Standardisation and Optimisation of Radiomic Techniques for the Identification of Robust Imaging Biomarkers in Oncology

by

Philip Whybra

A Thesis submitted to Cardiff University
for the degree of Doctor of Philosophy

July 2021

Thesis Abstract

Radiomics is a rapidly evolving field within oncology. It explores the extraction of quantitative features from medical scans to aid in diagnosis, prognosis and monitoring of disease. In effect, these features may act as imaging biomarkers. Radiomics is a potential piece of a multifaceted data puzzle, powering precision medicine approaches, where treatment strategies could be tailored more to the individual than relying on a one-size-fits-all strategy. However, there are crucial challenges within the field regarding reproducibility and reliability of many common radiomic features. This thesis explored the hypothesis that varying but valid approaches to engineered feature extraction can cause discrepancy that harms identification and validation of potential radiomic biomarkers. As a research community, we require guidelines, standards and references to see forward progression and to avoid a replication crisis. Through development of radiomics software, results from this work significantly contributed to a large collaborative consensus benchmarking effort to address this standardisation need. This work investigated the effect of implementation choices on compliance to this new standard. Alongside this, the role of interpolation in radiomics became a key focus through the lens of robustness. Optimal feature extraction should avoid redundancy and utilise robust features. Finally, benchmarking methodology and tools were developed in an effort to standardise the application of filters in image processing steps prior to feature extraction. Discrepancies between radiomics software were identified and evaluated using these tools. The uncertainties in developing optimal and robust radiomic imaging biomarkers that result in clinically useful models are discussed.

Acknowledgments

Firstly, special thanks must go to my primary supervisor Dr Emiliano Spezi. Your advice, assistance and guidance has been instrumental in my ability to reach the finish line. I will forever be grateful for your support through both academic and personal challenges. I feel I have grown so much in the last few years, thank you.

Secondly, I would like to thank my secondary supervisor, Prof John Staffurth. Your advice and clinical perspective were illuminating, particularly when I was getting started. I admire all of the work you and your team do at Velindre.

Next, my deepest gratitude must go to the researchers that make up the IBSI core group. I am so proud to have been a part of this work, and each time we convene online I gain some new insight and knowledge. Alex and Martin especially, thank you for the opportunities provided to me to contribute.

To all of the past and present CIDA team, it was lovely to share an office with such a friendly and talented collection of people. Special thanks to Craig for the coffee and chinwags, I'll pay you back for those lattes I swear... Salvo and Concetta, thanks for taking me dancing and teaching me Italian swear words, I'm sorry I cannot role my r's and that I like pineapple on pizza. Kieran, thank you for all your input, and for helping when I lost my luggage in France. Emily, thank you for re-assuring me and being supportive in particular times of doubt. Iona and Emad, I really appreciated the cheery online team catchups during the dreaded lockdowns.

Special thanks must go to Brodies Coffee Co for producing the caffeinated brew that has fuelled much of the words that follow. Thanks also to those who maintain Alexandra gardens, where I so often went to think. Also, appreciation to the Pen & Wig and Flute & Tankard for being fine establishments to drown some research sorrows in on a Friday.

To my first year flat mate Josh, our late nights spent setting the world to rights and watching terrible movies are especially cherished memories (mostly, some of those movies were truly awful). Rosie, thank you for the jams! No better stress reliever can be had than trying to relearn *freedom* for the 1000th time. We should probably get around to recording that album. Also, I'd like to thank Sarah, John and Paul for being the catalysts for many Cardiff adventures outside of work.

To my family, it goes without saying that I wouldn't have got far at all without you all. Mum, Dad, Jo, Matt, Xanthe, Nikki, Vinay, Laura, Sally and Ben: you're the best family one could ever hope for and you mean everything to me. Special thanks to Lau and Mum for listening to me rant on the phone *a lot* when I didn't know what to write.

On that note, I'll take the opportunity to specifically NOT thank COVID19. Lockdown on top of trying to write up a thesis has been a harrowing experience. However, it is nothing compared to those who have worked on the front lines in this pandemic (such as Nikki, Sally and Ben), to which I am in awe.

Lastly, and most important of all, Alice. Whatever happens, meeting you made everything worth it. I can't put into words how lucky I am that I get to spend each day with you. You're the best human *bean*, and I love you so much.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/N509449/1 (Doctoral Training Partnership).

Publications and Output

Output from this project has contributed to work that has been published in peer reviewed journals.

Key publications

- **P. Whybra**, C. Parkinson, K. Foley, J. Staffurth, and E. Spezi, ‘Assessing radiomic feature robustness to interpolation in ^{18}F-FDG PET imaging’, Sci Rep, vol. 9, no. 1, Jul. 2019. https://doi.org/10.1038/s41598-019-46030-0
- A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. G. Pfaehler, A. Rahmim, A. U. K. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, E. Spezi, R. J. H. M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E. G. C. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. P. van Velden, **P. Whybra**, C. Richter, and S. Löck, ‘The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping’, Radiology, vol. 295, no. 2, pp. 328–338, May 2020. https://doi.org/10.1148/radiol.2020191145

Additional co-authored

- K. G. Foley, Z. Shi, **P. Whybra**, P. Kalendralis, R. Larue, M. Berbee, M. N. Sosef, C. Parkinson, J. Staffurth, T. D. L. Crosby, S. A. Roberts, A. Dekker, L. Wee, and E. Spezi, ‘External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer’, Radiotherapy and Oncology, vol. 133, pp. 205–212, Apr. 2019. https://doi.org/10.1016/j.radonc.2018.10.033
- C. Piazzese, K. Foley, **P. Whybra**, C. Hurt, T. Crosby, and E. Spezi, ‘Discovery of stable and prognostic CT-based radiomic features independent of contrast administration and dimensionality in oesophageal cancer’, PLoS ONE, vol. 14, no. 11, p. e0225550, Nov. 2019. https://doi.org/10.1371/journal.pone.0225550
- C. Parkinson, K. Foley, **P. Whybra**, R. Hills, A. Roberts, C. Marshall, J. Staffurth, and E. Spezi, ‘Evaluation of prognostic models developed using standardised image features from different PET automated segmentation methods’, EJNMMI Res, vol. 8, no. 1, Apr. 2018. https://doi.org/10.1186/s13550-018-0379-3

- M. Mori, P. Passoni, E. Incerti, V. Bettinardi, S. Broggi, M. Reni, **P. Whybra**, E. Spezi, E. G. Vanoli, L. Gianolli, M. Picchio, N. G. Di Muzio, and C. Fiorino, 'Training and validation of a robust PET radiomic-based index to predict distant-relapse-free-survival after radio-chemotherapy for locally advanced pancreatic cancer', Radiotherapy and Oncology, Jul. 2020. https://doi.org/10.1016/j.radonc.2020.07.003

Pre-print

- A. Depeursinge, V. Andrearczyk, **P. Whybra**, J. van Griethuysen, H. Müller, R Schaer, M. Vallières, A. Zwanenburg, 'Standardised convolutional filtering for radiomics.' *arXiv*: https://arxiv.org/abs/2006.05470, Jun,. 2020.

Posters

- **P. Whybra**, K. Foley, C. Parkinson, J. Staffurth, and E. Spezi, 'EP-2117: Effect of Interpolation on 3D Texture Analysis of PET Imaging in Oesophageal Cancer', Radiotherapy and Oncology, vol. 127, pp. S1164–S1165, Apr. 2018. https://doi.org/10.1016/s0167-8140(18)32426-5
- **P. Whybra**, C. Parkinson, K. Foley, J. Staffurth, and E. Spezi, 'PO-0963 A novel normalisation technique for voxel size dependent radiomic features in oesophageal cancer', Radiotherapy and Oncology, vol. 133, pp. S523–S524, Apr. 2019. https://doi.org/10.1016/S0167-8140(19)31383-0
- A. Zwanenburg, M. A. Abdalah, A. Apte, S. Ashrafinia, J. Beukinga, M. Bogowicz, C. V. Dinh, M. Götz, M. Hatt, R. T. H. Leijenaar, J. Lenkowicz, O. Morin, A. U. K. Rao, J. Socarras Fernandez, M. Vallières, L. V. Van Dijk, J. Van Griethuysen, F. H. P. Van Velden, **P. Whybra**, E. G. C. Troost, C. Richter, and S. Löck, 'PO-0981: Results from the Image Biomarker Standardisation Initiative', Radiotherapy and Oncology, vol. 127, pp. S543–S544, Apr. 2018. https://doi.org/10.1016/S0167-8140(18)31291-X
- Z. Shi, L. Wee, K. Foley, E. Spezi, **P. Whybra**, T. Crosby, J. Pablo de Mey, J. Van Soest, and A. Dekker, 'PV-0318: External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients', Radiotherapy and Oncology, vol. 127, p. S168, Apr. 2018. https://doi.org/10.1016/S0167-8140(18)30628-5
- C. Parkinson, K. Foley, **P. Whybra**, R. Hills, A. Roberts, C. Marshall, J. Staffurth, and E. Spezi, 'PO-0931: Dependency of patient risk stratification on PET target volume definition in Oesophageal cancer', Radiotherapy and Oncology, vol. 127, pp. S503–S504, Apr. 2018. https://doi.org/10.1016/S0167-8140(18)31241-6
- C. Piazzese, **P. Whybra**, R. Carrington, T. Crosby, J. Staffurth, K. Foley, and E. Spezi, 'EP-2141: Evaluation of 2D and 3D radiomics features extracted from CT images of oesophageal cancer patients', Radiotherapy and Oncology, vol. 127, pp. S1180–S1181, Apr. 2018. https://doi.org/10.1016/S0167-8140(18)32450-2
- C. Piazzese, **P. Whybra**, E. Qasem, D. Harris, R. Gtaes, K. Foley, and E. Spezi, 'EP-1926 Radiomics in rectal cancer: prognostic significance of 3D features extracted from diagnostic MRI', Radiotherapy and Oncology, vol. 133, p. S1048, Apr. 2019. https://doi.org/10.1016/S0167-8140(19)32346-1
- C. Piazzese, **P. Whybra**, R. Carrington, T. Crosby, J. Staffurth, K. Foley, and E. Spezi, 'PO-0964 Stability and prognostic significance of CT radiomic features from oesophageal cancer patients', Radiotherapy and Oncology, vol. 133, pp. S524–S525, Apr. 2019. https://doi.org/10.1016/S0167-8140(19)31384-2

Contents

List of Figures

(Note that these are abbreviated Figure captions.)

List of Tables

List of Abbreviations

ACD:	Annihilation Coincidence Detection
AGO:	Align Grid Origins
AGC:	Align Grid Centres
AJCC:	American Joint Committee on Cancer
CBCT:	Cone-Beam Computed Tomography
CAD:	Computed Aided Diagnosis
CERR:	Computation Environment for Radiotherapy Research
CIDA:	Cancer Imaging and Data Analytics
CNN:	Convolutional Neural Network
CT:	Computed Tomography
CRM:	Consensus Response Map
DICOM:	Digital Imaging and Communications in Medicine
DNA:	Deoxyribonucleic acid
FBN:	Fixed Bin Number
FBS:	Fixed Bin Size
FBW:	Fixed Bin Width
GLCM:	Grey Level Co-occurence Matrix
GLDZM:	Grey Level Distance Zone Matrix
GLRLM:	Grey Level Run Length Matrix
GLSZM:	Grey Level Size Zone Matrix
GTV:	Gross Tumour Volume
HU:	Hounsfield units
IBSI:	Image Biomarker Standardisation Initiative
IGRT:	Intensity Guided Radiotherapy
IH:	Intensity Histogram
IMRT:	Intensity Modulated Radiotherapy
IVH:	Intensity Volume Histogram
LoG:	Laplacian of Gaussian
LOR:	Line of Response
MR:	Magnetic Resonance

MRI:	Magnetic Resonance Imaging
MSE:	Mean Squared Error
MTV:	Metabolic Tumour Volume
NGLDM:	Neighbourhood Grey level Difference Matrix
NGTDM:	Neighbourhood Greytone Difference Matrix
NaN:	Not a Number
NHS:	National Health Service
NSCLC:	Non-Small Cell Lung Cancer
OC:	Oesophageal Cancer
OAR:	Organs At Risk
PC:	Principle Component
PCA:	Principle Component Analysis
PET:	Positron Emission Tomography
RF:	Radiofrequency
ROI:	Region Of Interest
RQS:	Radiomics Quality Score
RT:	Radiotherapy
SSIM	: Structural Similarity index
SPAARC:	Spaarc Pipeline for Automated Analysis and Radiomics Computing
SPECT:	Single Photon Emission Computed Tomography
SUV:	Standardised Uptake Value
TNM:	Tumour Node Metastasis
TRIPOD:	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
UICC:	Union for International Cancer Control
VOI:	Volume Of Interest

1

Introduction

> *"Onkos was the Greek term for a mass or a load, or more commonly a burden; cancer was imagined as a burden carried by the body."*
>
> — Siddhartha Mukherjee,
> *The Emperor of All Maladies*

1.1 Preview

Oncology is a branch of medicine that focuses on the diagnosis and treatment of cancer. It is a large field, as *Cancer* is an umbrella term for a range of different diseases that can originate from almost any tissue and cell type. In general however, the defining characteristic is unregulated cell growth and subsequent metastases throughout the body. Cancer is a leading cause of death worldwide with approximately 18.1 million new cases and 9.6 million mortalities reported in 2018 [1]. Consequently, there is an ever-present search for novel methods to detect, diagnose and manage cancers more effectively, an instance of which - *Radiomics* - is the subject area of this thesis.

RADIOMICS: The transformation of medical imaging into *quantitative* measures, referred to as *features*, to aid in disease diagnosis, prognosis and monitoring [2].

"High-throughput extraction of quantitative features that result in the conversion of images into mineable data and the subsequent analysis of these data for decision support." [3]

Over the past decade, the emerging radiomics literature has hinted at the potential benefits and current underutilisation of automated computational image analysis within a clinical setting [2–5]. This is particularly true in oncology, as medical imaging is already a vital component of treatment management [6]. Tumour imaging analysis is expected to be pivotal in the movement to a more individualistic, data-driven and personalised approach to treatment in oncology, especially as we embrace the age of *Big Data* in medicine and harness the power of *high-throughput* computing [7].

Yet, despite the potential, the current challenges and limitations of radiomics have so far impeded any real translation from research projects to clinical practice [8]. A key challenge identified has been the lack of reporting standards for feature extraction to ensure study replication and sufficient validation, another, is the robustness and stability of potential features to the processing steps of extraction [9, 10].

Radiomic studies generally focus on a particular primary disease site, such as lung [11], brain [12], head & neck [13], breast [14], or the oesophagus [15], though generalisability of models across multiple cancer types have also been explored [16]. Interest within our research group at Cardiff University have been focused on the clinical potential of radiomics in the treatment of oesophageal cancer [17]. However, further rigorous standardisation and validation of the feature extraction and model development methods are needed before any hope of clinical application is realised. Herein lies the areas of study for this body of work. This introductory chapter provides a general overview of oncology, imaging and the field of radiomics, before outlining the thesis aims and content of subsequent chapters.

1.2 An Overview of Cancer Care

1.2.1 Defining a Tumour

A *tumour* refers to any abnormal mass of tissue within the body and is predominantly a result of uncontrolled cell growth due to mutated or damaged DNA (*deoxyribonucleic acid*). This damage causes deviation from the usual cell cycle, suppressing normal programmed cell death in favour of uncontrolled and unregulated replication accumulating in a mass of tumour tissue. Naively, tumours can be classified into two main types, *malignant* or *benign*; The former are considered

Table 1.1: Examples of key cancer and tumour categories by cell origin, according to the National Cancer Institute [18].

Type	Description
Carcinoma	Originates from epithelial cells covering the inner and outer surfaces of the body.
Sarcoma	Develop in supportive or connective tissues, such as bone, muscle, fat, blood or lymph vessels, tendons and ligaments.
Lymphoma	Originates in white blood cells (immune system T or B cells) known as lymphocytes.
Multiple Myeloma	Immune system cancer originating from plasma cells
Leukemia	Starts in bone marrow tissue which forms blood cells. These cancers do not develop into solid masses.
Melanoma	A cancer that forms from melanocytes, the cells that produce melanin. Most often develops in skin cells.
Brain / Spinal Cord Tumours	Originate in central nervous system from brain and spinal tissues. (e.g. Brain stem glioma, forming for glial cells)
Germ cell Tumours	Develop in germ cells that produce sperm or eggs.
Neuroendocrine Tumours	Form from cells that produce hormones that are released into the blood stream.

cancerous as they develop by invading new areas of the body, the latter being non-cancerous as their current state does not present a risk of spreading [19]. Tumour metastasis is thought to be a leading cause of death from cancer [20]. Clearly, efficient methods for characterisation of tumour types is vital for patient care.

MALIGNANT: Tumours metastasise as they develop: cells break off from the primary site and travel through the bloodstream or the lymphatic system, settling in areas and promoting secondary growth. Considered cancerous.

BENIGN: Do **not** spread to surrounding tissue. However, benign tumours can still grow to a significant size and can interfere with normal bodily function. Not considered cancerous.

Numerous cancers can form solid tumours. There are hundreds of unique cancers which occur with varying degrees of rarity within the population [18]. Cancers are named after the tissue or cells in which they originate; oesophageal cancer begins with a primary tumour in the oesophagus, prostate cancer starts in the prostate, melanoma arises in skin cells called melanocytes, and so forth. Tumours are broadly categorised by the types of cells they stem from, examples of which are outlined in Table 1.1 [18].

Cancerous tumours are thought to develop through key *hallmark* deviations from usual cellular function, which include: evading growth suppressors, activation of invasion and metastasis, enabling replicative immortality, inducing angiogensis, resisting cell death, sustaining proliferative signalling, reprogramming of energy metabolism, and evading immune destruction [21].

The first step in cancer management is to achieve an accurate tumour diagnosis by identifying the current extent of the disease, referred to as *stage* [22]. Importantly for radiomics, staging utilises medical imaging extensively.

1.2.2 Diagnosis & Staging

Cancer diagnosis comprises a combination of clinical and pathological observations that review the disease progression within a patient. Typically, this involves both medical imaging and biopsy procedures to verify the cancer type. Cancer *staging* aims to group patients based on similar tumour characteristics ascertained from these procedures, together with their prognostic outlook (i.e. chance of survival). The most widely used clinical staging system is *Tumour Node Metastasis* (TNM), developed by the *American Joint Committee on Cancer* (AJCC) in collaboration with the *Union for International Cancer Control* (UICC) [22]. To summarise: the primary tumour (T) component evaluates tumour aspects such as size and contiguous growth into surrounding tissue, the nodal (N) component defines the extent of cancer present in regional lymph nodes, and the distant metastasis (M) component is a binary measure of the absence or presence of metastases. Precise TNM characterisation is unique to each cancer site [22], though in general the value beside each component rises (e.g. T0-T4, N0-N3, M0-M1) with the increasing extent and severity of the cancer.

As part of staging, the cancer can be given a histologic grade using a biopsy, via an assessment of tumour cell *differentiation* [22]. This is usually a grading from 1-4, with grade 1 being *well* differentiated, meaning the cells resemble the surrounding tissue at the primary site, down to grade

3-4, *poorly* or *un-* differentiated, where the cells are so different there is an inability to identify the site of origin from the biopsy alone [22]. The combination of T, N and M components allow a patient to be placed into an overall prognostic group identified with Roman numerals (I-IV), which can be further subdivided with characters (e.g. A-C). As with the TNM stage mentioned above, higher stage groups relate to increasingly poor prognosis and disease severity.

Accurate staging is critical for treatment planning. Different treatment options are suitable for different disease stages. Pretreatment staging is a deciding factor between invasive (e.g. surgical), non-invasive (e.g. radiotherapy), or combined treatment approaches.

1.2.3 Conventional Treatments

There are many treatment options available for cancer management, with different applications and success rates depending on the current extent of the disease. Treatments can be given with either curative or palliative intent, and clinicians strive to select the most appropriate techniques based on a number of patient factors (e.g. staging, age), and drawing from previous clinical results. Before a treatment methodology is generally adopted, its effectiveness has been assessed using clinical trials, where different treatments are compared across separate arms of the trial. This is a continuous process in the pursuit of new treatment opportunities.

Often, treatment combinations are administered in carefully managed rounds and at multiple time-points. For example, a therapy is described as *neoadjuvant* if given before the main treatment modality (e.g. neoadjuvant radiotherapy scheduled prior to surgery). Conversely, *adjuvant* therapy is applied after the main treatment (e.g. adjuvant chemo-therapy scheduled after primary radiotherapy to mitigate potential relapse). Terminology is also often integrated, such as *chemoradiotherapy*, which refers to the joint delivery of radiotherapy and chemotherapy within a treatment plan. As an introduction, some predominant and more conventional categories of treatment; surgery, radiotherapy, chemotherapy and hormone therapy, are described below.

- **Surgery** - Surgical resection of tumour masses is one of the oldest and most established forms of treatment. It can have many advantages, as a focused operation at the site of the tumour mass is usually less damaging to healthy tissues compared to other indirect techniques such as chemotherapy. Nonetheless, surgery can include the aggressive removal of a quantity of healthy surrounding tissue, such as nearby lymph nodes, if the intent is curative. The predicted success of surgical procedure is highly dependant on the current extent of the cancer. Surgery is usually not possible or is ineffective at more advanced stages of disease.

- **Radiotherapy** - Radiotherapy is the delivery of high-energy *ionising* radiation to destroy tumour cells and to prevent further replication by damaging their genetic make up. This is done externally, via machine delivered beam radiation, or internally, via a physical implant or injection of radioactive liquid [23]. Linear accelerators (LINACs) are a common choice for external beam radiotherapy. These sophisticated systems deliver precision doses of high-energy x-rays to targeted tumour regions. Advances in delivery techniques are centred on improving the dose to tumours whilst decreasing what is received by surrounding healthy tissue and organs. Fractionation is one such technique, whereby the treatment is scheduled over a number of days or weeks to allow for the recovery of healthy tissue, but also to ensure high-dose delivery to tumour cells that may have previously been in a radio-resistant phase [24]. Technical advances have also led to more accurate dose delivery, with the invention of treatment modalities such as 3D conformal radiotherapy (3D-CRT), intensity-modulated radiotherapy (IMRT), and image guided radiotherapy (IGRT). With 3D-CRT, the use of multi-leaf collimators within LINAC systems enable the beam shape to be adjusted to better match, or *conform*, to the tumour profile. This is facilitated by computed tomography (CT) imaging, discussed in Section 1.3.1, to determine the precise morphology of the tumour. IMRT allows for further shaping of the dose delivery across the tumour by temporally *modulating* the radiation fluence [25]. These fall under IGRT, which to generalise, is the extensive use of imaging throughout radiation therapy to guide the treatment delivery for increased precision.

- **Chemotherapy** - Chemotherapy treatment is the delivery of powerful cytotoxic drugs to destroy malignant cells by disrupting replication and growth [26]. It is a systemic treatment - usually administered either intravenously or with oral tablets - where the drugs circulate throughout the body via the bloodstream. This systemic approach ensures the drugs reach all disease sites, but the broad delivery is a cause of notable toxicity side effects [26]. Chemotherapy is effective due to a definitive feature of malignant cells: cell proliferation. It works through complex interactions with the cell cycle, preventing division and activating apoptotic pathways leading to cell death [26].
- **Hormone Therapy** - Hormone therapy aims to reduce the presence of certain hormones within the body that signal growth in several cancer types [27]. Susceptible cancers are so-called hormone sensitive, such as those that develop in the prostate, breast, ovaries or womb. As an example, prostate cancer cell receptors respond to testosterone with increased cell division, and anti androgen drugs are used to intercept this by attaching to these receptors to block testosterone and suppress this response [27].

These conventional treatments are often administered to patients by considering how an "average" patient with similar symptoms would respond. However, cancer has proven to be a dynamic and heterogeneous disease [28]; each cancer patient in reality is a distinct combination of genetic and environmental factors. These differences manifest in the clinic, as patients with the same cancer type and staging often respond differently to the same treatment. Over the past few decades this has inspired a movement to shift away from a *one size fits all* treatment approach, to a new era of finding increasingly tailored strategies using more data from each individual [29]. This notion of a data driven approach is the basis of the *National Health Service* (NHS) effort to improve patient outcome, coined *Personalised Medicine* [29], also often referred to as *Precision Medicine* [30].

1.2.4 Personalised medicine in oncology

In essence, the hope of personalised or precision medicine in oncology is to use an array of personal data from a patient - principally the tumour biology - to select the optimal treatment strategy [30]. Perhaps ironically, there is sometimes confusion over the precise meaning of these terms, and what personalised care might entail. Berman [31] states that precision medicine cannot be the development of completely unique treatments for each individual, as sometimes imagined, because *"..treatments must be tested for safety and efficacy on groups of people. The best we can ever do is to assign patients to a group that has been fitted to a pre-approved treatment..."*. Rather than continuously developing unique treatments for each unique patient, the emphasis here is rather on understanding, dissecting and identifying the sequence of steps (pathways) a disease can take. If the critical biological characteristics of each step of a disease can be found and quantified clinically, precision medicine proposes that successful targeting and elimination of such critical steps might be an effective treatment strategy [31]. No less important, this approach could also identify which treatments would show little benefit. It is about delivering the optimal therapy first time from those available, avoiding "trial and error" treatment regimes, and better managing unwanted and unnecessary treatment side effects [29].

Of course, the use of tumour data for guiding clinical decisions is not a revolutionary notion: tissue biopsy and blood analysis are examples of common data points utilised for decades in attempts to better diagnose and treat patient cancers. Rather, it is via the recent developments in high-throughput *omics* analysis that further insights and new therapeutic strategies could emerge [32]. Examples of *-omics* [32] analysis that show promise for precision medicine include *genomics* (understanding the genome), *epigenomics* (understanding genome expression), *proteomics* (understanding protein expression), and potentially, the subject matter of this thesis: *radiomics* (understanding tumour phenotypes through imaging). Notably, Precision Medicine has been hindered by a major obstacle of omics strategies that depend on single tumour biopsy: intratumoural heterogeneity [33], introduced in the next section.

1.2.5 Tumour Heterogeneity

Heterogeneity and *homogeneity* are antonyms of the same observation: a description of high dissimilarity or high uniformity, respectively. Cellular tumour heterogeneity is a major challenge for clinical treatment. Cancerous tumours normally become more heterogeneous in nature as they develop, and display both inter- and intra- tumour heterogeneity; refering to variation between patients with the same histological subtype, and variation within the sub-populations of cells that make up a single tumour mass [28, 33].

Intra-tumoural variation within the sub-population of cells is caused by different characteristics of these sub-populations. For example, the varying genetic ability of sub-populations to metabolise or create vascular structure can lead to necrotic regions - areas of dead cells - within a tumour. As introduced earlier, of interest in this thesis is the potential identification of these differences in medical imaging as a useful quantitative measure.

Intra-tumour heterogeneity is thought to be a primary mechanism for tumour adaptation to targeted therapies as Darwinian principles of selection foster drug resistance [34]. For a highly heterogeneous tumour, therapies specifically targeted to a particular mutation identified from a single biopsy may only be successful at providing an environment for the cancerous cells without that mutation to thrive [34] - the tumour as a whole then survives and becomes resistant to the treatment as a result. As such, there is growing evidence that patients with high intra-tumour heterogeneity have a reduced treatment response and inferior outcomes [35]. Measures of heterogeneity may, therefore, be useful for clinical decision making - any such useful measure would be categorised as a *Biomarker*.

1.2.6 Biomarkers

Biomarker: *"A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention"* [36]

Clincial Endpoint: *"A characteristic that reflects how a patient feels, functions, or survives"* [36]

Surrogate Endpoint: *"A biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit..."* [36]

A biomarker is an objective indication (a *marker*) of either regular or irregular biological process [36–38]. Within oncology, biomarkers offer value in disease screening, staging, prognosis, prediction and monitoring [36]. They are most often acquired from sources such as clinical records, blood work and biopsies [39]. Biomarkers derived from tumour biopsies are currently a major criteria in oncology for accurate staging [22].

Biomarkers that can accurately predict benefit from therapies will be the essential cogs in the machinery of Precision Medicine. They aim to mark a clinical end point such as survival, or be used as a surrogate endpoint, which is effectively a substitute predicting clinical benefit [36]; For example, a shrinking tumour volume could be a surrogate endpoint to indicate a treatment is working. In general, radiomic research for imaging biomarkers focuses on extracting many tumour characteristics from scans that can be linked to clinical endpoints with a goal to provide prognostic or predictive benefit if incorporated into clinical practice [8]. There is a subtle though important difference between these: a prognostic biomaker informs of a patient's likely outcome independently of administered treatment and is therefore linked to overall survival; a predictive biomarker conveys the likely benefits (or not) of therapeutic intervention [40]. Medical imaging provides a bountiful avenue for biomarker discovery due to its intricate role in all areas of cancer management, as introduced next.

1.3 Imaging in Oncology

Anatomical imaging: Reveals tissue structure and geometry.

Functional imaging: Provides information on biological processes and functions.

Medical imaging reveals the hidden anatomical structures and biological functions of the human body. It is an incredible advantage to peer past the skin to assess a condition without an invasive surgical procedure for the patient, and imaging techniques are now indispensable in modern medicine as a result. There are a variety of distinct image acquisition *modalities* that utilise different physical phenomenon to produce a functional or anatomical image, including; high energy x-rays, radioactive decay, and magnetic resonance.

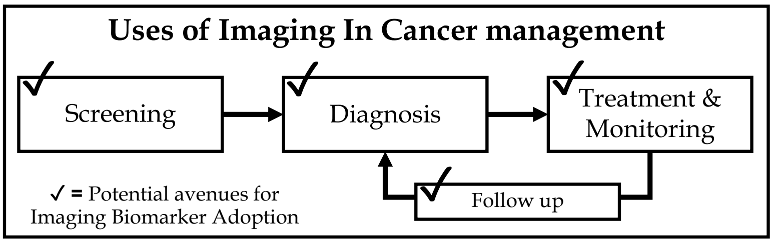

Uses of Imaging In Cancer management
✓
Screening
✓
Diagnosis
✓
Treatment &
Monitoring
✓
Follow up
✓ = Potential avenues for
Imaging Biomarker Adoption

Figure 1.1: Role of imaging in cancer management.

For a particular aliment, a clinician utilises the modality (or modalities) that best maximises the image clarity needed for assessment whilst minimising any potential side effects or discomfort for the patient. The collection of modalities at their disposal spans a broad measurement scale, ranging from microscopic analysis of cell structure, to macroscopic whole-body assessment of organ and bone.

Many of these modalities have potential for quantitative characterisation of disease using radiomic techniques, with a clear application in oncology as focused on in this body of work. Particularly, these include: *Computed Tomography* (CT), *Positron Emission Tomography* (PET), *Single Photon Emission Computed Tomography* (SPECT), *Magnetic Resonance Imaging* (MRI), and *Ultrasound* (US). This thesis utilises three of these key quantitative modalities in particular, CT, PET, and MRI, and as such they are introduced in more detail in the following Sections 1.3.1, 1.3.2, and 1.3.1 respectively. These main three are prevalent in cancer management, offering 3D volumetric assessment. The most common way to view these volumetric images is to scroll through slices in the axial, sagittal, and coronal planes at selected positions.

Modern oncology and medical imaging are now inseparable. Management of cancer involves many technical aspects that can be separated into four areas as outlined in Figure 1.1 : screening, diagnosis, treatment and/or monitoring, and follow up. [6]. Patient imaging is interwoven into every stage, and as a result, each offers fertile ground for imaging biomarker research which seeks to improve cancer management.

Different imaging techniques are preferable for different cancer sites and stages of treatment, often dependent on the type of tissue being examined. As a naive example, a clinician might prefer MRI for assessing brain tissue, yet favour CT for lung tumour diagnosis - though machine availability, radiation risk and scan time are examples of other factors that must be considered. Image quality is not the only concern, and there are health risks inherent to the use of ionising radiation that one must consider. Non-invasive imaging does not mean hazard free. High-energy radiation that penetrates deeply into tissue can cause damage to healthy tissue, so the *dose* of radiation delivered and its effects on tissue and organs must therefore be deliberated and mitigated accordingly. For ionising modalities, although an improved image quality might help with clinical assessment - which might yield more informative image feature extraction - the higher dose needed for this improved clarity would increase the likelihood of damage to healthy tissue. As such, non-ionising modalities such as MRI have a lower associated risk.

Due to the complexity of cancer, no single modality can provide a complete overview, hence the prevalence of multimodal approaches that combine imaging techniques for a more complete tumour characterisation [41]. Multimodal imaging is the merging of single modalities to provide complementary information. Usually, it involves a composite of functional and anatomical approaches, such as PET/CT or PET/MRI that provide an overview greater than the sum of its parts. Indeed, integrated machines designed to acquire PET and CT simultaneously were first pioneered at the cusp of the 21st century by Townsend and Beyer [42], and have since become an invaluable device used in routine patient care throughout the world. The following sections provide an introduction to CT, PET and MRI, which are the major modalities utilised in the field of radiomics.

1.3.1 Computed Tomography

CT is one of the most widely used modalities for diagnostic imaging. To give context of its prevalence, over 5.5 million scans were reportedly undertaken in England between March 2018 - March 2019 [43]. CT imaging techniques are valuable in a clinical setting as they produce a three-dimensional overview of body tissue composition. The scanners are built to measure attenuation of x-ray beams, where attenuation is the decrease in radiation intensity observed due to scattering and absorption of photons as they pass through a material. For a given photon beam energy, a material has a corresponding *linear attenuation coefficient* μ, that is related to the material density. The *Beer–Lambert* law describes the resulting intensity I based on the initial intensity I_0 passing through a material of length *ds* as [44]:

$$I = I_0 e^{-\mu \cdot ds} \tag{1.1}$$

However, human bodies are of course composed of many tissues, and not one constant material. Different tissues have different μ. Equation 1.1 thus becomes a summation of all the different attenuation coefficients μ_i along the beam path [44]:

$$I = I_0 e^{\int \mu(s) ds} \approx I_0 e^{-\Sigma_i \mu_i \cdot ds}. \tag{1.2}$$

By knowing the x-ray beam intensity before, and measuring the intensity after passing through the body, one can calculate this integral of attenuation coefficients, which is a sum when considering discrete portions, by rearranging equation 1.2 to [44]:

$$\Sigma_i \mu_i \cdot ds = \ln \frac{I_0}{I}. \tag{1.3}$$

By measuring the total attenuation of beams at many orientations, CT *reconstruction* techniques can be used to determine the attenuation occurring at set points along beams (i.e. at each voxel). As such, a typical CT scanner is a combination of an x-ray tube and detector, positioned opposite one another in a circular assembly that revolves during image acquisition [45], as shown in Figure 1.2. As the tube-detector rotates around the patient – who is translated horizontally through the scanner - it records the change in beam intensities as an electric signal within the detector. Utilising this spiral acquisition, advanced tomography techniques such as *backprojection* [44, 46] can reconstruct a 3D map of attenuation measured from the projections captured on the detector. These attenuation maps are viewed as volumetric images to visualise the distribution of tissue densities. Conventionally, within CT imaging the attenuation values assigned to each voxel are

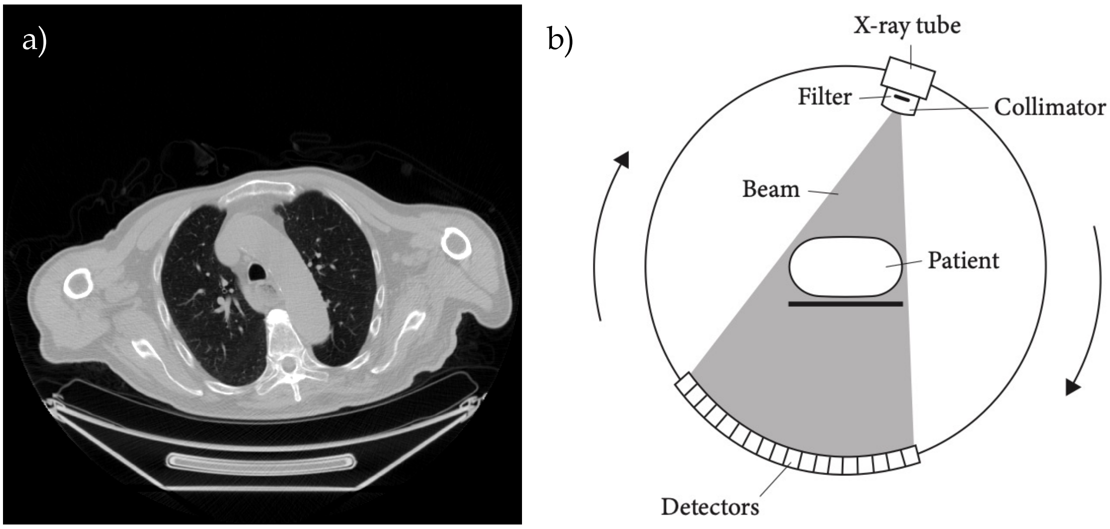


Figure 1.2: An introduction to Computed Tomography. Sub-figure **a)** shows an axial CT slice (range [-1000 500] HU), taken from STAGE dataset discussed in Chapter 4. Sub-figure **b)** is a simple schematic of a CT scanner, based on Reference [44].

replaced with integer values referred to as *Hounsfield units* (HU), a scale centred on the attenuation of water at 0 HU [46].

CT Units

CT images are a matrix of voxels with integer grey levels corresponding to Hounsfield units: a calibrated scale centred around water, given a value of 0 HU, and air, given a value of -1000 HU. This calibration makes CT an excellent candidate for quantitative image analysis techniques. Calculating HU depends on the measured attenuation coefficient μ at each voxel, via:

$$\text{CT}_{\text{number}}(HU) = 1000 \times \frac{\mu_{voxel} - \mu_{water}}{\mu_{water} - \mu_{air}}. \tag{1.4}$$

As this section offers only condensed introduction, more information on CT can be found in the following references [44–46].

1.3.2 Positron Emission Tomography

PET functions effectively in the opposite way to CT, where the detector measures radiation emitted from inside the patient, rather than from an outside source such as an x-ray tube [47]. Patients are administered a radioactive emitting material, usually via injection into the bloodstream, called a tracer (or radiopharmaceutical), which normally contains a combination of chemical compounds that target metabolic functions [48]. As a typical example, a tracer could be an analog for glucose uptake, which is a fuel of metabolic tissue. The most common positron emitter utilise for PET imaging is ^{18}F-Fluorodeoxyglucose (^{18}F-FDG) with a half-life of 110 minutes [48]. As the tracer circulates the body, it accumulates in areas of high metabolic uptake, which is particularly helpful in oncology to pinpoint tumours that exhibit high glucose metabolism as a consequence of rapid, malignant growth [47].

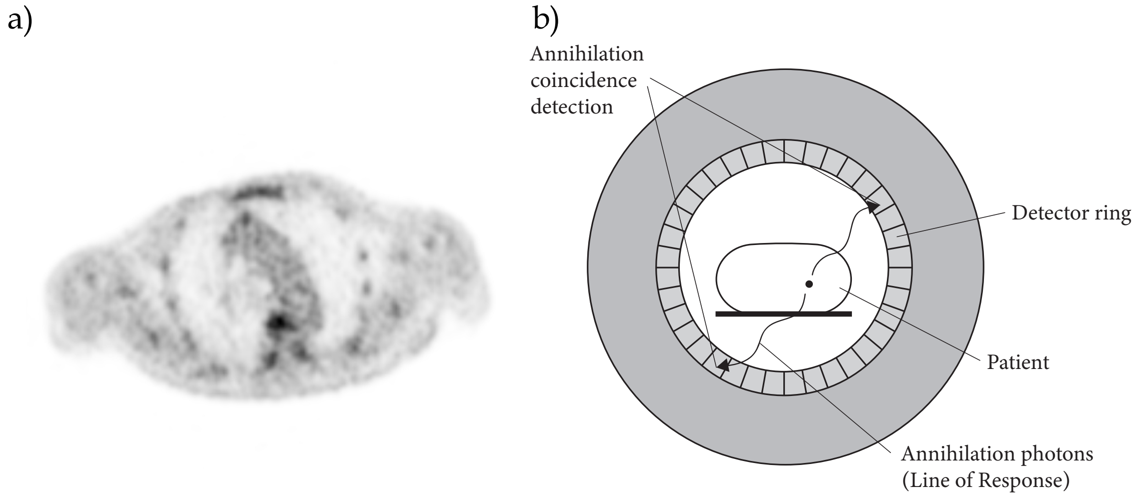

a)
b)
Annihilation coincidence detection
Detector ring
Patient
Annihilation photons (Line of Response)

Figure 1.3: An introduction to Position Emission Tomography. Sub-figure **a)** shows an axial PET slice, taken from STAGE dataset discussed in Chapter 4. The image is reverse grey-scale, with the darkest voxels representing the highest uptake of FDG. Sub-figure **b)** is a simple schematic of a PET scanner, adapted from Reference [48].

In essence, PET images reveal the spatial position of tracer decay, where the voxel intensities are some measure of the number of decays at that position in the image over a given time. As the name suggests, the radiopharmaceutical tracer utilised in this imaging contains positron emitters. These consist of proton abundant isotopes that undergo $\beta+$ decay [47]:

$$p \mapsto n + e^{+} + \nu, \tag{1.5}$$

where a proton p, decays into a neutron n, positron e^{+}, and neutrino ν. However, PET scanners do not directly detect positrons for $\beta+$ decay. Positrons are an antimatter particle and will travel a maximum of only a couple of millimetres before colliding with an electron e^{-} and annihilating into two photons γ of equal energy (e.g. 511 KeV) [47]:

$$e^{+} + e^{-} \mapsto 2\gamma. \tag{1.6}$$

A consequence of annihilation is that the photon-pair travel at near anti-parallel to each other, and the scanner is tuned to detect the *co-incidence* of these particles hitting opposing detectors, as demonstrated in Figure 1.3. This is referred to as *annihilation coincidence detection* (ACD). Using geometry of the anti-parallel photons, a *line of response* (LOR) connects the two points of detection, signalling that the decay must have occurred somewhere along the LOR.

The PET scanner is essentially a large ring of many detectors surrounding the patient to capture the coincidence of photon-pairs resulting from annihilation. The detectors are crystal scintillators, which convert the high energy photons into pulses of light that are picked up by photomultiplier tubes, that can measure and amplify these pulses as an electronic signal [48]. The scanner software is tuned to ignore detections that do not appear to relate to a photon-pair coincidence along a LOR. Analysis of many ACD across many LORs can be used to reconstruct the position of the annihilation events within the body. Also to be considered, when reconstructing a PET image, *attenuation correction* is required as the photons-pair travel through different tissues on route to the detectors. The combination of PET/CT scanners facilitate this as the CT can be utilised to calculate an attenuation map for each LOR [47]. Modern PET scanners with high time resolution are now capable of detecting the difference in arrival time of two coincidence photons, and thus determine the position of the event along a LOR: a process referred to as *time of flight* [47].

