop, S., Cervenansky, F., Girard, P., Glatard, T., Kain, M., Yao, Y., Barillot, C., . . . Visvikis, D. (2018b). The first MICCAI challenge on PET tumor segmentation. *Medical Image Analysis*, *44*, 177–195.

Hatt, M., Le Rest, C. C., Turzo, A., Roux, C., & Visvikis, D. (2009). A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE transactions on medical imaging*, *28*(6), 881–893.

Hatt, M., Lee, J. A., Schmidtlein, C. R., Naqa, I. E., Caldwell, C., De Bernardi, E., Lu, W., Das, S., Geets, X., Gregoire, V., et al. (2017). Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Medical physics*, *44*(6), e1–e42.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al. (2021). 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.

Holschneider, M., Kronland-Martinet, R., Morlet, J., & Tchamitchian, P. (1990). A real-time algorithm for signal analysis with the help of the wavelet transform. *Wavelets* (pp. 286–297). Springer.

Huang, B., Chen, Z., Wu, P.-M., Ye, Y., Feng, S.-T., Wong, C.-Y. O., Zheng, L., Liu, Y., Wang, T., Li, Q., et al. (2018). Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: A dual-center study. *Contrast media & molecular imaging*, *2018*.

Huynh, B.-N., Ren, J., Groendahl, A. R., Tomic, O., Korreman, S. S., & Futsaether, C. M. (2022). Comparing deep learning and conventional machine learning for outcome prediction of head and neck cancer in PET/CT. *Lecture Notes in Computer Science (LNCS) Challenges.*

Iantsen, A., Ferreira, M., Lucia, F., Jaouen, V., Reinhold, C., Bonaffini, P., Alfieri, J., Rovira, R., Masson, I., Robin, P., Mervoyer, A., Rousseau, C., Kridelka, F., Decuypere, M., Lovinfosse, P., Pradier, O., Hustinx, R., Schick, U., Visvikis, D., & Hatt, M. (2021). Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting. *European Journal of Nuclear Medicine and Molecular Imaging*, 1–13.

Iantsen, A., Visvikis, D., & Hatt, M. (2021). Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. *Lecture Notes in Computer Science (LNCS) Challenges.*

Iliadis, G., Selviaridis, P., Kalogera-Fountzila, A., Fragkoulidi, A., Baltas, D., Tselis, N., Chatzisotiriou, A., Misailidou, D., Zamboglou, N., & Fountzilas, G. (2009). The importance of tumor volume in the prognosis of patients with glioblastoma. *Strahlentherapie und Onkologie*, *185*(11), 743–750.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203–211.

Jemaa, S., Fredrickson, J., Carano, R. A., Nielsen, T., de Crespigny, A., & Bengtsson, T. (2020). Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. *Journal of digital imaging*, *33*, 888–894.

Jreige, M., Oreiller, V., Letovanec, I., Schaefer, N., Depeursinge, A., & Prior, J. O. (2020). PET/CT Radiomics predict Pulmonary Lymphangitic Carcinomatosis (PLC) in Non-Small Cell Lung Cancer (NSCLC). *Journal of Nuclear Medicine*, *61*(supplement 1), 1311–1311. http://jnm.snmjournals.org/cgi/content/short/61/supplement%7B%5C_%7D1/1311

Juanco-Müller, Á. V., Mota, J. F. C., Goatman, K., & Hoogendoorn, C. (2022). Deep supervoxel segmentation for survival analysis in head and neck cancer patients. *Lecture Notes in Computer Science (LNCS) Challenges*.

Junttila, M. R., & De Sauvage, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, *501*(7467), 346–354.

Kakarala, R. (2012). The bispectrum as a source of phase-sensitive invariants for fourier descriptors: a group-theoretic approach. *Journal of Mathematical Imaging and Vision*, *44*(3), 341–353.

Kakarala, R., Kaliamoorthi, P., & Li, W. (2011). Viewpoint invariants from three-dimensional data: the role of reflection in human activity understanding. *CVPR 2011 WORKSHOPS*, 57–62.