PET Units

Radioactive decay is often defined in units of Becquerel (Bq), where 1Bq = 1 disintegration per second. In the context of PET imaging this is normally measured as a concentration of radioactivity, such as Bq/mL (Becquerel per millilitre, where 1mL = 1 cubic centimetre). This can be further converted into a semi-quantitative unit referred to as the *standard uptake value* (SUV) that attempts to remove the variability between patients that occurs due to differences in body weight W and injected FDG [49]. The basic form of SUV is :

$$\mathrm{SUV} = \frac{C_{\mathrm{img}}[\mathrm{Bq/ml}] \times W[\mathrm{g}]}{D_{\mathrm{inj}}[\mathrm{Bq}]}, \tag{1.7}$$

where C_{img} is the radioactivity concentration measured by the scanner and D_{inj} is the injected dose. Assuming FDG was uniformly distributed within the body this would lead to an SUV = 1 g/ml everywhere, and would be dimensionless under the assumption that 1ml of tissue weights 1 gram [49].

As this section offers a condensed introduction, more information on PET imaging can be found from the following references [42, 47, 48].

1.3.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging makes use of a physical phenomenon known as *nuclear magnetic resonance* (NMR). Hydrogen is an element that displays NMR properties - which is ideal for medical imaging purposes- as the tissues of the human body are largely composed of it (e.g. water and fats) [50]. Hydrogen atoms, with a nuclei of a single proton, have two properties necessary for NMR: 1) non-zero charge and 2) non-zero spin, which give rise to *nuclear magnetism* and *spin angular momentum* [51]. It is the interaction of these two characteristics with an externally applied magnetic field and radio-frequency pulses that facilitate MRI. In essence, the contrast in signal intensity seen within typical MR images result from the density of hydrogen atoms and characteristic *relaxation times* of different tissues after excitation.

To begin, the patient lies within a strong magnetic field, $\mathbf{B}_0$, running straight down the cylindrical tube of the scanner. The presence of this field causes the hydrogen nuclei within the body to orient themselves in two states with respect to $\mathbf{B}_0$: parallel (spin up, lower energy), or anti-parallel (spin down, higher energy) [50]. As it requires slightly less energy, it is statistically more likely that the hydrogen nuclei end up in parallel (spin up) state compared to spin down, which results in a small net magnetic vector $\mathbf{M}_z$ in the direction of $\mathbf{B}_0$ [50]. $\mathbf{M}_z$ is hard to measure as it is dwarfed in magnitude by the static $\mathbf{B}_0$ to which it is aligned.

Within a static magnetic field $\mathbf{B}_0$, the hydrogen nuclei also precess in a complex motion around the field at particular frequency due to their spin angular momentum; This frequency of preces-

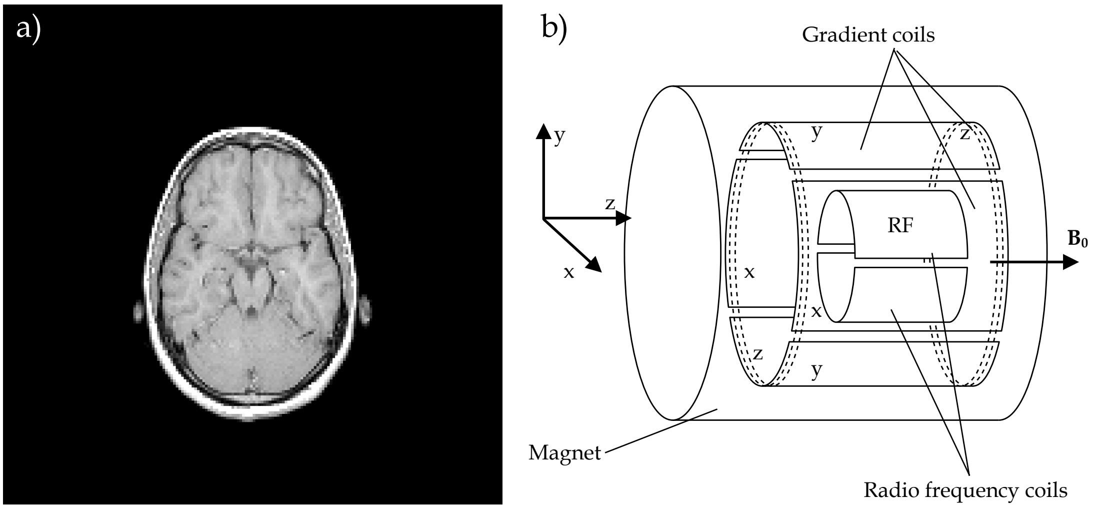


Figure 1.4: A brief introduction to MRI. Sub-figure **a)** shows an example MRI image of the brain [52]. Sub-figure **b)** is a simplified schematic of a MRI scanner coil configuration. This schematic is based on [48].

sion is called the *Larmor frequency*, which is proportional to the strength of the applied field [50]. Processing protons thus have a magnetic vector with a longitudinal and transverse component. As mentioned, it is hard to measure the net *longitudinal* magnetisation caused by aligned hydrogen nuclei along the static field as it points in the same direction as $\mathbf{B}_0$. As the hydrogen nuclei precess independently and in different phases, the net *transverse* magnetisations ($\mathbf{M}_{xy}$) sums to zero as they all cancel out. To cause a measurable in-phase magnetisation in $\mathbf{M}_{xy}$, energy needs to be added to the system. Generally in physics, energy is readily absorbed in systems when applied at a *resonance* frequency. A radiofrequency (RF) pulse tuned to the Larmor frequency applied in the transverse plane ($\mathbf{B}_1$, perpendicular to the external magnetic field $\mathbf{B}_0$) can tip the net magnetisation vector towards the x-y plane to be measured [50]. The RF pulse timings dictate the angle through which this net magnetisation vector moves. This rotating magnetisation is measured in the transverse plane as it induces a voltage in receiver coils, aligned at right angles to the transverse plane within the scanner.

After each RF pulse, the system *relaxes*, returning to a lower energy state. The *longitudinal* magnetisation $\mathbf{M}_z$ recovers with a characteristic T_1 relaxation time (through *spin-lattice* interactions), while the *transverse* magnetisation $\mathbf{M}_z$ decays with a a characteristic T_2 relaxation time (through *spin-spin* interactions) [50].

To produce an MR image requires the spatial localisation of these MR signals. This is accomplished through generating short term variation in the magnetic field across the patient in addition to the main field $\mathbf{B}_0$, with a set of *gradient* coils for each direction x,y,z [53]. As the Larmor frequency is proportional to the magnetic field, these gradients alter the frequency and phase of the processing protons, depending on their location within the scanner. Thus, utilising complex RF pulse sequences with application of these small gradients, higher or lower frequencies in an obtained MR signal distinguish different positions in space [53]. An inverse Fourier transform is used to transfrom the sampled MR frequency signals into a spatial image [53].

Figure 1.4 shows an example MRI image and a simple schematic of magnets and gradient coils, alongside the orientation of the coordinate system for a typical MRI scanner. MRI image generation is nuanced and complex, and this section offers only a brief introduction. The author encourages consulting the cited references for more insight [47, 50, 53, 54].

1.4 Measuring Heterogeneity in Imaging

When visually assessing diagnostic imaging it is evident that cancerous tumours are rarely simple homogeneous masses (as discussed in Section 1.2.5). Imaging of a single lesion can reveal complex voxel intensity patterns that signal underlying genetic *intra*-tumoural diversity. As an example, PET imaging voxel patterns indicate varying metabolic activity. Different genetic mutations, cell expression and cell metabolism across each tumour results in phenotypic diversity within patient cohorts, and this is thought to be exhibited in the imaging [16]. As discussed, identifying and quantifying this phenotypic diversity may have tangible clinical benefit, as it has been reported that increased tumour heterogeneity is associated with worse outcomes, including higher risk of treatment resistance, recurrence and metastasis [35].

If underlying heterogeneous characteristics can be detected and defined quantitatively with imaging, the appeal is clear; imaging biomarkers offer non-invasive, global measures of the tumour. A reliable non-invasive imaging test is highly desirable as they are repeatable (e.g. a sample is not destroyed during the test), and there is no additional associated risk to the patient beyond the radiation dose, which is mitigated if the imaging has already been taken as part of routine clinical care. With regards to heterogeneity, analysis encompassing the entire tumour would also be advantageous over routine single biopsies as these are subject to the area in which the biopsied is taken [33], and not having complete information is likely to hinder the ability to accurately model disease development as a result. In the current role of diagnostic imaging in the clinic, the notion of heterogeneity has not extended much beyond a qualitative assessment, though medical imaging analysis is an extremely active research domain that has intensified under the banner of *Radiomics*, which took hold as a term in the literature around 2012, and has seen an exponential rise in use as highlighted in Figure 1.5.

One of the most prominent techniques within radiomics for assessing tumour heterogeneity is *Texture Analysis*. It usually involves quantifying the spatial relationships between voxel intensities within a region of interest (ROI) through the use of *texture matrices* to summarise voxel

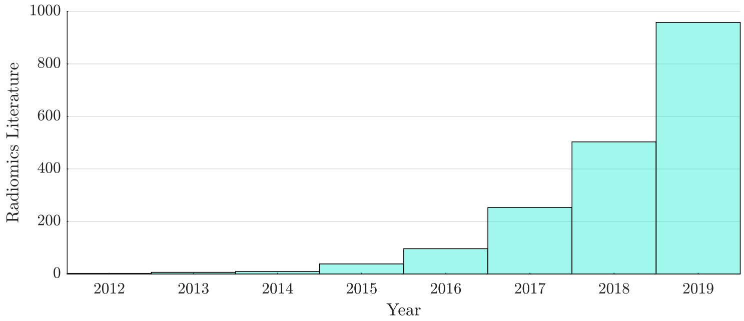


Figure 1.5: Literature documents with *Radiomics* in the Title, Abstract or as a Keyword; Found using a Scopus search [55] with the following criteria set: Publication between 2012-2019; Document types: articles, conference papers and reviews. There are 1,865 documents found matching these criteria, plotted by year of publication.

distributions. In radiomics studies the analysis techniques utilising texture matrices include the *Grey Level Co-occurence Matrix* (GLCM) [56], the *Grey Level Run Length Matrix* (GLRLM) [57], the *Grey Level Size Zone Matrix* (GLSZM) [58], the *Grey Level Distance Zone* (GLDZM) [59], the *Neighbourhood Greytone Difference Matrix* (NGTDM) [60], and the *Neighbouring Grey Level Dependence matrix* (NGLDM) [61]. Mathematical definitions of these texture matrices are described in detail in the following chapter. By constructing texture matrices, hundreds of so called "higher-order" quantitative features can be extracted as potential inputs for clinical models. These higher-order measures are named as such, as they incorporate the spatial relationships of pixels / voxels within the image.

As a demonstration, Figure 1.6 shows explicitly how first-order techniques, such as analysis of the intensity histogram, are not enough to capture intensity patterns that could represent heterogeneity within a tumour. In this simulated example, both images have the same intensity values within the contoured region, the pixels have just been spatially rearranged. Example (B) is clearly more heterogeneous in appearance as a result, and this is shown in the disorder of the GLCM. In this case, features extracted from the GLCM would easily discern the more heterogeneous tumour, unlike those extracted from the intensity histogram.

Texture analysis is a well established concept within the field of computer vision, and methods such as the GLCM and GLRLM first date back to the mid-1970s [56, 57]. These techniques had interesting early use cases in e.g. the assessment of terrain from aerial and satellite imaging [56]. It did not take long to consider these options for medical imaging. However, the unique nature of medical imaging offers many challenges compared to texture extraction on other types of 2D digital photography [2]. Particularly, 3-dimensional (3D) assessment of the overall heterogeneity of tumours in patients with cancer may benefit from recently extended 3D texture analysis techniques. With the ever decreasing computational cost and increasing computational

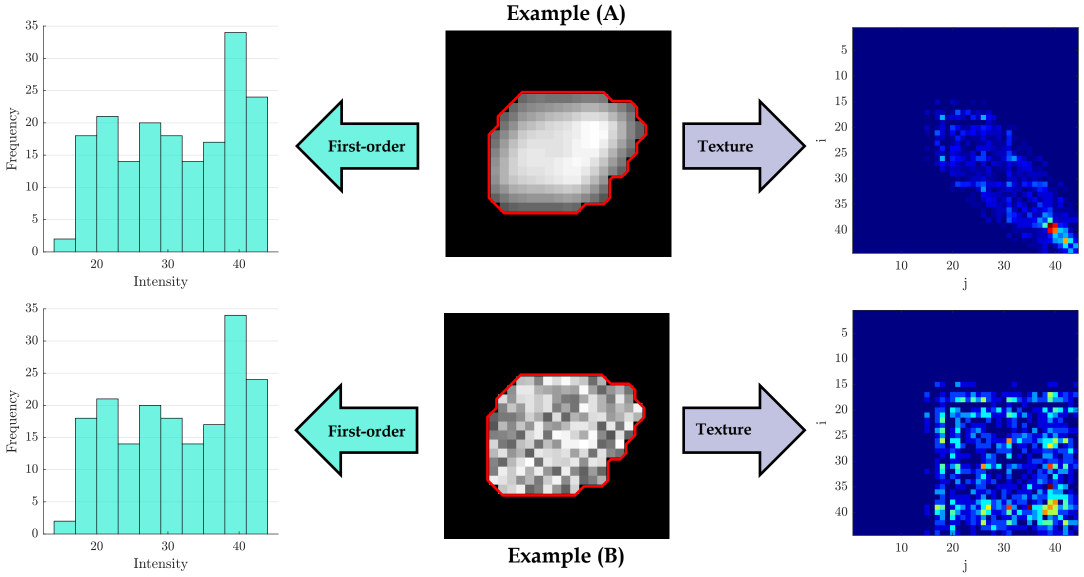


Figure 1.6: Visual comparison of first-order and texture matrix-based analysis. Within the contoured regions of Example (A) and (B) there are the same intensity values, but they are in different spatial positions. To the left of each image is the intensity histogram distribution, which remains unchanged between examples. On the right for each image is a visual representation of the corresponding *grey level co-occurrence matrix* (GLCM) (calculated in 2D with direction merging, see Section 2.4.6). Clearly, Example (B) is much more heterogeneous and this is visually evident in the disorder of the GLCM.

power, the combination of many of these older techniques with newer variations to conduct a "high-throughput" extraction of many features, has led to this resurgent interest in the further utilisation of imaging under *Radiomics*, as demonstrated in Figure 1.5. Exploring the strengths, weaknesses, pitfalls and future of these biomedical texture analysis techniques within radiomics studies is at the core interest of this thesis.

1.5 Radiomics Overview

Although the term radiomics has taken off in the last decade (Figure 1.5), medical imaging analysis has a longer history utilising other terms such as *Computer Aided Diagnosis* (CAD) [62]. In its inception, what appears to set the early radiomics studies apart from previous research is the workflow emphasis on high-throughput feature extraction: i.e., to collect a great number of features from many patient scans [3]. Several earlier CAD studies could be considered "radiomic studies" without explicit use of this newer terminology. Instead, they might emphasise the use of a particular image analysis technique (e.g. a *texture analysis* study) [2, 62]. In general, a typical radiomics study is attempting to develop a clinically useful model from imaging data and other clinical measures using machine learning techniques that group patients based on expected outcomes [63].

Briefly, it should be noted here the further recent separation of radiomics into two domains, which has occurred within the field throughout the course of this project. This thesis focuses on the first domain: the potential of *engineered* features to act as imaging biomarkers. These features have been mathematically defined to measure certain aspects of an image region. The second domain concerns the unprecedented and overwhelming serge of deep-learning (DL) techniques using neural networks (e.g. convolutional neural networks (CNNs)), trained to self-learn patterns directly from imaging that are relevant to a given task [64]. This project set out to address the key challenges within the first domain, though discussion of both domains and the rapidly evolving field of radiomics can be found in Chapter 6. This section introduces the development and challenges of this first domain.

The traditional radiomics workflow was first presented as a progression through 4 major stages:

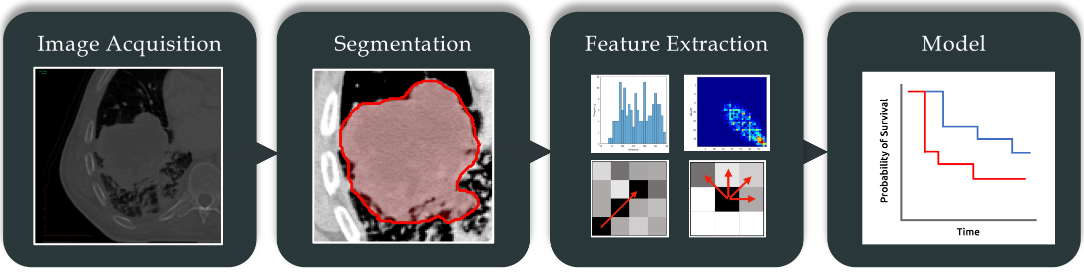


Figure 1.7: The basic components of a traditional radiomics workflow as outlined in seminal works such as Lambin *et. al.* [4], Aerts *et. al.* [16] & Gillies *et. al.* [3]. As explored in Chapter 2, these key stages in fact consists of many complex sub-stages.

1) *Acquisition*, 2) *Segmentation*, 3) *Feature extraction*, and 4) *Model development* [3, 4, 16]. In reality this is a simplification of the complex pipeline needed to convert raw imaging into potentially useful data. Each stage contains many sub-stages with challenges to address and overcome. In particular, the steps one uses to *process* an image between acquisition and feature extraction are critical for consistent feature values. For an introduction, this section provides a brief overview of these defining stages of radiomics. This thesis will then focus particularly on the critical challenges within the many sub-stages of feature extraction.

1.5.1 Acquisition

Before any image analysis can take place, clearly one must obtain an image. Medical imaging is acquired with medical scanners, such as CT, PET or MRI introduced in Section 1.3, via complex protocols and machine settings, where tweaking any number of parameters (e.g. CT tube voltage) can adjust the contrast and quality of the final image. More over, continued technical advances in hardware and software [65] (e.g. refining reconstruction algorithms) - with commercial companies competing for hospital business - have led to ever improving imaging acquisition. Importantly, the effect of acquisition on feature extraction in radiomics should be well understood.

To explore feature consistency to repeated acquisition, studies [66–69] have utilised a test-retest approach, where patients or phantoms are scanned twice in short intervals to obtain two images for comparison. With a fixed acquisition and reconstruction protocol, features with minimal variation between the test-retest scans can be identified as potentially stable. As such, Van Timmeren *et*. *al*. [67] recommends that test-retest analysis should be performed whenever possible for prospective radiomics studies. This will verify if potential radiomic signatures are robust to the acquisition protocol. It is feasible a dedicated modality dependent acquisition protocol for radiomics could emerge to standardise analysis across many centres.

1.5.2 Segmentation

Segmentation (or *delineation*, or *contouring*) is the act of spatially outlining a *region of interest* (ROI) or *volume of interest* (VOI) in an image that will undergo the analysis. In an oncological setting, VOIs typically define tumour lesions or the organs-at-risk (OAR) that are susceptible to treatment delivery. It is a deceptively hard and tedious task when done manually. More over, accurate contours are critical in radiomics as they have a direct impact on the features extracted [68] - perhaps intuitively, what is included (or not) in analysis will have an effect on the analysis. Segmentation defines the region in which features are extracted. How sensitive a radiomic feature is to marginal changes in tumour definition is likely an indication on usefulness as clinical measure.

Expert manual contouring is normally considered the ground truth. However, this can be problematic as manual segmentation is susceptible to inter and intra-reader variability [68, 70–73]. This is exacerbated when considering 3D tumour volumes which are usually manually segmented on the axial plane, slice by slice. This time consuming and monotonous task can in fact be impractical for large datasets that are not already segmented. To alleviate these problems, semi-automatic and automatic segmentation methods are usually recommended in radiomics work-flows as the resulting delineations are significantly more robust and reproducible [70]. Some of the simplest techniques for segmentation automation include region growing, clustering and thresholding algorithms, though there are many open source and commercially available tools with more advanced approaches [74].

One such example is a tool for semi-automatic segmentation developed at Cardiff University called ATLAAS (*Automatic decision Tree-based Learning Algorithm for Advance Segmentation*) [75, 76]. For PET imaging, through the use of decision trees, ATLAAS has been trained to select the most accurate contour from a range of segmentation algorithms. In the work in Chapter 4 of thesis, ATLAAS was utilised to efficiently batch segment the primary tumours from PET imaging of a large cohort of patients with Oesophageal Cancer. A key benefit of accurate contour automation is that it can allow radiation experts to concentrate on more technical tasks, yet presently, an appropriate specialist must still approve any delineations produced in this way.

1.5.3 Feature extraction

Radiomic features are used to quantitatively describe the characteristics of a segmented VOI of an image, and each extracted feature is a single-value representation of a particular attribute of that VOI. Alongside *texture* feature families mentioned in Section 1.4; *morphological* (shape based), *statistical* and *first-order Histogram* families make up a sizeable proportion of the hundreds of features that see wide-spread use within radiomics studies. These feature families are implemented from scratch and explored within this thesis, with full methodology and mathematical definitions left to Chapter 2. This work will discuss how image processing - through re-segmentation, interpolation, discretisation, filtering etc. - is integral and inherent to the feature extraction, and a lack of reporting implemented techniques impedes the ability to validate promising radiomics models.

1.5.4 Modelling

Uses for Radiomic Models

Diagnosis: Such as the ability to determine malignancy (or benign) status, tissue histology or tumour stage.

Overall survival: Giving a prognostic outlook on a patient's survival chance.

Tumour aggression: The chance of progression, re-occurrence, or relapse.

Distant metastases: Predicting the presence of metastases based on analysis of the primary tumour.

Treatment response E.g. the likelihood a patient will response to a given treatment regime (predictive models).

Genetics: Linking underlying tumour genetics to the imaging tumour phenotype (*Radiogenomics* studies).

Once features have been obtained, the goal of radiomics is collating them with other clinical variables to develop a prognostic or predictive model with a tangible clinical benefit. These vary in methodology considerably depending on the hypothesis of the study. Examples model areas are listed in the box above.

A seminal radiomics model often cited is that of Aerts *et. al.* [16], who extracted 440 quantitative features from CT imaging of 1019 patients with lung or head-and-neck cancer to develop a prognostic "radiomics signature". The Aerts *et. al.* [16] signature appeared to have prognostic power, and showed an apparent underlying association with gene expression and primary tumour stage. This study stimulated many other investigations, and was one of the influential publications responsible for the popularity increase of radiomics literature as shown in Figure 1.5. However, further recent investigation of the reported Aerts *et. al.* radiomics signature by Vallières *et. al.* [77] suggested that an underlying correlation with tumour volume was a prominent artefact of the features being expressed in the signature, rather than an actual measures of tumour heterogeneity. Nonetheless, they suggested that by adopting modified versions of the features that attempted to correct for this apparent volume dependence, it was possible to still obtain a higher prognostic power with the same features compared to volume [64, 77]. This is a clear example of the potential use, and the subsequent challenges, that can arise in pursuit of clinically useful imaging biomarkers. The following section summaries key challenges within radiomic feature extraction that became the focus of this thesis project.

1.5.5 Key Challenges in Radiomic Feature Extraction

Coinciding with the beginning of this project, Hatt *et. al.* [10] published a comprehensive review paper that summarised the current state of radiomics research in PET/CT imaging. Despite a focus on PET/CT, many of the issues discussed are in fact valid for radiomic analysis of any imaging modality. They identified a number of open challenges within the field which became a catalyst, focus and further justification for the work carried out in this project. To see forward progression in radiomics, greater pooling, comparison, replication and validation of studies is required [10] . Of particular relevance to this thesis, four key issues inhibiting this forward progression are:

1. **Variation in feature nomenclature & definition** - Studies can become difficult to agregate and replicate if they use individual feature nomeclature and defintions, and this becomes almost impossible if the extraction methodology is sufficiently under-reported. Hatt *et. al.* [10] highlighted examples within the literature of apparent confusion or discrepancy in feature naming, such as models presented as using *texture* features when only first-order statistical measures are utilised (E.g. [78, 79]), or features with same mathematical definition but referred to by different names. Conversely, there are also many features with the same name but derived from different feature families, (e.g. GLCM *entropy* / Intensity Histogram *entropy*), which creates ambiguity in discussion when these features are not reffered to precisely [80]. Not using a common nomenclature and definition facilitates confusion and harms study replication.

2. **Radiomic workflow complexity** - A clear challenge in radiomics lies in the complexity of extraction. There are many processing steps which require methodological decisions that can quickly become distinct sources of variability. Variability in processing steps - such as resampling voxel size (interpolation), re-segmentation, binning of intensity values (dis-

cretisation), texture matrix aggregation, or the application of image filtering - can cascade down the radiomics pipeline, where even small differences are likely to be amplified and accumulate into greater discrepancy. Hatt *et. al.* [10] strongly states the need for consensus in methodology and benchmarks for software to help tackle the challenge of workflow complexity causing replication issues.

3. **Feature Robustness, Reproducibility and Repeatability** - Clinically useful imaging biomarkers will need to be consistent. As introduced in the previous section, evaluating the repeatability, reliability and robustness of promising features have been a major consideration within the growing radiomics literature [10, 80]. Repeatability refers to test-retest style studies that evaluate the precision under controlled near-identical experiments, where as reproducibility and robustness studies evaluate the resilience of features to different extraction scenarios [10]. As examples, studies have evaluated reproducibility of features due to segmentation [68, 70–72, 81], discretisation [68, 71, 81], and image reconstruction algorithms [82, 83]. However, it should be noted these studies mentioned have been performed under the burden of challenges 1 and 2. E.g. the software and precise work-flow utilised in the study by Leijenaar *et. al.* [68] is different to that of the image reconstruction study by Galavis *et. al.* [82]. Understanding of feature robustness to many other image processing steps, such as interpolation and filtering, is also far from complete.

4. **Feature Redundancy** - If a complex feature is highly correlated with a simpler or more intuitive metric, the complex feature is likely a redundant measure. By its very nature a "high-throughput" approach like radiomics is likely to result in many correlated and therefore redundant features as 100s to 1000s can be obtained. Removing redundant features must be a preliminary step of any radiomics modelling. This process falls under *feature selection*, which is a much studied area of computer vision and a core aspect of machine learning techniques [63]. Many of these techniques have been explored in radiomic analysis to reduce redundancy prior to modelling. When identifying useful clinical biomarkers it remains a key challenge.

   As an example, Hatt *et. al.* [10] argues any radiomic feature quantifying heterogeneity is likely to only add additional benefit if it is not highly correlated with the tumour volume, and that there are two subtleties here concerning the potential causes of correlation: a) the algorithm itself may be highly dependent on the number of voxels within the tumour, and thus correlate with larger tumours as they contain more voxels, or 2), biologically, larger tumours by their nature have the potential to exhibit more heterogeneity compared to a smaller tumour. As larger tumours intrinsically have more mass, they have more capacity to contain different cell and tissue types. It is not easy to determine whether a potential correlation is due to limitation of a particular feature definition, or just a biological aspect of large tumours. The latter may mean that, although correlated, a feature might offer additional benefit over volume.

With hundreds of radiomic features - that can be tweaked with many processing choices - it is paramount that a standardised approach to extraction is established. Even as radiomics continues to develop, a common methodology is needed as an anchor to assess any evolution or improvement. A standard reference for nomenclature and definitions, alongside recommendations,

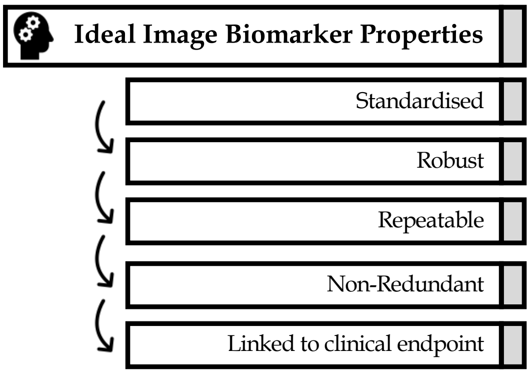


Figure 1.8: Ideal biomarker properties

guidelines, and benchmarks for implementation, is urgently required [10]. From there, one can be more confident in identifying those features both robust to the processing steps of radiomics, as well as non-redundant. As summarised in Figure 1.8: it is robust, non-redundant, *standardised* features that stand the best chance of adoption into clinically useful models in oncology.

1.6 Thesis Aims

The preceding sections provided an introduction to contextualise the following hypotheses and aims explored in this work.

The primary working hypothesis of this thesis is that undiscovered and unreported variation in radiomic techniques is harming the reproducibility of radiomic features that have the potential to act as image biomarkers in oncology. Image biomarkers could provide tangible clinical benefit by acting as prognostic and predictive measures that quantify tumour heterogeneity. Yet, the quantitative values of potential tumour biomarkers are likely to significantly diverge without precise definition, methodology and reporting of techniques used.

No clear guidelines or recommendations is significantly affecting replication and validation studies of promising radiomics models. There is also a lack of benchmarks for radiomics that properly consider the intricacies of medical image processing. Recommendations need to be agreed on as a research community for radiomics to progress. Tools and methods for reaching consensus are required.

Furthermore, this work explores the hypothesis that many prominent radiomic features currently used in the literature are too sensitive to necessary processing steps of extraction and would not be stable enough to provide clinical benefit as a result. A standardised feature would provide no benefit if it is not stable. Removing these features and identifying stable biomarker candidates would lead to more optimal modelling. In this context, the *optimisation* of radiomic extraction refers to the process of determining stable, useful features.

As such, this research project had three overarching primary aims:

Aim-1): Produce benchmarks and recommendations for radiomic feature extraction that will support clinical adoption of radiomics techniques.

Aim-2): Develop analysis methods that enable evaluation of multiple radiomics software to reach standardised benchmarks through consensus.

Aim-3): Identify features that are robust to prominent image processing steps to facilitate more optimal and generalisable radiomics modelling.

1.7 Thesis Content

This thesis has been structured into 6 chapters.

- **Chapter** 1 provides an introduction to this project through a general discussion of oncology, medical imaging, tumour heterogeneity, biomarkers and the field of radiomics. Through a brief overview of influential literature, this chapter presents some of the key challenges and open areas of research within radiomics that underpin the work carried out in the rest of this thesis
- **Chapter** 2 details further background information on radiomic feature extraction and summarises the collaborative development of a standardised image processing scheme by the *image biomarker standardisation initiative* (IBSI). This chapter introduces the SPAARC (*Spaarc Pipeline for Automated Analysis and Radiomics Computing*) radiomics package, the software developed alongside this thesis. The main aim of this chapter is to describe the technical task of implementing a standardised radiomics pipeline through rigorous description of the processing scheme and feature families. A secondary aim of the chapter was to introduce common methods used to measure feature reproducibly, repeatability and robustness.
- **Chapter** 3 further explores the standardisation of radiomic extraction and is split into three main sections. The first section reports on the results of a large international collaborative effort to determine reference values for many prominent radiomic features, via an iterative consensus based study by the IBSI. Results from this work were published in the journal *Radiology* [84]. For this thesis, Cardiff's software progression and the author's contributions are further explored. The second section extends on this study by identifying, quantifying and evaluating some key causes of discrepancy that occurred within the initiative identified by the author's analysis. The final section offers an overall discussion of this work and the impact on radiomic studies moving forward. The main aim was to produce a set of consensus based reference values - for features found frequently in radiomics studies - that enable software and studies to be more effectively validated and reproducible.
- **Chapter** 4 narrows the focus with a study assessing the robustness of IBSI standardsied radiomic features to isotropic voxel interpolation, an often necessary processing step of 3D feature extraction. This work utilises a large PET imaging dataset of oesophageal tumours to eventuate the stability of features to extraction at different voxel sizes. This study also assessed how different interpolation methods affected the feature categorisation. The aim

of this study was to identify features robust to interpolation to facilitate feature reduction and optimisation techniques. This study was published in *Scientific Reports* [85].

- **Chapter 5** further explores the challenges of reproducible radiomics by considering standardisation of another major processing step: image filtering. The key aims of this chapter were to identify pitfalls and challenges of filter application, and to develop and evaluate a methodology to determine consensus-based reference *response maps* for further benchmarking. The author developed a method to assess potential variation in different software that can apply filters, to move towards a standardised approach to filter-based radiomics features. This chapter contains preliminary work for the next instalment of the IBSI, which aims to extend the number of reference values to include many common filter-based features. This work has contributed to a pre-print of the next IBSI instalment available on *arXiv* [86], with the main study ongoing.
- **Chapter 6** provides a further discussion of this body of research. The significant findings and contributions are summarised alongside brief discussion of other co-authored published studies that have utilised the developed SPAARC radiomics software. Following this, there is a general discussion and critical reflection of radiomics research, including: commentary on the rise of deep learning and the potential advantages and disadvantages compared to the traditional engineered feature approach, the challenge of high dimensional feature extraction and modelling, and the need for data availability. Finally, there is a summary of the potential future work leading from this project and final conclusions.

1.8 The Image Biomarker Standardisation Initiative

A core aspect of this project fed into a large international collaboration known as the *Image Biomarker Standardisation Initiative* (IBSI). This initiative was led by Dr Alex Zwanenburg and Dr Martin Vallières, and the governance of the IBSI comprises an inner core-group of 6 researchers (see Table 1.2). The author is one of these core members in the IBSI endeavour, providing significant input to the project direction and design as a result. The author's contributions to the IBSI discussed in this thesis are highlighted in the next section. Participation to the IBSI (Appendix A) remains open to any institution or research group that develops radiomics software to encourage adoption of this standard.

The main novel aspect of the IBSI was to achieve collaboratively determined benchmarks through exhaustive implementation. The more teams that can provide matching results for a benchmark, the stronger the consensus. Many different groups are needed to independently develop functionalities for medical image processing and feature extraction in their preferred programming languages. As such, this exhaustive approach produces benchmarks that the wider research community can have strong confidence in, as many prominent institutions and software provided values to validate the benchmarks through consensus. A key aspect is that this approach to benchmarking uncovers many unknown discrepancy causes, software bugs, and processing steps that need clarification of methodology.

With the exponential increase in radiomics (Figure 1.5), at this pivotal moment for the field, the

IBSI is well positioned to addressed this pressing need for standards.

1.9 Contributions

Chapters within this thesis include published material in which the author was a lead or co-author. This thesis is in the author's own words. How published content is distributed within the subsequent chapters and the author's main contributions are as follows:

- **Chapter 2** and **Chapter 3** discuss and contain material that was published in *Radiology* as a manuscript titled *"The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping"* [84]. The main study is discussed in Section 3.2.
  - The author provided 1776 benchmark values for a baseline set of 165 radiomic features. This data directly led to the identification of several key issues, namely, 1) clear discrepancy arising from the interpolation grid generation and 2), the effect of combining multiple re-segmentation methods.
  - The recommendations for these image processing tasks were refined as a result of this work.
  - The author performed further independent data analysis for these two discrepancy issues in this thesis.
  - As one of the top contributing teams, data provided yielded valid benchmarks for rare feature variants (e.g. All 2.5D texture features) that would not have been standardised otherwise.
  - As co-author on the published manuscript, the author of this thesis provided editing and revision.
- The study in **Chapter 4** was published in *Scientific Reports* as a manuscript titled *"Assessing radiomic feature robustness to interpolation in ^{18}F-FDG PET imaging"* [85].
  - The author primarily designed and led the study.
  - The author performed all the experimental data analysis.
  - The author drafted the original manuscript.
- **Chapter 5** discusses material from further work conducted by the *Image Biomarker Standardisation Initiative* that has been made available as a pre-print titled *"Standardised convolutional filtering for radiomics"* on *arXiv* [86]. This study remains ongoing to develop consensus benchmarks for filter-based radiomics extraction.
  - The author designed the methodology to determine consensus response maps for benchmarking software.
  - The author developed the data analysis pipeline and performed the experimental analysis.

  - The author identified key causes of discrepancy in the use of filters through this analysis, namely, these were errors arising from padding, filter orientation, and orientation pooling techniques for *odd* compared to *even* filter kernels. Recommendations for these image processing tasks are based directly on this work.
  - The author implemented 21/25 of the filter tests assessed in this work.

- **Chapter 6** highlights further impact resulting from this work and contains a brief discussion of published material ([76, 87–89]) in which the software developed by the author was used to conduct the radiomic analysis.
  - As a co-author for these articles, the primary contribution of the author of this thesis was to provide software guidance to facilitate and perform the radiomic feature extraction.

Table 1.2: IBSI core team (December 2020). * = Lead investigators. Also see https://theibsi.github.io/contact/.

Name	**Institution(s)**
Alex Zwanenburg*	OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany. National Center for Tumour Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, and; Helmholtz Association / Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
Martin Vallières*	Department of Computer Science, University of Sherbrooke, Sherbrooke, Québec, Canada. GRIIS, University of Sherbrooke, Sherbrooke, Québec, Canada
Adrien Depeursinge	Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Switzerland. Service of Nuclear Medicine and Molecular Imaging, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
Vincent Andrearczyk	Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Switzerland
Philip Whybra	Medical Engineering Research Group, School of Engineering, Cardiff University, Cardiff, United Kingdom
Joost van Griethuysen	Department of Radiology, the Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands. GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands. Department of Radiation Oncology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.
Henning Müller	Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Switzerland. University of Geneva, Geneva, Switzerland.
Roger Schaer	Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Switzerland.

2

Developing Radiomic Techniques & Tools

"The act of making something will force you to learn more deeply than reading ever will."

— James Clear

2.1 Preview

This chapter provides technical information on the core processing steps of feature extraction in medical imaging. The following contains an adapted, condensed summary of the pipeline developed by the *Image Biomarker Standardisation Initiative* (IBSI) [84], to which the author is a core contributor. Prior to this initiative, there was little published consensus on the precise handling of medical imaging for feature extraction, and no available verified benchmarks for many prominent features utilised in the radiomics literature. Throughout the course of this project, the IBSI have produced a comprehensive scheme for feature computation from medical imaging, which aims to be a key reference and standard for the radiomics community moving forward [84]. The radiomics software package that was developed by the author alongside this thesis project to participate in the standardisation initiative is introduced in this Chapter in the context of each processing step along the pipeline. Following on from this is additional background information on the statistical analysis utilised for feature robustness testing. Useful clinical features need to be both standardised and robust.

2.2 SPAARC: Radiomic Feature Extraction Package

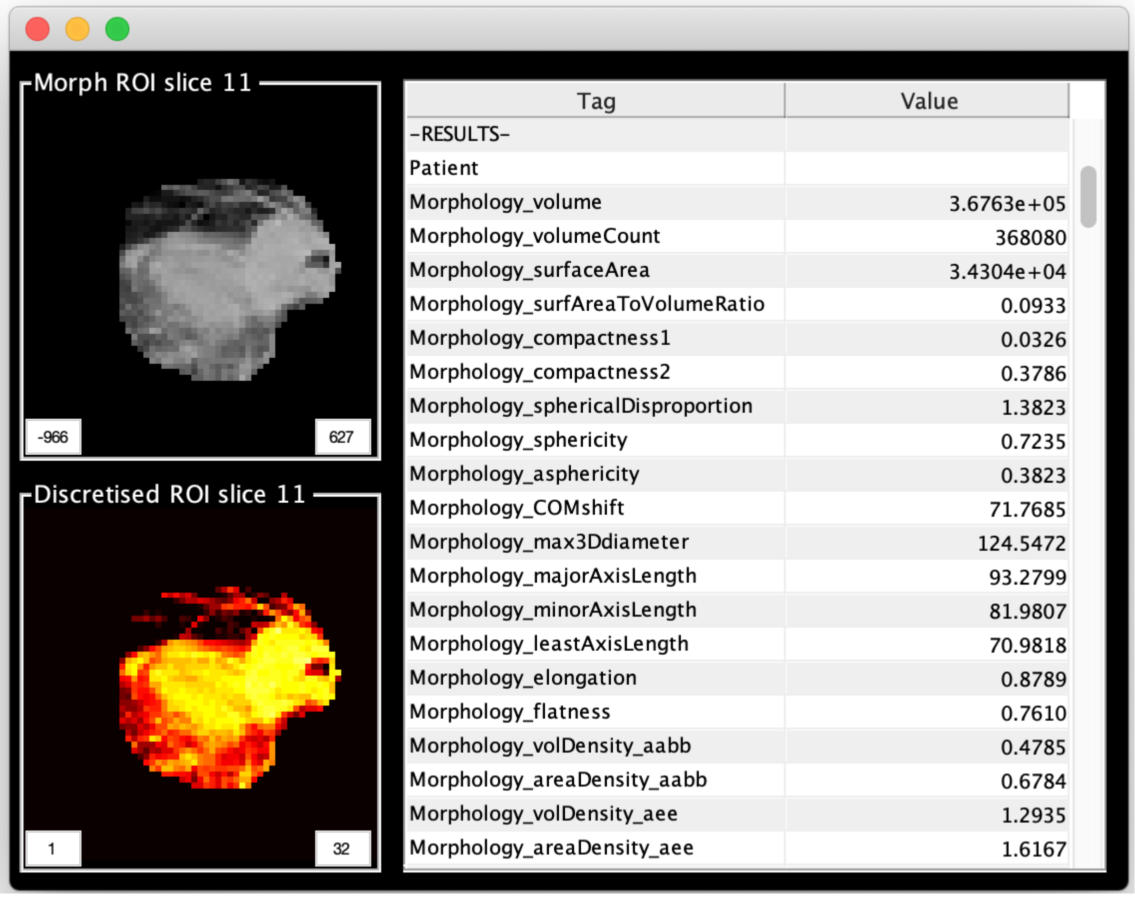
SPAARC, a recursive acronym standing for *Spaarc Pipeline for Automated Analysis and Radiomics Computing*, comprises a suite of tools that implements a wide range of solutions for processing imaging and radiotherapy data. It is a collection of applications developed within our research group CIDA (*Cancer Imaging and Data Analytics*) at Cardiff University. For this project, the author independently developed the entire radiomic feature extraction pipeline for SPAARC. This software was built and utilised for Cardiff's benchmark contributions in the IBSI [84, 90], detailed extensively in Chapter 3. A number of published radiomics studies from our research group have used the SPAARC radiomics package [76, 85, 88, 89].

The SPAARC radiomics package is built with the Matlab programming language [52]. All key functions written for feature extraction have been implemented within the SPAARC pipeline from scratch, introduced throughout this Chapter. To handle the import of DICOM files into Matlab, the author opted to use existing functionality from the *Computational Environment for Radiotherapy Research* (CERR) open source software [91]. This is discussed in more detail in Section 2.3.1.2. Code Box 2.1 illustrates a basic example script one might run to extract features using SPAARC. This is introduced now as it will be referenced in the following sections where the feature extraction process is examined through each step of the radiomics image processing scheme that became the standard recommended by the IBSI [90]. The different settings one has to consider and set for radiomic feature extraction are contained within this script.

Code Box 2.1: Basic SPAARC feature extraction run script with key option settings shown.

```
% SPAARC Feature Extraction Example Script
% -------------------------------------------------------------------------
% @uthor: PWhybra
% -------------------------------------------------------------------------
pathname   = '.\Data';
filelist   = {'.'}; % Cell array of  data

% Build inputP With settings
inputP.metric             = 'RadiomicAnalysis';
inputP.prefix             = 'Example_';
inputP.savedir            = './Results';
inputP.options.saveROIs   = true;

% Name of ROI to analyse
% -------------------------------------------------------------------------
inputP.struct_name        = 'GTV';

% Feature Families
% -------------------------------------------------------------------------
inputP.metric_types       = {'Morphology','Statistical','3D','2D'};

% Interpolation settings
% -------------------------------------------------------------------------
%If inputP.options has a field 'Interp', code will run interpolation
inputP.options.Interp.Method      = 'linear'; % linear or spline
inputP.options.Interp.newVoxelDcm = [0.2 0.2 0.2];
inputP.options.Interp.rounding    = 1; % use 0 for PET SUV

% Re-segmentation options. Set to [-inf inf] if not utlised
% -------------------------------------------------------------------------
inputP.options.scanRange        = [-500 500];   % range resegmentation
% inputP.options.intensityOutlierFiltering = 1; % intensity outlier option

% Select discretisation method  (Choose From FBN or FBS)
% -------------------------------------------------------------------------
% Fixed Bin Number
% ---------------
% inputP.options.rescaleMethod = 'FixedBinNumber';
% inputP.options.NumLevels     = 64; %  number of intensity levels

% Fixed Bin Size
% ---------------
inputP.options.rescaleMethod = 'FixedBinSize';
inputP.options.BinWidth      = 10;

% TA Settings (use defaults unless advanced user)
% -------------------------------------------------------------------------
inputP = defaultRadiomicsSettings(inputP);

% Run Analysis
% -------------------------------------------------------------------------
run_SPAARC_Radiomics(pathname,filelist,inputP);
```

The SPAARC radiomics package has many visualisation options that can occur alongside extraction, including quick montages for regions of interest and 3D morphological assessment. Although radiomics requires batch processing, there is also value in an ability to quickly evaluate and visualise any individual regions of interest. An example of an implemented summary GUI for a single region of interest is shown in Figure 2.1.



Tag	Value
-RESULTS-	
Patient	
Morphology_volume	3.6763e+05
Morphology_volumeCount	368080
Morphology_surfaceArea	3.4304e+04
Morphology_surfAreaToVolumeRatio	0.0933
Morphology_compactness1	0.0326
Morphology_compactness2	0.3786
Morphology_sphericalDisproportion	1.3823
Morphology_sphericity	0.7235
Morphology_asphericity	0.3823
Morphology_COMshift	71.7685
Morphology_max3Ddiameter	124.5472
Morphology_majorAxisLength	93.2799
Morphology_minorAxisLength	81.9807
Morphology_leastAxisLength	70.9818
Morphology_elongation	0.8789
Morphology_flatness	0.7610
Morphology_volDensity_aabb	0.4785
Morphology_areaDensity_aabb	0.6784
Morphology_volDensity_aee	1.2935
Morphology_areaDensity_aee	1.6167

Figure 2.1: Example of SPAARC (*Spaarc Pipeline for Automated Analysis and Radiomics Computing*) summary box and GUI which can be used to visualise the *region of interest* (ROI) and displays radiomic results (and extraction settings). The two scrollable panels visualise both the morphological ROI and the discretised ROI simultaneously (see Section 2.3.6 for more information) alongside extraction settings and feature results.

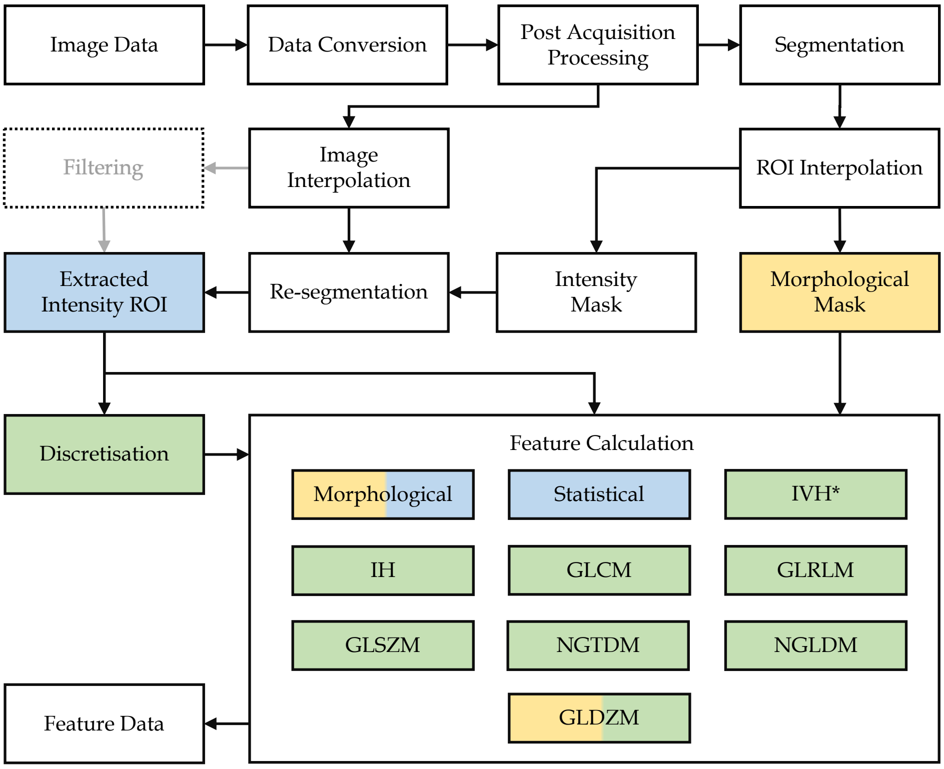


Figure 2.2: Image processing scheme to extract radiomic features based on Zwanenburg *et. al.* [90]. For feature calculation, the families have been colour coded according to which pre-processing step they utilise. Each step is expanded on in the following sections. The pipeline is as follows: Image Data 2.3.1, Data Conversion & Post Acquisition Processing, Segmentation 2.3.2, Interpolation 2.3.3, Splitting of ROI and Re-segmentation 2.3.4, Extraction of Intensity ROI volume 2.3.5 , Intensity Discretisation 2.3.6, and Feature Calculation 2.4. This chapter does not discuss filtering options, which is a significant additional processing step that is the focus of Chapter 5.

2.3 Radiomics Image Processing Scheme

A radiomics *image processing scheme* is the set of computational steps to get from acquired images to extracted features. The precise step number can of course vary depending on the image modality and desired analysis. For example, a “PET specific” post acquisition processing step might be conversion of the scan intensities into SUV. However, radiomics standardisation requires a clear and rigorous methodology. The IBSI aimed to define, developed and refine a generalisable, overarching processing scheme for software implementations to follow that expands on previous radiomics literature. Each of the steps within the scheme in Figure 2.2 are described in the following sections within this chapter, alongside information on the SPAARC implementation.

2.3.1 Imaging Data

Intuitively, the processing scheme first begins with image data. Before looking at the specific format of medical imaging it is useful to define what images are generally, along with the notation used in this thesis. All digital images are comprised of a discrete grid of elements known as pixels (picture elements) or voxels (volume elements) in 2D and 3D respectively. Each element in the grid is assigned an intensity value and it is the arrangement of these intensities that creates an image. When the intensity at each position is a single number (or channel) as is the case with most medical imaging, the intensity values are often referred to as *grey levels*, and the image visualised in *grey scale*, on a gradient from the lowest intensity as black to the highest as white.

Medical scanners usually acquire 3D volumetric data via stacking of image slices taken at regular intervals. The imaging is acquired at a particular resolution such that each voxel in the image has the same dimensions. As such, each voxel has a corresponding coordinate position, $\boldsymbol{k}$ and the coordinate grid of the image is regularly spaced starting from some defined origin. A single channel, grey scale image can be described with a discrete function $I[\boldsymbol{k}]$, where

$$\text{Intensity value} = I[\boldsymbol{k}]. \tag{2.1}$$

The discrete spatial position $\boldsymbol{k} = (k_1, k_2)$ in 2D and $\boldsymbol{k} = (k_1, k_2, k_3)$ in 3D respectively locates the corresponding intensity output (or grey value) in the image function. An image $\mathbf{I}$ can also be represented in matrix notation. For example, a 2D image of size $d \times n$ would have the form:

$$\mathbf{I} = \begin{bmatrix} I_{11} & I_{12} & I_{13} & \dots & I_{1n} \\ I_{21} & I_{22} & I_{23} & \dots & I_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ I_{d1} & I_{d2} & I_{d3} & \dots & I_{dn} \end{bmatrix} \tag{2.2}$$

where I_{rc} or $I(r, c)$ gives the grey level value for matrix indicies r and c (and s for 3D imaging). In 3D the indices are used to give the intensity value at a particular row, column and slice. When looping through voxels for analysis, it is often more straight forward to use the matrix indices over the physical spatial coordinates associated with a medical image, and feature extraction algorithms utilise the indices to navigate the image matrix in this regard. Traditional radiomic feature extraction, in essence, is applying computational operations to a matrix and arriving at a scalar representation for that matrix. Finally, for more compact index notation it can be useful to collapse an image containing a number (N) of voxels into elements in a 1D vector, i.e.

$$\mathbf{I}_k = (I_1, I_2, ..., I_N). \tag{2.3}$$

2.3.1.1 DICOM Format

Digital images are store in a variety of different file types. One of the most commonly used file formats for medical imaging is known as DICOM (*Digital Imaging and Communications in Medicine*) [92]. Most commercial scanners utilise the DICOM standard, in part due to its ability to manage the slice-by-slice image acquisition, as well as its wide spread integration with PACS (*Picture Archiving and Communications Systems*), which are the infrastructure for storing and viewing a variety of patient imaging in clinical settings. The official documentation for the DICOM standard is vast and complex [92]. What is important to note here is that the structure of DICOM files are separated into a header and subsequent image data. The wealth of information stored in the DICOM header is its defining feature. Each file can be configured to contain patient information (name, hospital ID, date of birth etc.) and acquisition protocol (equipment, series, study, patient coordinates, scan units etc.) alongside hundreds of other potential attributes. The header entries are stored and retrieved via use of an 8 character *tag*.

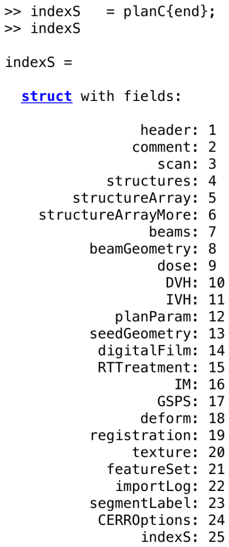
The DICOM standard library documents which tag is associated with a particular attribute [92]. As an example, the patient ID is stored next to the following tag: (0010,0020). Each modality has a minimum number of tags that should be reported to meet the required standard. DICOM RTSTRUCT files are a special sub-type of DICOM file that contain the coordinate points of contours used to define regions within images. Radiomic analysis of tumours requires these files to define the area of the image to analyse. Utilising the Image Position and Image Orientation tags, software can orient the coordinate points to the associated image. Radiomics software must first extract and store the relevant information from the DICOM tags during data import to be used for feature extraction.

2.3.1.2 SPAARC Data import

To handle the import of DICOM files, SPAARC uses functionality from the *Computational Environment for Radiotherapy Research* (CERR) software [91]. CERR is free and open source, developed in MATLAB for viewing and sharing research results. CERR functionality converts patient DICOM files into a data format known as a *planC*. The *planC* is a Matlab cell-array that has unpackaged the DICOM contents into elements highlighted in Figure 2.3, with access to the scan arrays, structures (i.e. contours) and scan information from the DICOM tags. There is one *planC* for each unique patient, which can contain many different studies, scans and associated structures.

2.3.2 ROI Segmentation

As outlined in Section 1.5.2, precise delineation of the tumour determines the boundaries of analysis. In oncology, manual or semi-automatic contours of the primary tumour are usually the main ROIs that are analysed for radiomics models.



```
>> indexS   = planC{end};
>> indexS

indexS =

  struct with fields:

               header: 1
              comment: 2
                 scan: 3
           structures: 4
       structureArray: 5
   structureArrayMore: 6
                beams: 7
        beamGeometry: 8
                 dose: 9
                  DVH: 10
                  IVH: 11
            planParam: 12
        seedGeometry: 13
          digitalFilm: 14
          RTTreatment: 15
                   IM: 16
                 GSPS: 17
               deform: 18
         registration: 19
              texture: 20
           featureSet: 21
            importLog: 22
         segmentLabel: 23
           CERROptions: 24
               indexS: 25
```

Figure 2.3: The CERR [91] planC structure.

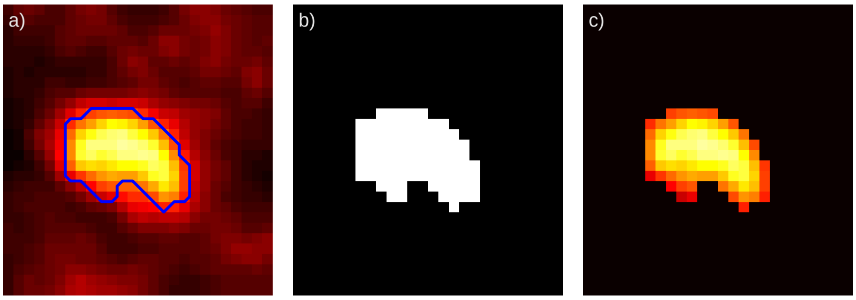


Figure 2.4: ROI (*region of interest*) segmentation example of oesophageal tumour from a PET image. a) Image with delineated tumour boundary; b) visualisation of the corresponding binary ROI mask; c); extracted tumour region.

Segmentation, computationally, becomes a voxel-wise assessment where each individual voxel is assigned as either being inside or outside the region of interest. This results in a binary segmentation *mask* the same dimensions as the original image, so for an image **I** there is an associated ROI **R** of the same size where each element R_k is either 1 or 0 [90]. i.e.

$$R_k = \begin{cases} 1 & \text{inside the ROI} \\ 0 & \text{outside} \end{cases} \tag{2.4}$$

Figure 2.4 provides a visualisation of this process with 3 images; 2.4 a) is a cropped 2D slice of an OC PET image containing a segmented tumour, 2.4 b) is a visualisation of the segmentation binary ROI mask, and 2.4 c) is the resulting extracted region of the image to be analysed.

Some medical imaging formats store segmented structures as ROI masks explicitly (e.g. NifTi [93]) and can be simply loaded alongside the accompanying scan, where as others (e.g. DICOM RTSTRUCT) store a set of points defining closed polygons for each image slice. These sets of points can be converted to ROI masks by determining if the centre of each voxel lies inside the closed polygon(s) or not with methods such as the *crossing number* algorithm described in the IBSI documentation [90] .

2.3.2.1 SPAARC ROI Mask Retrieval

As mentioned in Section 2.3.1.2, the SPAARC software utilises the CERR *planC* data format. During import, segmentations are stored in the "structures" cell position (see Figure 2.3) and converted to a ROI binary mask with CERR functionality. Using a structure's name, its position in the *planC* can be identified with `getStructNum` function and the binary mask retrieved with the `getUniformStr` function. If there are multiple structures in a *planC*, each should have a unique name. For SPAARC radiomics batch processing, the structures need to have a consistent naming convention (e.g. GTV *Gross Tumour Volume*). Notably, there are global efforts to have a consistent nomenclature in radiation oncology [94] which should be followed.

2.3.3 Image & ROI Interpolation

As introduced in Section 2.3.1, medical imaging is acquired at a particular resolution, and as such, each voxel centre sits on a grid at that resolution within a coordinate system. With image interpolation, an algorithm generates new voxel values at a new desired grid spacing, artificially changing the resolution of the image. Radiomics studies in the literature utilise scan interpolation for several reasons, including: the comparison and analysis of different dataset with different acquisition protocols [95]; the resampling of multimodal imaging to the same dimensions (such as PET/CT) [71]; and to obtain isotropic voxel dimensions for 3D feature extraction [96].

Isotropic voxels - i.e. square voxels with equal dimensions $(\Delta x = \Delta y = \Delta z)$ - are recommended for 3D feature extraction to remove directional bias and maintain rotational invariance [90, 96]. In routine clinical image acquisition, the voxels are often anisotropic, where the thickness between axial slices is large compared to the in-plane resolution $((\Delta z > (\Delta x, \Delta y))$, e.g. a typical CT scan has a sub millimetre axial resolution compared to a 3-5mm slice thickness. As interpolation is needed to obtain isotropic voxels, it is a critical step to map out in the radiomics image processing scheme.

For some medical imaging one might consider rounding the image values after interpolation. For example, CT Hounsfield units are integer (Section 1.3.1), so after interpolation, scans can be rounded back to nearest integer values to remain "meaningful". When interpolating an image where the ROI has been defined (Section 2.3.2), one must also interpolate the associated ROI mask (R_k) to the same coordinate system. Depending on the algorithm (e.g. tri-linear, spline [97]), this may result in a non-binary mask, and as such, it must be re-binarised using a set threshold δ. An intuitive threshold is $\delta = 0.5$, where $R_k \geq \delta$ is set to 1, and $R_k < \delta$ is set to 0. The effect of interpolation on radiomic features is discussed and explored in great detail in Section 3.3.2 and Chapter 4. Here, just the methodology is described.

2.3.3.1 Defining a New Grid

The first step to interpolation is to generate a new grid with a new spacing. There are two main ways to map a new grid to an old one; *align the grid origins*, or *align the grid centres* [90]. Figure 2.5 demonstrates these two methods visually. In this 2D toy example, the origin of the original grid is at (0,0) and the original spacing is 3 units between each axis point; New grids are created at a spacing of 2 units using the two methods.

2.3.3.2 Align Grid Origins

To create a new grid with aligned origins, one takes the origin point of the old coordinate system and generates the new grid by systematically adding at increments of the desired new spacing, for the 3 axis directions separately. For a given axis direction, let x_o be the origin point of the old grid axis and s_n be the new desired spacing. Using the grid aligned method, the new grid axis would be at positions of $(x_o, x_o + s_n, x_o + 2s_n, x_o + 3s_n, ...)$, stopping when a new grid axis point surpasses the end point of the original axis, as shown in Figure 2.5 b).

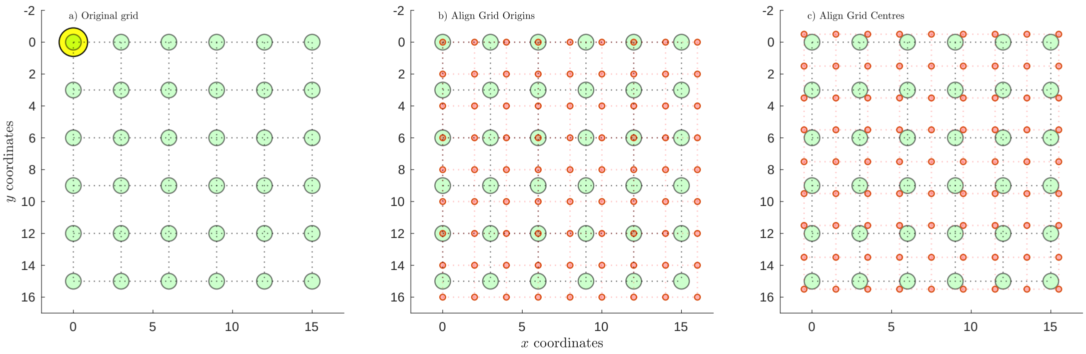


Figure 2.5: 2D example of interpolation grid generation; a) is the original grid with a separation of 3 units in both axis direction and origin of (0,0); b) is the *align grid origins* method for a desired grid spacing of 2 units and c), *align grid centres* method for a desired grid spacing of 2 units.

2.3.3.3 Align Grid Centres

The advantage of the *align grid centres* method is that it does not matter which corner of the image is defined as the origin. This negates any possible software differences in determining image origin, which is advantageous to reduce discrepancy (see Section 3.3.2). For a given axis, the new centre aligned grid coordinates can be calculated in the following way [90]: let n_o be the number of points on the old grid axis and s_o be the spacing of the old grid. The number of points on the new grid axis n_n is calculated via

$$n_n = \left\lceil \frac{n_o s_o}{s_n} \right\rceil, \tag{2.5}$$

where $\lceil , \rceil$ is a ceiling bracket rounding up to the nearest integer. To be centre aligned, the starting position of the new grid axis x_n, can be calculated via

$$x_n = x_o + \frac{s_o(n_o - 1) - s_n(n_n - 1)}{2}. \tag{2.6}$$

The new grid axis positions would then be at $(x_n, x_n + s_n, x_n + 2s_n, x_n + 3s_n, ..., x_n + (n_n - 1)s_n)$. A demonstration of a centre aligned grid system is shown in Figure 2.5 generated using these algorithms as implemented in SPAARC.

2.3.3.4 Interpolation Methods

There are a variety of interpolation algorithms one can use to generate new voxel values at a newly constructed grid spacing. In Matlab 2018b, key options available include; *nearest neighbour*, *linear*, and *spline* [97]. These approaches have an associated dimensional implementation; for example, linear interpolation in 2D is referred to as bi-linear, and in 3D referred to as tri-linear. Although for medical imaging the 2D and 3D versions are usually utilised, the example in Figure 2.6 demonstrates these methods using a 1D example for easier interpretability.

With the *nearest neighbour* method, each point on the new grid is given the same intensity value as the closest neighbour on the old grid. For *linear* interpolation, new intensity values are assigned

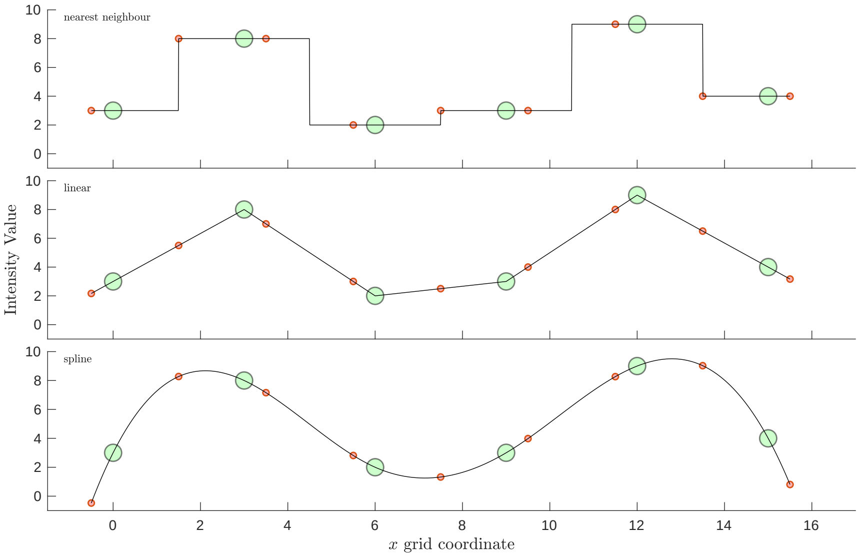


Figure 2.6: A 1D example of *nearest neighbour*, *linear*, and *spline* interpolation methods respectively. The axis positions were taken from x grid coordinates of the example in Figure 2.5 c). The intensity values were chosen for demonstration purposes for each point on the original grid spacing (green), and calculated via the designated algorithm for the new grid spacing (red). This image illustrates that different interpolation methods lead to different intensity values, which may affect radiomic analysis.

between points on the old grid by sampling along a linear regression connecting those points; 1D linear interpolation uses the 2 closest points, 2D bi-linear uses 4 points, and 3D tri-linear uses 8 points. For cubic *spline* interpolation, a piecewise set of third order polynomials smoothly connect each point on the original grid, and intensity values for the new grid are sampled from these fits [98].

2.3.3.5 SPAARC Interpolation

Within the run script of the SPAARC radiomic package, one can activate the option to interpolate imaging by defining the algorithm to use and the desired voxel dimensions in centimetres, as highlighted in Code Box 2.1, lines 25-26. If 3 voxel dimensions are specified a full 3D interpolation is performed. If 2 voxel dimensions are specified, image interpolation is performed slice by slice. One can also set the option to round the image back to integer values prior to feature extraction, as shown on line 27 in Code Box 2.1. Both *align grid origins* and *align grid centres* grid generation methods are implemented, though SPAARC defaults to the latter, which became the recommendation discussed more in 3.3.2 of the next chapter. SPAARC utilises Matlab in-built functionality for the interpolation algorithms [52, 97]. Both the *linear* or *spline* methods are recommended options. The ROI mask is interpolated alongside the image to the same dimensions. Though changeable within default settings, linear interpolation is always used for the mask and rounded with a threshold $\delta = 0.5$.

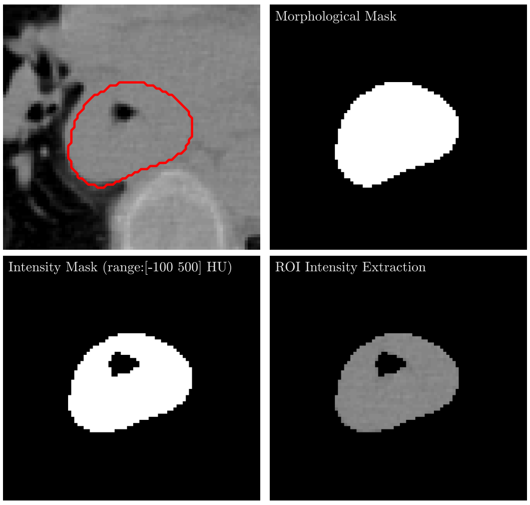


Figure 2.7: An example of the morphological & Intensity mask after re-segmentation. Top left: CT image with segmentation of oesophageal cancer (cropped). Top right: the corresponding morphological mask. Bottom left: the intensity mask after range resegmentation between [-100 500] HU. Bottom right: the extracted intensity volume. Although inside the original segmentation, voxels corresponding to air will not be included in the texture analysis.

2.3.4 ROI Re-segmentation: Intensity & Morphological Masks

After interpolation of the scan and ROI mask, the next step in the *image processing scheme* is separation and re-segmentation of the ROI. Here, if desired, the ROI can be further refined based on the corresponding intensity values in the image; i.e. the removal of voxels with intensities either outside a certain range, or as a result of being considered "outliers". This is utilised to fine tune analysis to a particular tissue, and often referred to as thresholding in the literature (e.g. [99]). A common reason for thresholding prior to feature extraction is to remove air within a segmentation of CT imaging, as highlighted in Figure 2.7.

To manage resegmentation, the IBSI processing scheme evolved from this point to define two distinct masks; a *morphological* mask and an *intensity* mask [90]. The *morphological* mask remains the same segmentation defined by either an expert or semi-automatically, where as the *intensity* mask is re-segmented (see Figure 2.7) . Intuitively, with no resegmentation option set, the intensity and morphological masks are identical. The Morphological and GLDZM feature families make use of both segmentation masks, as highlighted in Figure 2.2.

2.3.4.1 Re-segmentation methods

Two re-segmentation approaches were evaluated in the IBSI; *range* re-segmentation and *intensity outlier filtering* [90]. With *range* re-segmentation, one simply removes from the ROI intensity mask voxels that fall outside a certain range. Figure 2.7 highlights this with a re-segmentation range of [-200 500] HU, effectively removing voxels corresponding to air from the intensity mask. Here -200 and 500 are the lower and upper limit respectively, and voxels with those values are included within the range. When specifying the range for re-segmentation, it is a valid approach to set only an upper or lower value, e.g. $[-200, \infty)$ HU. Range re-segmentation is reserved for known scan units such as HU or SUV. For some imaging, such as MRI data, scan intensity values are arbitrary and a range re-segmentation is thus not appropriate.

As implied, using the *intensity outlier filtering*, voxels defined as outliers are removed from the intensity mask. One way to do this is to calculate the mean μ and standard deviation σ of the intensities in the ROI and set a lower a and upper b limit such that $[a, b] = [\mu - s_n\sigma, \mu + s_n\sigma]$ where s_n is the selected number of standard deviations from the mean. Vallières *et. al.* [100] first suggested this approach with $s_n = 3$, i.e. $[\mu - 3\sigma, \mu + 3\sigma]$.

2.3.4.2 SPAARC Re-segmentation

Both range and intensity outlier options have been implemented in SPAARC. As highlighted on line 31-32 in Code Box 2.1, the example has a range re-segmentation of [-500, 500] and will also do intensity outlier filtering. The default σ for intensity outlier filtering is set to 3 as utilised by Vallières *et. al.* [100]. Evidently, either one or both re-segmentation methods can be applied. If both are selected, they are first calculated independently and the intersection of the resulting masks used for the final intensity mask.

2.3.5 Extraction of ROI Intensity Volume

The intensity mask is used to isolate the relevant image voxels intensities for analysis. From the full image, only the intensities that align with the mask are included in the extracted intensity volume, as shown in the example in Figure 2.7. Using the intensity mask, all other image voxels outside the mask are replaced with a value to signal they should be ignored computationally. Most often, a *NaN* (Not a Number) value can achieve this purpose. Matlab uses a special value *NaN* to represent a number that is neither real or complex [101]. For an image $\mathbf{I}$, with an intensity mask of equal size $\mathbf{R}$, each element V_k in the extracted intensity volume $\mathbf{V}$ is thus:

$$V_k = \begin{cases} I_k & R_k = 1 \\ NaN & R_k = 0 \end{cases} \tag{2.7}$$

2.3.6 Discretisation

To populate texture matrices, voxel values need to be discrete integers as the intensity value is used as an index for the matrix. Any imaging without integer voxel values (e.g. PET with SUV units) must be converted via the process of discretisation for subsequent texture analysis. Discretisation is simply intensity binning. Even for imaging where the voxels are already integer (e.g. CT Hounsfield units), discretisation helps avoid large, sparsely populated texture matrices, thus making feature calculation more manageable and less memory intensive. Discretisation can also have the effect of suppressing noise within the image as they become coarser [90]. Following Figure 2.2, discretisation is conducted on the extracted intensity volume $\mathbf{V}$.

The most common method of image discretisation in radiomics is either a fixed bin size (FBS) or fixed bin number (FBN), with some studies examining the effect of these two methods on feature extraction in particular [102]. As such, the IBSI set out to standardise the methodology

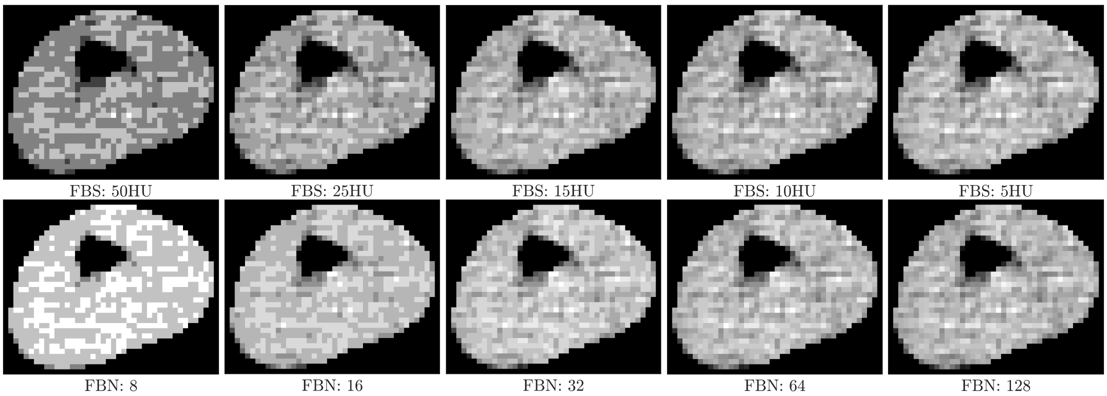


Figure 2.8: Demonstration of discretisation methods using the extracted intensity volume from Figure 2.7. **Top row**: *fixed bin size* method. As you decrease the bin size, the resulting image gets less coarse. **Bottom row**: *fixed bin number* method. As you increase the bin number, the resulting image gets less coarse.

for these two methods explicitly. For IBSI definitions, the lowest voxel intensity value possible after discretisation is 1, as this avoids division of 0 errors that can occur for some texture feature calculations. Figure 2.8 highlights discretisation of the ROI intensity mask extracted in Figure 2.7 using both the FBS and FBN methods for a variety of bin sizes and numbers respectively.

2.3.6.1 Fixed Bin Number

With FBN the intensities are discretised to a set number of bins. As an example, for a 32 bin discretisation ($N_g = 32$), the lowest intensity value in the ROI is set to 1 and the highest value 32, with every value in between scaled to an integer value from 1 to 32 via Eq. 2.8. Let $\mathbf{V}_{gl}$ be the extracted intensity volume with original grey level values, and $\mathbf{V}_d$ be the discretised result. To bin into N_g discrete values [90]:

$$V_{d,k} = \begin{cases} \left\lfloor N_g \frac{V_{gl,k} - V_{gl,\text{min}}}{V_{gl,\text{max}} - V_{gl,\text{min}}} \right\rfloor + 1 & V_{gl,k} < V_{gl,\text{max}} \\ N_g & V_{gl,k} = V_{gl,\text{max}} \end{cases} \tag{2.8}$$

Here voxel k with original intensity $V_{gl,k}$ is converted to the discrete intensity $V_{d,k}$ by scaling using the minimum $V_{gl,\text{min}}$ and maximum $V_{gl,\text{min}}$ of $\mathbf{V}_{gl}$. Values are rounded down to the nearest integer (where $\lfloor , \rfloor$ is a floor bracket) and 1 added so that the first bin has a value of 1.

By design with a FBN approach each resulting ROI has the same range, and are effectively normalised, thus losing any physical meaning of the units within the scan [90]. For example, consider a CT ROI tumour A which has a maximum intensity of 150 HU, and a CT ROI tumour B which has a maximum of 110 HU. After a 32 bin discretisation, both tumours would have a maximum intensity value of 32, and the knowledge of which tumour had the higher density value within it would be lost. FBN is useful to emphasize contrast within a given ROI, as well as for analysis of arbitrary units such as with MRI or with filtered imaging.

2.3.6.2 Fixed Bin Size

A FBS discretisation requires a minimum value, G_{min}, and a selected bin size w_b. Starting from G_{min}, each bin has a width equal to w_b, and discretised intensity values are assigned the value of the bin in which they fall. As an example using CT imaging, for $G_{\text{min}} = -1000\text{HU}$ and $w_b = 10$ HU, intensity values that range from [-1000,-991] are assigned a value of 1, from [-990,-981] a value of 2, from [-980,-971] a value of 3, and so on. For FBS discretisation [90]:

$$V_{d,k} = \left\lfloor \frac{V_{gl,k} - G_{min}}{w_b} \right\rfloor + 1 \tag{2.9}$$

Unlike FBN, a FBS discretisation maintains the relationship to the units of the original voxel value [90]. For bin values to correspond across analysis of different ROIs, the minimum value must be set globally, which can be done using the lowest value defined by the re-segmentation range. A notable study by Leijenaar *et*. *al*. [102] on PET imaging concluded that when possible, a FBS should be the preferred method due to maintaining relationship to scan units.

2.3.6.3 SPAARC Intensity Discretisation

Both FBN and FBS options for intensity discretisation are implemented in the SPAARC radiomics package based on equations 2.8 and 2.9 respectively. As highlighted on lines 38-44 in Code Box 2.1, once the discretisation method is set, one defines either the number of bins or the bin width. In this example the FBS approach is selected and the other commented out. The discretised ROI is also set to be saved for visualisation after extraction if required (line 12).

2.3.7 Feature Extraction

The last step in the image processing scheme is to convert the ROI into a series of single value quantitative measures via feature extraction. With these features, the goal of radiomics is to identify and characterise tumours in a clinically relevant manner, by hopefully deciphering the underlying tumour heterogeneity, and connecting it to outcomes. Radiomics is most effective with high-throughput of large cohorts [3]. As with most radiomics software, SPAARC was developed with efficient batch processing in mind. With each batch, the feature results are outputted in *.mat* and *.csv* formats. Features are calculated in groups or *families*, introduced comprehensively in the next section.

2.4 Feature Families

When considering feature extraction, related radiomic features belong to so called *families*. The families and subsequent features chosen for standardisation by the IBSI core group - and implemented by the author into SPAARC for this work - are extended and updated collections based on those proposed by Hatt *et*. *al*. [10] and Aerts *et*. *al*. [16], and are ubiquitous in the radiomics literature. Naturally, the work of standardisaiton cannot be exhaustive of every imaging feature

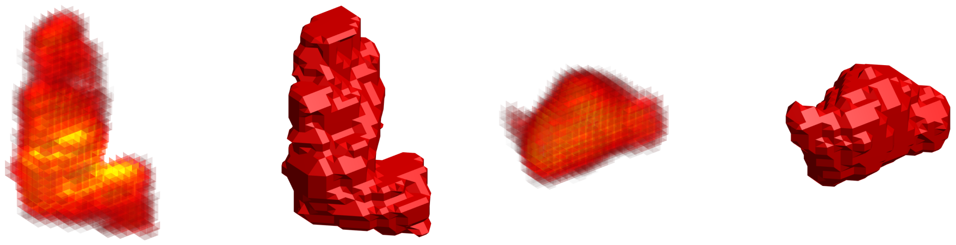
Table 2.1: Feature families and corresponding number of baseline features implemented in SPAARC.

Type	Family Name	Baseline Feat No.
Shape-based	Morphological 2.4.1	23
First-order	Intensity-Based Statistics 2.4.2	18
	Intensity Histogram 2.4.3	23
	Intensity-Volume Histogram 2.4.4	7
Texture	Grey Level Co-occurrence Matrix (GLCM) 2.4.6	25
	Grey Level Run Length Matrix (GLRLM) 2.4.7	16
	Grey Level Size Zone Matrix (GLSZM) 2.4.8	16
	Grey Level Distance Zone Matrix (GLDZM) 2.4.9	16
	Neighbourhood Grey Tone Difference Matrix (NGTDM) 2.4.10	5
	Neighbourhood Grey Level Dependence Matrix (NGLDM) 2.4.11	16

concivable, and the decision to exclude features such as those based on *fractals* [2], for example, is a clear limitation. Features previously utilised in published radiomics research were the main consideration for standardisation by the IBSI and those presented here represent a clear foundation of features to consider for radiomics modelling. The following sections introduce 10 of the feature families that were implemented in SPAARC, which essentially covers three main types of analysis as highlighted in Table 2.1: *Shape-based* analysis, *First-order* analysis, and *Texture* analysis (Second-order or higher). The individual features from each family can be found in Appendix B.

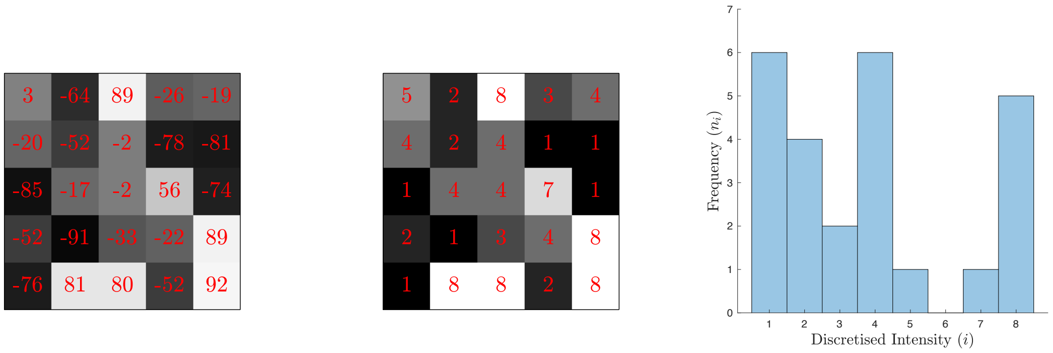
2.4.1 Morphological

Features within the morphological family relate to an ROIs volumetric shape. Intuitively, morphology features are defined for the whole 3D ROI and are not considered for 2D slices. For morphological feature calculation there are three ways a volume is represented [90]. 1) the volume is expressed as a set of voxels all with the same dimension; 2), the volume is expressed as a set of points corresponding to each voxel centre, and 3), the volume is expressed with a surface mesh. Figure 2.9 shows an example visualisation of differing tumour morphology from PET imaging of OC with both voxel and mesh representations.



(a) Example Tumour 1 **(b)** Example Tumour 2

Figure 2.9: Example of differences in tumour morphology of two oesophageal cancer tumours taken from PET imaging. The Tumour in **a)** and the Tumour in **b)** show clear deviations in shape, with the latter having a more spherical appearance. Both tumours have been visualised twice, the left side is a voxel representation of the tumour, coloured based on intensity value, and the right a surface mesh-based approach.



3	-64	89	-26	-19
-20	-52	-2	-78	-81
-85	-17	-2	56	-74
-52	-91	-33	-22	89
-76	81	80	-52	92

5	2	8	3	4
4	2	4	1	1
1	4	4	7	1
2	1	3	4	8
1	8	8	2	8

(a) Example Image

(b) Discretised Image

(c) Intensity Histogram

Figure 2.10: An illustration of intensity histogram calculation from a discretised image. a) Test image with intensity value of each voxel given in red, b) Image discretised using FBN = 8. c) Histogram visualisation **H** of frequency of bin values

2.4.1.1 SPAARC Morphology Feature Calculation

A combination of all 3 approaches mentioned above are used to calculate the morphological features in SPAARC. To generate a surface-mesh for the volume, SPAARC can be configured to use either an open-source implementation of the "Marching Cubes" algorithm [103], or a Matlab specific mesh implementation through the isosurface function [104]. See Appendix B for feature definitions.

2.4.2 Statistical (Intensity-Based)

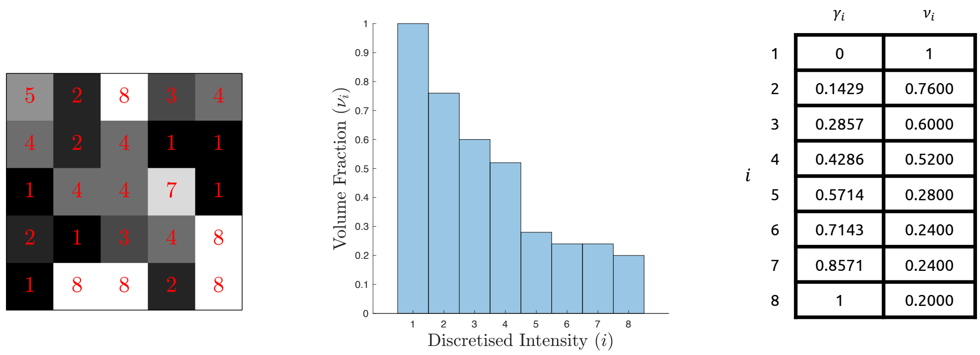
The intensity-based statistical family of features are derived using the ROI intensity volume with no discretisation, as highlighted in Figure 2.2. As such, the meaning is limited for scans not on a quantitative scale [90]. Definitions for the 18 statistical features implemented in SPAARC can be found in Appendix B. These features are extracted over the whole 3D volume by default.

2.4.3 Intensity Histogram

As highlighted in image processing scheme in Figure 2.2, intensity histogram features are extracted from the discretised intensity volume. They share many of the same mathematical definitions as the statistical family (Appendix B). A histogram gives the frequency of each discretised intensity value. Let n_i be the number of intensities with value i. The histogram is thus a set $\mathbf{H} = \{n_1, n_2, ..., n_{N_g}\}$, where N_g is the maximum intensity [90]. Figure 2.10 highlights this is what a standard histogram plot shows.