Kakarala, R., & Mao, D. (2010). A theory of phase-sensitive rotation invariance with spherical harmonic and moment-based representations. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 105–112.

Ke, Q., & Li, Y. (2014). Is rotation a nuisance in shape recognition? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4146–4153.

Kim, B., & Ye, J. C. (2019). Mumford-Shah loss functional for image segmentation with deep learning. *IEEE Transactions on Image Processing*, *29*, 1856–1866.

Koboldt, D. C., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., Mardis, E., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Kondor, R., Lin, Z., & Trivedi, S. (2018). Clebsch-Gordan nets: a fully Fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 10117–10126.

Kondor, R., & Trivedi, S. (2018). On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*.

Krizhevsky, A., Hinton, G. et al. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A., et al. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE transactions on medical imaging*, *38*(11), 2556–2568.

Kumar, A., Fulham, M., Feng, D., & Kim, J. (2019). Co-Learning Feature Fusion Maps from PET-CT Images of Lung Cancer. *IEEE Transactions on Medical Imaging.*

Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O. F., Tsougenis, E., Chen, H., Heng, P.-A., Li, J., Hu, Z., et al. (2019). A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, *39*(5), 1380–1391.

Lafarge, M. W., Bekkers, E. J., Pluim, J. P., Duits, R., & Veta, M. (2021). Roto-translation equivariant convolutional networks: application to histopathology image analysis. *Medical Image Analysis*, *68*, 101849.

Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., & Veta, M. (2019). Learning domain-invariant representations of histological images. *Frontiers in Medicine*, *6*, 162.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., et al. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, *14*(12), 749–762.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., de Jong, E. E., van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., van Wijk, Y., Woodruff, H., van Soest, J., Lustberg, T., Roelofs, E., van Elmpt, W., Dekker, A., Mottaghy, F. M., Wildberger, J. E., & Walsh, S. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, *14*(12), 749–762.

Lang, D. M., Peeken, J. C., Combs, S. E., Wilkens, J. J., & Bartzsch, S. (2022). Deep learning based GTV delineation and progression free survival risk score prediction for head and neck cancer patients. *Lecture Notes in Computer Science (LNCS) Challenges.*

Lapuyade-Lahorgue, J., Visvikis, D., Pradier, O., Cheze Le Rest, C., & Hatt, M. (2015). SPEQTACLE: An automated generalized fuzzy C-means algorithm for tumor delineation in PET. *Medical physics*, *42*(10), 5720–5734.

Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K., & O'Leary, A. (2019). Pywavelets: a python package for wavelet analysis. *Journal of Open Source Software*, *4*(36), 1237.

Lee, J., Kang, J., Shin, E. Y., Kim, R. E. Y., & Lee, M. (2022). Dual-path connected CNN for tumor segmentation of combined PET-CT images and application to survival risk prediction. *Lecture Notes in Computer Science (LNCS) Challenges.*

Leijenaar, R. T., Nalbantov, G., Carvalho, S., Van Elmpt, W. J., Troost, E. G., Boellaard, R., Aerts, H. J., Gillies, R. J., & Lambin, P. (2015). The effect of suv discretization in quantitative fdg-pet radiomics: the need for standardized methodology in tumor texture analysis. *Scientific reports*, *5*(1), 1–10.

Leseur, J., Roman-Jimenez, G., Devillers, A., Ospina-Arango, J., Williaume, D., Castelli, J., & et al. (2016). Pre- and per-treatment 18F-FDG PET/CT parameters to predict recurrence and survival in cervical cancer. *Radiother Oncol J Eur Soc Ther Radiol Oncol*, *120*(3), 512–8.

Li, L., Zhao, X., Lu, W., & Tan, S. (2019). Deep learning for variational multimodality tumor segmentation in PET/CT. *Neurocomputing.*

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: common objects in context. *European conference on computer vision*, 740–755.