2.4.4 Intensity Volume Histogram

The IVH represents the intensity data as a binned cumulative distribution. For each discrete intensity i, it gives the fraction of the voxels, ν_i, within the ROI with *at least* a value of i. Intuitively, the lowest intensity has a fractional volume $\nu = 1$, as every voxel is at least the minimum intensity.



i	γ_i	ν_i
1	0	1
2	0.1429	0.7600
3	0.2857	0.6000
4	0.4286	0.5200
5	0.5714	0.2800
6	0.7143	0.2400
7	0.8571	0.2400
8	1	0.2000

(a) Discretised Image

(b) Visualisation of IVH ν_i

(c) IVH values

Figure 2.11: Intensity Volume Histogram Example. *a)* Image discretised to 8 Bins. *b)* Discretised intensity plotted against volume fraction ν_i, c) summary of IVH values

IVH requires discrete values, and the intensity values are often discretised to a larger number of bins (FBN), or with a smaller bin width (FBS) than the discretisation for texture analysis [90]. For CT, HU units are already discrete. For imaging with calibrated continuous intensities (e.g. PET SUV units), using a FBS method, the values corresponding to the bin centres are used in the IBSI definition, whereas for arbitrary units (e.g. MRI), a FBN is recommended (1000 Bins) [90]. For each discrete intensity, i, the fractional volume, ν_i, in a discretised image volume, $\mathbf{V}_k$, with N_v valid voxels, is calculated by considering each element V_k, via:

$$\nu_i = 1 - \frac{1}{N_v} \sum_{k=1}^{N_v} \begin{cases} 1 & V_k < i \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

For features derived from an IVH, the discretisation range $\mathbf{G}$ is utilised to work out the intensity fraction γ [90]. $\mathbf{G}$ can be based on the minimum and maximum of the individual ROI, or the re-segmentation range (see Section 2.3.4). The lowest i has an intensity fraction $\gamma_{\text{min}} = 0$, and the highest $\gamma_{\text{max}} = 1$, with the others scaled between, via [90]:

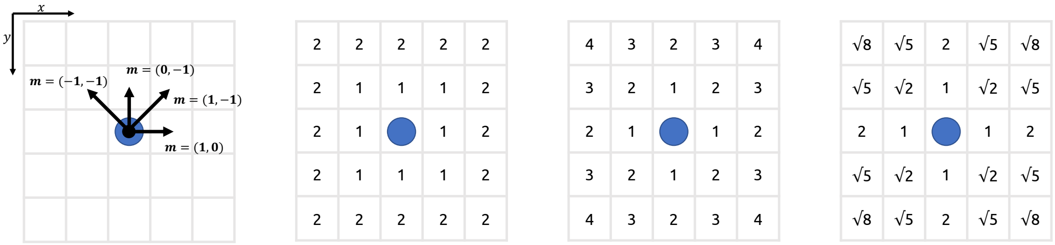
$$\gamma_i = \frac{i - \mathbf{G}_{\text{min}}}{\mathbf{G}_{\text{max}} - \mathbf{G}_{\text{min}}} \tag{2.11}$$

Figure 2.11 gives a visual example of the cumulative IVH and the corresponding IVH elements γ_i and ν_i. There are 5 features to extract from the IVH, listed in Appendix B.

2.4.5 Texture Families: Defining Distance and Direction Between Voxels

Texture analysis utilises the positional relationship between voxels in the image grid. There are several ways to quantify distance between points, namely: Euclidean distance, Manhattan distance, and Chebyshev distance. Examples of distance values from a central pixel with these approaches are shown in Figure 2.12.

Consider a central pixel in a 2D image. A pixel has 8 neighbours with a Chebyshev distance equal to 1, and requires 4 unique unit directional vectors to reach those neighbours (a negative unit vector goes in the opposite direction), as shown in Figure 2.12. This can be extended to a



2	2	2	2	2
2	1	1	1	2
2	1	●	1	2
2	1	1	1	2
2	2	2	2	2

4	3	2	3	4
3	2	1	2	3
2	1	●	1	2
3	2	1	2	3
4	3	2	3	4

√8	√5	2	√5	√8
√5	√2	1	√2	√5
2	1	●	1	2
√5	√2	1	√2	√5
√8	√5	2	√5	√8

(a) Unique directions **(b)** Chebyshev **(c)** Manhattan **(d)** Euclidean

Figure 2.12: 2D Example of navigating image direction and distance: a) unique vectors in 2D, b) Chebyshev distance: $max(|x_1 - x_2|, |y_1 - y_2|)$, c) Manhattan distance: $|x_1 - x_2| + |y_1 - y_2|$, d) Euclidean distance: $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

central voxel in a 3D image. i.e. a voxel has 26 neighbours with a Chebyshev distance equal to 1, and requires 13 unique directional vectors to reach those neighbours. Many texture analysis matrices utilise the 8-connected and 26-connected neighbourhood in 2D and 3D respectively.

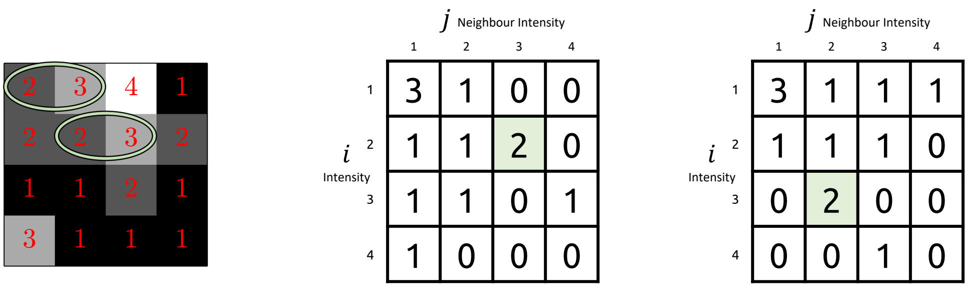
2.4.6 Grey Level Co-occurrence Matrix

The GLCM, first introduced back in 1973 by Haralick *et. al.* [56], represents a landmark in the conception of texture analysis of imaging, and common in many radiomics studies [2]. For a given direction and distance, a GLCM, **C**, collects the number of times a "voxel pair" has a certain intensity level relationship.

Here, **C** is an $N_g \times N_g$ square matrix where N_g is the highest voxel bin value. For example, for an ROI discretised with a FBN of 32, the GLCM would be 32 $\times$ 32. Each element (i,j) in **C** equates to the number of times a voxel of intensity i has a neighbour of intensity j, given a chosen direction vector **m**. This can be written as $\mathbf{C}_\mathbf{m}(i, j)$, or c_{ij}. The GLCM is populated iteratively by considering each voxel in turn.

A 2D example is provided for direction vectors $\mathbf{m}_+$=(1,0) and $\mathbf{m}_-$=(-1,0) in Figure 2.13. As a demonstration, in Figure 2.13 the number of times a voxel of intensity 2 has a neighbour to the right with an intensity 3, is stored in $\mathbf{C}_{(\mathbf{1},\mathbf{0})}(2, 3) = 2$.

When generating GLCMs for two opposite directions - as in Figure 2.13 - the matrices are trans-



2	3	4	1
2	2	3	2
1	1	2	1
3	1	1	1

	1	2	3	4
1	3	1	0	0
2	1	1	2	0
3	1	1	0	1
4	1	0	0	0

	1	2	3	4
1	3	1	1	1
2	1	1	1	0
3	0	2	0	0
4	0	0	1	0

(a) Grey level Image **(b)** $\mathbf{C}_{(\mathbf{1},\mathbf{0})}$ (i.e. →) **(c)** $\mathbf{C}_{(-\mathbf{1},\mathbf{0})}$ (i.e. ←)

Figure 2.13: GLCM Example. a) Grey level image with intensity values in red. Highlighted in green are voxels of intensity 2 with a neighbour of intensity 3. b) The resulting GLCM for direction $\mathbf{m}_+$=(1,0) . c) The resulting GLCM for direction $\mathbf{m}_-$=(-1,0).

posed versions of one another, hence the appearance of symmetric GLCMs. To save time computationally, one can compute the unique directions (see Section 2.4.5) and apply a transpose; A symmetric GLCM for a given unique direction $\mathbf{m}$ is thus $\mathbf{C_m} = \mathbf{C_{m+}} + \mathbf{C_{m-}} = \mathbf{C_{m+}} + \mathbf{C^T_{m+}}$ [90].

$\mathbf{C_m}$ is normalised to give a probability distribution $\mathbf{P_m}$ by dividing by a sum of the elements in $\mathbf{C_m}$. It is actually from this normalised probability matrix $\mathbf{P_m}$ that GLCM features are calculated. For the feature definitions each element in $\mathbf{P_m}$ is referred to as p_{ij}, which is the joint probability of a voxel of intensity i having a neighbour of intensity j (in direction m) . A total of 25 baseline GLCM features are implemented in SPAARC. The definitions are given in Appendix B.

2.4.6.1 Aggregation of features from directional texture matrices

Texture analysis utilises texture matrices to summarise voxel relationships within a ROI, and these matrices are computed either for a specific direction, or for 2D/3D neighbourhoods.

Take the GLCM introduced above. For 2D analysis (8-connectivity), 4 symmetric GLCMs are generated for each unique direction, for 3D (26-connectivity), 13 symmetric GLCMs are generated. The IBSI defined six potential ways to aggregate features calculated from directional texture matrices [90].

Directional texture matrices aggregation [90]

2D averaged: Feature extracted from each 2D directional matrix on each slice, then values averaged over each direction and slice.

2D merged: Feature extracted from one matrix after merging all 2D directional matrices on each slice, then the averaged over the slices

2.5D direction-merged: 2D directional matrices are merged per direction. Feature is extracted from each of the resulting merged directional matrices and averaged.

2.5D merged: Feature computed from single matrix after merging all 2D directional matrices

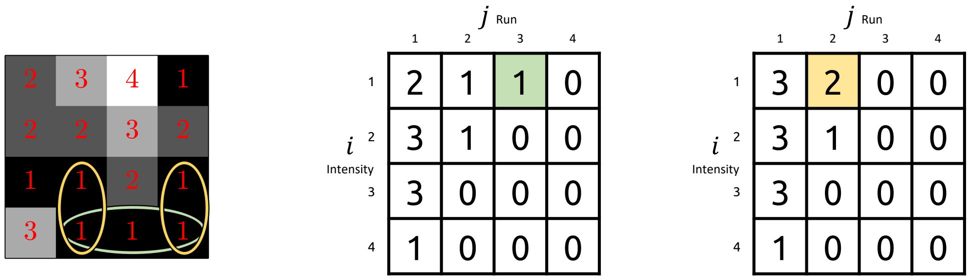
3D averaged: Feature extracted from each unique directional matrix, then values averaged.

3D merged: Feature extracted from one matrix after merging all 3D directional matrices

The difference in some of these aggregate approaches may appear subtle. However, different aggregation methods result in different final feature values, hence the necessity to define the aggregation approach thoroughly in the radiomics extraction pipeline. Each aggregation approach acts as a multiplier for the baseline number of features, so for the GLCM features, with 6 aggregation approaches and 25 features, there are 150 different features one could extract.

2.4.7 Grey Level Run Length Matrix

The GLRLM, first developed by Galloway *et. al.* [57] in 1975, collects within a matrix the size and number of "runs" of voxels with the same intensity value, for a given direction $\mathbf{m}$. Resulting features extracted from GLRLMs are popular texture measures. Again, as with all texture matrices, the ROI must first be discretised: i.e. it is calculated from the ROI intensity volume as described



2	3	4	1
2	2	3	2
1	1	2	1
3	1	1	1

(a) Grey level Image

	1	2	3	4
1	2	1	1	0
2	3	1	0	0
3	3	0	0	0
4	1	0	0	0

(b) $\mathbf{R}_{(\mathbf{1},\mathbf{0})}$ (i.e. →)

	1	2	3	4
1	3	2	0	0
2	3	1	0	0
3	3	0	0	0
4	1	0	0	0

(c) $\mathbf{R}_{(\mathbf{0},-\mathbf{1})}$ (i.e. ↑)

Figure 2.14: GLRLM Example. a) Grey level image with intensity values in red. Highlighted in green are zones with run length 3 in direction m=(1,0). Highlighted in yellow are zones with run length 2 in direction m=(0,-1). b) GLRLM for direction m=(1,0). c) GLRLM for direction m=(0,-1)

in Section 2.3.6. By definition, GLRLMs calculated in opposing directions are equivalent, so only unique directions are considered.

For a GLRLM, $\mathbf{R}_\mathbf{m}$, element (i, j) of the matrix, is the count of voxels with intensity value i, that have a run length of j, in direction $\mathbf{m}$. As such, the GLRLM size is first initiated using the highest intensity value, N_g, and N_r, the maximum possible run length; i.e. N_g X N_r.

In the example in Figure 2.14, the image has a maximum intensity of 4 and the maximum possible run length of 4, hence the resulting 4 x 4 matrices. For direction $\mathbf{m} = (1, 0)$ there is one case of a run length of 3 voxels with intensity 1, highlighted in green, i.e. $\mathbf{R}_{(\mathbf{1},\mathbf{0})}(1, 3) = 1$. When moving in direction $\mathbf{m} = (0, -1)$, there are 2 cases of a run length of 2 voxels with intensity 1, highlighted in yellow, i.e. $\mathbf{R}_{(\mathbf{0},-\mathbf{1})}(1, 2) = 2$.

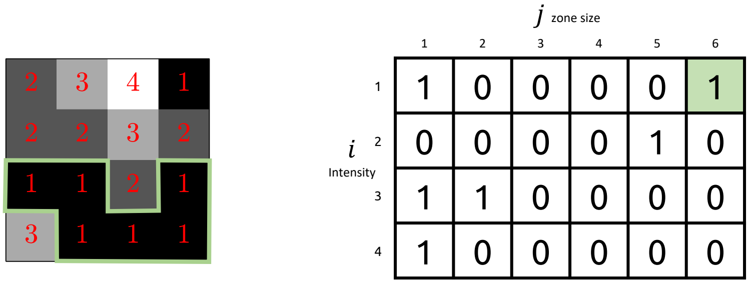
A total of 16 baseline features are calculated from GLRLMs [90], detailed in Appendix B. As GLRMs are directional matrices, features can be aggregated in the same way as described in Section 2.4.6.1.

2.4.8 Grey Level Size Zone Matrix

The GLSZM is a relatively new texture matrix, introduced by Thibault *et. al.* [58] in 2013. The GLSZM records the size of connected zones of voxels with the same intensity. In 2D, 8-connectivity, and in 3D, 26-connectivity is utilised [90], as described in Section 2.4.5. As with all texture matrices the image must first be discretised.

For a GLSZM, $\mathbf{S}$, element (i, j) is equal to the number of zones consisting of j voxels, with an intensity value of i. The matrix of size $N_g \times N_z$ is initialised using the highest intensity value N_g and the maximum possible zone size N_z. As one might imagine, for tumours with large numbers of voxels, the initialised GLSZM would also be large. As such, the column component N_z can be reduced to the maximum recorded zone size, after computation. As an example, in Figure 2.15, for the test image there is 1 zone of 6 voxels with intensity 1, highlighted in green, hence $\mathbf{S}(1, 6) = 1$. As N_z = 16, the matrix has been cropped after computation to the maximum zone size obsevered.

The GLSZM has 16 baseline features, as with the GLRLM. It is in a way an extension of the



2	3	4	1
2	2	3	2
1	1	2	1
3	1	1	1

(a) Grey level Image

	1	2	3	4	5	6
1	1	0	0	0	0	1
2	0	0	0	0	1	0
3	1	1	0	0	0	0
4	1	0	0	0	0	0

(b) S

Figure 2.15: GLSZM Example. a) Grey level image with intensity values in red. Highlighted in green is a zone of 6 voxels with intensity value of 1. b) The resulting GLSZM (**S**). Note the GLSZM has been cropped to the maximum recorded zone size.

GLRM, where GLRLMs consider only the number of 1D zones (runs) in a particular direction [58]. As such, the features derived from the GLSZM are based on the GLRM features and similar in design [90]. These can be found in Appendix B.

2.4.8.1 Aggregation of features from zone-based texture matrices

Zone-based texture matrices are calculated using connected voxels in 2D/3D, such as the GLSZM introduced above, rather than for a particular direction. The IBSI defined three ways to extract features from these matrices.

Zone-based texture matrices aggregation [90]

2D AVERAGED: Features calculated from textures matrices generated for each slice, then averaged.

2D MERGED: Texture matrices from each slice are merged, then feature extracted from resulting merged matrix.

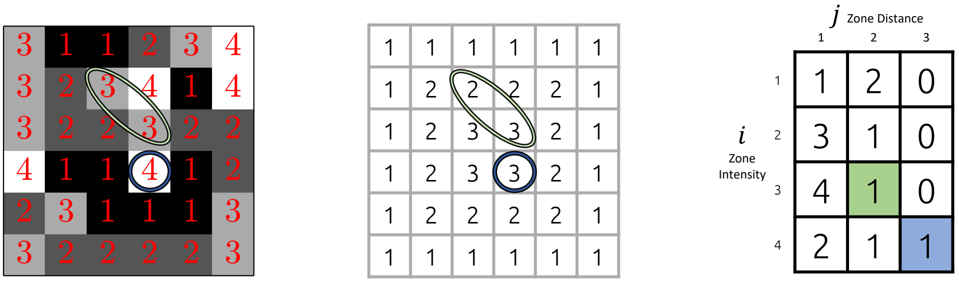
3D: Feature extracted from single texture matrix.

2.4.9 Grey Level Distance Zone Matrix

The GLDZM was first introduced by Thibault *et. al.* [59] in 2014. It is unique over GLCMs, GLRLMs and GLSZMs in that it incorporates a distance measurement to the boundary of the ROI.

For a GLDZM, **D**, element (i, j) yields the number of zones of equal intensity, i, which have a distance, j, to the ROI boarder. The IBSI defined an adapted, generalised version of GLDZMs that differs over Thibault *et. al.* [59], which considered only 2D imaging. As with GLSZMs, the zones have either 8-connectivity or 26-connectivity for 2D and 3D respectively. The matrix of size $N_g \times N_d$ is initialised using the highest intensity value, N_g, and the maximum possible distance, N_d.

GLDZMs require a distance map to determine the distance of each voxel to the ROI boundary [59]. The distance map is a matrix the same size as the ROI, but each position within the matrix corresponds to the minimum distance to the boundary at that point. There are several ways to



	1	2	3
1	1	2	0
2	3	1	0
3	4	1	0
4	2	1	1

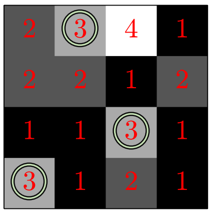
(a) Grey level Image **(b)** Distance Map **(c) D**

Figure 2.16: GLDZM Example. a) Grey level image with intensity values in red. Highlighted in green are zones with intensity 3 that are a minimum distance of 2 from the boarder. Highlighted in blue are zones with intensity 4 that are a minimum distance of 3 from the boarder. b) Corresponding distance map for image with zones still highlighted. c) The resulting GLDZM.

define distance to other voxels within an image (Section 2.4.5). The distance value needs to be an integer number so that it can be indexed in the GLDZM matrix, so either a Manhattan or Chebyshev representation is appropriate. In the IBSI [90] definition the minimum distance to the boarder is 1, and a *Manhattan* representation of distance is recommended.

Figure 2.16 contains a test image, the corresponding distance map, and the resulting GLDZM. Highlighted on 2.16a in green are zones with intensity 3 that have a corresponding minimum distance to the ROI boarder of 2, hence $\mathbf{D}(3,2) = 1$. Also highlighted in blue are zones of intensity 4 that have a minimum distance of 3 to the ROI boarder, hence $\mathbf{D}(4,3) = 1$.

As highlighted in the image processing scheme (Figure 2.2), to generate a GLDZM requires both the ROI morphological mask and the ROI intensity mask. The distance map is generated using the morphological mask, and the zones calculated using the intensity mask. A total of 16 baseline features can be extracted from GLDZMs, as listed in Appendix B. As GLDZMs are zone based matrices, features can be aggregated in the same way as described in Section 2.4.8.1



i Intensity	n_i	p_i	s_i
1	7	0.4375	6.9333
2	5	0.3125	1.2583
3	3	0.1875	4.4250
4	1	0.0625	2.2000

(a) Grey level Image **(b) TD**

Figure 2.17: NGTDM Example. a) Grey level image with intensity values in red. Highlighted in green are three voxels of intensity 3, with 8, 5 and 3 valid neighbours of $\delta = 1$. b) The resulting NGTDM with three columns, where n_i is number of voxels with intensity i that have at least 1 neighbour, p_i is the grey level probability, and s_i is the grey tone difference measure for intensity i.

2.4.10 Neighbourhood Grey Tone Difference Matrix

The NGTDM, first introduced by Amadasun *et. al.* [60] in 1989, is another popular and well utilised texture analysis technique in the radiomics literature. The core NGTDM gives the summation of the absolute differences between all grey levels in the image with intensity i, and the average value of their surrounding neighbourhood [60]. The IBSI definition [90] generalises to 3D and differs from Amadasun *et. al.* [60] in that the pixel or voxel does not need a complete neighbourhood to be included in the analysis. In other words, voxels at the edge of a ROI with incomplete neighbourhoods are still utilised.

Let **I** be a segmented image matrix with N_v voxels inside the ROI. For a valid voxel at position $\mathbf{k} = (k_x, k_y, k_z)$ in **I**, the average value of the surrounding neighbourhood ,A_k, with Chebyshev distance δ is:

$$A_k = \frac{1}{W_k} \sum_{m_z=-\delta}^{\delta} \sum_{m_y=-\delta}^{\delta} \sum_{m_x=-\delta}^{\delta} \mathbf{I}(k_x{+}m_x, k_y{+}m_y, k_z{+}m_z),$$
$$\text{where } (m_x, m_y, m_z) \neq (0,0,0), \text{ and } (k_x{+}m_x, k_y{+}m_y, k_z{+}m_z) \text{ is in ROI.} \tag{2.12}$$

Note that the neighbourhood excludes the central voxel, i.e. when $(m_x, m_y, m_z) \neq (0,0,0)$, and must be within the ROI to be included as a neighbour. Here, W_k is the number of valid neighbours of Chebyshev distance δ for the voxel at position **k**. The neighbourhood grey tone difference s_i for any intensity grey level i is then:

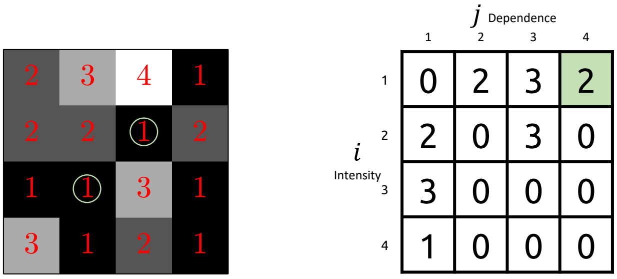
$$s_i = \sum_{k}^{N_v} \begin{cases} |i - A_k| & \mathbf{I}(\mathbf{k}) = i \text{ and } W_k \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

To solidify understanding, consider the calculation of s_3 for the simple image in Figure 2.17 for a neighbourhood of $\delta = 1$. There are three pixels with intensity 3, one voxel has a full neighbourhood, one has 5 neighbours, and one has 3 neighbours. Thus, the grey tone difference for intensity $i = 3$ is:

$$\begin{aligned} s_3 &= \left|3 - \frac{(2+1+2+1+1+2+1+1)}{8}\right| + \left|3 - \frac{(2+2+2+1+4)}{5}\right| + \left|3 - \frac{(1+1+1)}{3}\right| \\ &= 4.4250 \end{aligned}$$

In the IBSI definition [90], the NGTDM is an $N_g \times 3$ matrix where N_g is the maximum intensity in the discretised ROI. The first column n_i is the number of voxels with intensity i that have at least 1 neighbour. The second column p_i is the probability of selecting a voxel of intensity i with at least 1 neighbour, i.e. $p_i = n_i / \sum_{i=1}^{N_g} n_i$. The third column is the neighbouhood grey tone difference measure, s_i, described in the Eq. 2.13.

A total of 5 features can be extracted from a NGTDM, listed in Appendix B. NGTDM features follow the zone based approach for aggregation as described in Section 2.4.8.1.



2	3	4	1
2	2	1	2
1	1	3	1
3	1	2	1

(a) Grey level Image

i Intensity \ j Dependence	1	2	3	4
1	0	2	3	2
2	2	0	3	0
3	3	0	0	0
4	1	0	0	0

(b) M

Figure 2.18: NGLDM Example. a) Grey level image with intensity values in red. Highlighted in green are voxels with intensity 1 that have a neighbourhood dependence of 4 when $\alpha = 0$ and $\delta = 1$. b) The resulting NGLDM. Notes: The column of the NGLDM has been cropped to the maximum recorded dependence of 4. The calculation for dependence includes the central voxel.

2.4.11 Neighbourhood Grey Level Dependence Matrix

The NGLDM was first introduced by Sun *et*. *al*. [61] in 1983, as another way to assess texture coarseness, with an initial application classifying terrain types in satellite imaging. Similarly to the NGTDM mentioned in the previous section, the NGLDM analyses each valid voxel and their surrounding neighbourhood, but this time to assess neighbourhood *dependence*.

A voxel with intensity b that neighbours a central voxel with intensity a is considered *dependent* on that voxel if $|a - b| \leq \alpha$, where α is a non-negative integer coarseness parameter [61, 90]. Intuitively, for each voxel, the total dependence, j_k, is just the number of neighbours found to meet this criteria. As before, let $\mathbf{I}$ be a segmented image matrix. For a valid voxel at position $\mathbf{k} = (k_x, k_y, k_z)$, the total neighbourhood dependence is:

$$j_k = \sum_{m_z=-\delta}^{\delta} \sum_{m_y=-\delta}^{\delta} \sum_{m_x=-\delta}^{\delta} \begin{cases} 1 & \text{when } |\mathbf{I}(\mathbf{k}) - \mathbf{I}(\mathbf{k}+\mathbf{m})| \leq \alpha \text{ and in ROI.} \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

This time the definition includes the central voxel $(m_x, m_y, m_z) = (0, 0, 0)$, so that the minimum value of j_k is at least 1 [90].

For a NGLDM, M, element (i, j) is the number of voxels with intensity i that have a total neighbourhood dependence of j. The NGLDM can initialised with a size $N_g \times M_d$, where N_g is the maximum intensity number and M_d is the maximum possible neighbourhood dependency.

Consider the 2D example in Figure 2.18. The NGLDM is calculated for the image with a coarseness parameter of $\alpha = 0$ and $\delta = 1$. When $\alpha = 0$, a neighbouring voxel is dependant only if it is the same intensity level. Highlighted in green are two voxels with intensity 1, that have a total neighbourhood dependence of 4, hence $\mathbf{M}(1, 4) = 2$.

There are 16 baseline features for the NGLDM which can be found in Appendix B. As NGLDMs are based on local neighbourhoods, features can be aggregated with a zone based approach as described in Section 2.4.8.1.

2.5 Feature Robustness Measures

In any domain that utilises modelling, one needs robust inputs to get meaningful outputs and a generalisable model. As such, in the context of radiomics, robustness of features for clinical modelling is essential, and many studies within the literature have explored this topic [68, 70, 73]. As outlined in Chapter 1, in this thesis a number of experiments are performed to test the robustness of features to various parts of the image processing scheme described above. This section briefly introduces the measures used. They were selected based on previously published robustness studies for radiomics discussed in the review by Traverso *et. al.* [80] and the work by Hatt *et. al.* [105].

2.5.1 Percentage Error

A simple first measure to assess variation between one "ground truth" value and another is to calculate the percentage error. This can also be called percentage change. For two values A and B, where A is the ground truth, the percentage error is given via

$$\%error = \frac{|B - A|}{A} \times 100. \tag{2.15}$$

2.5.2 Spearman rank coefficient

Spearman rank correlation coefficient (ρ) is a widely used measure of correlation between variables, ranging from -1 to 1, where 1 indicates a perfect positive monotonic relationship. As ρ is a measure of rank order, this is useful in feature robustness testing to assess the stability of patient order in a cohort between, for example, two different extraction settings, or to assess redundancy. Spearman rank correlation is calculated via

$$r_s = \frac{\text{Cov}(r_x, r_y)}{\sigma_{r_x}\sigma_{r_y}}, \tag{2.16}$$

where $\text{Cov}(r_x, r_y)$ is the covariance in the ranks of the group x and group y, and σ_{r_x} and σ_{r_y} are the corresponding standard deviations of the rank values.

It is a good measure of stability for a feature, as a substantial change in rank order, i.e a low ρ, will likely lead to poor generalisable modelling results. For example, as Wang *et. al.* [106] argues, many radiomics studies involve the use of Kaplan-Meier (KM) analysis to dichotomise datasets, and a substantial change in rank order - if the KM analysis uses an unstable feature that is sensitive to a particular attribute of extraction - is likely to lead to patients moving between dichotomised sets. And as a consequence of using the unstable feature, the models will not generalise. In terms or redundancy, comparison between two different features that yield a high ρ suggests they may offer complimentary information, and one may be redundant to use in the modelling process. A review of studies on repeatability and reproducibly of radiomic features by Traverso *et. al.* [80] cited 5 studies that used ρ as the main form of analysis, though others utilised it along other measures such as the *intraclass correlation coefficient* (ICC).

2.5.3 Bland Altman Analysis

Bland-Altman analysis is often used within medical research to assess agreement between two measurements techniques. Giavarina [107] provides an excellent introduction to this type of analysis. Commonly, it is used to compare an established measurement technique and a new alternative. Correspondingly, it is well suited for test-retest studies and to identify potential bias.

A Bland-Altman plot is constructed where the X axis is the mean of the two measurement methods. i,e [107]:

$$X = \frac{\text{method A} + \text{method B}}{2}, \tag{2.17}$$

and the Y axis is the difference between the two measures (which can be scaled proportionally to the average magnitude of the measurements and expressed as a percentage) [107]:

$$Y = \text{method A} - \text{method B}, \tag{2.18}$$

or

$$Y = (\text{method A} - \text{method B})/\text{Average}(\%). \tag{2.19}$$

Ideally of course, for results in perfect agreement all measures from Eq. 2.18 or 2.19 would be zero. The range and distribution of the measured difference indicates stability. The mean of all the measured differences can be used to assess bias between the methods and an agreement interval is obtained that is an upper and lower bound that contains 95% of the difference measured [107]. Bland-Altman analysis has been used in several studies in radiomics robustness and repeatability, such as by Hatt *et*. *al*. [105] and by Desseroit *et*. *al*. [108].

2.5.4 Intraclass correlation coefficient

ICC is a common statistical measure of reliability, used in test-retest, intrarater and interrater reliability studies [109]. ICC is often discussed in terms of *raters* (the entity doing measurements) and *subjects* (the entity being measured) [109]. In essence, one is interested in the reliability of rater measurements of the subjects. In radiomics, different raters are typically the different extraction settings, or acquisition settings, or software. Each feature is a measure of the patient imaging, (i.e, each radiomic feature is a particular measurement of a subject). Following convention, to calculate ICC the data is arranged into a matrix, where each row is a subject and each column is the corresponding rater/measurement of the subject by each rater [110].

There are 10 versions of ICCs, with separate assumptions and interpretations for the results depending on which form is used [109, 110]. One must select the appropriate ICC based on the aim of the study. Koo and Lee [109] detail the *Model*, *Type*, and *Definition* for each form of ICC. They provide a guide for which ICC to use depending on study type, and how to interpret the result. Theoretically, ICC results range from 0 to 1. The closer to 1, the stronger the measure of reliability. Following the convention of McGraw and Wong [110], as reported in Koo and Lee [109], this thesis utilises two types of ICC: The *two-way random effects, agreement, singer rater/measurement* ICC, estimated by

$$\frac{MS_r - MS_e}{MS_r + (k-1)MS_e} \tag{2.20}$$

and the *two-way, random effects, consistency, singer rater/measurement* ICC, estimated by

$$\frac{MS_r - MS_e}{MS_r + (k-1)MS_e + \frac{k}{n}(MS_c - MS_e)}. \tag{2.21}$$

Here MS_r is the mean square for rows, MS_c is the mean square for columns, MS_e is the mean square for error, k is the number of rater/measurements, and n is the number of subjects. In this context, mean squares are an estimate of variance, calculated as the sum of squared differences, divided by the associated degrees of freedom. Alongside ICC estimates, the 95% confidence intervals should be calculated and reported [109].

ICC has been utilised in a number of radiomics studies on feature robustness. For example, Parmar *et. al.* [70] compared manual and semi-automatic GTV segmentation of NSCLC patients in the context of feature robustness. From this study, they recommenced semi-automatic segmentation as a better alternative to manual segmentation where possible for radiomics, based on higher overall reproducibility of features, which was determined with ICC. As another example, Bogowicz *et. al.* [111] assessed the reproducibility of radiomic features extracted from PET Imaging using two different implementations; They found low reproducibility of features between the two implementations using ICC analysis. The review of studies on repeatability and reproducibly of radiomic features by Traverso *et. al.* [80] identified 14 articles that used some form of ICC in their reproducibility analysis. Traverso *et. al.* [80] reported that the ICC threshold to identify "robust" features varied amongst these 14 studies, which can make it difficult to collate for an overall consensus of which features to use moving forward. Adhering to recommendations such as Koo and Lee [109] would help in this regard.

2.6 Concluding Remarks

The aim of this Chapter was to introduce the necessary technical details of image processing for radiomics, and to present the SPAARC radiomics pipeline developed to extract radiomic features. Also introduced are common statistical techniques used to measure feature robustness. The image processing scheme detailed in this chapter, to which SPAARC adheres, evolved through a collaborative effort by the IBSI to standardise common features found in the literature. Through refinement, the goal of the IBSI was to achieve valid consensus among teams and thereby improve reliability of future radiomics studies. This standardisation effort to provide a set of benchmarks based on a consensus of implementations is detailed in the next chapter.

Take home message

1. Radiomics is a rapidly evolving imaging analysis technique.
2. Extraction of engineered radiomic features requires many intricate steps.
3. To enable standardisation and facilitate reproducibility in radiomics research, a comprehensive image processing scheme must be implemented precisely.
4. The *SPAARC* radiomics package was developed to extract over 160 baseline radiomic features from medical imaging. Many different variants of the baseline features were implemented (such as the different texture matrix aggregation to conduct 2D/2.5D/3D analysis)
5. Robustness testing can be explored with many types of statistical analysis.

3

Standardisation of Radiomic Feature Extraction

“Without standards there can be no kaizen (change for the better) .”

— Taiichi Ohno.

3.1 Preview

This chapter contains work to obtain standardised radiomic feature extraction in medical imaging utilising the image processing scheme and software introduced in Chapter 2. Chapter 3 is split into three sections. **Section** 3.2 details a study on consensus based radiomics standardisation by the *Image Biomarker Standarisation Intiative* (IBSI). This study was a large multicentre collaboration, with the formation of 25 teams submitting feature values for consensus benchmarking. Theses results from the intiative have been published by Zwaneneburg *et*. *al*. [84]. Cardiff’s contributions and progression, through the implementation of SPAARC, are emphasised for this thesis. The author, representing Cardiff, was a core contributing team to the initiative. The study aimed to determine reference values for 174 features commonly found in the radiomics literature. The imaging datasets and archive of team submissions have been made available online [112]. **Section** 3.3 extends this work by independently evaluating key discrepancies found in the implementation of SPAARC (alongside other teams) identified through the iterative standardisation process of the study in Section 3.2. This section acts as a technical note, highlighting the individual impact and clinical implications of standardisation choices. Section 3.4 is an overall discussion of both sections of work.

3.1.1 Author Contribution

The author is a core contributor to the IBSI, as described previously in Sections 1.8 and 1.9. The key contributions for the work discussed in this chapter are summarised below. The author provided 1776 benchmark values for a baseline set of 165 implemented radiomic features. This data directly led to the identification of several key issues, namely, 1) clear discrepancy arising from the interpolation grid generation and 2), the effect of combining multiple re-segmentation methods. The recommendations and subsequent IBSI standard for these image processing tasks were refined directly by the author's input and analysis. The author performed further independent data analysis for these two identified discrepancy issues in this thesis chapter, detailed in Section 3.3. As a top contributor, the author's data yielded valid consensus benchmarks for rare feature variants (e.g. All 2.5D texture features) that would not have been standardised otherwise. The author provided edits and revision for the published manuscript.

3.2 Consensus Based Standardisation with the IBSI

3.2.1 Introduction

There is lack of benchmarks for common radiomic features [10]. When developing extraction software, the final feature values computed from an image can vary due to various implementation choices. No clear standard or reference for implementations leads to inconsistencies between software, and may hinder reproducibility and validation of promising studies, especially if authors underreport the studies design and extraction details. The *Image Biomarker Standardisation Initiative* is an international, collaborative endeavour. Concretely, the IBSI attempted to meet 4 key objectives [84]:

1. Establish feature nomenclature and definitions.
2. Define a standardised and generalisable processing scheme for feature calculation from common medical imaging modalities.
3. Provide datasets and reference values for software calibration and verification.
4. Provide reporting guidelines to help study reproducibility and validation.

Chapter 2 introduced the general image processing scheme, feature nomenclature and definitions that evolved in tandem with this study. What follows is the consensus based process used to determine reference values thorough conensus, offering effective software calibration.

3.2.2 Materials & Methods

3.2.2.1 Study Phases & Datasets

The IBSI collaboration evolved and expanded in iterative steps. To tackle the objectives listed above, the work became separated into 3 distinct phases, where each phase utilised a unique dataset.

Phase 1's focus was to define the feature nomenclature and determine reference values using `Dataset-1`. This phase did not assess any image processing steps prior to feature extraction. The algorithmic implementation of the features were in effect checked, without the overhead of handling any medical imaging data.

`Dataset-1` is a small *digital phantom*. This 3D volume is simply a matrix of 80 voxels and effective ROI selecting 74 of the voxels for analysis. The volume is sufficiently small such that most features could feasibly be computed by hand (though arduous). The digital phantom is visualised in Figure 3.1, slice by slice, with associated voxel intensities given in red.

Phase 2's focus was to develop a standardised image processing scheme, and to provide reference values for a range of processing settings, using a "typical" medical image. Five configurations of settings were tested with `Dataset-2`.

`Dataset-2` is a CT image of a patient with lung cancer, taken from a publicly available dataset [7]. The ROI for feature extraction is defined by a segmentation of the gross tumour volume (GTV), labelled GTV-1, that accompanies this image. This patient and GTV contour are visualised slice by slice in Figure 3.2. Formats used are publically available [112].

Phase 3 was a validation of the first two phases. Features considered standardised via phase 1 and 2 were assessed for reproducibility with `Dataset-3`. This dataset included new modalities to help determine generalisability of the study.

`Dataset-3` contains multimodal imaging from a cohort of 51 patients with soft-tissue sarcoma made publically avaliable on the Cancer Imaging Archive [100]. Each patient has a co-registered CT, ^{18}F-FDG PET and T1-weighted MRI imaging, all with a GTV segmentation. The original dataset was, checked, cleaned and pre-processed for the needs of the IBSI and re-uploaded [112].

3.2.2.2 Radiomic features

The nomenclature and definition for 174 base features were selected and defined by the IBSI for this work. With texture features, this number increases from the base, as there are several possible ways to extract them (for an example, refer back to Section 2.4.6.1).

Table 3.1 shows the final number of features the author implemented for team Cardiff (in SPAARC radiomics package) for each phase and configuration, alongside the total number assessed in this study. Cardiff were one of 5 teams to implement over 95 % of the possible number of features considered for benchmarking through phases 1 and 2. In total, the author provided a final set of 1776 benchmark values for these phases.

3.2.2.3 Consensus and Strength of Reference Values

To assess the consensus between teams, features were extracted for the first two phases using the settings outlined in Table 3.2.

Initially, for each feature, the result was taken to 3 significant figures and the modal value among teams used as a provisional reference. The final level of consensus was calculated based on two measures:

Phase 1: Determine reference values for features without image processing
Dataset-1: Digital Phantom

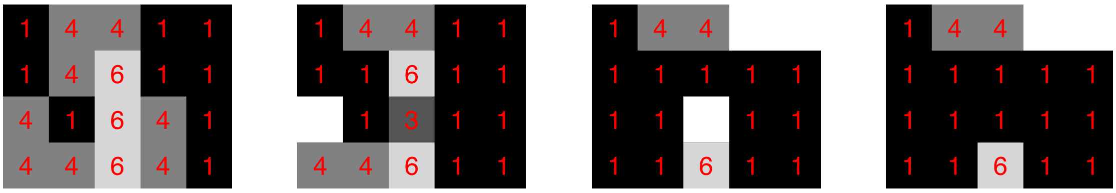


Figure 3.1: Visualisation of the digital phantom used in Phase 1. The phantom is a simple 5 × 4 × 4 matrix, with 6 voxels excluded, leaving 74 for analysis. This provides non-trivial values for morphology based features. Voxels are 2 × 2 × 2 mm. Intensity values given in red, ranging from 1 to 6, with none equal to 2 or 5. Imaging data available at [112].

Phase 2: Determine reference values for features with image processing
Dataset-2: 1 CT of Lung cancer

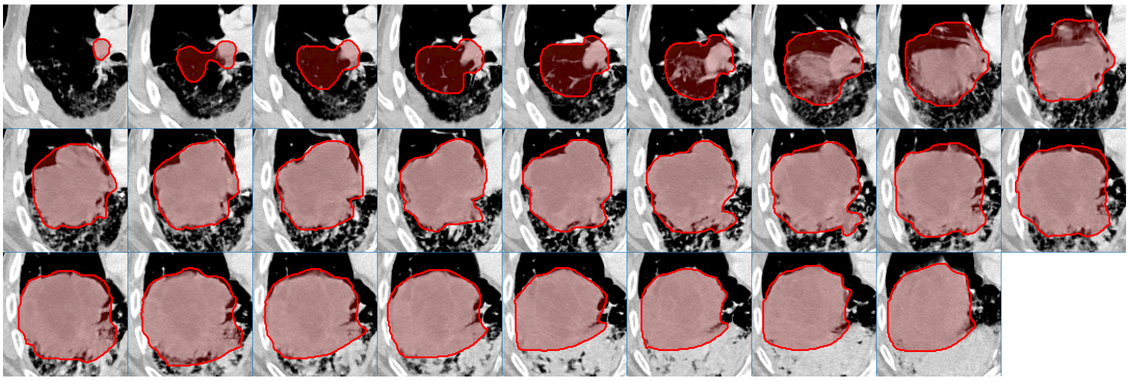


Figure 3.2: Montage visualisation of `Dataset-2` used in Phase 2. It is a single CT image from a non-small-cell lung carcinoma patient. The accompying GTV segmentation -the ROI - has been overlaid in red. Imaging data available at [112].