Liu, T., Su, Y., Zhang, J., Wei, T., & Xiao, Z. (2022). 3D U-net applied to simple attention module for head and neck tumor segmentation in PET and CT images. *Lecture Notes in Computer Science (LNCS) Challenges.*

Liu, Z., Liu, F., Chen, W., Liu, X., Hou, X., Shen, J., Guan, H., Zhen, H., Wang, S., Chen, Q., et al. (2020). Automatic segmentation of clinical target volume used for post-modified radical mastectomy radiotherapy with a convolutional neural network. *Frontiers in oncology*, *10*, 3268.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, J., Lei, W., Gu, R., & Wang, G. (2022). Priori and posteriori attention for generalizing head and neck tumors segmentation. *Lecture Notes in Computer Science (LNCS) Challenges.*

Ma, B., Guo, J., De Biase, A., Sourlos, N., Tang, W., van Ooijen, P., Both, S., & Sijtsema, N. M. (2022). Self-supervised multi-modality image feature extraction for the progression free survival prediction in head and neck cancer. *Lecture Notes in Computer Science (LNCS) Challenges.*

Ma, J. (2021). Cutting-edge 3D medical image segmentation methods in 2020: Are happy families all alike? *arXiv preprint arXiv:2101.00232.*

Ma, J., & Yang, X. (2021). Combining CNN and hybrid active contours for head and neck tumor segmentation. *Lecture Notes in Computer Science (LNCS) Challenges.*

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., & Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110.

Mackin, D., Fave, X., Zhang, L., Fried, D., Yang, J., Taylor, B., Rodriguez-Rivera, E., Dodge, C., Jones, A. K., & Court, L. (2015). Measuring ct scanner variability of radiomics features. *Investigative radiology*, *50*(11), 757.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, *9*(1), 1–13.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al. (2020). BIAS: Transparent reporting of biomedical image analysis challenges. *Medical Image Analysis*, 101796.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Martinez-Larraz, A., Asenjo, J. M., & Rodríguez, B. Á. (2022). PET/CT head and neck tumor segmentation and progression free survival prediction using deep and machine learning techniques. *Lecture Notes in Computer Science (LNCS) Challenges*.

Meng, M., Peng, Y., Bi, L., & Kim, J. (2022). Multi-task deep learning for joint tumor segmentation and outcome prediction in head and neck cancer. *Lecture Notes in Computer Science (LNCS) Challenges*.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, *34*(10), 1993–2024.

Miller, T. R., & Grigsby, P. W. (2002). Measurement of tumor volume by pet to evaluate prognosis in patients with advanced cervical cancer treated by radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, *53*(2), 353–359.

Moe, Y. M., Groendahl, A. R., Mulstad, M., Tomic, O., Indahl, U., Dale, E., Malinen, E., & Futsaether, C. M. (2019). Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. *Medical Imaging with Deep Learning*.

Murugesan, G. K., Brunner, E., McCrumb, D., Kumar, J., VanOss, J., Moore, S., Peck, A., & Chang, A. (2022). Head and neck primary tumor segmentation using deep neural networks and adaptive ensembling. *Lecture Notes in Computer Science (LNCS) Challenges*.

Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. *International MICCAI Brainlesion Workshop*, 311–320.

Naser, M., van Dijk, L., He, R., Wahid, K., & Fuller, C. (2021). Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. *Lecture Notes in Computer Science (LNCS) Challenges*.

Naser, M. A., Wahid, K. A., Mohamed, A. S. R., Abdelaal Abobakr, M., He, R., Dede, C., van Dijk, L. V., & Fuller, C. D. (2022). Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET-CT imaging data. *Lecture Notes in Computer Science (LNCS) Challenges*.

Naser, M. A., Wahid, K. A., van Dijk, L. V., He, R., Abobakr Abdelaal, M., Dede, C., Mohamed, A. S. R., & Fuller, C. D. (2022). Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in PET-CT images. *Lecture Notes in Computer Science (LNCS) Challenges*.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al. (2021). 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.

Nioche, C., Orlhac, F., Boughdad, S., Reuzé, S., Goya-Outi, J., Robert, C., Pellot-Barakat, C., Soussan, M., Frouin, F., & Buvat, I. (2018). Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer research*, *78*(16), 4786–4789.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray–scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallières, M., Zhu, S., Xie, J., Peng, Y., et al. (2021). Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. *Medical Image Analysis*, 102336.

Paeenafrakati, M. S., Hajianfar, G., Rezaeijo, S. M., Ghaemi, M., & Rahmim, A. (2022). Advanced automatic segmentation of tumors and survival prediction in head and neck cancer. *Lecture Notes in Computer Science (LNCS) Challenges.*

Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2005). Global cancer statistics, 2002. *CA: A Cancer Journal for Clinicians*, *55*(2), 74–108.

Perona, P. (1992). Steerable-scalable kernels for edge detection and junction analysis. *European Conference on Computer Vision*, 3–18.

Pietikäinen, M., & Zhao, G. (2015). Two decades of local binary patterns: a survey. *Advances in independent component analysis and learning machines* (pp. 175–210). Elsevier.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49–70.

Puspoki, Z., Fageot, J., Amini, A., Ward, J. P., & Unser, M. (2019). Angular accuracy of steerable feature detectors. *SIAM Journal on Imaging Sciences*, *12*(1), 344–371.

Qayyum, A., Benzinou, A., Mazher, M., Abdel-Nasser, M., & Puig, D. (2022). Automatic segmentation of head and neck (H&N) primary tumors in PET and CT images using 3D-Inception-ResNet model. *Lecture Notes in Computer Science (LNCS) Challenges.*

Rabbani, M. (2002). JPEG2000: Image Compression Fundamentals, Standards and Practice. *Journal of Electronic Imaging*, *11*(2). https://doi.org/10.1117/1.1469618

Rao, C., Pai, S., Hadzic, I., Zhovannik, I., Bontempi, D., Dekker, A., Teuwen, J., & Traverso, A. (2021). Oropharyngeal tumour segmentation using ensemble 3D PET-CT fusion networks for the HECKTOR challenge. *Lecture Notes in Computer Science (LNCS) Challenges.*

Reiazi, R., Abbas, E., Famiyeh, P., Rezaie, A., Kwan, J. Y., Patel, T., Bratman, S. V., Tadic, T., Liu, F.-F., & Haibe-Kains, B. (2021). The impact of the variation of imaging parameters on the robustness of computed tomography radiomic features: a review. *Computers in Biology and Medicine*, *133*, 104400.

Ren, J., Eriksen, J. G., Nijkamp, J., & Korreman, S. S. (2021). Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. *Acta Oncologica*, 1–8.

Ren, J., Huynh, B.-N., Groendahl, A. R., Tomic, O., Futsaether, C. M., & Korreman, S. S. (2022). PET normalizations to improve deep learning auto-segmentation of head and neck in 3D PET/CT. *Lecture Notes in Computer Science (LNCS) Challenges.*

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Saeed, N., Al Majzoub, R., Sobirov, I., & Yaqub, M. (2022). An ensemble approach for patient prognosis of head and neck tumor using multimodal data. *Lecture Notes in Computer Science (LNCS) Challenges.*

Segmentation of neuronal structures in EM stacks challenge - ISBI 2012 [Accessed: 2021-07-29]. (2012).

Sepehri, S., Tankyevych, O., Iantsen, A., Visvikis, D., Cheze Le Rest, C., & Hatt, M. (2021). Accurate tumor delineation vs. rough volume of interest analysis for 18F-FDG PET/CT radiomic-based prognostic modeling in non-small cell lung cancer. *Frontiers in oncology*, *292*(2), 365–373.

Shorten, C., & Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

Simard, P., Steinkraus, D., & Platt, J. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 958–963. https://doi.org/10.1109/ICDAR.2003.1227801

Smith, S. W. et al. (1997). *The scientist and engineer's guide to digital signal processing.* California Technical Pub. San Diego.

Song, Q., Bai, J., Han, D., Bhatia, S., Sun, W., Rockey, W., Bayouth, J. E., Buatti, J. M., & Wu, X. (2013). Optimal co-segmentation of tumor in PET-CT images with context information. *IEEE Transactions on Medical Imaging*, *32*(9), 1685–1697.