Phase 3: Validation; Assessing reproducibility of standardised radiomics features
Dataset-3: Multimodal imaging of 51 patients with soft-tissue sarcoma

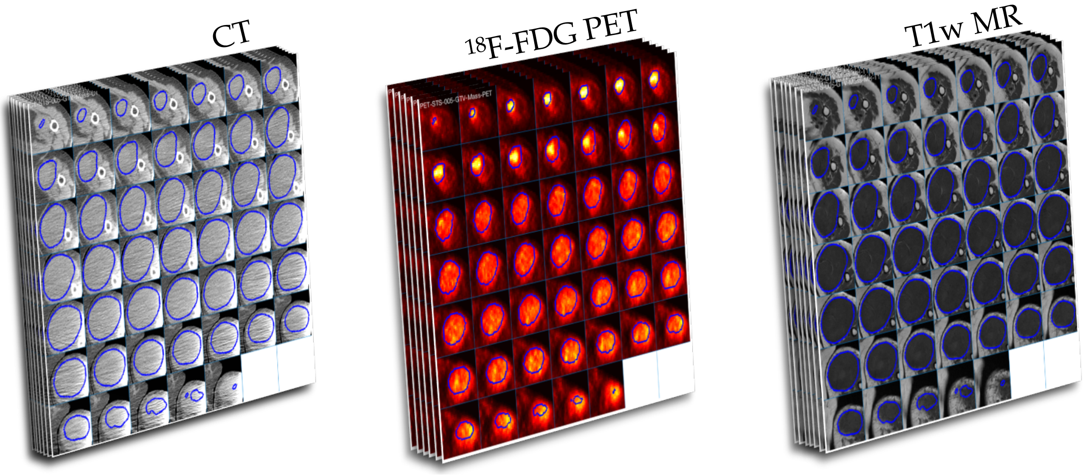


Figure 3.3: Summary of Dataset-3, a cohort of 51 patients with registered multi-modal imaging of soft-tissue sarcoma . The GTV is segmented for each image type. Represented here is a montage for each modality (there would be a stack of 51 montages). Dataset-3 was used for validating reproducibility of standardised features in Phase 3. Imaging data is available at [112].

Measure-1: The number of teams that matched the provisional reference value within a given tolerance.
Measure-2: The number of teams that matched the above provisional reference value against the total number of teams that submitted a value. (%)

The number of teams that matched a provisional reference indicated the strength of the consensus. Agreement between teams of ≥ 10 : very strong; 6-9: strong; 3-5: moderate, <3: weak. To become a valid reference value, the feature was required to have at least a moderate consensus (Measure-1 ≥ 3), for an absolute majority of teams that submitted a value for that feature (Measure-2$> 50\%$).

3.2.2.4 Standardisation Methodology

The groups in the IBSI converged towards references for features in an iterative fashion: submission, analysis, review, update, then repeat. Table 3.3 outlines key collection dates for submissions. Alongside are key notes of the updates to Cardiff's code occurring at that time point. To simply figures, each time point in the table is represented as successive arbitrary time unit starting from 1. The initiative was open and voluntary to join at any point. Teams were eligible to participate if they developed their own software for radiomics studies. Teams were also free to participate in any phase of the study. As outlined in Table 3.3, the concept of the IBSI was formulated in July 2016 and the initial time point for contributions began in September 2016; Cardiff joined the initiative and first made a contribution of feature values for the December 2016 update (arbitrary time point 4), and from there became a core, active member of the group.

At each submission time point, feature values were collected and processed to determine the strength of consensus. At each time point teams liaised to identify, clarify and improve the processing pipeline. Ambiguities regarding the descriptions and processing schemes in the reference manual were queried, and iterative updates made to reflect. Potential coding errors and dis-

Table 3.1: Summary of feature numbers for consensus benchmarking. There were 174 features considered in the baseline set. For each feature family, the total feature number is shown alongside the number Cardiff implemented in brackets (e.g. Cardiff implemented 23/29 Morphology features). As texture features have a number of settings for extraction, the total number of computed features available for benchmarking increases from the baseline (e.g. GLCM has 6 aggregation approaches, for a maximum of 6.*25=150 feature variants), see Table 3.2 for the extraction settings used for each Phase).

	Number of Features									
Feature Family	*Base Definition*		*Phase 1*		*Phase 2 config A-B*		*Phase 2 config C-E*		*Phase 3*	
Morphology	(23)	29	(23)	29	(23)	29	(23)	29	(23)	29
Local Intensity	(0)	2	(0)	2	(0)	2	(0)	2	(0)	2
Statistical	(18)	18	(18)	18	(18)	18	(18)	18	(18)	18
IH	(23)	23	(23)	23	(23)	23	(23)	23	(23)	23
IVH	(7)	7	(7)	7	(7)	7	(7)	7	(7)	7
GLCM	(25)	25	(150)	150	(100)	100	(50)	50	(25)	25
GLRLM	(16)	16	(96)	96	(64)	64	(32)	32	(16)	16
GLSZM	(16)	16	(48)	48	(32)	32	(16)	16	(16)	16
GLDZM	(16)	16	(48)	48	(32)	32	(16)	16	(16)	16
NGTDM	(5)	5	(15)	15	(10)	10	(5)	5	(5)	5
NGLDM	(16)	17	(48)	51	(32)	34	(16)	17	(16)	17
Total	(165)	174	(476)	487	(341)	351	(206)	215	(165)	174

crepancies that required refinements to the image processing scheme were identified. Teams then made updates to implementations and submitted new feature values. The iterative cycle repeated in this manner. Phase 2 did not begin until Phase 1 had 70% of features with at least a moderate or better consensus on the validity of their reference values, which occured at time point 10. From there, both phases ran concurrently. The standardisation process ended on time point 25, after reaching strong or better consensus on valid reference values for over 90% of features for both phases [84].

3.2.2.5 Validation of Standardised Features

To test the standardisation effort, the IBSI conducted a final validation phase. Unlike with the previous phases, this was a one-time submission using `Dataset-3`. Reproducibility was evaluated between teams with an intraclass correlation coefficient (two-way random effects, single rater, absolute agreement). For teams that submitted to phase 3, features were removed prior to ICC calculation if that teams software had not matched the reference value for the corresponding feature in phase 2. Using the categories provided by Koo an Li [109], the lower boundary of the 95% confidence interval (ICC-LB) categorised feature reproducibility across difference software implementations as follows: Excellent: $0.9 \leq$ ICC-LB ; good: $0.75 \leq$ ICC-LB < 0.9; moderate: $0.5 \leq$ ICC-LB < 0.75; poor: ICC-LB< 0.5.

3.2.2.6 Diagnostic Reference Features and Tolerance Margins

For comparison of implementations in Phase 2, alongside the main set of radiomic features proposed for standardisation, a set of *diagnostic* measurements extracted at different points along the processing pipeline were also recommended to be reported by teams. Like radiomic features, these are single value results, yet not designed for potential clinical merit. Rather, they are useful for calibration, recording how the image, VOI, and masks alter at different stages of the pipeline. For example, several diagnostic features report the size of ROI in terms of a bounding box dimensions in each direction, both before and after re-segmentation is applied. As with the radiomic features, these diagnostic feature were also benchmarked. For each configuration (labelled *config. A-E*) in Phase 2, 60 additional very simple diagnostic features were benchmarked (See Cardiff results table in Appendix C for full list of features). When troubleshooting software, the diagnostic features were extremely helpful to pin point sources of discrepancy, such as grid alignment differences discussed later in Section 3.3.2.

Very minor differences between implementations can cause small digression from a reference value. For example, this could include small float rounding errors, or slight deviations in voxel inclusion in the ROI during image processing. As such, a tolerance was utilised to bound each preliminary reference within a certain small range. A first preliminary reference and tolerance was to use the modal result $\pm 0.5\%$ (to 3 s.f.). Within the IBSI, the MIRP team generated a more sophisticated tolerance on top of this by perturbing the image and ROI mask, prior to interpolation, by planar translation, rotation, growing and shrinking of the ROI [84, 113]. Features were extracted for each perturbation, and the updated tolerance was taken as 5% of the interquartile range for each feature [84].

Table 3.2: Summary of the image processing settings used at each Phase of the consensus study [84]. When not explicitly stated, default settings in SPAARC are used, following the image processing scheme discussed in the previous Chapter.

	Phase 1	**Phase 2**					**Phase 3**		
Parameter	*Digital Phantom*	*config. A*	*config. B*	*config. C*	*config. D*	*config. E*	*CT*	*PET*	*MRI*
slice-wise (2D) or volume (3D)	2D & 3D	2D	2D	3D	3D	3D	3D	3D	3D
interpolation	no	no	yes	yes	yes	yes	yes	yes	yes
new voxel spacing (mm)			2×2 (axial)	$2 \times 2 \times 2$	$2 \times 2 \times 2$	$2 \times 2 \times 2$	$1 \times 1 \times 1$	$3 \times 3 \times 3$	$1 \times 1 \times 1$
method			bilinear	trilinear	trilinear	spline	spline	spline	spline
intensity rounding			nearest int.	nearest int.	nearest int.	nearest int.	nearest int.	no	no
ROI interpolation method			bilinear	trilinear	trilinear	trilinear	trilinear	trilinear	trilinear
ROI partial mask threshold			0.5	0.5	0.5	0.5	0.5	0.5	0.5
re-segmentation									
range		$[-500, 400]$	$[-500, 400]$	$[-1000, 400]$	none	$[-1000, 400]$	$[-200, 200]$	$[0, \infty)$	$[0, \infty)$
outlier filtering		none	none	none	3σ	3σ	none	none	none
discretisation									
texture and IH		FBS: 25	FBN: 32	FBS: 25	FBN: 32	FBN: 32	FBS: 10	FBS: 0.25	FBS: 10
IVH		none	none	FBS: 2.5 HU	none	FBN: 1000	none	FBS: 0.1	FBS:1
texture aggregation									
GLCM, GLRLM									
2D averaged	✓	✓	✓						
2D merged	✓	✓	✓						
2.5D direction-merged	✓	✓	✓						
2.5D merged	✓	✓	✓						
3D averaged	✓			✓	✓	✓			
3D merged	✓			✓	✓	✓	✓	✓	✓
GLSZM, GLDZM									
NGTDM NGLDM									
2D averaged	✓	✓	✓						
2D merged	✓	✓	✓						
3D	✓			✓	✓	✓	✓	✓	✓

Table 3.3: Outline of key events in the IBSI consensus study. Dates of team submissions were given an arbitrary time unit starting from 1.The timeline as first published in Zwanenburg *et*. *al*. [84] is included alongside additional information of Cardiff progression.

Date	Time	Description of timeline as presented in Zwanenburg *et*. *al*. [84]	Cardiff Notes
June 8, 2016	.	*A draft study proposal was formulated and shared with initial participants.*	
June 30, 2016	.	*Final study proposal was formulated and shared. The digital phantom was created and shared, together with the first version of the work document. Phase I was initiated.*	
September 14, 2016	1	*Initial contributions for the digital phantom are shared.*	Project began
October 9, 2016	2	*Contributions were updated and shared.*	
October 24, 2016	3	*The IBSI was presented at the Radiomics meeting in Clearwater, Florida, USA. Contributions were updated and shared.*	Cardiff join the IBSI.
December 6, 2016	4	*Contributions were updated and shared.*	Shared first results for Digital Phantom. Families: Statistical, All current GLCM varients (2D / 3D / merged / average).
December 8, 2016	.	*A major update to the work document was shared with the research teams. Several new features were added, based on requests. Volume and surface area features were redefined based on meshing algorithms. The general radiomics image processing scheme was drafted.*	Feedback on working document.
December 23, 2016	5	*Contributions were updated and shared. Sections of the work document were posted to arXiv to provide a reference for radiomics features. The dataset for phase II was identified.*	Added results for all current GLRLM varients (2D / 3D / merged / average).
January 24, 2017	6	*Contributions were updated and shared.*	-
January 30, 2017	.	*The image processing configurations were defined. Phase II was initiated.*	Began implementing additional processing steps into software in prep for phase II
February 10, 2017	7	*Contributions were updated and shared.*	-
February 24, 2017	8	*Contributions were updated and shared.*	-
March 10, 2017	9	*Contributions were updated and shared.*	-
April 14, 2017	10	*Contributions were updated and shared, including initial results for phase II.*	Features added: Morphological, Intensity Histogram, GLSZM varients (2D and 3D). Added Intial phase II results for current config A-D
April 21, 2017	.	*Segmentation of the RT structure set and image interpolation were identified as major sources of divergence.*	Feedback on working document. Discrepency between teams found for many features in Phase II.
May 6, 2017	.	*Meeting of several IBSI teams during the ESTRO 36 conference, where an electronic poster for IBSI was presented.*	Attended meeting at Estro event (Coauthor of Poster).
May 19, 2017	11	*Contributions were updated and shared. The description of interpolation is made more precise, and the concept of morphologic and intensity ROI masks was introduced.*	Identifed interpolation grid as potential discrepency source. Added AGO approach. Added seperation of ROIs into morphological and intensity masks.
June 26, 2017	12	*Contributions were updated and shared.*	Fixed small bug which resulted in double discretisation for some texture feature calculation.
July 24, 2017	13	*Contributions were updated and shared. The arXiv document was updated with a new image processing section.*	Added results for NGTDM (2D/3D). Added Diagnostic features. Minor updates to image processing pipeline following new recomendations.
August 11, 2017	14	*Contributions were updated and shared.*	Feedback on working document. Added GLDZM features (2D/3D) and corresponding config E results for all features implemented so far in SPAARC.
August 31, 2017	15	*Contributions were updated and shared.*	-
October 11, 2017	16	*Contributions were updated and shared. First use of tolerance in determining reference values.*	-
October 23, 2017	.	*Progress of IBSI was presented at the Radiomics meeting in Clearwater, Florida, USA.*	-

Table 3.3: Outline of key events in the consensus study continued.

Date	Time	Description of timeline as presented in Zwanenburg *et. al.* [84]	Cardiff Notes
November 16, 2017	17	*Contributions were updated and shared. The arXiv document was updated with a guidelines section, as well as all prior changes to sections of the IBSI work document included in the arXiv document. Configurations C and D were revised. Moreover, the section describing the Intensity-Volume Histogram was extensively revised.*	-
December 4, 2017	18	*Contributions were updated and shared.*	Added results for IVH features. Shared updated configuration C and D results.
January 5, 2018	19	*Contributions were updated and shared.*	-
January 17, 2018	.	*A draft version of the manuscript was prepared and shared with several coauthors.*	-
February 1, 2018	.	*A revised version of the manuscript was shared with all coauthors.*	Comments and feedback provided.
February 13, 2018	20	*Late contributions were updated and shared.*	
February 20, 2018	.	*Manuscript was sent out for peer-review.*	-
August 22, 2018	.	*Manuscript was returned with reviewer comments.*	-
August 30, 2018	.	*Discretization definitions were updated.*	-
October 1, 2018	21	*Contributions were updated and shared. The arXiv document was updated to include the improvements to discretization.*	Updated results using new FBW discretisation.
October 5, 2018	.	*Configuration E was updated to reflect new resegmentation definitions. 2.5D texture features were added.*	Updated pipeline such that re-segmentation performed in parellel when mutiple methods selected. Cong E results updated to reflect change.
October 16, 2018	.	*Progress of IBSI was presented at the Radiomics meeting in Clearwater, Florida, USA.*	-
November 22, 2018	22	*Contributions were updated and shared.*	Implemented 2.5D feature extraction and shared results.
January 4, 2019	23	*Contributions were updated and shared.*	Fixed small bug in calculation of 2.5D features for GLRLM family.
February 1, 2019	24	*Contributions were updated.*	-
March 1, 2019	25	*Contributions were updated and shared. Consensus on the validity of reference values was found to be sufficient to halt the iterative standardization process.*	-
April 4, 2019	.	*A completely revised version of the manuscript was shared with all coauthors.*	Detailed comments and feedback provided. Part of regular teleconference discussions with core group to finalise manuscript.
May 16, 2019	.	*The arXiv document was updated to include tables of reference values.*	-
May 23, 2019	.	*Manuscript was sent out for peer-review.*	-
August 6, 2019	.	*Review comments were received.*	-
September 4, 2019	.	*The validation phase (III) was started using new datasets deriving from CT, PET and MRI.*	Discussed potential datasets and provided preliminary checks on chosen validation dataset with core group before sent out to all teams.
October 14, 2019	.	*All validation results were collected and parsed.*	Shared Cardiff validation results for phase III.
October 22, 2019	.	*The revised manuscript was submitted for peer-review.*	-
December 9, 2019	.	*A second revision was submitted.*	-

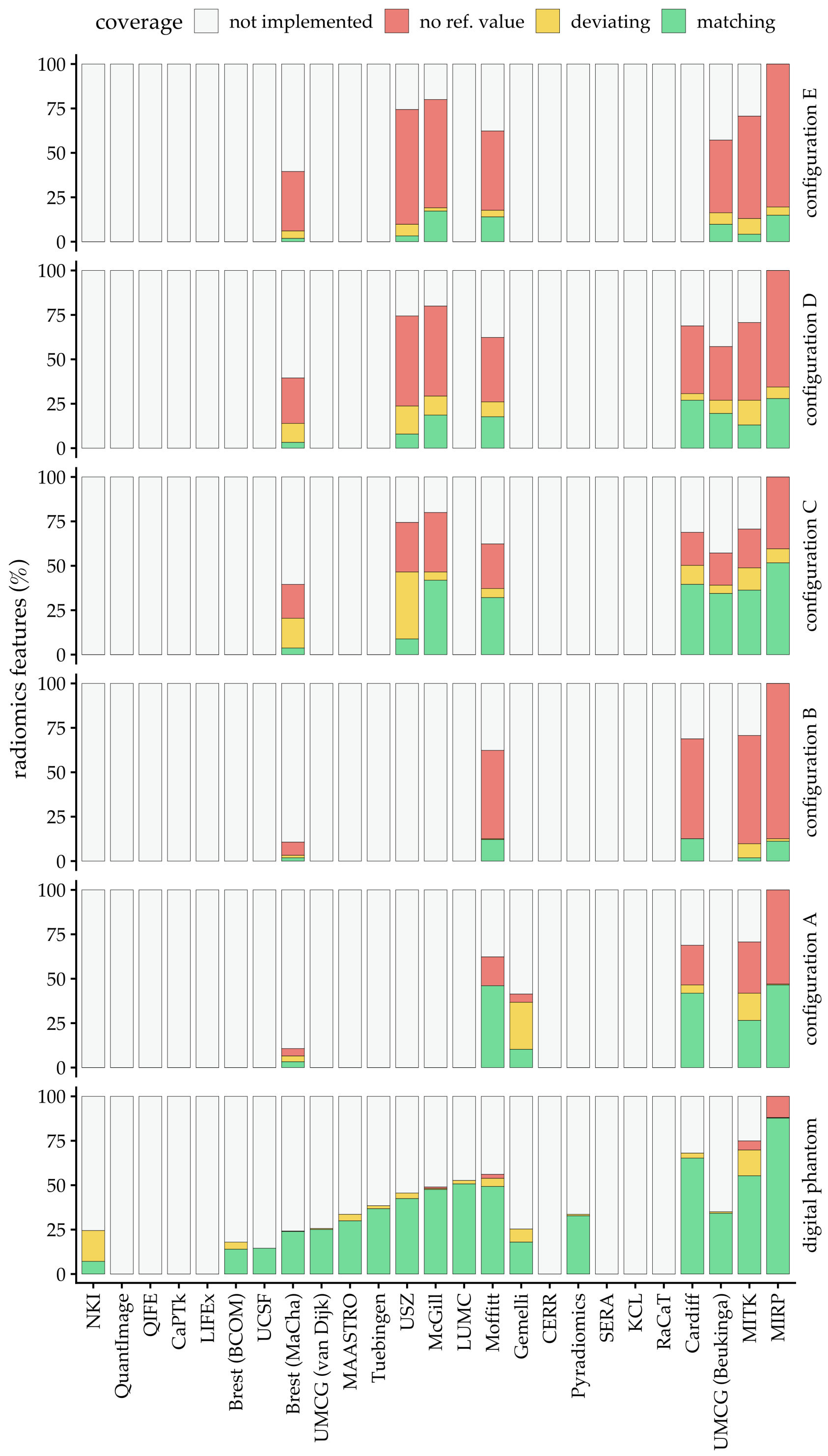


Figure 3.4: Feature coverage for each IBSI team at timepoint 10, which was the introduction of phase-2. Results shown separately for the digital phantom of phase-1, and the 5 configurations of phase-2. Expressed as a percentage is the amount of features that (at this timepoint): matched the provisional reference (green); deviated from it (yellow); did not achieve a valid reference (red), or were not implemented (grey). For consistency, teams are ordered by coverage at the final time point. Note that 8 teams had not joined the initiative yet. This Figure was created from data published by Zwnanenburg *et. al.* [84].

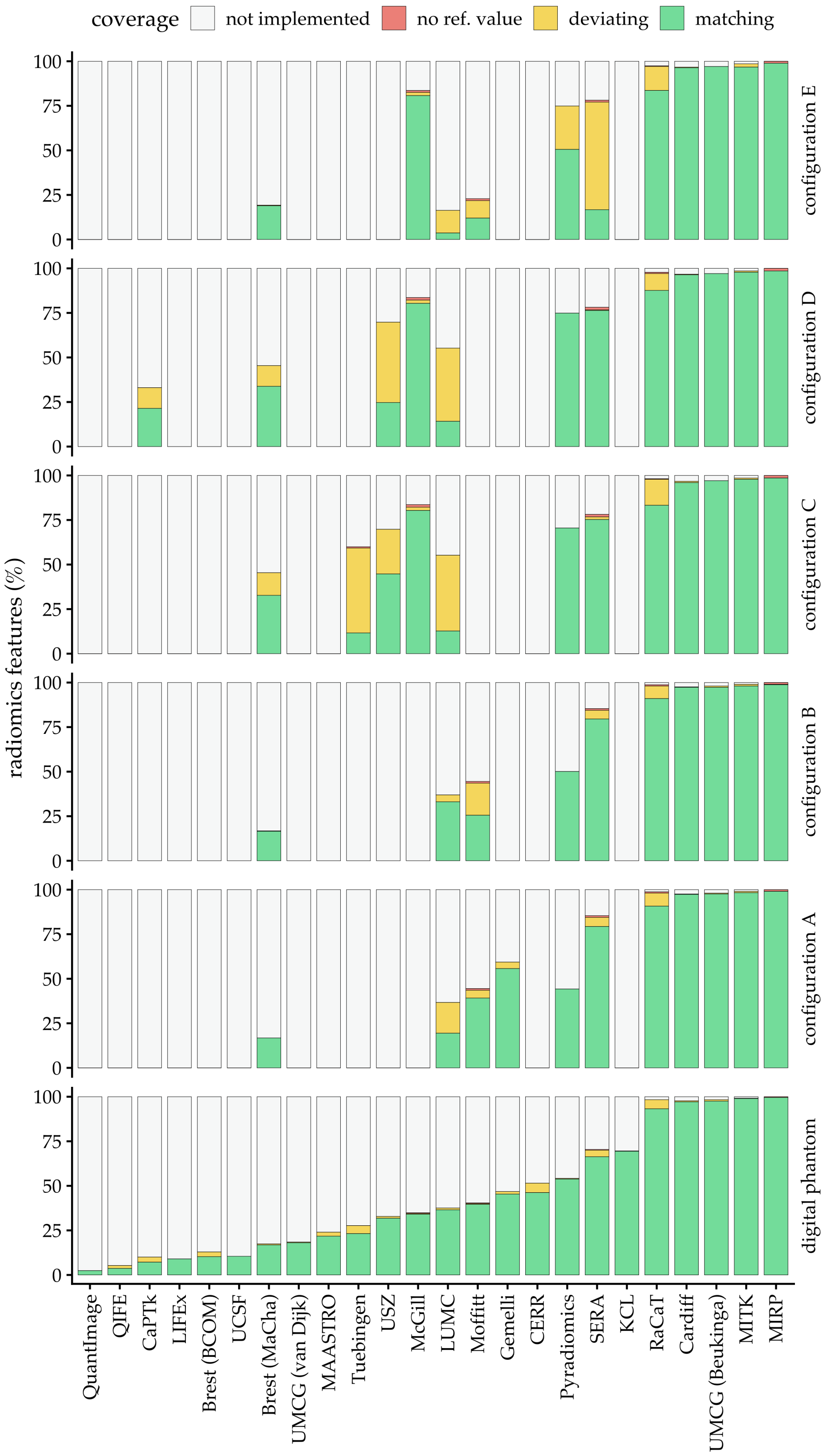


Figure 3.5: Feature coverage for each IBSI team at the final time point. Results shown for the digital phantom of phase-1, and the 5 configurations of phase-2, taken from the final time point of the initiative. Expressed as a percentage is the amount of features that: matched the provisional reference (green); deviated from it (yellow); did not achieve a valid reference (red), or were not implemented (grey). Note there are 24 teams, as one team retired (NKI). This Figure was created from data published by Zwnanenburg *et. al.* [84].

3.2.3 Results

In total, 25 teams participated throughout the course of the initiative. There was no requirement to implement every feature or image processing option, and as a consequence, there was broad variation in the number of features covered by each team. Of course, the more teams that could provide a matching value for a given feature, the stronger the consensus (`measure-1`). In the following results, teams are referred to by an abbreviation, see Appendix A for full team names alongside the software programming language that was utilised.

Figure 3.4 and Figure 3.5 summarise the total feature coverage of teams, at time point 10 and 25 (the final time point) respectively. Time point 10 marked a major update in the initiative with the initiation of phase 2. The large deviation and discrepancy in feature values between teams submitting results highlighted the challenges of image processing that needed to be resolved. Comparison with the final time point illustrates the marked progression for the majority of teams. Coverage is expressed as a percentage of features, and separated into the digital phantom of Phase 1, and the 5 configurations of Phase 2. These Figures detail, for each team at that timepoint, the percentage of features values submitted that: matched the provisional reference (green), deviated from it (yellow), did not achieve a valid reference (red), or were not implemented (grey). Cardiff were one of the top 5 contributing teams that implemented at least 95% of the total feature variants. Nine teams implemented over 50% of the total feature variants. Figure 3.5 shows results from 24 teams, as one retired their software (NKI).

Figure 3.7 isolates and summarises the individual progress of Cardiff at each submission time point. More features were added and processing steps implemented as the initiative continued. Table 3.3 give a brief overview of implementation details for Cardiff at each time point in the initiative alongside the reported timeline [84]. As mentioned, time point 10 saw the introduction of phase 2. Here, differences between teams due to the processing approach became immediately apparent, and with it a need for further refinement and clarification to the developing image processing scheme. In particular, generation of the interpolation grid was identified as a potential cause of discrepancy. Section 3.3 explores this in more detail, alongside other identified causes of deviation. As evident in Cardiff's progression in Figure 3.7, these discrepancies in image processing were ultimately resolved, aiding in the strong consensus and impact of the study.

At time point 22, 2.5D texture features were introduced to the initiative. These features were assessed for the digital phantom and configurations A and B. Here, Cardiff's contributions for the 2.5D GLRM features deviated from the mode results amongst teams that provided values, due to a subtle coding error in the direction summation of the run length matrices. Essentially, this amounted to a minus sign misplacement. This error was identified quickly once the deviation was reported, and rectified for the next update, as shown in time point 23 in Figure 3.7. Without benchmarks, errors like this can easily go unnoticed, and this serves as an excellent example of the benefit and necessity of this work.

The strength of consensus, and therefore the validity of previsional reference values, improved throughout the course of the initiative. Figure 3.6 highlights this progression. In summary, as reported by Zwanenburg *et. al.* [84], at the final time point 463/487 features in Phase 1, and 1220/1347 features in Phase 2, combining all configurations, reached a stronger or better con-

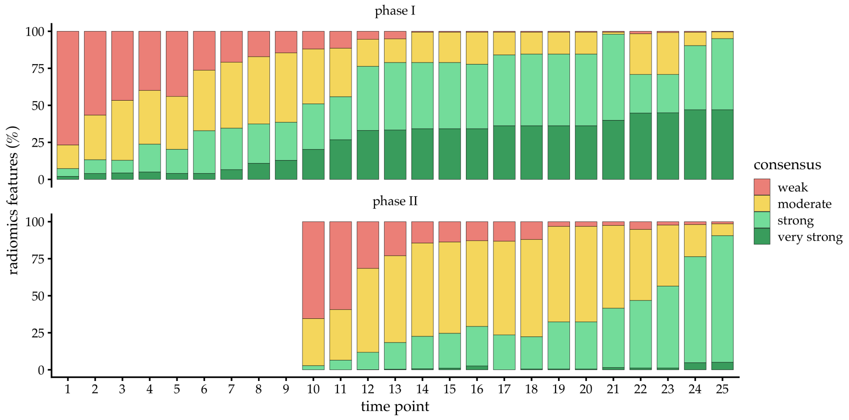


Figure 3.6: An overview of the strength of consensus between teams at each submission time point. This was based on the number of teams that produced the same provisional reference value. Agreement between <3 teams: weak (red); 3-5 teams: moderate (yellow); 6-9 teams: strong (light green); ≥ 10 teams: very strong (dark green). Highlighted here is the general iterative trend toward strong or better consensus for over 90% of the feature variants in each phase. Figure adapted from Zwnanenburg *et. al.* [84].

sensus (i.e. measure-1>5, measure-2>50%).

At the final time point, of the 174 baseline features, 169 had at least a moderate or better consensus for all variations assessed in Phase 1 and 2, and had a majority consensus (i.e. measure-2>50%). These features were considered standardised. Five features did not reach a valid reference and remained unstandardised.

Cardiff implemented 165/174 of baseline features, and of these, 164 were standardised. Of the 5/174 unstandardised baseline features, 4 were morphological features that were not implemented in SPAARC (due to a trade-off between computational time and complexity, compared to the benefit of having these features in the SPAARC pipeline). The one baseline feature in the Cardiff implementation that remains unstandardised between all teams is the IVH area under curve.

3.2.3.1 Validation Results

Nine teams provided results for the validation phase by extracting features from `Dataset-3` using the extraction settings in Table 3.2. These extraction settings gave 1 result per baseline feature (174), as only one type of texture aggregation was selected to represent a typical model development scenario in radiomics. Excellent reproducibility was achieved between team implementations for the vast majority of features across all 3 imaging modalities [84]. In CT, MRI and PET respectively, 166/174, 164/174, and 164/174 features were excellently reproducible; In addition, 1/174, 3/174 and 3/174 achieved good reproducibility. In each modality, 2 features were computed by only 1 team in the validation (these were morphological features that Cardiff did not implement) so ICC could not be determined. As 5/174 features were unable to be standardised in Phase 1 and 2, they were not assessed in Phase 3.

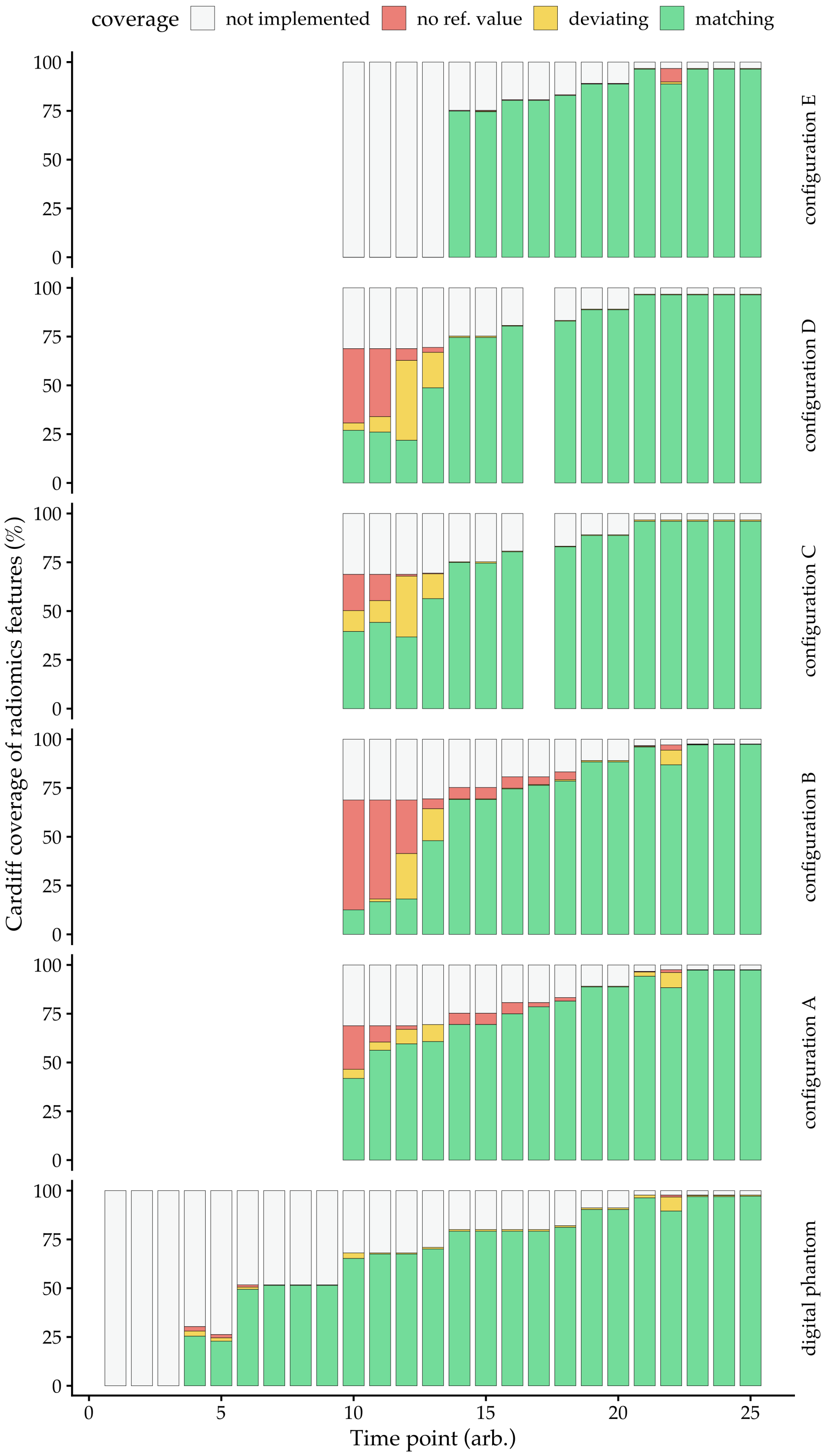


Figure 3.7: Cardiff's progression and contributions towards standardised reference values at each time point. Results shown for the digital phantom of Phase 1, and the 5 configurations of Phase 2, which began at time point 10. Expressed as a percentage is the amount of features that: matched the current provisional reference at that timepoint (green); deviated from it (yellow); had not yet achieved a valid reference (red), or not implemented (grey). At time point 17, settings for configurations C & D were updated (see Table 3.3) and so were not included for analysis. This Figure was generated from the data published by Zwnanenburg *et. al.* [84].

3.3 Evaluation of Implementation Discrepancies

3.3.1 Introduction

The study in Section 3.2 confirmed a concern of previous radiomics reviews [10], and a stated hypothesis for this thesis, that subtle differences in the implementation of image processing techniques can cause feature values to deviate. This became evident in the early stages of Phase 2, as shown in Figure 3.6 and 3.7, with low initial consensus amongst teams that submitted values. In some cases, the lack of consensus originated from a bug (a coding mistake). To resolve these types of discrepancies, clearly one must identify and fix the mistake. A significant output of the study in Section 3.2 is a set of benchmarks to now safe guard against future implementation errors.

However, rather than just identifying coding error, in the interest of standardisation, several decisions concerning the way one *could* implement, had to become more rigorous recommendations on how one *should* implement. Most discrepancies were not from mistakes. It became clear that the image processing scheme needed to be more refined, to offer in several places one recommended (i.e. standardised) option amongst several arguably acceptable alternatives.

Through the iterative process of the IBSI, several discrepancies were identified amongst teams that were valid interpretations of the image processing scheme at that time point. The discussion and analysis in the following section aimed to assess how key discrepancies impacted the radiomic features, and arguments for the standard that was chosen. This work acts as a technical note, and evaluates several processing steps needed to reach compliance.

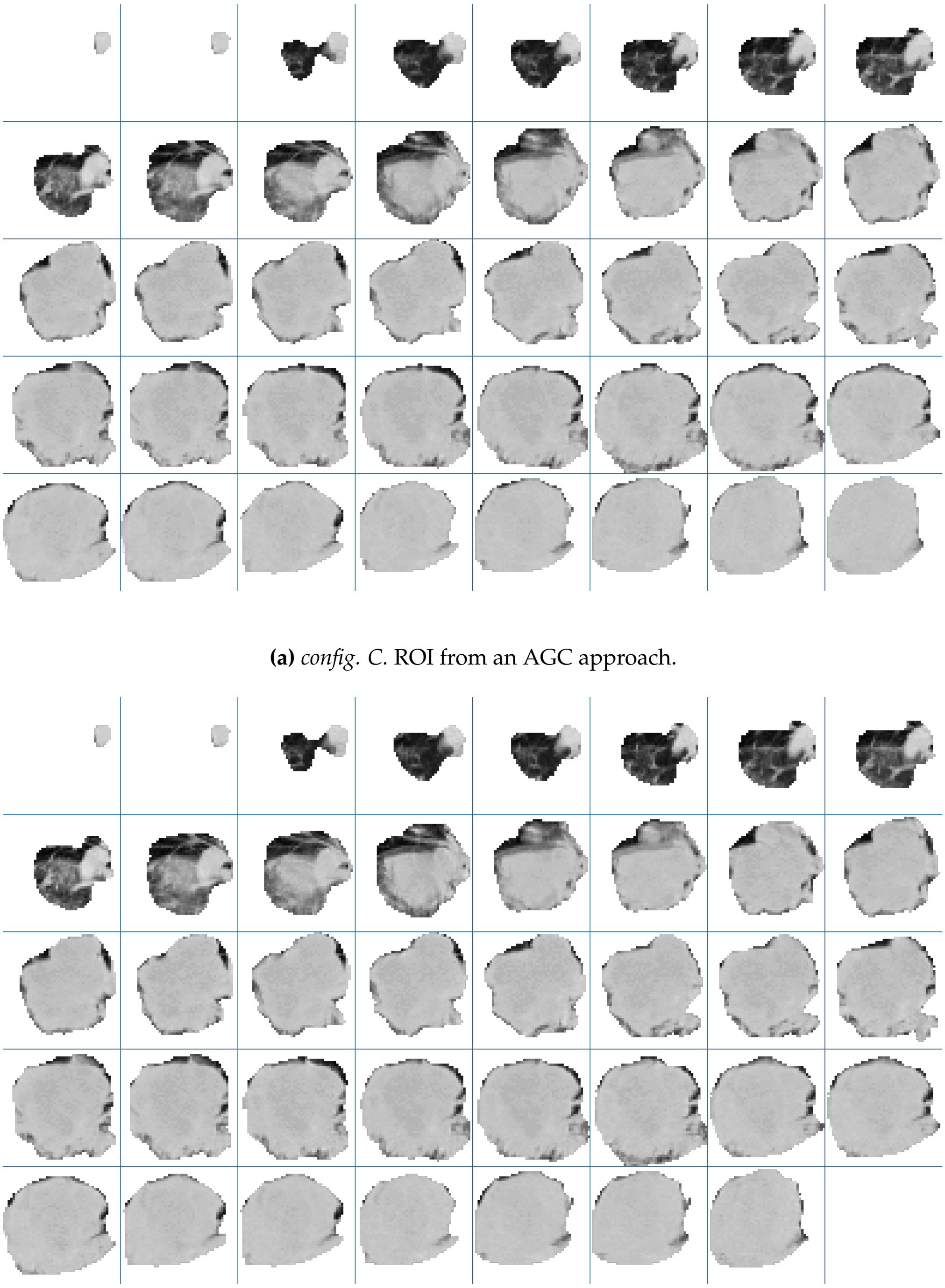
3.3.2 Interpolation Grid Generation Discrepancy

A key image processing technique is interpolation to new voxel dimensions. The author identified early into its introduction in the IBSI that it was a likely source of deviation. During the interpolation step, between teams, it was discovered that there are two main methods for generating a new grid coordinate system at a new desired spacing: *aligning grid origins* (AGO) and *aligning grid centres* AGC, the theory of which were discussed in detail in Section 2.3.3. Cardiff alongside some other teams originally implemented using the AGO approach. AGC became the standard. Utilising a now fully standardised pipeline, this section isolates and demonstrates the degree to which these methods impacted the feature extraction process.

3.3.2.1 Methods

Imaging `Dataset-2` and `Dataset-3` described in Section 3.2.2.1 were used once again for this analysis. Only the set of features implemented by Cardiff are considered in this work (See Table 3.1). As before, features were extracted from the relevant datasets with the corresponding setting configurations outlined in Table 3.2.

For comparison, features were extracted twice, once with AGO, and once with an AGC approach. As *config A* does not include interpolation, it was omitted from the analysis here. Feature variation was quantified in two ways. First, for each feature, the absolute percentage change from AGC to



(a) *config. C.* ROI from an AGC approach.

(b) *config. C.* ROI from an AGO approach.

Figure 3.8: A visual comparison of the discretised intensity ROIs using a) *align grid centres* (AGC) and b) *align grid origins* (AGO) methods for interpolation grid generation. These were extracted with configuration C settings (Table 3.2) from `Dataset-2`. On closer inspection, the differences from re-sampling at a slightly shifted coordinate systems become apparent for each slice. The most notable divergence in this visualisation is that the AGO approach results in an additional slice (40) compared to AGO (39).

AGO was determined. To visualise and report the results, features were categorised into 3 groups: $< 0.5\%$, $0.5\% - 10\%$, and $\geq 10\%$. The use of 0.5% as a category cut off was due to it being utilised as an approach to measure similarity early on in the IBSI. Secondly, the AGO features were assessed against the new IBSI reference values and sorted into 3 categorises based on IBSI compliance: within tolerance, outside tolerance, or no valid reference value.

Using `Dataset-3`, robustness of features across a patient cohort between AGO and AGC methods was measured with ICC, the cut-off for categories is the same as defined in Section 3.2.2.5.