Starke, S., Thalmeier, D., Steinbach, P., & Piraud, M. (2022). A hybrid radiomics approach to modeling progression-free survival in head and neck cancers. *Lecture Notes in Computer Science (LNCS) Challenges.*

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209–249.

Tang, C., Hobbs, B., Amer, A., Li, X., Behrens, C., Canales, J. R., Cuentas, E. P., Villalobos, P., Fried, D., Chang, J. Y., et al. (2018). Development of an immune-pathology informed radiomics model for non-small cell lung cancer. *Scientific reports*, *8*(1), 1–9.

Thibault, G., Fertil, B., Navarro, C., Pereira, S., Cau, P., Levy, N., Sequeira, J., & Mari, J. (2009). Texture indexes and gray level size zone matrix application to cell nuclei classification. *Pattern Recognition and Information Processing*, 140–145.

Traverso, A., Wee, L., Dekker, A., & Gillies, R. (2018). Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics*, *102*(4), 1143–1158.

Unser, M., & Chenouard, N. (2013a). A unifying parametric framework for 2D steerable wavelet transforms. *SIAM Journal on Imaging Sciences*, *6*(1), 102–135.

Unser, M., & Chenouard, N. (2013b). A unifying parametric framework for 2d steerable wavelet transforms. *SIAM Journal on Imaging Sciences*, *6*(1), 102–135.

Unser, M., Chenouard, N., & Van De Ville, D. (2011). Steerable Pyramids and Tight Wavelet Frames in $L_2(\mathbb{R}^d)$. *IEEE Transactions on Image Processing*, *20*(10), 2705–2721.

Vallieres, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J., Khaouam, N., Nguyen-Tan, P. F., Wang, C.-S., Sultanem, K., et al. (2017). Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific reports*, *7*(1), 1–14.

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., & Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, *77*(21), e104–e107.

Varma, M., & Zisserman, A. (2005a). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, *62*(1-2), 61–81. https://doi.org/10.1007/s11263-005-4635-4

Varma, M., & Zisserman, A. (2005b). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, *62*(1-2), 61–81.

Varshalovich, D., Moskalev, A., & Khersonskii, V. (1988). *Quantum theory of angular momentum*. World Scientific.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open science now: a systematic literature review for an integrated definition. *Journal of business research*, *88*, 428–436.

Vivaldi, F. (2006). The arithmetic of discretized rotations. *AIP Conference Proceedings*, *826*(1), 162–173.

Wahid, K. A., He, R., Dede, C., Mohamed, A. S. R., Abobakr Abdelaal, M., van Dijk, L. V., Fuller, C. D., & Naser, M. A. (2022). Combining tumor segmentation masks with PET/CT images and clinical data in a deep learning framework for improved prognostic prediction in head and neck squamous cell carcinoma. *Lecture Notes in Computer Science (LNCS) Challenges*.

Wahl, R. L., Jacene, H., Kasamon, Y., & Lodge, M. A. (2009). From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *Journal of nuclear medicine*, *50*(Suppl 1), 122S–150S.

Wang, G., Huang, Z., Shen, H., & Hu, Z. (2022). The head and neck tumor segmentation in PET/CT based on multi-channel attention network. *Lecture Notes in Computer Science (LNCS) Challenges.*

Wang, J., Peng, Y., Guo, Y., Li, D., & Sun, J. (2022). CCUT-Net: pixel-wise global context channel attention UT-Net for head and neck tumor segmentation. *Lecture Notes in Computer Science (LNCS) Challenges.*

Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, *23*(7), 903–921.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., & Cohen, T. S. (2018). 3d steerable cnns: learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 10381–10392.

Weiler, M., Hamprecht, F. A., & Storath, M. (2017). Learning steerable filters for rotation equivariant CNNs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 849–858.

Wiesenfarth, M., Reinke, A., Landman, B. A., Eisenmann, M., Saiz, L. A., Cardoso, M. J., Maier-Hein, L., & Kopp-Schneider, A. (2021). Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports 2021 11:1*, *11*(1), 1–15.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Winkels, M., & Cohen, T. S. (2019). Pulmonary nodule detection in CT scans with equivariant CNNs. *Medical image analysis*, *55*, 15–26.

Worrall, D., & Brostow, G. (2018). CubeNet: Equivariance to 3D rotation and translation. *ECCV, Lecture Notes in Computer Science*, *11209*, 585–602.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., & Brostow, G. J. (2016). Harmonic networks: deep translation and rotation equivariance. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7168–7177.

Wu, X., Bi, L., Fulham, M., & Kim, J. (2020). Unsupervised positron emission tomography tumor segmentation via GAN based adversarial auto-encoder. *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 448–453.

Xie, H., Zhang, X., Ma, S., Liu, Y., & Wang, X. (2019). Preoperative differentiation of uterine sarcoma from leiomyoma: comparison of three models based on different segmentation volumes using radiomics. *Mol Imaging Biol*, *21*(6), 1157–64.

Xie, J., & Peng, Y. (2021). The head and neck tumor segmentation using nnU-Net with spatial and channel 'squeeze & excitation' blocks. *Lecture Notes in Computer Science (LNCS) Challenges.*

Xie, J., & Peng, Y. (2022). The head and neck tumor segmentation based on 3D U-Net. *Lecture Notes in Computer Science (LNCS) Challenges.*

Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., Piraud, M., Buck, A., Shi, K., & Menze, B. H. (2018). Automated whole-body bone lesion detection for multiple myeloma on 68Ga-Pentixafor PET/CT imaging using deep learning methods. *Contrast Media & Molecular Imaging.*

Xue, Z., Li, P., Zhang, L., Lu, X., Zhu, G., Shen, P., Shah, S. A. A., & Bennamoun, M. (2021). Multi-Modal Co-Learning for Liver Lesion Segmentation on PET-CT Images. *IEEE Transactions on Medical Imaging.*

Yousefirizi, F., Janzen, I., Dubljevic, N., Liu, Y.-E., Hill, C., MacAulay, C., & Rahmim, A. (2022). Segmentation and risk score prediction of head and neck cancers in PET/CT volumes with 3D U-Net and cox proportional hazard neural networks. *Lecture Notes in Computer Science (LNCS) Challenges.*

Yousefirizi, F., & Rahmim, A. (2021). GAN-based bi-modal segmentation using Mumford-Shah loss: application to head and neck tumors in PET-CT images. *Lecture Notes in Computer Science (LNCS) Challenges.*

Yu, H., Caldwell, C., Mah, K., Poon, I., Balogh, J., MacKenzie, R., Khaouam, N., & Tirona, R. (2009). Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. *International Journal of Radiation Oncology* Biology* Physics*, *75*(2), 618–625.

Yuan, Y. (2021). Automatic head and neck tumor segmentation in PET/CT with scale attention network. *Lecture Notes in Computer Science (LNCS) Challenges.*

Yuan, Y., Adabi, S., & Wang, X. (2022). Automatic head and neck tumor segmentation and progression free survival analysis on PET/CT images. *Lecture Notes in Computer Science (LNCS) Challenges.*

Zhao, G., Ahonen, T., Matas, J., & Pietikainen, M. (2011). Rotation-invariant image and video description with local binary pattern features. *IEEE transactions on image processing*, *21*(4), 1465–1477.

Zhao, T., & Blu, T. (2020). The Fourier-Argand representation: An optimal basis of steerable patterns. *IEEE Transactions on Image Processing*, *29*, 6357–6371.

Zhao, X., Li, L., Lu, W., & Tan, S. (2018). Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Physics in Medicine & Biology*, *64*(1), 015011.

Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., & Wu, X. (2018). 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 228–231.

Zhou, T., Ruan, S., & Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 100004.

Zhu, S., Dai, Z., & Ning, W. (2021). Two-stage approach for segmenting gross tumor volume in head and neck cancer with CT and PET imaging. *Lecture Notes in Computer Science (LNCS) Challenges.*

Zucchelli, M., Deslauriers-Gauthier, S., & Deriche, R. (2020). A computational Framework for generating rotation invariant features and its application in diffusion MRI. *Medical Image Analysis*, *60*, 101597.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*(2), 328–338.