3.3.2.2 Results

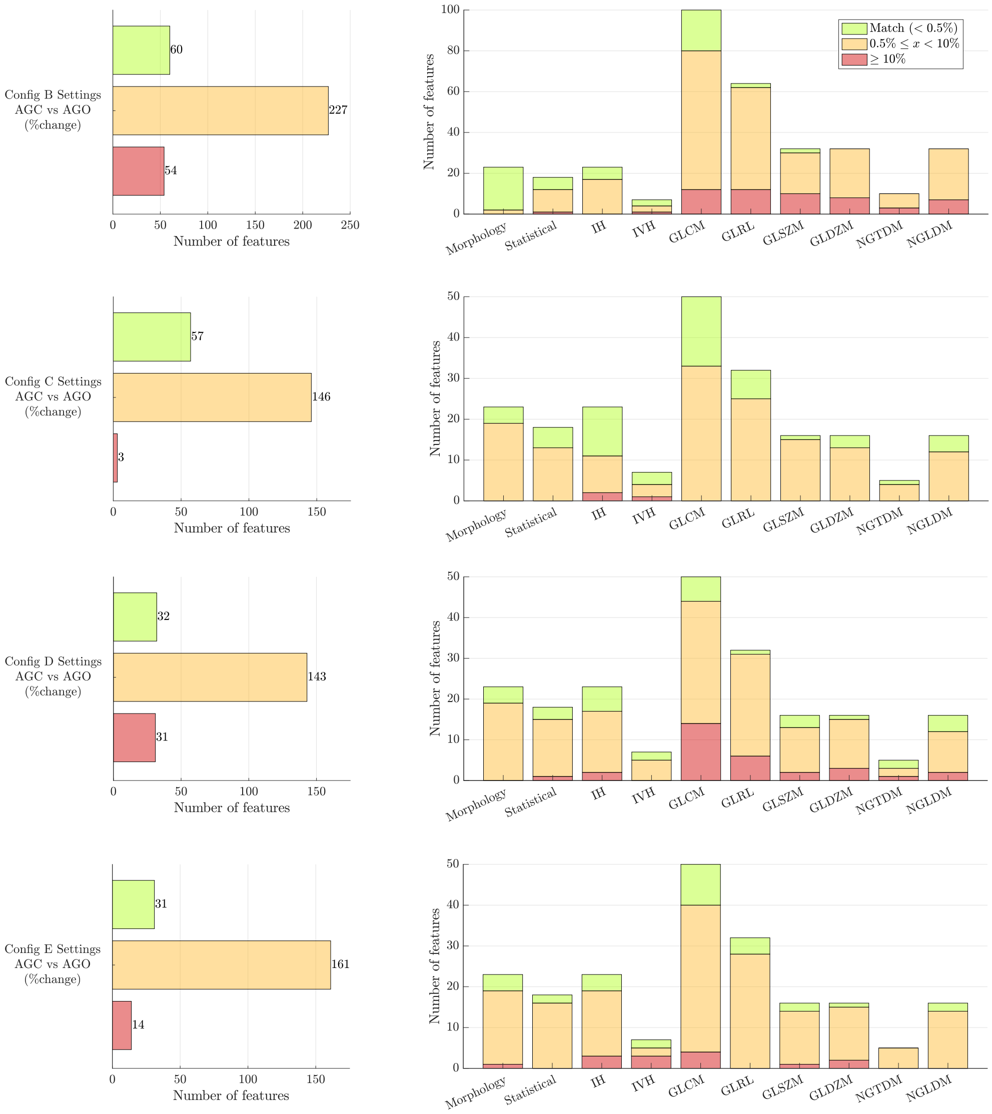


Figure 3.9: Summary of the absolute percentage change between features extracted using the AGC method compared to AGO, using `Dataset-2` and the *config B-E* settings from Table 3.2. For this plot, features were split into 3 groups: green ($< 0.5\%$), orange ($0.5\% - 10\%$), and red ($\geq 10\%$). Each row corresponds to a configuration, the left plot is the total distribution of features, the right is this distribution separated into the feature families, where the 3 groups are presented in a stacked bar chart. The majority of features varied between $0.5\% - 10\%$.

Two discretised intensity ROIs, obtained from different grid alignment methods using configuration C, are visualised in Figure 3.8. Through closer qualitative inspection one can begin to see the variation due to resampling at slightly different grid coordinates. In this instance, the most obvious difference is that the AGC method happens to result in an additional slice, though subtle differences between every slice.

Figure 3.9 contains results of the absolute percentage change in feature values for each configuration. The results are displayed in total as well as divided into feature families. In total, the following number of features had less than a 0.5% change when extracted with an AGO compared to AGC: *config B* 60/341 (17.6%); *config C* 57/206 (27.7%); *config D* 32/206 (15.5%), and *config E* 31/206 (15%). The majority saw variation between 0.5% and 10%.

Correspondingly, shown in Figure 3.10 is the number of features that still fell within compliance when considering the additional tolerance. The generation of this tolerance was discussed in Section 3.2.2.6. Evidently, the compliance of many features is adversely effected when using an AGO approach. In isolation, using the AGO method is enough to severely effect IBSI compliance of software.

Despite the adverse effect on IBSI compliance illustrated in Figure 3.10, this did not translate to a low measure of robustness with the ICC analysis of a cohort of patients (`Dataset-3`). In fact, as shown in Figure 3.11, nearly all features were excellently robust in this case for all 3 imaging modalities (CT: 163/165; PET:162/165; and MRI: 164/165).

Figure 3.12 highlights a plot of results from a single feature, for both methods, for all patients. It is an attempt to visualise why differences from the two methods for any 1 patient did not lead to a low robustness score across a cohort, using as an example the *GLDZM large distance emphasis* features extracted from the CT modality. For this particular feature, the percentage change from AGC to AGO had a median of 1.9% across the 51 patients, though went as high as 23% in one case. However, the ICC was 0.9987 (lb=0.9978 ub=0.9993), an excellent robustness score.

3.3.2.3 Discussion of AGC vs AGO

Early in the IBSI study, with the introduction of Phase 2, a likely source of discrepancy was identified in the interpolation step with the generation of a new resampling grid. Until this point, to the best of the author's knowledge, within the literature this had not been previously considered as a potential source of error. Teams within the IBSI had approached this with 2 main methods, aligning the new grid with the centre of the previous grid, or aligning the origins. Cardiff were one of several teams that originally implemented the AGO approach.

In the interest of standardisation the AGC method became the IBSI recommendation. Both are valid methods, though the main argument for AGC is that it is rotation invariant. With any axis arrangement, the centre of the image is always the centre, and the new grid generated will be the same. Where as with AGO, when implementing, one picks a "corner" of the image volume and generates a new grid from that position (refer back to Figure 2.5). As a result, flipping the volume and picking another corner leads to slight grid differences, and therefore different resampled volumes as confirmed in this analysis. However, one could make the counter argument that in the case of medical imaging, using the DICOM header information, one should be able to orientate the image and pick the same corner consistently. Despite this, AGC is seemingly the more elegant solution as there is no potential ambiguity.

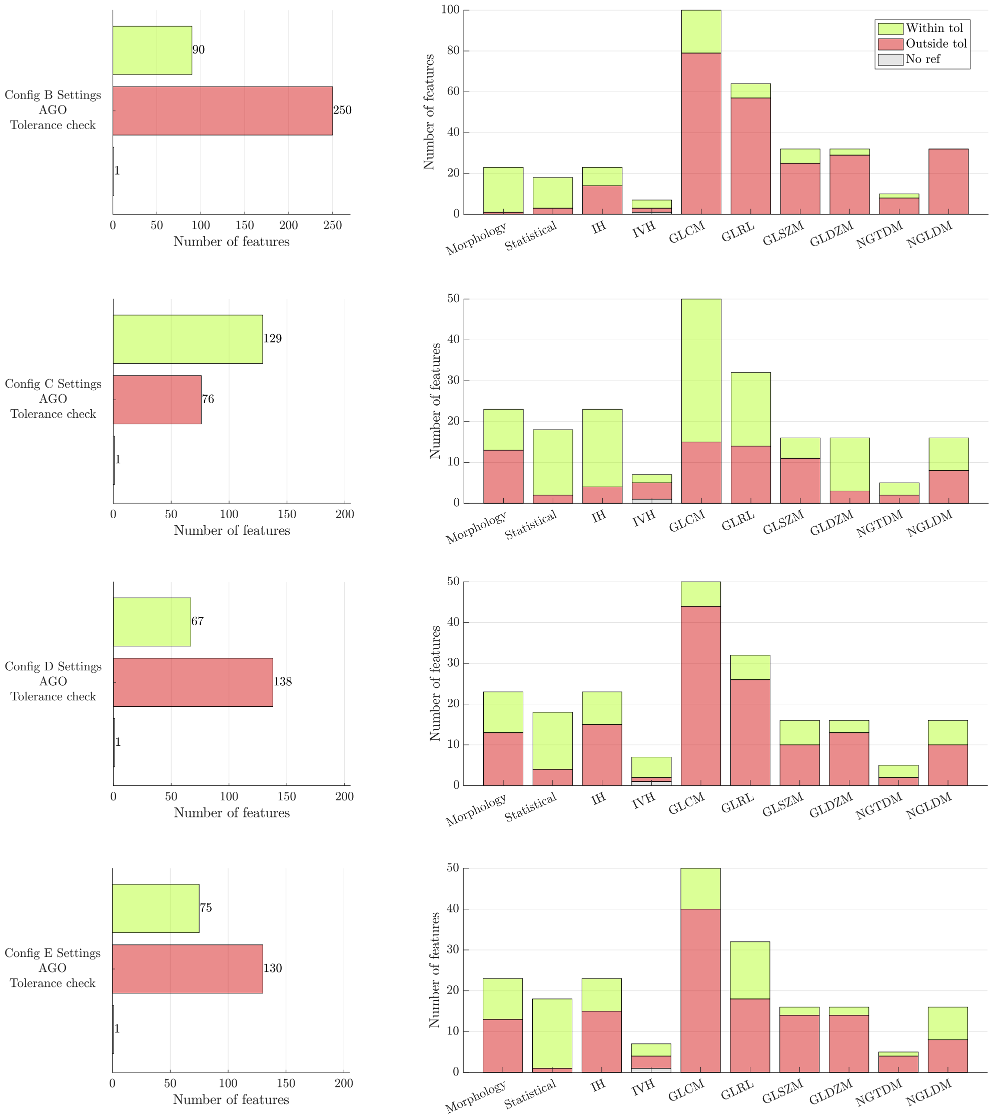


Figure 3.10: Summary of the compliance of features to the IBSI standard when extracted using the AGO method. Extracted using `Dataset-2` and the *config B-E* settings from Table 3.2. For this plot, features were split into 3 groups: within tolerance (green) , outside tolerance (red) , and no valid reference value (grey). Each row corresponds to a configuration, the left plot is the total distribution of features, the right is this distribution separated into the feature families, where the 3 groups are presented in a stacked bar-chat.

As such, AGC became the standard and for completeness this section attempted to quantify how the alternate method affected compliance. Based on teams in the IBSI that originally used AGO, one can infer that many previous radiomics studies have utilised the AGO approach. Figure 3.9 and 3.10 show that variation was significant enough to cause many features to deviate to the point that they were no longer considered standardised. The leniency provided by the tolerance in some cases allowed features that varied more than 0.5% to still be compliant. Hence, more features are seen in the top category in Figure 3.10, compared to Figure 3.9, for each configuration.

Importantly, a key takeaway for this study is that when considering feature robustness across a patient cohort with `Dataset-3`, small discrepancies caused by the grid difference appear not significant. This was highlighted with an example feature in Figure 3.11. For the settings and imaging utilised in this case, the difference between grid methods did not harm reproducibility

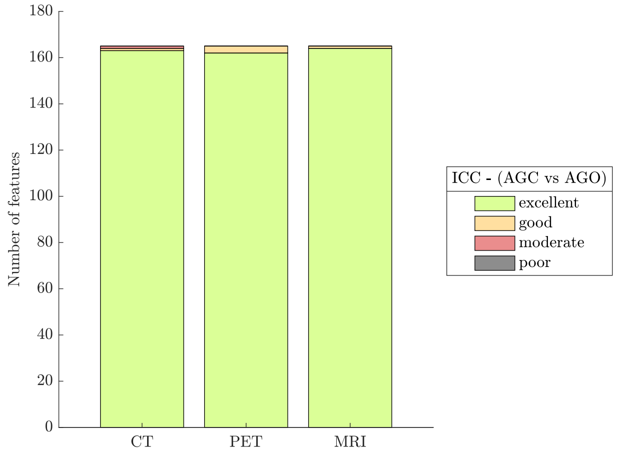


Figure 3.11: Examining the robustness of 164 standardised radiomics features to AGO and AGC grid generation methods when interpolating. Using both methods, features were extracted twice from `Dataset-3`. Features from the two methods were compared using ICC (two-way, random effects, single rater, absolute agreement). Each feature was categorised based on the lower bound of the 95% confidence interval. Excellent: $0.9 \leq$ ICC-LB ; good: $0.75 \leq$ ICC-LB < 0.9; moderate: $0.5 \leq$ ICC-LB < 0.75; poor: ICC-LB< 0.5.

of features across the cohort. One can conclude here that a radiomics study should not be affected by the grid generation method and that for a robust radiomics model, the same signal would likely arise had either AGO or AGC methods been used. However, one cannot be sufficiently compliant without utilising the AGO approach.

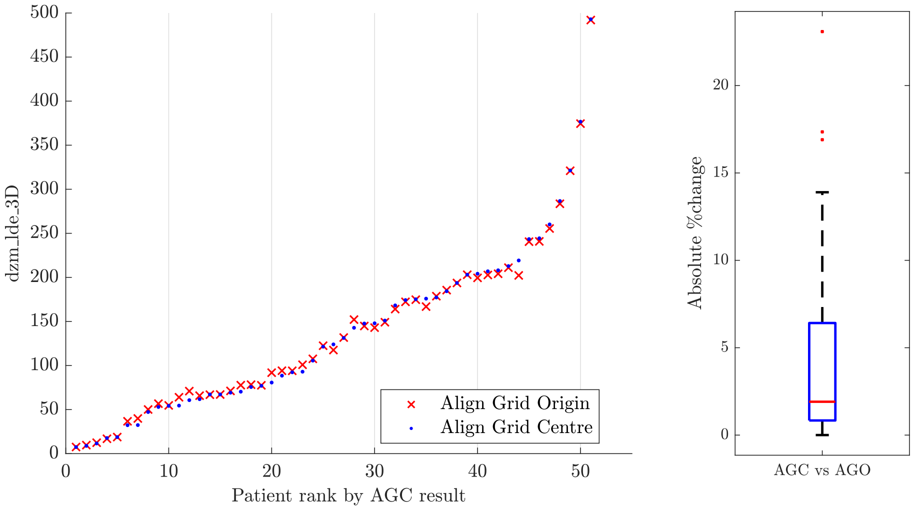


Figure 3.12: Comparing the results for AGO vs AGC for the feature *GLDZM large distance emphasis*. This example is using features from CT scans of `Dataset-3`. Left plot: Patients were assigned a rank based on the AGC result. Both results for each patient were plotted against the rank. Right plot: a box plot of measured %Error from AGC to AGO (AGC is taken to be the ground truth).

3.3.3 Re-segmentation Discrepancy

Re-segmentation of the VOI was another processing technique that required more rigorous definition to solve discrepancies between implementations. The technical details behind re-segmentation were introduced in Section 2.3.4 of the previous chapter. To summarise, it is an option to further refine the voxels within the VOI by removing those outside a range, or that are considered outliers based on standard deviation and mean intensity. As discussed, the image processing scheme evolved with the introduction of two distinct masks explicitly for re-segmentation: the morphological mask and the intensity mask. The following Section 3.3.3.1 expands on and assesses the case for 2 masks defined to settle the discrepancies. Assumptions now made by radiomic implementations that split the masks are also discussed.

Furthermore, as there are multiple re-segmentation methods, how an implementation combined approaches - e.g when range and outlier re-segmentation are used together - also proved to be another discrepancy point. The Cardiff implementation originally applied the methods in sequential order if more than one method was selected. With the observations from this analysis, other potential quality checks for future radiomics studies that utilise re-segmentation options are also discussed.

3.3.3.1 Mask separation: 2 Masks vs 1 Mask

Discrepancy arose between implementations - before the image processing scheme was redefined with the separate masks - when some implementations tried to fill small gaps that appeared within the VOI and others did not [84]. The Cardiff implementation originally followed the latter, with no attempt to patch these holes.

As re-segmentation can remove voxels, potentially this alters the shape of the resulting VOI. The more aggressive the re-segmentation, the more dramatic the change in shape. This is demonstrated with `Dataset-2` for a selection of different re-segmentation ranges ([-1000, 400] HU, [-500, 400] HU and [-200, 400] HU) as shown in Figure 3.13. The resulting VOIs are visualised with both the mesh-based and voxel based representation. Note these ranges are merely for illustration and do not necessarily represent realistic or useful re-segmentation ranges for radiomics modelling. Figure 3.13 shows the appearance of disjointed small groups of voxels as more aggressive re-segmentation to the lower densities is applied. Clearly re-segmentation effect is VOI specific; dependant on the tumour type and quality of the initial segmentation, but in this case with `Dataset-2`, which is a manually contoured GTV of NSCLC, the segmentation contains some air and lower density tissue, so the disjointed removal of air can be seen clearly.

Disjointed VOIs with internal holes in general do not often represent a logical shape to define with single value morphological features, and rather than attempting to refill some of these small holes once again after re-segmentation, which adds further complexity, the IBSI recommendation settled on assuming that the input mask already accurately defines the shape, and that re-segmentation is for refining analysis of the internal grey level intensity distribution. Hence, the introduction of two masks to remove any confusion. It should be clearly understood that standardised radiomics pipelines for morphological and GLDZM features assume accurate segmentations at the input, and if a researcher wants to adjust the segmentation for the shape based

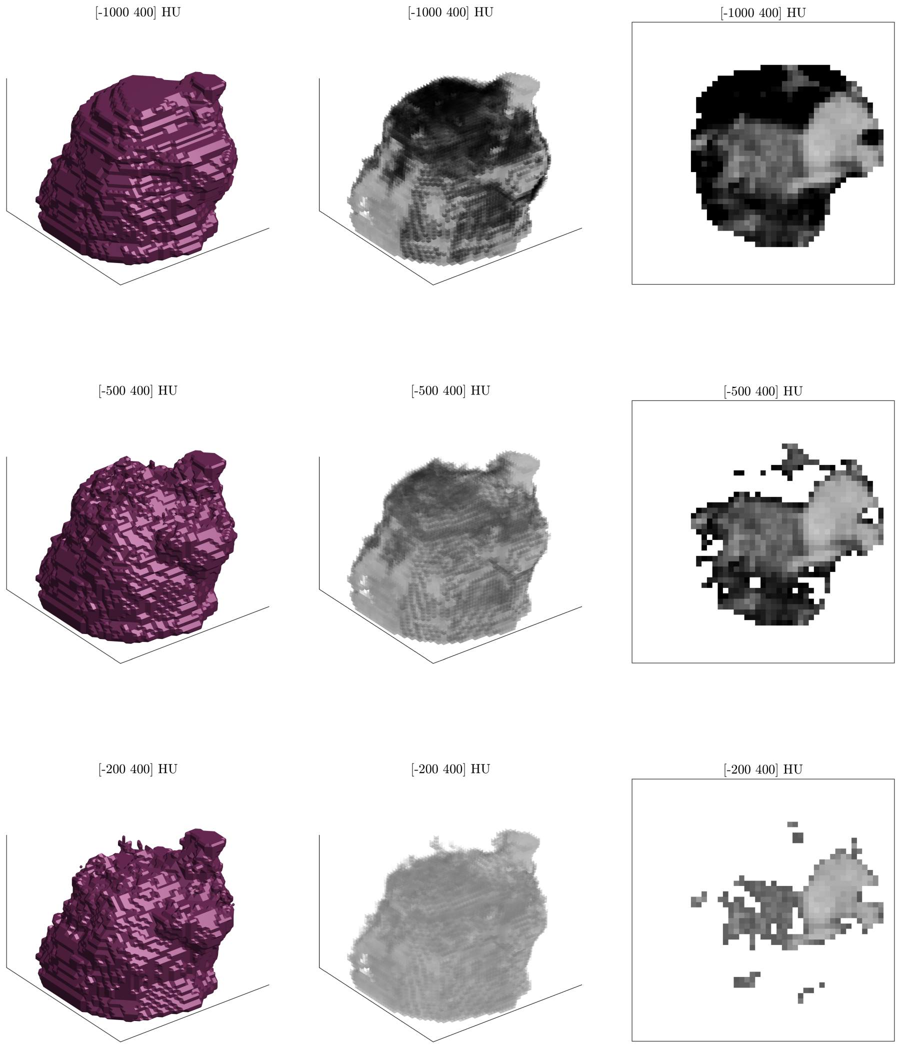
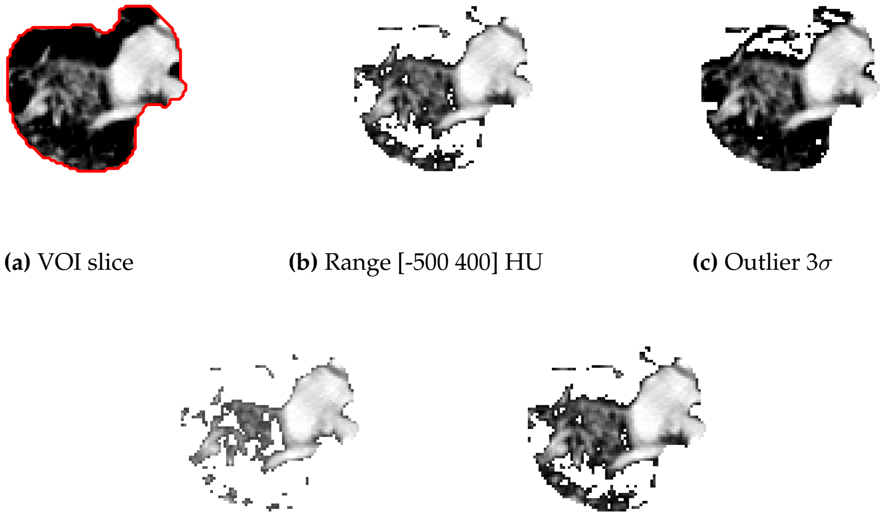


Figure 3.13: Visualising the effect of different re-segmentation ranges on the shape of the VOI from `Dataset-2`. Top row: [-1000 400] HU ,middle row [-500 400] HU , bottom row [-200 400] HU. The left column is the 3D mesh-based representation, the centre column is the 3D voxel based representation, and the right column is then a single cross sectional slice (no. 10) from the VOI on each row. Before re-segmentation was applied the VOI was linearly interpolated to isotropic 2mm voxel dimensions. The appearance of holes and disjointed voxels with re-segmentation can be seen clearly, and this can make morphological and GLDZM features less interpretable. As such, the IBSI introduced a separate morphological mask that is not re-segmented.

features, there needs to be an additional step as part of pre-possessing images, which of course needs to be clearly reported.

3.3.3.2 Using Multiple Re-segmentation Methods

In principle one can stack different re-segmentation methods. This was benchmarked with *config E* which tested both range and outlier re-segmentation on the same extraction. During participation in the IBSI collaboration, an ambiguity arose as it was unclear if software should apply them in a sequential order, or in parallel and then take the intersection. The effect is illustrated in Figure 3.14. As sequential application of methods changes the result depending on the order, the



(a) VOI slice **(b)** Range [-500 400] HU **(c)** Outlier 3σ

(d) Range, then outlier **(e)** Intersection

Figure 3.14: Visualising the result of range and outlier re-segmentation methods if both are used. Cardiff implementation was one of several in the initiative that originally applied the methods sequentially. Shown is: **(a)** a single slice from the VOI of `Dataset-2`; **(b)** the VOI re-segmenting to a range [-500 400] HU; **(c)** the VOI after removing outliers above/below 3σ ; **(d)** the VOI after applying the range then outlier methods sequentially; **(e)** the VOI when considering the intersection of both methods.

parallel approach was selected as the standard. This is the default action in SPAARC.

3.4 Further Discussion

Through a large international collaborative effort of 25 teams that formed the IBSI, a set of consensus based radiomic reference values were developed that enable calibration of different software to a new standard. Benchmarks were determined for 169/174 prominent features. Many of these features are commonly utilised in radiomic studies, including morphological, first-order and texture-based approaches to the analysis of radiological imaging. Afterwards, the reproducibility of these now standardised features was validated between software implementations with a separate batch extraction on a multi-modal dataset. As described in Chapter 1, a primary challenge of radiomic studies has been insufficient study reproducibility and inconsistent methodology [10]. This standardisation effort represents a major step towards resolving some of the issues surrounding this reproducibility concern, and addressed a key hypothesis and objective of this thesis project.

This study offers any future researcher or clinician the ability to validate the software they are using by extracting features - with the relevant settings and datasets - and comparing to the now published benchmarks [84]. It also acts as a principle unit test for software as it is expanded, patched and improved. For completeness, all of Cardiff's final results for each feature variant and configuration are included in Appendix C and are used as unit tests in SPAARC's continuing development. It is important that implementations *remain* standardised as they are updated.

Prior to any radiomics study software should always be re-checked for compliance, and this preliminary check confirmed in the literature moving forward.

There were six unique programming languages used between the teams (see Appendix A). Matlab was the most popular, and the one utilised for the author's implementation. The other programming languages used by teams were C++, Python, R, Java, and IDL. As this was a consensus based study, it was important to ensure there was not a language bias in reference values. For example, as discussed in Section 2.3.3, SPAARC uses *interp3*, an inbuilt Matlab function [97] for interpolation. It is likely that most Matlab radiomics implementations utilised this convenient function available readily within the language. Therefore, one could imagine that a reference value has the potential to be linked to a particular language's inbuilt functionality (such as *interp3*), and that other teams only deviated due to using another "standard" inbuilt function within the respective language they utilised. Another potential for error could arise when using already defined functions within a language, if they are not properly tested, as they may not align with the IBSI definition. As an example, the kurtosis feature from both the statistical and intensity histogram feature families standardised here is actually formally defined as *excess* kurtosis [84], which adds a -3 correction to the definition to centre the value on 0 for a normal distribution, where as Matlab's inbuilt kurtosis implementation does not incorporate this correction by default.

Importantly for this study, there was no dependence found between standardised reference values and any one programming language, as every feature variant that achieved a moderate or better consensus (`Measure-1`>3, `Measure-2`>50%) - and therefore considered standardised - had at least 2 teams using different languages that matched the consensus. This is also somewhat visually intuitive by assessing the coverage (Figure 3.5) and programming languages of the top 5 teams at the final time point. (MIRP = Python, MITK = C++, UMCG (Beukinga) = Matlab, Cardiff = Matlab, RaCaT = Python).

It should also be considered that within the IBSI there were two institutions that had two different teams submitting values (Brest (BCOM) and Brest (MaCha), and, UMCG (van Dijk) and UMCG (Beukinga)). Both teams within each institution utilised the same language, so feasibly they could have shared code, which would have undermined the consensus based approach to this research. This is unlikely to be the case, as observation of evolving contributions by these teams show clear differences in feature values. These teams submitted different numbers of feature variants across different phases and configurations. Even if these teams had used the same code for some features, this would have had a very limited effect on overall consensus and standardised values. Every feature variant with moderate or better consensus had results from at least 3 different institutions. The 5 top submitting teams were from 5 different institutions.

Cardiff was a major contributing team to this work, as highlighted in Figure 3.5 and Figure 3.7, joining the other core members of the group. In the course of implementing processing options and providing values for over 95% of the feature variants for all phases of the initiative, SPAARC results in particular boosted the lacking strength of consensus on many of the more uncommon features variants and setting configurations, such as the 2.5D texture-based results. Cardiff were one of the top submitting teams from early on in the initiative as highlighted in Figure 3.4. The initial weak agreement between software arose from a variety of discrepancies, particularly surrounding the image processing prior to the actual feature extraction. This led to further necessary revision of the processing scheme for increased clarity. It is valuable to understand the impact of these discrepancies on feature values, as they were likely present in many prior radiomics studies reported in literature. As teams often addressed multiple discrepancies at once, or addressed them at different points in the initiative, it was difficult to assess their individual severity. Rather, the goal was simply to identify differences and update methods to harmonise the outputs of different implementations.

Section 3.3 further evaluates some key discrepancies in isolation that were encountered when developing the radiomics extraction pipeline of SPAARC. In a comparison of features extracted from both interpolation grid generation methods, analysis showed that, alone, grid choice was a major factor affecting compliance with final reference values developed with `Dataset-2`. However, it did not appear to affect feature reproducibility across `Dataset-3` significantly, suggesting the choice of AGO or AGC would have had little consequence on the outcome of any radiomics model developed on that cohort. Individual variation in patients depended on a their tumour location and size. Yet, analysis found any individual variation was not significant compared to the variation amongst the patients within the cohort, at least for the validation dataset that was utilised for the consensus study. As a result of this sub-study, one can suggest that a simple test of model generalisability can be performed with software that can extract with both grid methods. If the features, and therefore the model, changes significantly when the only difference is the alignment of the resampling grid, it is unlikely that the model will be very robust or generalisable.

Furthermore, Section 3.3 detailed the discrepancy caused by re-segmentation approach, and why the strict definition of 2 masks in the image processing scheme was required to reach consensus between teams. Additionally, this also visually emphasised the impact of different re-segmentation thresholds (Figure 3.13). Particularly for `Dataset-2`, which is a lung tumour, the re-segmentation range selected introduces disjointed voxels in the VOI. When performing radiomic analysis and utilising a re-segmentation range, one should consider if subtle differences in a selected range (e.g. [-195 205] HU instead of [-200 200] HU) are enough to significantly alter the predictive nature of any radiomics model developed on that cohort. In each case, this is highly dependent on the quality of the segmentation and tissue type being examined.

SPAARC includes every baseline texture feature bar one from the NGLDM family, as shown in Table 3.1. This missing feature is named *dependence count percentage*. As discussed in Chapter 2, Section 2.4.11, the definition of NGLDM used here considers all voxel neighbourhoods, not just those that are complete. Essentially, the feature *dependence count percentage* is a fractional measure of the number of neighbourhoods over the maximum number of potential neighbourhoods. Under the adapted definition of NGLDM, as partial neighbourhoods are included, this feature always evaluated to 1. As such, it can never offer any descriptive information. This feature was thus removed from SPAARC by default, though it remained within the initiative (and was standardised) for continuity in the study. In general, this feature should just be omitted from other implementations moving forward.

The extraction settings that were tested throughout this Chapter (Table 3.2) are not necessarily optimal choices for radiomic analysis, but in their variety help to identify different types and causes of discrepancy. Much work still needs to be done concerning the optimal extraction settings, which is at the very least modality and tumour site specific. This work enables more confidence in the robustness studies performing this type of analysis moving forward. In this regard, Chapter 4 explores an aspect of this further with an investigation of interpolation method and voxel size dependence of features.

When assessing this standardisation effort, there are several limitations to consider. Firstly, the scope of the feature set was not exhaustive. Effort was made to ensure the common texture analysis techniques were chosen, though features such as those derived from *fractal*-based analysis [2, 96], as an example, were not included. Also not currently considered was image filtering, and filter-based measures such as wavelet features (e.g [111]). Filtering is a major processing component of radiomic analysis, and Chapter 5 investigates the unique challenges of filter based radiomic features, which remains an ongoing standardisation aim.

As shown, even within the feature set considered in this study, the coverage varied greatly amongst teams. This speaks to the challenge of implementing many of these features and processing steps required for the variants of each feature. Clearly, a greater coverage from all teams would have led to potentially stronger consensus for many more features. It should be noted that several teams only provided values in the first stage of the initiative, and as evident in Phase 2, this is not sufficient in itself to meet the standard. Reporting compliance to the standard in literature necessarily requires testing software with the relevant configurations from Phase 2 as well. Alongside this, although extraction from multiple imaging modalities was evaluated in Phase 3, certain modality specific processing steps - for example, PET conversion from BQ/ML to SUV - were not tested between implementations. Potential discrepancy could arise here and this would be mitigated with appropriate conversion tests between teams on a designated reference phantom.

3.5 Concluding Remarks

Through a large collaborative effort with the IBSI, a set of validated reference values for a wide variety of radiomic features and extraction variants were developed. This work addressed a critical need within the field of radiomics for benchmarks, and will greatly facilitate study reproducibility moving forward. Furthermore, this chapter explored in depth several key underlying discrepancy causes between software, and analysed potential impact on standardisation compliance.

Take home message

1. Stardardisation was necessary for lacking study reproducibility in radiomics.
2. The IBSI developed strong consensus based reference values for a large number of feature variants and extraction methods.
3. Different grid definitions for interpolation affect compliance to the IBSI standard. Aligning the grid centres is recommended.
4. Thresholding techniques can dramatically alter extraction. The order of multiple thresholding methods is critical.

4

Feature Response to Isotropic Interpolation in ^{18}F-FDG PET Imaging

"All models are wrong, but some are useful"

— George E. P. Box

4.1 Preview

Work from this Chapter was published by Whybra *et. al.* [85]. This study utilised SPAARC's standardised radiomics pipeline to assess feature robustness to isotropic voxel interpolation using a large cohort of patients with oesophageal cancer who underwent ^{18}F-FDG PET imaging with the same protocol, acquired during routine staging. For 3D radiomic analysis, interpolation is a key component of the image processing pipeline. The goal of this study was to assess, categorise and potentially model feature response to interpolation, to further analyse feature stability in PET radiomics. The feature categorisations obtained aim to inform feature reduction techniques for future studies to produce more generalisable models.

4.1.1 Author Contribution

To reiterate the contributions described in Section 1.9, the author primarily designed the study discussed in this chapter, performed all the experimental data analysis, and drafted the original published manuscript [85].

4.2 Introduction

It is vital in radiomics to ensure features that may show prognostic or predictive value are first suitably robust to image processing steps. As discussed in both Chapter 2 and 3, a major step of most radiomics analysis is interpolation (also known as *voxel size resampling*) of the scan to a new coordinate system. Three key uses of interpolation in medical imaging are [85]: (1) to compare and combine datasets that come from multiple centres with varying protocols that result in different voxel sizes [95], (2) to resample multi-modal imaging to a common voxel size (e.g. such as with PET/CT) [71], and (3) to acquire isotropic voxels for 3D radiomic analysis [96].

Previous studies [95, 114–116] have looked at image interpolation effects on radiomics, primarily in the context of CT, and primarily when resampling to anisotropic voxel sizes. This study focused on isotropic resampling of a large PET dataset using newly standardised radiomic features. As discussed in Section 2.3.3, it is routine for medical scanners to produce anisotropic imaging. Typically, after reconstruction the axial resolution is higher than the slice thickness (i.e. $(\Delta z > (\Delta x, \Delta y))$. This is a concern for 3D texture analysis as it introduces a directional bias [90, 96]. To establish the same scale in all 3 image axis, it is recommended for texture analysis to resample with 3D interpolation such that $\Delta x = \Delta y = \Delta z$ [96]. The scale at which one extracts the features may also impact predictive power [117, 118].

Texture features quantify the spatial variation in voxel intensities. From the reconstructed voxel size, with interpolation one can either up-sample to a higher spatial resolution, or down-sample to a lower one. Down-sampling leads to information loss, whereas up-sampling creates artificial information to attain the higher resolution. At the voxel level, heavily down-sampling creates a lower quality image with less texture information, whereas heavy up-sampling creates local homogeneity as the image is smoothed out. This is demonstrated with an example in Figure 4.1, by visualising the central slice of a VOI extracted after re-sampling the scan different isotropic voxel sizes. When the goal is to measure tumour heterogeneity, understanding the effect of interpolation is critical. With each radiomics study, when interpolating, the selected voxel size is discretionary for the researcher, and the appropriate re-sampling size compared to the original reconstructed voxel size, remains an open question. It may be the case that optimal modelling in radiomics utilises features extracted from imaging interpolated to a range of voxel sizes with a process refered to as texture optimisation [118].

With medical scans the image texture is linked to the image acquisition scale, and interpolation is an estimate of the image intensity distribution at a new scale. As interpolation is a major component of the processing pipeline, it is necessary to understand how features vary with it. In particular, this study assessed if features remained consistent, or responded in a systematic or unstable way to up-sampling the image. If the feature response is consistent over a range of voxel-sizes, then there would be redundancy in extracting these features at multiple resampled scales, and if feature extraction is only performed at one voxel dimension, then features that are stable to the process are preferable.

The main purpose of this study was to measure the response of standardised radiomic features to isotropic interpolation, for a range of voxel sizes, using a large oesophageal cancer (OC) patient cohort who all underwent the same scanning protocol. Feature values extracted after both linear

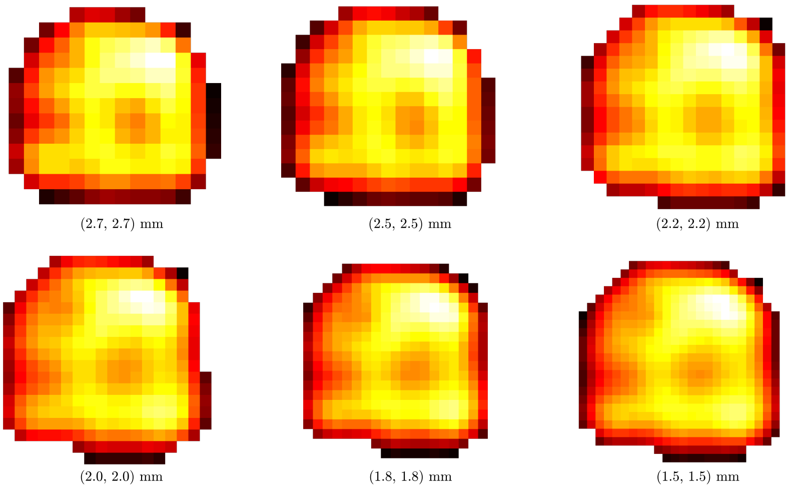


Figure 4.1: Example showing the effect of linear interpolation on an extracted volume of interest (an oesophageal tumour) in PET imaging. (Example is from patient ID 1052 from this study.) Six different isotropic voxel sizes (1.5 mm, 1.8mm, 2.0mm, 2.2mm 2.5mm, 2.7mm) are shown. This visually demonstrates the smoothing effect of up-sampling the image with interpolation.

and spline interpolation methods were investigated and compared to see if the stability differed. Features were categorised based on the responses - i.e. how the values varied when extracted at each voxel dimension - and apparent systematic tendencies were modelled to explore potential correction factors.

4.3 Materials and Methods

4.3.1 Imaging

This work used a retrospective highly curated dataset of OC patient imaging, that was collated by the author's research group as part of previous studies [17, 76, 87]. This cohort of patients have either adenocarcinoma or squamous cell carcinoma biopsy OC, and had undergone PET/CT imaging as part of the diagnostic staging pathway. In collection of the original dataset there were several exclusion criteria: patients were excluded if the primary tumour was classed as poorly FDG-avid (SUV_{max} <3); if they had a primary synchronous tumour on PET/CT (i.e. multiply primary tumours); an oesophageal stent *in situ*; or incomplete histology (or that differed from adenocarcinoma or SCC). All imaging was acquired with the same PET protocol. Voxel size after PET reconstruction was 2.73 × 2.73 × 3.27 mm. Before segmentation, there were 465 patients considered for inclusion in this study.

4.3.2 Segmentation

ATLAAS, as introduced in Section 1.5.2, was used to outline the metabolic tumour volume (MTV) on the PET for each patient. As a semi-automatic approach, a bounding box containing the tumour is first decided manually as a guide for contour generation. All contours were created on the original image dimensions. With a binary score, these segmentations were then rated by an expert radiologist as either acceptable (1) or unacceptable (0). Of all patients segmented, 441/465 were deemed acceptable representations of the MTV and used as the dataset for this study.

4.3.3 Interpolation and feature extraction settings

With the SPAARC software, standardised feature extraction was carried out following the new IBSI recommendations. To study response of features to interpolation, 12 unique extractions were performed on the dataset, using 6 different voxels sizes (2.7mm, 2.5mm, 2.2mm, 2.0mm, 1.8mm, 1.5mm), for 2 different interpolation methods (linear and spline). Once selected, all other extraction settings were kept constant. Standardised features analysed in the main study include: Morphological (22), Statistical (18), Intensity Histogram (23), GLCM (25), GLRLM (16), GLSZM (16), GLDZM (16), and NGTDM (5). Only 3D features variants were considered as they require isotropic interpolation. For the GLRLM and GLCM features, the aggregation method was set to merging (see Section 2.4.6.1). For discretisation, a FBW of 0.5 SUV was used.

4.3.4 Statistical Analysis

ICC analysis (Section 2.5.4) was used to assess the reproducibility of features when extracted at different voxel dimensions. A *2-way random-effects, single rater, absolute agreement* version of ICC was calculated. For ICC, absolute agreement was selected over consistency for categorisation purposes as a feature was considered perfectly stable to interpolation if the same value was reproduced when extracted at different voxel sizes. ICC> 0.9 was used as the threshold for robustness. The 95% confidence intervals were also reported alongside the results. ICC was calculated utilising the R language with the statistical package irr (version 0.84.1) [119].

For each feature, the consistency of patient ranking after interpolation was also assessed using a spearman rank correlation coefficient (Section 2.5.2), similar to the method utilised by Leijenaar *et*. *al*. [102]. Here, the results from the 2.7mm extraction were taken as the ground truth and compared pairwise with the other 5. The 2.7mm results were selected as the ground truth as that voxel size was the closest to the original axial plane dimension of the images, and the target dimension for rescaling models (introduced in Section 4.3.5). Using the mean result of all pairwise tests, a threshold of $\rho_{\text{mean}} > 0.95$ was used to indicate that high ranking consistency remained after interpolation to different voxel sizes.

Using both statistical tests, features were categorised using the method summarised in Figure 4.2. Robust (R) features were those above the threshold for both tests. Limited Robustness (LR) features had high ICC, yet were below the threshold for ρ_{mean}, suggesting some limited discriminative value for these feature as the patient ranking was variable. Potentially correctable (C) features scored below the threshold for ICC, yet above the threshold for $\rho_{\text{mean}} > 0.95$. The fact that the patient ranking remained consistent suggested features in this category might have a dependence on interpolation that could be modelled. Finally, features that were below the threshold for both statistical tests were categorised as not robust (NR). The robustness to voxel size resampling was performed twice, once with the set that were linear interpolated, and once with the

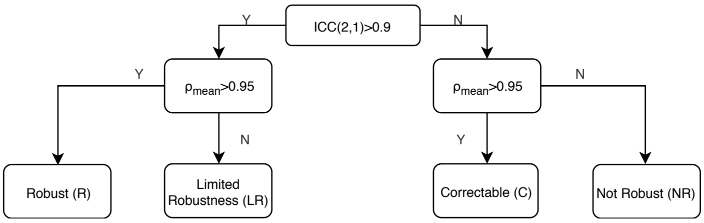


Figure 4.2: A flow diagram of categorisation criteria used in this study. Features were grouped with two statistical tests: intraclass correlation coefficient (ICC) > 0.9, and patient ranking correlation with spearman rank correlation coefficient $\rho_{\text{mean}} > 0.95$. This flow chart of categoristaion methodology was published by Whybra *et. al.* [85].

set that were spline interpolated, to see if features were categorised the same way irrespective of the interpolation method used. To compare feature values *between* linear and spline methods, a Bland-Altman (BA) [(Method A - Method B)/average %] style analysis [107] was performed for each feature using the 2.7mm results. An individual BA plot was produced for each feature in but condensely summarised with box plots.

4.3.5 Modelling Potential Systematic Response To Interpolation

Features categorised based on Figure 4.2 as potentially correctable were considered for response modelling. These features were likely showing systematic change in value linked to voxel size. If responses could be modelled, the goal was to see if they were sufficient to re-scale features back to a selected voxel size, in this case the 2.7mm results. To establish a correction factor model, fits were produced linking the change in feature value and voxel size. The percentage change for each feature compared to the 2.7mm result was calculated. A polynomial surface was fitted representing the percentage change from the 2.7mm result, based on the current feature value and voxel size. A surface fit (S_f) could be used to re-scale a feature to a new, *surface shifted* value, that should correspond to the 2.7mm result, via the equation

$$f_{ss} = f - (S_f(f, V_s) \times f) \tag{4.1}$$

where f is the value of the feature extracted at voxel size V_s in mm^3, and f_{ss} is the corrected feature value. In this work the corrected feature was given the additional suffix "surfaceShifted". Matlab's curve fitting toolbox [52], was utilised to plot and visualise the surfaces in this work. This method is visualised and discussed more in Section 4.4.2.

4.3.6 Splitting the Cohort

As a secondary aim was to see if response to interpolation could be modelled, the patients with acceptable contours (n=441) were divided randomly (80%/20%) into *Robustness Testing* (n=353) and *Validation* (n=88) datasets. Each scan was interpolated to the 6 different voxels sizes stated above. Explicitly, there were $6 \times 353 = 2118$ and $6 \times 88 = 528$ VOIs analysed in the *Robustness Testing* and *Validation* sub-datasets respectively (for each interpolation method).

Intuitively, the main robustness analysis here was performed on the *Robustness Testing* dataset. With this dataset, for features that were deemed potentially correctable, the surface fitting procedure, described above, was used to model the relationship between feature value and voxel size. The ability of surfaces models to correct feature values, via Eq. 4.1, was then examined using the *Validation* dataset. I.e. The models were tested on data that they hadn't been generated on. Once the corrections using the surface fits were applied to the *Validation* dataset, features were assessed for reproducibility once again to see if the adjusted values were more stable between the different voxel sizes.

4.4 Results

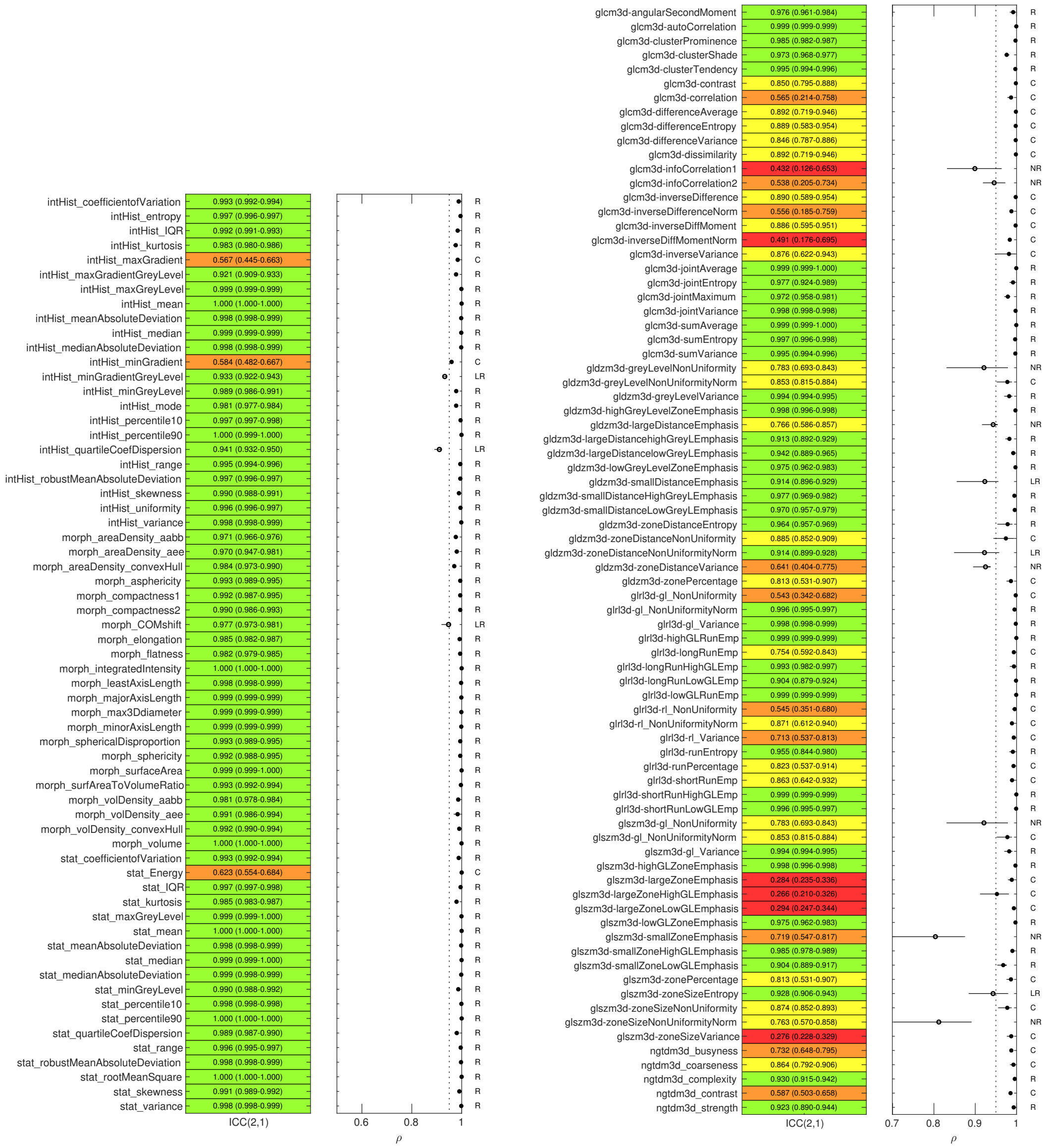
First, considering linear interpolation, utilising the criteria specified in Figure 4.2, features were categorised overall as: 93/141 Robust, 6/141 Limited Robustness, 34/141 Potentially Correctable, and 8/141 Not Robust. These results are summarised in Figure 4.3a and 4.3b by dividing the features into 2 groups. Group-1 (63/141) in Figure 4.3a contains results for the Morphological, Intensity Histogram, and Statistical family of features (categorised as: 57 R, 3 LR, 3 C, 0 NR). Group-2 (78/141) in Figure 4.3b contains the texture feature results (categorised as: 36 R, 3 LR, 31 C, 8 NR). These Figures colour code the ICC score as red (0.5<), orange ($0.5 - 0.75$), yellow ($0.75 - 0.9$), and green (>0.9). For each feature, the range of pairwise ρ is highlighted with a line, and the ρ_{mean} shown with a closed marker if $\rho_{\text{mean}} > 0.95$, and an open marker if below this threshold.

This robustness testing was then repeated using a spline interpolation method. Although slight variation in statistical test values were found for features, interestingly, the categorisation of all features responses remained the same between linear and spline. Figure 4.4 highlights these corresponding robustness results with the spline interpolation method.

4.4.1 Visualising Feature Variation Due to Interpolation

To visualise feature response, the feature values extracted at each voxel dimension were plotted against the patient ranking obtained using the 2.7mm result. In these plots, the 2.7mm results thus appear monotonically ascending, with the patient with the lowest value for a given feature at rank 1, the second lowest value at rank 2, and so on. Intuitively, for a robust feature where the extracted value is the same at each voxel size, results from all the other voxel sizes will overlap with the 2.7mm result. With this approach, one can qualitatively assess if the values appear to vary significantly or not when extracted at a different voxel sizes, and if this variation has a systematic quality. Here, these graphs are referred to as *response plots*. Exhaustive response plots of this nature were generated for every feature and provided in the supplementary materials of Whybra *et. al.* [85] for completeness.

Figure 4.5 shows specific examples of these visualisations for six features with a variety of responses to linear interpolation: a) *glcm-sumAverage*, categorised as Robust, b) *gldzm3d-smallDistanceEmphasis*, categorised as having limited Robustness, c) *glcm3d-correlation* categorised as potentially correctable, d) *glszm3d-smallZoneEmphasis* categorised as not robust, e) *glrl3d-rl_NonUniformity* categorised as potentially correctable, and f) the normalised version of *glrl3d-rl_NonUniformity*



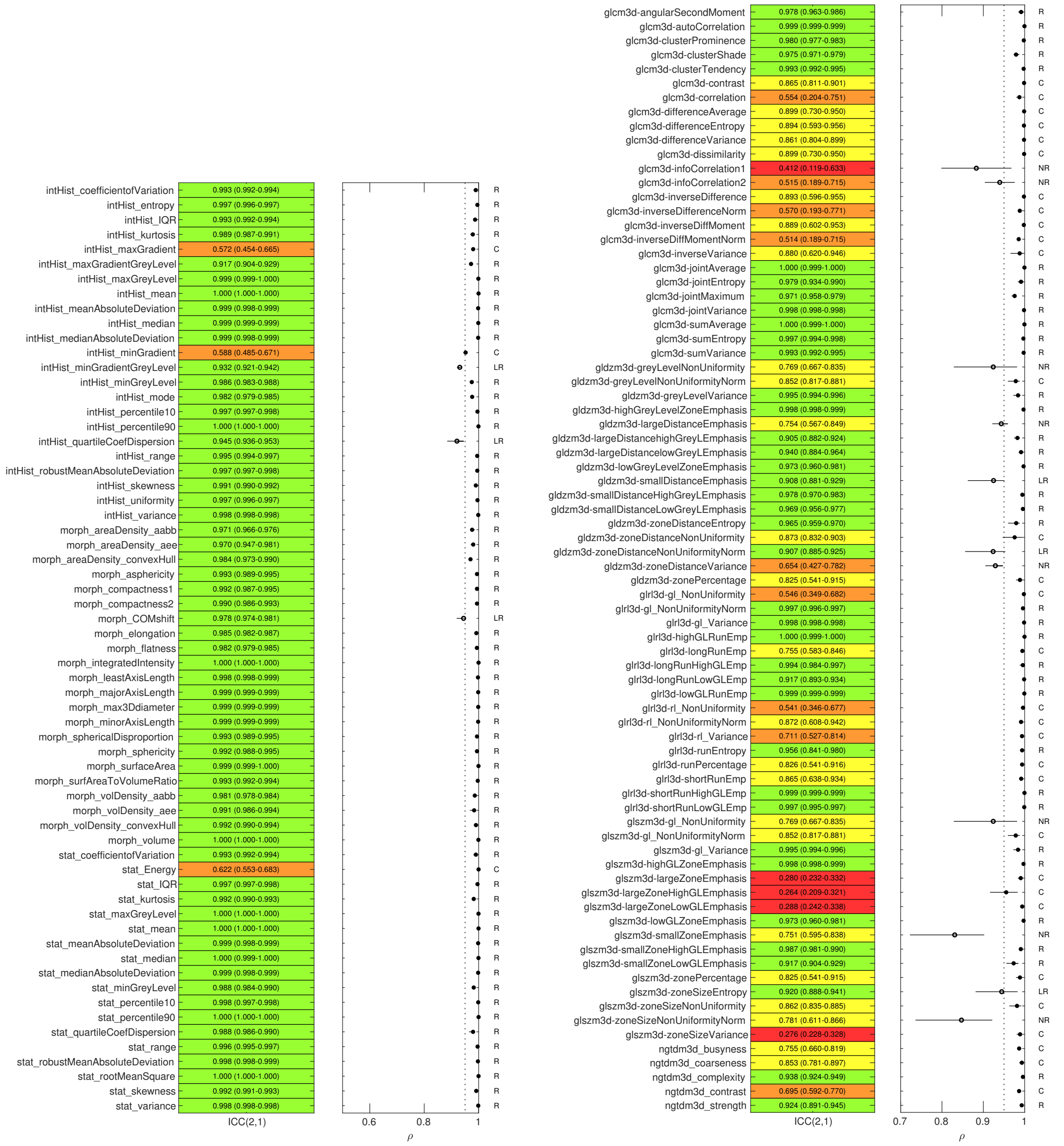
	ICC(2,1)	
intHist_coefficientofVariation	0.993 (0.992-0.994)	R
intHist_entropy	0.997 (0.996-0.997)	R
intHist_IQR	0.992 (0.991-0.993)	R
intHist_kurtosis	0.983 (0.980-0.986)	R
intHist_maxGradient	0.567 (0.445-0.663)	C
intHist_maxGradientGreyLevel	0.921 (0.909-0.933)	R
intHist_maxGreyLevel	0.999 (0.999-0.999)	R
intHist_mean	1.000 (1.000-1.000)	R
intHist_meanAbsoluteDeviation	0.998 (0.998-0.999)	R
intHist_median	0.999 (0.999-0.999)	R
intHist_medianAbsoluteDeviation	0.998 (0.998-0.999)	R
intHist_minGradient	0.584 (0.482-0.667)	C
intHist_minGradientGreyLevel	0.933 (0.922-0.943)	LR
intHist_minGreyLevel	0.989 (0.986-0.991)	R
intHist_mode	0.981 (0.977-0.984)	R
intHist_percentile10	0.997 (0.997-0.998)	R
intHist_percentile90	1.000 (0.999-1.000)	R
intHist_quartileCoefDispersion	0.941 (0.932-0.950)	LR
intHist_range	0.995 (0.994-0.996)	R
intHist_robustMeanAbsoluteDeviation	0.997 (0.996-0.997)	R
intHist_skewness	0.990 (0.988-0.991)	R
intHist_uniformity	0.996 (0.996-0.997)	R
intHist_variance	0.998 (0.998-0.999)	R
morph_areaDensity_aabb	0.971 (0.966-0.976)	R
morph_areaDensity_aee	0.970 (0.947-0.981)	R
morph_areaDensity_convexHull	0.984 (0.973-0.990)	R
morph_asphericity	0.993 (0.989-0.995)	R
morph_compactness1	0.992 (0.987-0.995)	R
morph_compactness2	0.990 (0.986-0.993)	R
morph_COMshift	0.977 (0.973-0.981)	LR
morph_elongation	0.985 (0.982-0.987)	R
morph_flatness	0.982 (0.979-0.985)	R
morph_integratedIntensity	1.000 (1.000-1.000)	R
morph_leastAxisLength	0.998 (0.998-0.999)	R
morph_majorAxisLength	0.999 (0.999-0.999)	R
morph_max3Ddiameter	0.999 (0.999-0.999)	R
morph_minorAxisLength	0.999 (0.999-0.999)	R
morph_sphericalDisproportion	0.993 (0.989-0.995)	R
morph_sphericity	0.992 (0.988-0.995)	R
morph_surfaceArea	0.999 (0.999-1.000)	R
morph_surfAreaToVolumeRatio	0.993 (0.992-0.994)	R
morph_volDensity_aabb	0.981 (0.978-0.984)	R
morph_volDensity_aee	0.991 (0.986-0.994)	R
morph_volDensity_convexHull	0.992 (0.990-0.994)	R
morph_volume	1.000 (1.000-1.000)	R
stat_coefficientofVariation	0.993 (0.992-0.994)	R
stat_Energy	0.623 (0.554-0.684)	C
stat_IQR	0.997 (0.997-0.998)	R
stat_kurtosis	0.985 (0.983-0.987)	R
stat_maxGreyLevel	0.999 (0.999-1.000)	R
stat_mean	1.000 (1.000-1.000)	R
stat_meanAbsoluteDeviation	0.998 (0.998-0.999)	R
stat_median	0.999 (0.999-1.000)	R
stat_medianAbsoluteDeviation	0.999 (0.998-0.999)	R
stat_minGreyLevel	0.990 (0.988-0.992)	R
stat_percentile10	0.998 (0.998-0.998)	R
stat_percentile90	1.000 (1.000-1.000)	R
stat_quartileCoefDispersion	0.989 (0.987-0.990)	R
stat_range	0.996 (0.995-0.997)	R
stat_robustMeanAbsoluteDeviation	0.998 (0.998-0.999)	R
stat_rootMeanSquare	1.000 (1.000-1.000)	R
stat_skewness	0.991 (0.989-0.992)	R
stat_variance	0.998 (0.998-0.999)	R

(a) Feature Group-1.

	ICC(2,1)	
glcm3d-angularSecondMoment	0.976 (0.961-0.984)	R
glcm3d-autoCorrelation	0.999 (0.999-0.999)	R
glcm3d-clusterProminence	0.985 (0.982-0.987)	R
glcm3d-clusterShade	0.973 (0.968-0.977)	R
glcm3d-clusterTendency	0.995 (0.994-0.996)	R
glcm3d-contrast	0.850 (0.795-0.888)	C
glcm3d-correlation	0.565 (0.214-0.758)	C
glcm3d-differenceAverage	0.892 (0.719-0.946)	C
glcm3d-differenceEntropy	0.889 (0.583-0.954)	C
glcm3d-differenceVariance	0.846 (0.787-0.886)	C
glcm3d-dissimilarity	0.892 (0.719-0.946)	C
glcm3d-infoCorrelation1	0.432 (0.126-0.653)	NR
glcm3d-infoCorrelation2	0.538 (0.205-0.734)	NR
glcm3d-inverseDifference	0.890 (0.589-0.954)	C
glcm3d-inverseDifferenceNorm	0.556 (0.185-0.759)	C
glcm3d-inverseDiffMoment	0.886 (0.595-0.951)	C
glcm3d-inverseDiffMomentNorm	0.491 (0.176-0.695)	C
glcm3d-inverseVariance	0.876 (0.622-0.943)	C
glcm3d-jointAverage	0.999 (0.999-1.000)	R
glcm3d-jointEntropy	0.977 (0.924-0.989)	R
glcm3d-jointMaximum	0.972 (0.958-0.981)	R
glcm3d-jointVariance	0.998 (0.998-0.998)	R
glcm3d-sumAverage	0.999 (0.999-1.000)	R
glcm3d-sumEntropy	0.997 (0.996-0.998)	R
glcm3d-sumVariance	0.995 (0.994-0.996)	R
gldzm3d-greyLevelNonUniformity	0.783 (0.693-0.843)	NR
gldzm3d-greyLevelNonUniformityNorm	0.853 (0.815-0.884)	C
gldzm3d-greyLevelVariance	0.994 (0.994-0.995)	R
gldzm3d-highGreyLevelZoneEmphasis	0.998 (0.996-0.998)	R
gldzm3d-largeDistanceEmphasis	0.766 (0.586-0.857)	NR
gldzm3d-largeDistancehighGreyLEmphasis	0.913 (0.892-0.929)	R
gldzm3d-largeDistancelowGreyLEmphasis	0.942 (0.889-0.965)	R
gldzm3d-lowGreyLevelZoneEmphasis	0.975 (0.962-0.983)	R
gldzm3d-smallDistanceEmphasis	0.914 (0.896-0.929)	LR
gldzm3d-smallDistanceHighGreyLEmphasis	0.977 (0.969-0.982)	R
gldzm3d-smallDistanceLowGreyLEmphasis	0.970 (0.957-0.979)	R
gldzm3d-zoneDistanceEntropy	0.964 (0.957-0.969)	R
gldzm3d-zoneDistanceNonUniformity	0.885 (0.852-0.909)	C
gldzm3d-zoneDistanceNonUniformityNorm	0.914 (0.899-0.928)	LR
gldzm3d-zoneDistanceVariance	0.641 (0.404-0.775)	NR
gldzm3d-zonePercentage	0.813 (0.531-0.907)	C
glrl3d-gl_NonUniformity	0.543 (0.342-0.682)	C
glrl3d-gl_NonUniformityNorm	0.996 (0.995-0.997)	R
glrl3d-gl_Variance	0.998 (0.998-0.999)	R
glrl3d-highGLRunEmp	0.999 (0.999-0.999)	R
glrl3d-longRunEmp	0.754 (0.592-0.843)	C
glrl3d-longRunHighGLEmp	0.993 (0.982-0.997)	R
glrl3d-longRunLowGLEmp	0.904 (0.879-0.924)	R
glrl3d-lowGLRunEmp	0.999 (0.999-0.999)	R
glrl3d-rl_NonUniformity	0.545 (0.351-0.680)	C
glrl3d-rl_NonUniformityNorm	0.871 (0.612-0.940)	C
glrl3d-rl_Variance	0.713 (0.537-0.813)	C
glrl3d-runEntropy	0.955 (0.844-0.980)	R
glrl3d-runPercentage	0.823 (0.537-0.914)	C
glrl3d-shortRunEmp	0.863 (0.642-0.932)	C
glrl3d-shortRunHighGLEmp	0.999 (0.999-0.999)	R
glrl3d-shortRunLowGLEmp	0.996 (0.995-0.997)	R
glszm3d-gl_NonUniformity	0.783 (0.693-0.843)	NR
glszm3d-gl_NonUniformityNorm	0.853 (0.815-0.884)	C
glszm3d-gl_Variance	0.994 (0.994-0.995)	R
glszm3d-highGLZoneEmphasis	0.998 (0.996-0.998)	R
glszm3d-largeZoneEmphasis	0.284 (0.235-0.336)	C
glszm3d-largeZoneHighGLEmphasis	0.266 (0.210-0.326)	C
glszm3d-largeZoneLowGLEmphasis	0.294 (0.247-0.344)	C
glszm3d-lowGLZoneEmphasis	0.975 (0.962-0.983)	R
glszm3d-smallZoneEmphasis	0.719 (0.547-0.817)	NR
glszm3d-smallZoneHighGLEmphasis	0.985 (0.978-0.989)	R
glszm3d-smallZoneLowGLEmphasis	0.904 (0.889-0.917)	R
glszm3d-zonePercentage	0.813 (0.531-0.907)	C
glszm3d-zoneSizeEntropy	0.928 (0.906-0.943)	LR
glszm3d-zoneSizeNonUniformity	0.874 (0.852-0.893)	C
glszm3d-zoneSizeNonUniformityNorm	0.763 (0.570-0.858)	NR
glszm3d-zoneSizeVariance	0.276 (0.228-0.329)	C
ngtdm3d_busyness	0.732 (0.648-0.795)	C
ngtdm3d_coarseness	0.864 (0.792-0.906)	C
ngtdm3d_complexity	0.930 (0.915-0.942)	R
ngtdm3d_contrast	0.587 (0.503-0.658)	C
ngtdm3d_strength	0.923 (0.890-0.944)	R

(b) Feature Group-2.

Figure 4.3: Summary of the analysis of feature robutsness to linear interpolation. Features split into two groups: Group-1 contains morphological and first-order features. Group-2 contains texture features. For each subplot, the left is the ICC result wth 95% confidence intervals. The right plot is the range of pairwise ρ between the 2.7mm rankings and the rankings at other voxel sizes, with ρ_{mean} highlighted with a closed dot if >0.95. Features categorised based of the criteria outlined in Figure 4.2 as Robust (R), Limited Robustness (LR), Potentially Correctable (C) or Not Robust (NR). These results were first published by Whybra *et*. *al*. [85].



(a) Feature Group-1.

(b) Feature Group-2.

Figure 4.4: Summary of the analysis of feature robutsness to Spline interpolation. The methodology is the same as described in Figure 4.3, but with Spline interpolation used instead. All features were categorised the same as the linearly interpolated dataset. These results were first published by Whybra *et*. *al*. [85].

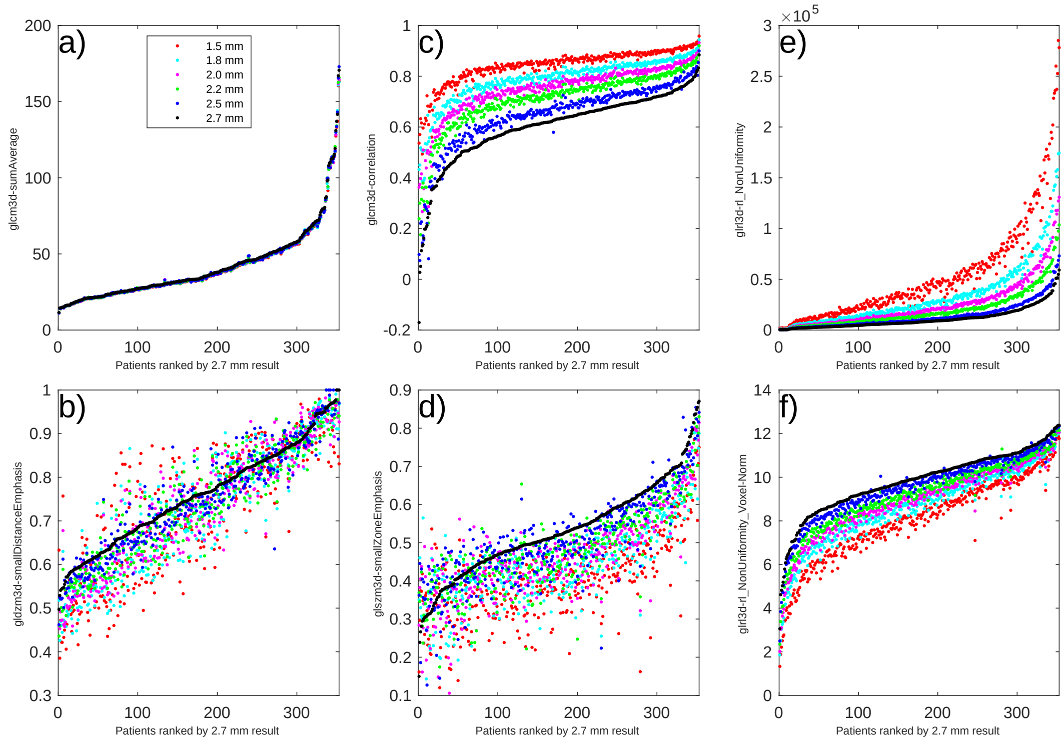


Figure 4.5: Linear interpolation response plots for a selection of 6 features. For each feature, the patient ranking based on the 2.7mm result was plotted against feature values extracted from all voxel sizes. a) *glcm-sumAverage*, categorised as Robust, b) *gldzm3d-smallDistanceEmphasis*, categorised as having limited Robustness, c) *glcm3d-correlation* categorised as potentially correctable, d) *glszm3d-smallZoneEmphasis* categorised as not robust, e) *glrl3d-rl_NonUniformity* categorised as potentially correctable, and f) the normalised version of *glrl3d-rl_NonUniformity* modified to include voxel number, as described by Shafiq-ul-Hassan *et. al.* [115]. This case study of features was first published by Whybra *et. al.* [85]. Exhaustive plots for all features can be found in supplementary materials of [85].

that was modified to include voxel number within its definition, as described by Shafiq-ul-Hassan *et. al.* [115], that was still categorised as potentially correctable.

4.4.2 A Surface Re-scaling Method for Response Correction

For the 34 features categorised as Potentially Correctable via the criteria of Figure 4.2, the author attempted to model the response to interpolation with a surface fit as described in Section 4.3.5. To help understanding, this section walks through an example of this process with results for the feature *glcm3d-inverseDifference*, shown in Figure 4.6.

Using the robustness testing dataset, a surface was generated that linked feature value and voxel size to the percentage change in value when compared to the 2.7 mm result. The voxel size, V_s, is plotted in mm^3 (e.g. 2.7 mm × 2.7 mm × 2.7 mm = 19.683 mm^3). For the 2.7 mm VOIs, the percentage change in feature value is clearly zero, hence all the black markers on the zero line in **c)** of Figure 4.6. Best fit polynomial surfaces were generated with the *fit* functionality from Matlab's curve fitting toolbox. For this example, the best surface fit of feature *glcm3d-inverseDifference* visualised in **c)** of Figure 4.6 is described by the polynomial:

$$S_f(x,y) = 44.8 - 10.9x - 5.25y - 18.7x^2 + 0.750xy + 1.20x^2y - 0.0212xy^2 - 0.00588y^3, \quad (4.2)$$

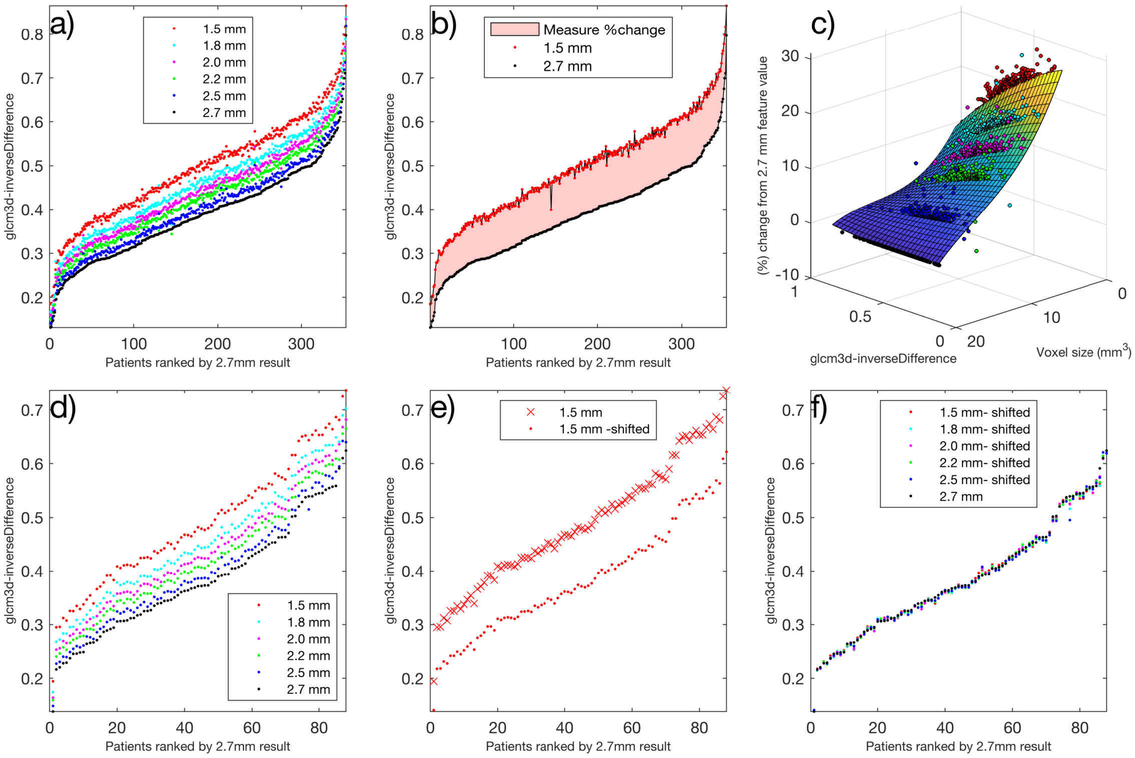
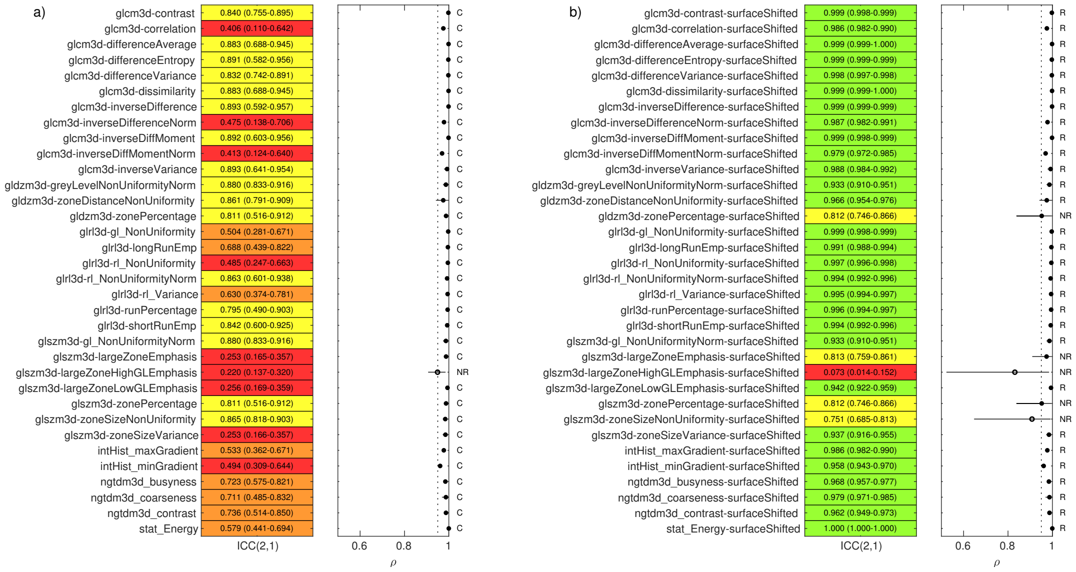


Figure 4.6: Surface fit re-scaling example for *glcm3d-inverseDifference* feature. **Top row:** Modelling interpolation response with *Robustness Testing* dataset (n=353). **a)** Response plot. **b)** Highlight of measured %change between 2.7 mm and 1.5 mm value. **c)** Best fit polynomial surface linking voxel size, feature value and %change in feature value. **Bottom row:** Exploring correction to *Validation* dataset (n=88). **d)** Response plot without correction. **e)** 1.5mm result before and after correction. **f)** Response plot after applying corrections.



a)	ICC(2,1)	
glcm3d-contrast	0.840 (0.755-0.895)	C
glcm3d-correlation	0.406 (0.110-0.642)	C
glcm3d-differenceAverage	0.883 (0.688-0.945)	C
glcm3d-differenceEntropy	0.891 (0.582-0.956)	C
glcm3d-differenceVariance	0.832 (0.742-0.891)	C
glcm3d-dissimilarity	0.883 (0.688-0.945)	C
glcm3d-inverseDifference	0.893 (0.592-0.957)	C
glcm3d-inverseDifferenceNorm	0.475 (0.138-0.706)	C
glcm3d-inverseDiffMoment	0.892 (0.603-0.956)	C
glcm3d-inverseDiffMomentNorm	0.413 (0.124-0.640)	C
glcm3d-inverseVariance	0.893 (0.641-0.954)	C
gldzm3d-greyLevelNonUniformityNorm	0.880 (0.833-0.916)	C
gldzm3d-zoneDistanceNonUniformity	0.861 (0.791-0.909)	C
gldzm3d-zonePercentage	0.811 (0.516-0.912)	C
glrl3d-gl_NonUniformity	0.504 (0.281-0.671)	C
glrl3d-longRunEmp	0.688 (0.439-0.822)	C
glrl3d-rl_NonUniformity	0.485 (0.247-0.663)	C
glrl3d-rl_NonUniformityNorm	0.863 (0.601-0.938)	C
glrl3d-rl_Variance	0.630 (0.374-0.781)	C
glrl3d-runPercentage	0.795 (0.490-0.903)	C
glrl3d-shortRunEmp	0.842 (0.600-0.925)	C
glszm3d-gl_NonUniformityNorm	0.880 (0.833-0.916)	C
glszm3d-largeZoneEmphasis	0.253 (0.165-0.357)	C
glszm3d-largeZoneHighGLEmphasis	0.220 (0.137-0.320)	NR
glszm3d-largeZoneLowGLEmphasis	0.256 (0.169-0.359)	C
glszm3d-zonePercentage	0.811 (0.516-0.912)	C
glszm3d-zoneSizeNonUniformity	0.865 (0.818-0.903)	C
glszm3d-zoneSizeVariance	0.253 (0.166-0.357)	C
intHist_maxGradient	0.533 (0.362-0.671)	C
intHist_minGradient	0.494 (0.309-0.644)	C
ngtdm3d_busyness	0.723 (0.575-0.821)	C
ngtdm3d_coarseness	0.711 (0.485-0.832)	C
ngtdm3d_contrast	0.736 (0.514-0.850)	C
stat_Energy	0.579 (0.441-0.694)	C

b)	ICC(2,1)	
glcm3d-contrast-surfaceShifted	0.999 (0.998-0.999)	R
glcm3d-correlation-surfaceShifted	0.986 (0.982-0.990)	R
glcm3d-differenceAverage-surfaceShifted	0.999 (0.999-1.000)	R
glcm3d-differenceEntropy-surfaceShifted	0.999 (0.999-0.999)	R
glcm3d-differenceVariance-surfaceShifted	0.998 (0.997-0.998)	R
glcm3d-dissimilarity-surfaceShifted	0.999 (0.999-1.000)	R
glcm3d-inverseDifference-surfaceShifted	0.999 (0.999-0.999)	R
glcm3d-inverseDifferenceNorm-surfaceShifted	0.987 (0.982-0.991)	R
glcm3d-inverseDiffMoment-surfaceShifted	0.999 (0.998-0.999)	R
glcm3d-inverseDiffMomentNorm-surfaceShifted	0.979 (0.972-0.985)	R
glcm3d-inverseVariance-surfaceShifted	0.988 (0.984-0.992)	R
gldzm3d-greyLevelNonUniformityNorm-surfaceShifted	0.933 (0.910-0.951)	R
gldzm3d-zoneDistanceNonUniformity-surfaceShifted	0.966 (0.954-0.976)	R
gldzm3d-zonePercentage-surfaceShifted	0.812 (0.746-0.866)	NR
glrl3d-gl_NonUniformity-surfaceShifted	0.999 (0.998-0.999)	R
glrl3d-longRunEmp-surfaceShifted	0.991 (0.988-0.994)	R
glrl3d-rl_NonUniformity-surfaceShifted	0.997 (0.996-0.998)	R
glrl3d-rl_NonUniformityNorm-surfaceShifted	0.994 (0.992-0.996)	R
glrl3d-rl_Variance-surfaceShifted	0.995 (0.994-0.997)	R
glrl3d-runPercentage-surfaceShifted	0.996 (0.994-0.997)	R
glrl3d-shortRunEmp-surfaceShifted	0.994 (0.992-0.996)	R
glszm3d-gl_NonUniformityNorm-surfaceShifted	0.933 (0.910-0.951)	R
glszm3d-largeZoneEmphasis-surfaceShifted	0.813 (0.759-0.861)	NR
glszm3d-largeZoneHighGLEmphasis-surfaceShifted	0.073 (0.014-0.152)	NR
glszm3d-largeZoneLowGLEmphasis-surfaceShifted	0.942 (0.922-0.959)	R
glszm3d-zonePercentage-surfaceShifted	0.812 (0.746-0.866)	NR
glszm3d-zoneSizeNonUniformity-surfaceShifted	0.751 (0.685-0.813)	NR
glszm3d-zoneSizeVariance-surfaceShifted	0.937 (0.916-0.955)	R
intHist_maxGradient-surfaceShifted	0.986 (0.982-0.990)	R
intHist_minGradient-surfaceShifted	0.958 (0.943-0.970)	R
ngtdm3d_busyness-surfaceShifted	0.968 (0.957-0.977)	R
ngtdm3d_coarseness-surfaceShifted	0.979 (0.971-0.985)	R
ngtdm3d_contrast-surfaceShifted	0.962 (0.949-0.973)	R
stat_Energy-surfaceShifted	1.000 (1.000-1.000)	R

Figure 4.7: The 34 features categorised as *Potentially Correctable* in the *Robustness Testing* dataset were further analysed with the *Validation* dataset. The plots of ICC and ρ analysis presented are the same layout as described in Figure 4.3. Here, **a)** is analysis before correction, and **b)** is after applying the corrections using the surface fits generated on the *Robustness Testing* dataset. This result was first published by Whybra *et. al.* [85].

where x is the feature value and y is the current voxel size. $S_f(x, y)$ models the percentage change required to shift the feature to the 2.7mm result. This is then used via Eq. 4.1 to correct values for this feature extracted at different voxel sizes. The results before and after applying the correction

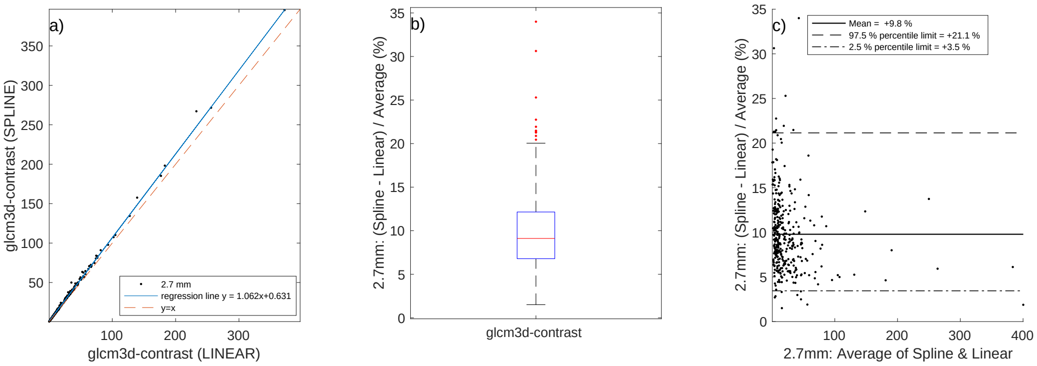


Figure 4.8: Linear vs Spline analysis performed for each feature using the 2.7mm results, demonstrated in this figure with the feature *glcm3d-contrast*. Sub-figure **a)** is a plot of the spline against linear results, **b)** is a boxplot of the measured difference [(Spline Value - Linear value)/average (%)], and **c)** is the corresponding Bland-Altman plot.

are shown on the validation dataset for *glcm3d-inverseDifference* in the bottom row of plots in Figure 4.6. Corresponding plots for all 34 features that were deemed potentially correctable can be found in the supplementary materials in Whybra *et*. *al*. [85].

Figure 4.7 summarises the robustness results for all 34 features, prior to and post correction. Analysis of the validation dataset before correction found 33/34 remained categorised as potentially correctable. Post correction, 29/34 features were re-categorised as robust. For 5/34 features the correction did not improved robustness, and thus, the surface fits did not appear to model the response found in the validation dataset.

4.4.3 Linear vs Spline

As shown in Figure 4.3 and Figure 4.4, the categorisation of responses for features remained constant when either a linear or spline interpolation method was used. Further analysis was performed to measure the difference in extracted feature values with the two different interpolation methods, using the 2.7 mm results. Individual plots analysing these differences are shown with an example in Figure 4.8, and for each feature in the supplementary materials of Whybra *et*. *al*. [85] . For a number of features, the choice of interpolation method caused large variation in the extracted value. Figure 4.9 summarises the variation between methods for the texture features (Group-2). The features are ordered from lowest to highest based on the interquartile range of the variation shown by the box plots.

4.5 Discussion

For imaging biomarkers that provide prognostic or predictive potential to gain traction in a clinical setting, they should first be vetted as robust and reproducible prior to inclusion in any model. As isotropic resampling of a scan is a necessary step in 3D extraction, it is also imperative to understand how this processing might alter features that become the model discriminators. Indeed, the boundary values used to discriminate, and the very ability of a feature to discriminate, may change with voxel size. Many studies use interpolation, yet the optimal method (e.g. linear vs

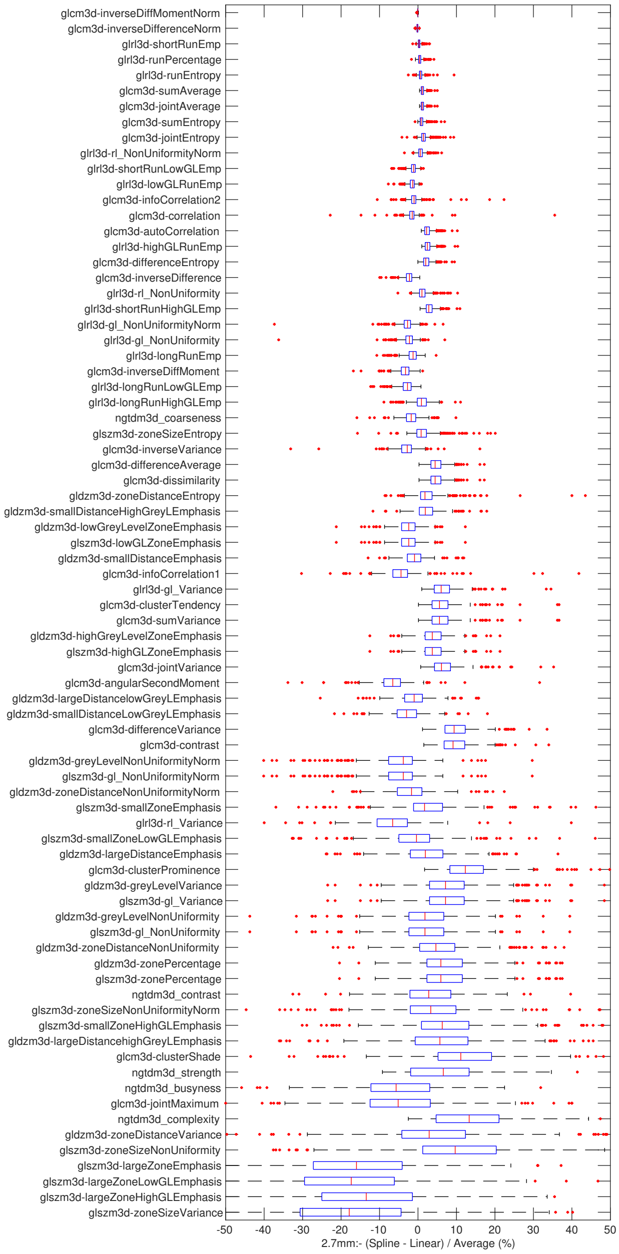


Figure 4.9: Box plots summarising the measured variation in feature values when using linear compared to spline interpolation. Results are for the 2.7 mm extraction of the *Robustness Testing* dataset. Difference measured as [(Spline value - Linear value)/average (%)]. The x axis is restricted between +/-50% for readability, and the features have been ordered based on the box plot interquartile range.

spline), or best resampling size based on the original dimensions (e.g. 2.7 mm vs 1.5 mm), is not known. Evidently, one should be concerned with feature robustness to these extraction choices.

This study focused entirely on feature stability due to isotropic interpolation of PET, analysing the response seen as the imaging was upsampled to an effective higher resolution. To the best of the author's knowledge, this study was the first to use this modality to specifically categorise features by interpolation response, and to identify strong systematic variation alongside stable and unstable responses. The study explored a novel correction technique utilising surface models to re-scale the systematically varying response seen for some features to a selected voxel size. The feature categorisations determined here should help guide future studies, both as a feature selection tool, and as a way to reduce redundancy if extraction is performed at multiple voxel sizes. These results have been summarised in Table 4.1.

There were several trends identified when evaluating the response of features to interpolation. Importantly, it is evident that morphological and first-order measures appeared much more robust to increasing the image resolution (i.e. when comparing Group-1 results to Group-2). In general, these Group-1 features will remain good candidates for radiomics modelling when there is a need to interpolate PET imaging to a uniform, isotropic voxel size. As an example, all statistical measures apart from 1 were shown to be robust. The only statistical feature not robust, *stat-Energy*, responded in a systematic way that could be modelled. This feature is an example of one that is directly correlated to the number of voxels in the ROI, so as the number of voxels increased by interpolating to a smaller resolution, the feature value increased accordingly. In general, features with this characteristic responded to correction modelling.

Trends in robustness of higher-order texture features at a family level are not as clear from the analysis of this dataset. Notably though, each texture family had a subset of at least some stable features that remained constant when extracted at increasing voxel resolutions.

In some cases, the type of response that a texture feature displays should be intuitive. Interpolating to higher resolutions smooths an image, flattening the intensity gradients between neighbouring voxels (as illustrated in Figure 4.1). Texture features that emphasise sudden contrast changes in neighbours are sure to be affected as a result. Take as example the feature *glcm3d-inverseDifference*, that was used to demonstrate the surface correction technique in Figure 4.6. This feature gives less weight to large differences in the GLCM (hence the name *inverse* difference), and so the higher the feature value, the more homogeneous the measured region is. From this definition, one would expect this feature value to increase from smoothing. This trend is shown in the data of this study (Figure 4.6 **a)**), where the feature values increase for each patient with decreasing voxel size (i.e. the 1.5mm results are consistently above the 2.7mm results). However, increasingly higher-order texture features, such as zone-based GLSZM measures, become too complex to intuit if the response will be stable or not.

4.5.1 Reflection on Statistical Tests

Two statistical tests were selected to categorised the features: ICC and patient ranking analysis with ρ. These tests have seen use in several previous robustness studies for radiomic features as highlighted by Traverso *et. al.* [80]. In particular, the combination of these tests presented in this study allowed identification of different types of variation that were also shown qualitatively with extensive plotting of feature response, as demonstrated in the case studies in Figure 4.5, and for all features in the supplementary materials of Whybra *et. al.* [85]. In itself, these plots represent a useful method for visualising feature variation that the author would recommend for any robustness study. The binary threshold selected for reproducibility in feature value was an ICC score >0.9, following other studies in the literature [72, 120]. Reported alongside is the 95% confidence intervals of these results. The form of ICC used in the main analysis emphasised absolute agreement, based on the author's definition that a robust feature would have the same value, independent of resampled voxel size, but for completeness, the consistency version of ICC was also calculated. The reliability of patient ranking based on the feature values was assessed pairwise with the 2.7 mm result and all other voxel sizes, with ρ_{mean} >0.95, a high threshold, used to determine reliable patient ranking. This approach was selected as the 2.7mm results are framed as a ground truth to measure against and correct back to in the case of systematically varying features. This study found the patient ranking for the majority of features remained highly correlated when resampling to different voxel sizes. Another appropriate method would be to assess *all* pairwise combinations and instead of looking at the mean result, consider the minimum value of the range as the threshold. Leijenaar *et. al.* [102] utilised a similar analysis technique to assess the effect of discretisation methods on radiomic features, using the range instead of a mean value, but with a threshold of $\rho > 0.9$. A particularly high threshold is used in the work here to determine if a feature is categorised as "potentially correctable", as the surface modelling correction technique is only logical if the patient ranking remained highly consistent.

Based on the statistical testing, 34 features were identified as potentially correctable. A novel surface fitting correction technique was explored on the linear interpolated results, first by modelling response in the main *Robusteness testing* dataset, and then correcting those same features in the *validation* dataset. As shown in Figure 4.6 and 4.7, on a feature-by-feature basis, this method was successful in adjusting 29/34 features showing a dependence on interpolation by rescaling the values to closer match results from the ground truth voxel size (in this case the 2.7 mm results).

4.5.2 Limitations of Modelling Interpolation Response

There are a number of limitations to the surface model correction approach explored in this work. One key limitation is that this analysis was conducted holding all other extraction settings constant. For example, only one discretisation setting was used: a FBW of 0.5 SUV. A FBW was chosen as it is thought to provide a more meaningful inter- and intra-patient comparison for PET imaging in SUV units [102]. The limitation here is that the surfaces derived are likely to be specific to this discretisation method and bin width. Furthermore, each feature had a unique surface fit, rather than one overall systematic response that could be generalised with one model. The corrections are confined to the range of voxel sizes the models were built on, in this case between 2.7 and 1.5mm. For each feature, all data from the *Robustness testitng* dataset was used to create the polynomial fit. Better models might be achieved if potential outliers were identified and ignored in the fitting process.

Although a large dataset was used, with separate validation data to test models, a clear further extension to this study would be to assess if the general correction trends found in this dataset can be replicated in other datasets and imaging modalities, such as CT.

For the features where this correction method succeeded, it strongly suggested that although the feature values changed due to interpolation, the change was predictable. Moreover, these correction models were built on real imaging data, and do not address any underlying mathematical definition that could explain the systematic variation that is seen.

4.5.3 Modified Feature Definitions

An update to the feature definition to improve robustness would be preferable to any correction shift generated from experimental data. A feature that has a dependence on the number of voxels in the VOI will show a dependence to interpolation, and by normalising the feature using the number of voxels, one might improve robustness. In fact, this has already been utilised for many of the standardised features explored in this work. Multiple features have a normalised variant that adds a factor in the mathematical definition to attempt to reduce dependence on the number of voxels within the VOI. For example, *glcm3d-inverseDiffMoment* and *glcm3d-inverseDiffMomentNorm*, or *gldzm3d-greyLevelNonUniformity* and *gldzm3d-greyLevelNonUniformityNorm*. However, in this study, many of these normalised features are categorised as potentially correctable, rather than robust. By adding a mathematical factor to normalise a feature by the number of voxels, the author suggests from this work that this very correction factor can evidently become the dominant contribution to these feature values when interpolating to higher and higher resolutions.

In contrast to the author's findings, a relevant study by Shafiq-ul-Hassan *et. al.* [114] found that these modifications to mathematical definitions to remove, as they describe, *intrinsic voxel-size dependency*, improved the robustness of 10 features significantly. They assessed the impact of pixel size and slice thickness on features acquired on 116 CT images of texture phantoms with different acquisition and reconstruction parameters. They resampled all of the images to a common voxel size ($1 \times 1 \times 2$ mm) using linear interpolation to determine if this improved the features robustness. The same group developed on this in another study [115] to validate these 10 normalised features using phantom images alongside a set of 18 patients with non-small cell lung cancer CT imaging that was interpolated to 4 different pixel sizes and 6 slice thicknesses. Using spearmans rank analysis, they found that eight of these features showing high correlation with voxel number ($\rho > 0.9$), had low correlation after modifying the feature definitions to include voxel number ($\rho < 0.5$) [115].

To provide an example of the contrast in findings, the feature set in this work includes both the original and normalised version of the feature *glrl3d-rl_NonUniformity*, which is modified as described by Shafiq-ul-Hassan *et. al.* [115]. These results are visualised in the case studies in Figure 4.5 **e)** and **f)**. The author finds that systematic response due to isotropic interpolation was present in the normalised version of the feature, showing that the updated definition can still exhibit a dependence on interpolation, even when adapted to incorporate the voxel number explicitly. This trend is evident in other normalised features that were also categorised as correctable.

4.5.4 Linear vs Spline

This study demonstrated that feature values can vary significantly when extracted using a linear interpolation method versus spline (Figure 4.9). However, the robustness categorisations did not change at all between these methods. If a feature was categorised robust when the dataset was linearly interpolated, it was categorised robust when using a spline method instead (Figure 4.3 vs Figure 4.4). In some cases the features did not vary at all. This is because the extraction followed the IBSI recommendations, and when interpolating, the recommendation is to always use the linear method for the binary mask defining the VOI (refer back to Section 2.3.3.5). Thus, several morphological features that utilised only the morphological mask remain exactly the same, as they do not depend on interpolation algorithm used for the scan.

There is no strong consensus over the most appropriate interpolation algorithm to use in ^{18}F-FDG PET imaging. A study by Yip *et. al.* [117] reported to investigate the effect of various extraction settings on predicting somatic mutation status of *Non-small cell lung cancer*, including a range of voxel sizes, interpolation methods and discretisation bin widths. By testing the predictive performance of 66 radiomic features individually, they found 29 features remained predictive for a range of settings, though stress that the combined effects of extraction settings substantially alter predictive performance and should be optimised [117]. However, they provide limited details of the actual variation between interpolation methods and voxels sizes on a feature- by-feature level.

Despite the measured variability in many features between linear and spline, the robustness categorisation was the same for both methods when upsampling PET images. This variability suggests that model thresholds could be specific to the interpolation method selected, though either method should still yield the predictive or prognostic signal, *if* there is one to be found. However, as addressed previously, thorough reporting of feature extraction settings including interpolation method and voxel size is a necessity for reproducibility and validation with every study.

4.5.5 Feature Stability and Clinical Relevance

For any feature, strong stability to isotropic interpolation does not inherently translate to any clinically useful application. That said, features that do become clinically relevant are likely to be a subset of those that have a predictable or robust interpolation response as the models utilising these features should be more generalisable. There are copious features in radiomics, with a need for feature reduction techniques to reduce overfitting during model development. Removing feature that have not shown adequate interpolation stability, as presented in this study, is one of several prerequisite selection steps to optimise future radiomics studies that utilise resampling of imaging from multi-centre datasets to a common voxel size.

4.5.6 Robustness Testing Before Standardisation

A strength of this study is that it was conducted with fully standardised extraction algorithms, so the findings can be more readily integrated in follow up meta-analysis and reviews. The Traverso *et. al.* [80] review article of repeatability and reproducibility of radiomic features reported no emergent pattern or consensus for highly reproducible textural features. As no clear standards were around at the time of many of the studies in this review, it is difficult to determine if feature definitions and methodology have been consistent enough to compare the results. As shown in Chapter 3, initial consensus was low between different software implementations, so there is uncertainty in the ability to combine feature robustness findings of past studies. With the development of standardised algorithms, this will improve for future research.

Traverso et al. [80] commented that for texture features, coarseness and contrast appeared among the least reproducible. Interestingly, the standardised features with these names (*ngtdm3d_coarseness*, *ngtdm3d_contrast*, *glcm3d-contrast*) were found not robust to interpolation, but all showed a systematic response that was potentially correctable and the response could be modelling for this dataset. However, it should be clearly stated that a feature robust to interpolation may still be very unstable to other extraction settings.

Table 4.1: Final overview of feature categorisations for response to interpolation. R = Robust, C = Potentially correctable systematic response, LR = Limited Robustness, NR= Not Robust. Y= successful correction in validation dataset, N = correction did not work for validation dataset.

Family	Feature Name	Code	Categorisation	Successful Surface Shift
GLCM	Angular Second Moment	glcm3d-angularSecondMoment	R	
	Auto Correlation	glcm3d-autoCorrelation	R	
	Cluster Prominence	glcm3d-clusterProminence	R	
	Cluster Shade	glcm3d-clusterShade	R	
	Cluster Tendency	glcm3d-clusterTendency	R	
	Contrast	glcm3d-contrast	C	Y
	Correlation	glcm3d-correlation	C	Y
	Difference Average	glcm3d-differenceAverage	C	Y
	Difference Entropy	glcm3d-differenceEntropy	C	Y
	Difference Variance	glcm3d-differenceVariance	C	Y
	Dissimilarity	glcm3d-dissimilarity	C	Y
	First measure of information correlation	glcm3d-infoCorrelation1	NR	
	Second measure of information correlation	glcm3d-infoCorrelation2	NR	
	Inverse difference	glcm3d-inverseDifference	C	Y
	Inverse difference normalised	glcm3d-inverseDifferenceNorm	C	Y
	Inverse difference moment	glcm3d-inverseDiffMoment	C	Y
	Inverse difference moment normalised	glcm3d-inverseDiffMomentNorm	C	Y
	Inverse Variance	glcm3d-inverseVariance	C	Y
	Joint Average	glcm3d-jointAverage	R	
	Joint Entropy	glcm3d-jointEntropy	R	
	Joint Maximum	glcm3d-jointMaximum	R	
	Joint Variance	glcm3d-jointVariance	R	
	Sum Average	glcm3d-sumAverage	R	
	Sum Entropy	glcm3d-sumEntropy	R	
	Sum Variance	glcm3d-sumVariance	R	
GLDZM	Grey Level Non-uniformity	gldzm3d-greyLevelNonUniformity	NR	
	Grey Level Non-uniformity Normalised	gldzm3d-greyLevelNonUniformityNorm	C	Y
	Grey Level Variance	gldzm3d-greyLevelVariance	R	
	High Grey Level Zone Emphasis	gldzm3d-highGreyLevelZoneEmphasis	R	
	Large Distance Emphasis	gldzm3d-largeDistanceEmphasis	NR	
	Large Distance High Grey Level Emphasis	gldzm3d-largeDistancehighGreyLEmphasis	R	
	Large Distance Low Grey Level Emphasis	gldzm3d-largeDistancelowGreyLEmphasis	R	
	Low Grey Level Zone Emphasis	gldzm3d-lowGreyLevelZoneEmphasis	R	
	Small Distance Emphasis	gldzm3d-smallDistanceEmphasis	LR	
	Small Distance High Grey Level Emphasis	gldzm3d-smallDistanceHighGreyLEmphasis	R	
	Small Distance Low Grey Level Emphasis	gldzm3d-smallDistanceLowGreyLEmphasis	R	
	Zone Distance Entropy	gldzm3d-zoneDistanceEntropy	R	
	Zone Distance Non-Uniformity	gldzm3d-zoneDistanceNonUniformity	C	Y
	Zone Distance Non-Uniformity Normalised	gldzm3d-zoneDistanceNonUniformityNorm	LR	
	Zone Distance Variance	gldzm3d-zoneDistanceVariance	NR	
	Zone Percentage	gldzm3d-zonePercentage	C	N
GLRL	Grey Level Non-uniformity	glrl3d-gl_NonUniformity	C	Y
	Grey Level Non-uniformity Normalised	glrl3d-gl_NonUniformityNorm	R	
	Grey Level Variance	glrl3d-gl_Variance	R	
	High Grey Level Run Emphasis	glrl3d-highGLRunEmp	R	
	Long Runs Emphasis	glrl3d-longRunEmp	C	Y
	Long Run High Grey Level Emphasis	glrl3d-longRunHighGLEmp	R	
	Long Run Low Grey Level Emphasis	glrl3d-longRunLowGLEmp	R	
	Low Grey Level Run Emphasis	glrl3d-lowGLRunEmp	R	
	Run Length Non-uniformity	glrl3d-rl_NonUniformity	C	Y
	Run Length Non-uniformity Normalised	glrl3d-rl_NonUniformityNorm	C	Y
	Run Length Variance	glrl3d-rl_Variance	C	Y
	Run Entropy	glrl3d-runEntropy	R	
	Run Percentage	glrl3d-runPercentage	C	Y
	Short Run Emphasis	glrl3d-shortRunEmp	C	Y
	Short Run High Grey Level Emphasis	glrl3d-shortRunHighGLEmp	R	
	Short Run Low Grey Level Emphasis	glrl3d-shortRunLowGLEmp	R	

Table 4.1: continued

Family	Feature Name	Code	Categorisation	Successful Surface Shift
GLSZM	Grey Level Non-uniformity	glszm3d-gl_NonUniformity	NR	
	Grey Level Non-uniformity Normalised	glszm3d-gl_NonUniformityNorm	C	Y
	Grey Level Variance	glszm3d-gl_Variance	R	
	High Grey Level Zone Emphasis	glszm3d-highGLZoneEmphasis	R	
	Large Zone Emphasis	glszm3d-largeZoneEmphasis	C	N
	Large Zone High Grey Level Emphasis	glszm3d-largeZoneHighGLEmphasis	C	N
	Large Zone Low Grey Level Emphasis	glszm3d-largeZoneLowGLEmphasis	C	Y
	Low Grey Level Zone Emphasis	glszm3d-lowGLZoneEmphasis	R	
	Small Zone Emphasis	glszm3d-smallZoneEmphasis	NR	
	Small Zone High Grey Level Emphasis	glszm3d-smallZoneHighGLEmphasis	R	
	Small Zone Low Grey Level Emphasis	glszm3d-smallZoneLowGLEmphasis	R	
	Zone Percentage	glszm3d-zonePercentage	C	N
	Zone Size Entropy	glszm3d-zoneSizeEntropy	LR	
	Zone Size Non-uniformity	glszm3d-zoneSizeNonUniformity	C	N
	Zone Size Non-uniformity Normalised	glszm3d-zoneSizeNonUniformityNorm	NR	
	Zone Size Variance	glszm3d-zoneSizeVariance	C	Y
NGTDM	Busyness	ngtdm3d_busyness	C	Y
	Coarseness	ngtdm3d_coarseness	C	Y
	Complexity	ngtdm3d_complexity	R	
	Contrast	ngtdm3d_contrast	C	Y
	Strength	ngtdm3d_strength	R	
IH	Coefficient of Variation	intHist_coefficientofVariation	R	
	Entropy	intHist_entropy	R	
	Interquartile range	intHist_IQR	R	
	Kurtosis	intHist_kurtosis	R	
	Maximum histogram gradient	intHist_maxGradient	C	Y
	Maximum gradient grey level	intHist_maxGradientGreyLevel	R	
	Maximum	intHist_maxGreyLevel	R	
	Mean	intHist_mean	R	
	Mean absolute deviation	intHist_meanAbsoluteDeviation	R	
	Median	intHist_median	R	
	Median absolute deviation	intHist_medianAbsoluteDeviation	R	
	Minimum histogram gradient	intHist_minGradient	C	Y
	Minimum gradient grey level	intHist_minGradientGreyLevel	LR	
	Minimum	intHist_minGreyLevel	R	
	Mode	intHist_mode	R	
	10th percentile	intHist_percentile10	R	
	90th percentile	intHist_percentile90	R	
	Quartile coefficient of dispersion	intHist_quartileCoefDispersion	LR	
	Range	intHist_range	R	
	Robust mean absolute deviation	intHist_robustMeanAbsoluteDeviation	R	
	Skewness	intHist_skewness	R	
	Uniformity	intHist_uniformity	R	
	Variance	intHist_variance	R	
Morphology	Area density (AABB)	morph_areaDensity_aabb	R	
	Area density (AEE)	morph_areaDensity_aee	R	
	Area density (convex hull)	morph_areaDensity_convexHull	R	
	Asphericity	morph_asphericity	R	
	Compactness 1	morph_compactness1	R	
	Compactness 2	morph_compactness2	R	
	Centre of mass shift	morph_COMshift	LR	
	Elongation	morph_elongation	R	
	Flatness	morph_flatness	R	
	Integrated intensity	morph_integratedIntensity	R	
	Least axis length	morph_leastAxisLength	R	
	Major axis length	morph_majorAxisLength	R	
	Maximum 3D diameter	morph_max3Ddiameter	R	
	Minor axis length	morph_minorAxisLength	R	
	Spherical disproportion	morph_sphericalDisproportion	R	
	Sphericity	morph_sphericity	R	
	Surface area	morph_surfaceArea	R	
	Surface to volume ratio	morph_surfAreaToVolumeRatio	R	
	Volume density (AABB)	morph_volDensity_aabb	R	
	Volume density (AEE)	morph_volDensity_aee	R	
	Volume density (convex hull)	morph_volDensity_convexHull	R	
	Volume (mesh-based)	morph_volume	R	
Statistical	Coefficient of Variation	stat_coefficientofVariation	R	
	Energy	stat_Energy	C	Y
	Interquartile range	stat_IQR	R	
	Kurtosis	stat_kurtosis	R	
	Maximum	stat_maxGreyLevel	R	
	Mean	stat_mean	R	
	Mean absolute deviation	stat_meanAbsoluteDeviation	R	
	Median	stat_median	R	
	Median absolute deviation	stat_medianAbsoluteDeviation	R	
	Minimum	stat_minGreyLevel	R	
	10th percentile	stat_percentile10	R	
	90th percentile	stat_percentile90	R	
	Quartile coefficient of dispersion	stat_quartileCoefDispersion	R	
	Range	stat_range	R	
	Robust mean absolute deviation	stat_robustMeanAbsoluteDeviation	R	
	Root mean square	stat_rootMeanSquare	R	
	Skewness	stat_skewness	R	

Table 4.1: continued

Family	Feature Name	Code	Categorisation	Successful Surface Shift
	Variance	stat_variance	R	

4.6 Conclusion

This study evaluated the robustness and variation seen in radiomic features when interpolating PET imaging, from a large cohort of patients with oesophageal cancer, to a range of isotropic voxel dimensions. Analysis of these standardised features revealed a majority that were robust to this resampling process. Thirty-four features showed varying feature values, yet retained highly correlated patient rankings; they were deemed to be varying in a potentially systematic way and possibly correctable. Surface fits modelling this variation were explored, feature by feature, and used as a correction factor, which performed well for 29 features in a validation dataset. However, the correction models are specific to each feature and extraction setting. Furthermore, there were 8 features identified as not robust that behaved in an unstable manner, with both feature value and patient ranking varying widely between voxel sizes. These should be used with caution in radiomics studies that resample imaging from different protocols to one common voxel size. Two standardised interpolation methods, linear and spline, were assessed in this work. For many features, there was significant variation between the two methods, yet the overall categorisations of feature response remained consistent.

Take home message

1. Many features remain robust when isotropically upsampling PET imaging to a new voxel dimension. Interpolation is often required in radiomics studies for 3D analysis.
2. The response of features that appear to vary systematically can potentially be modelled and shifted to correct the variation measured.
3. By assessing both linear and spline interpolation methods, this study found that this choice often affects the quantitative values of features, but did not change their robustness categorisation.
4. As many features are stable, extracting features at different voxel sizes will lead to large amounts of redundancy.
5. For radiomic analysis, stable features are recommended when interpolating datasets acquired at different voxel sizes to one common resolution for analysis.

5

Towards Reproducible Convolutional Filter-Based Imaging Features in Radiomics: Challenges and Methodology For Consensus Benchmarking

"A genuine leader is not a searcher for consensus but a molder of consensus."

— Martin Luther King Jr

5.1 Preview

This chapter explores further challenges of reproducible radiomics by considering standardisation of another major image processing step: *image filtering*. A large array of filtering techniques are indispensable for medical image analysis, helping to enhance and characterise various image properties (e.g. through noise reduction, sharpening, edge and spot detection). This chapter examines common filtering techniques within radiomic feature extraction that are potentially fraught with replication issues. Through experimental examples, several implementation decisions that significantly affect reproducibility are discussed and demonstrated. Key preliminary work by the author to develop and evaluate a methodology to determine consensus-based reference *response maps* for further benchmarking with the *Image Biomarker Standardisation Initiative* (IBSI) [86] is introduced. This methodology is first designed with simulated examples and then results presented for an initial set of submitted filter tests from participating IBSI teams. The aim is to determine tolerances for arriving at acceptable consensus for radiomic analysis. This work has contributed to a pre-print available on *arXiv* to guide the second instalment of the IBSI [86].

5.1.1 Author Contribution

The author is a core contributor to the IBSI, as described previously in Sections 1.8 and 1.9. The key contributions for the work discussed in this chapter are summarised below. The author designed the methodology used by the IBSI to determine consensus response maps for benchmarking software. The author developed the data analysis pipeline and performed the experimental analysis. The author primarily identified key causes of discrepancy in the use of filters through this analysis, namely, these were errors arising from padding, filter orientation, and orientation pooling techniques for *odd* compared to *even* filter kernels. Recommendations for these image processing tasks will be based directly on this preliminary work. The author implemented filtering methods into SPAARC to provide results for 21/25 of the filter tests assessed in this work.

5.2 Introduction

Filtering techniques are a cornerstone of image analysis: their application transforms imaging, and these transformations offer new ability to assess, enhance, and emphasise many underlying characteristics. Different filter kernels are selected and convolved with the image to produce a new output, referred to in this work as a *response map*. Naturally, image biomarkers can be extracted from response maps as readily as the original image, thus, filtering in radiomics is another major processing step that requires careful consideration and standardisation to facilitate potential clinical adoption. Indeed, it has been indicated that many filter-based features may have poor reproducibility when extracted using different software [111]. There are numerous hyper-parameters - unique to different filter designs - that make reproducibility difficult if not extensively reported and applied in the same way. Under-reporting of methodology and sub-optimal use of filters harms the usefulness and generalisability of potential imaging biomarkers derived using these techniques.

Filtering and filter-based radiomic features were not addressed in the extensive standardisation effort discussed throughout Chapter 3 to keep the scope of that task manageable. As such, this chapter details necessary work to tackle this additional complex standardisation challenge. First, this chapter provides an overview of the general principles of convolution for image biomarker extraction. Practical considerations for filtering in radiomics are outlined with an assessment of the workflow, including a discussion of key factors to consider such as directional sensitivity and rotational dependence. Several of the filters that are the current target of standardisation within the IBSI are then introduced (with a focus on those presently implemented in our SPAARC software). Prominent examples of these filtering techniques used for feature extraction from radiomics literature are discussed.

To produce benchmarks for common filtering techniques, the IBSI core group have designed a set of filter tests that utilise various digital phantoms. The aim here is to first evaluate the application of filters without any other processing considerations using a number of these digial phantoms.

This will begin with technical assessment and validation of these response maps directly, without any other processing and prior to any aggregation into a single feature. In this instalment the goal is to produce valid *consensus based* response maps, not just single radiomic feature values.

The key aim and contribution of the work presented in Section 5.6 was to develop and test a methodology for this technical assessment of response maps to enable better evaluation of the variation that may occur between different software, and to design the criteria for arriving at a stronger consensus in the presence of discrepancy. This work will significantly aid the reproducibility of radiomic studies which examine filter-based features as potential imaging biomarkers. Section 5.7 presents the initial results from the first submissions from currently contributing teams. These preliminary results showcase some initial discrepancies, which further justify the need for benchmarks and the potential to expand with further additional tests.

5.2.1 Overview of Filter-based Image Biomarker Extraction

The following section introduces the basic extraction process for features based on filtering. In essence, one can consider the process in three stages: *padding* → *convolution* → *aggregation*. An overview of this process is summarised in Figure 5.1.

Padding - It is important to decide before filtering how the process will be handled at the image boundaries. In principle, the image is extended by various padding techniques (e.g. *constant value*, *mirror*) to facilitate calculation of the response map intensities at the image edge. As such, different padding techniques can affect the intensity values of voxels within the response maps. Intuitively, for reproducibility the same padding will need to be used if the VOI is near to the boundary. Some form of padding is needed for the response map to be the same size as the original image. Otherwise, one can imagine that for edge voxels the kernel will be applied to an undefined region outside of the image.

Convolution - A selected filter is convolved with the padded image to produce a response map that is the size of original image. Convolution is defined in seciton 5.2.2. One can consider filtering in both the *spatial* and *frequency* domains via the *convolution theorem*. As will be discussed, some common filtering techniques are directionally and rotationally sensitive, which can have consequences for consistent analysis of medical imaging. Rotation invariance strategies [86] can be used to minimise this as discussed in Section 5.3.2.

Aggregation - After the application of a given filter to produce a response map, there is effectively a new image, which can be aggregated (summarised) within a volume of interest (VOI) with any radiomic feature discussed throughout this thesis. Indeed, filtering fits within the workflow as described in Figure 2.2. As such, by considering different filter configurations alongside radiomic analysis on an unfiltered scan, one can vastly increase the number of features that can be extracted. This ability to rapidly increase feature numbers necessarily requires caution.

5.2.2 Convolution

This section briefly introduces relevant theory for convolution. First consider the continuous case [86]: a D-dimensional convolution of a filter $g(\boldsymbol{x})$ with an image $I(\boldsymbol{x})$ produces a response map

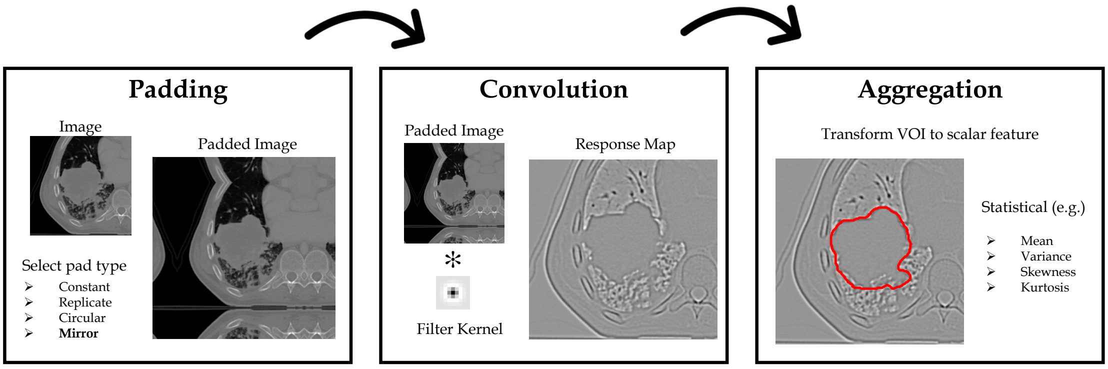


Figure 5.1: Summary of filter-based image biomaker extraction.

$h(\boldsymbol{x})$ of the same size, where:

$$h(\boldsymbol{x}_0) = (g * I)(\boldsymbol{x}_0) = \int_{\mathbb{R}^D} g(\boldsymbol{x}) I(\boldsymbol{x}_0 - \boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{5.1}$$

Here the spatial coordinates are continuous, i.e. $\boldsymbol{x} = (x_1, x_2, ...x_D) \in \mathbb{R}^D$, and $I(\boldsymbol{x})$ is the intensity at any given position.

However, with digital imaging the discrete case is considered, where the spatial coordinates are given via $\boldsymbol{k} = (k_1, k_2, ...k_D) \in \mathbb{Z}^D$: in practice, a discrete convolution of a filter $g[\boldsymbol{k}]$ of size $M^D = M \times .. \times M$ with an image $I[\boldsymbol{k}]$ of size $N^D = N \times .. \times N$ gives a response map $h[\boldsymbol{k}]$ defined by [86]:

$$h[\boldsymbol{k}_0] = (g * I)[\boldsymbol{k}_0] = \sum_{\boldsymbol{k} \in M^D} g[\boldsymbol{k}] I[\boldsymbol{k}_0 - \boldsymbol{k}]. \tag{5.2}$$

To be compact, this equation assumes the filter and image are square ($D = 2$) in $2D$, cubic ($D = 3$) in $3D$, and so on. The convolution is the operation $(g * I)$. In essence, the filter kernel $g[\boldsymbol{k}]$ slides across the image $I[\boldsymbol{k}]$ over all positions. Then at each fixed position $\boldsymbol{k}_0$ the intensity of the response map $h[\boldsymbol{k}_0]$ is the scalar product between the filter kernel, centred at that position, and the image.

The convolution operation can also be performed in *frequency* space. One can convert between the spatial and frequency domains of a function with a *Fourier transform*. Let the Fourier transform of a function $I(\boldsymbol{x}) \in \mathbb{R}$ be noted $\hat{I}(\boldsymbol{u}) \in \mathbb{C}$, where $\boldsymbol{u} = (u_1, ..u_D) \in \mathbb{R}^D$ is the frequency coordinates vector. The two operations to perform a transformation from the spatial to the frequency domain and back are known as the *Fourier transform pair* [121]. If $\mathcal{F}$ is the Fourier transform operator, then the transform pair can be defined as:

$$\hat{I}(\boldsymbol{u}) = \mathcal{F}\{I(\boldsymbol{x})\} = \int_{\mathbb{R}^D} I(\boldsymbol{x}) e^{-2\pi \mathrm{j} \boldsymbol{u} \cdot \boldsymbol{x}} \mathrm{d}\boldsymbol{x}. \tag{5.3}$$

$$I(\boldsymbol{x}) = \mathcal{F}^{-1}\{\hat{I}(\boldsymbol{u})\} = \int_{\mathbb{R}^D} \hat{I}(\boldsymbol{u}) e^{2\pi \mathrm{j} \boldsymbol{u} \cdot \boldsymbol{x}} \mathrm{d}\boldsymbol{u}. \tag{5.4}$$

where j is the imaginary symbol for complex numbers.

Again, for computation with digital imaging (i.e. pixels and voxels) one must consider a discrete version. The *Discrete Fourier Transform* (DFT) and inverse are given respectively in 1D via [121]:

$$\hat{I}[\nu] = \sum_{k=0}^{N-1} I[k]e^{-\mathrm{j}2\pi(\nu k/N)}, \tag{5.5}$$

$$I[k] = \frac{1}{N}\sum_{\nu=0}^{N-1} \hat{I}[\nu]e^{\mathrm{j}2\pi(\nu k/N)}. \tag{5.6}$$

where k and ν are the discrete coordinates in the spatial and frequency domains, for N samples. This extends to multiple dimensions via:

$$\hat{I}[\nu_1, .., \nu_D] = \sum_{k_1=0}^{N_1-1} ... \sum_{k_D=0}^{N_D-1} I[k_1, .., k_D]e^{-\mathrm{j}2\pi(\nu_1 k_1/N_1+...+\nu_D k_D/N_D)}, \tag{5.7}$$

$$I[k_1, .., k_D] = \frac{1}{N_1 \times ... \times N_D}\sum_{\nu_1=0}^{N_1-1} ... \sum_{\nu_D=0}^{N_D-1} \hat{I}[\nu_1, .., \nu_D]e^{\mathrm{j}2\pi(\nu_1 k_1/N_1+...+\nu_D k_D/N_D)}. \tag{5.8}$$

In practice, the DFT is computed with an efficient algorithm known as a *Fast Fourier Transform* (FFT). Nearly every programming language and software used for any signal or image processing will have the FFT implementation built in [121]. As an example, the SPAARC software is developed in Matlab, and makes use of the signal processing tool box which has functions, such as *imfilter* or *fftn*, that utilise FFT [52].

The key insight of the *convolution theorm* is that convolution can be performed in the frequency domain as a simple product:

$$\mathcal{F}\{(g * I)(\boldsymbol{x})\} = \mathcal{F}\{g(\boldsymbol{x})\} \cdot \mathcal{F}\{I(\boldsymbol{x})\}. \tag{5.9}$$

In other words, in the D-dimensional continuous case, the equivalent convolution defined in Eq. 5.1 can also be computed in the Fourier domain as:

$$(g * I)(\boldsymbol{x}) \overset{\mathcal{F}}{\longleftrightarrow} \hat{g}(\boldsymbol{u})\hat{I}(\boldsymbol{u}), \tag{5.10}$$

which remains true for the discrete case [86].

There are many computation advantages to filtering within the frequency domain compared to the spatial domain. As the size of the filter kernels increases, utilising FFT to calculate a convolution in the fourier domain will become more computationally efficient [121]. Some filters are also directly defined within the frequency domain, thus require use of an FFT by design.

5.2.2.1 Separable filters

Several filter types used in radiomic studies (e.g. *Gaussian*, *Laws*, *Wavelets*) have a quality know as *separability* [122]. As the name suggests, a filter kernel of any dimension is said to be separable if it can be obtained from the outer product of simpler 1D kernels [86]. For example, a 2D filter could be separated into two 1D filters, e.g. g_a and g_b. By convention, the combined 2D filter would be g_{ab}. Continuing this 2D example, this would mean convolution of an image $I[\boldsymbol{k}]$ with g_a in the k_1

direction, and then another convolution of the intermediate image with g_b applied in the k_2 direction, is the same as a single 2D convolution of g_{ab} with $I[\boldsymbol{k}]$. The resulting response map would be referred to as $h_{ab}[\boldsymbol{k}]$. If the 1D filters are not identical, there is of course a different convolution along each image axis. This introduces a directional bias which can have consequences for medical imaging with the need to potentially introduce rotational invariance strategies for certain separable filters.

5.2.2.2 Padding Types

Consider a spatial convolution. For a filter kernel centred on a given voxel, if the distance to the boundary of the image is less than half the width of the kernel, then some part of the filter will fall outside the image. To compute the value of the response map at that position requires estimating the value of voxels outside of the image boundary. This involves extending the image with an arbitrary padding decision. The choice of padding clearly will affect the response map intensities at these boundary positions. Depending on the size of the kernel, this can potentially affect a substantial number of voxels in the image. For analysis of a lot of medical imaging the ROIs are far from the boundaries (e.g. tumours centred in the image) so the padding choice would presumably not be a concern. Often padding decisions are not reported in radiomics literature, though it is clear they should be for consistency and reproducibility. There are four main types of padding:

Constant value - The image is padded with a constant value. Zero is often selected (*zero padding*) as the default padding for many applications.

Replicate / Nearest - The image is padded by repeating the nearest intensity found at the image boarder.

Circular / Periodic / Tiling - The image is padded by repeating the image along every dimension.

Mirror / Symmetric - The image is padded with reflections along each boarder. This method is a recommended choice for radiomics - if the ROI is near the boarder and could be influenced by padding - as it avoids sharp transitions that could be associated with *constant value* or *circular* methods.

5.3 Practical Matters for Filtering In Radiomics

5.3.1 Image Direction and Patient Orientation

With convolution, separable filters are applied to grids of voxels in set directions along each axis (i.e. along each row, column, and slice). This is the image coordinate system as shown in Figure 5.2. However, with medical imaging it is intuitive for filters to be applied in consistent *anatomical* directions for any particular dataset. This is often not reported or considered fully. Different software and medical applications can orientate 3D image volumes in various ways. The image coordinate system and patient coordinate system need not be aligned, but the patient should be *consistently* orientated in every image of the dataset. In other words, for maximum reproducibility the patient frame of reference should be orientated the same way in all images in a study. Note, this will usually be the case for a set of patients scanned with the same protocol, and often for CT imaging the patient coordinates are already aligned with the image coordinate system. If the dataset contains imaging with a mix of protocols and patient orientations (e.g. some oblique scans), it is recommended that all scans should be rotated and re-sampled appropriately such that patient orientation is the same in each image, using an additional preprocessing step.

Consistent patient orientation within image [86]

1. Report common orientation of image reference frame compared to the patient reference frame (e.g. in DICOM, the *Image Orientation Patient* (0020 0037) field)
2. Report rotations required for each image to align to common orientation (if necessary)

5.3.2 Directional Dependency

An argument for consistent orientation for patients discussed in the previous section is that many filters are directionally sensitive. Thus, it is important for reproducibility within a study to apply these types of filters in the same way anatomically for every patient. Directional sensitivity is a necessary characteristic that allows for differentiation of structures through the varying interaction of the filter kernel at different edge orientations and tissue boundaries in the image.

However, this directional sensitivity leads to variation in the response map depending on the global rotation of the input image, which is a draw back as rotations of the same structure result in different filter responses. These different responses will likely aggregate into different feature values. Ideally, for medical image analysis features would be robust to rigid rotation and translation. If an image volume is the same yet globally rotated in some manner - for example when viewed from a sagittal plane instead of axial - then ideally, the same filter-based image biomarker should be aggregated from the ROI through analysing either orientation of the image.

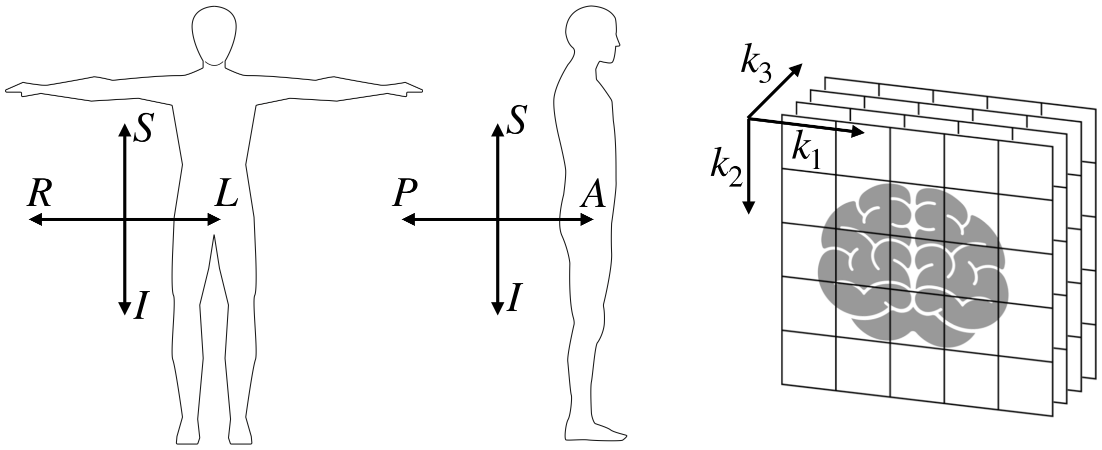


Figure 5.2: Illustration of the anatomical coordinate system directions and the image coordinate directions. Here, S = Superior, I = Inferior, A = Anterior, P = Posterior, R = Right, L = Left. For the image coordinate system, the k_1 direction refers to moving along the image matrix rows, k_2 refers to down the image matrix columns, and k_3 is down though the image slices. For radiomics, filters should be applied in a consistent anatomical direction for the entire dataset and reported, as software can orientate the images in various ways.

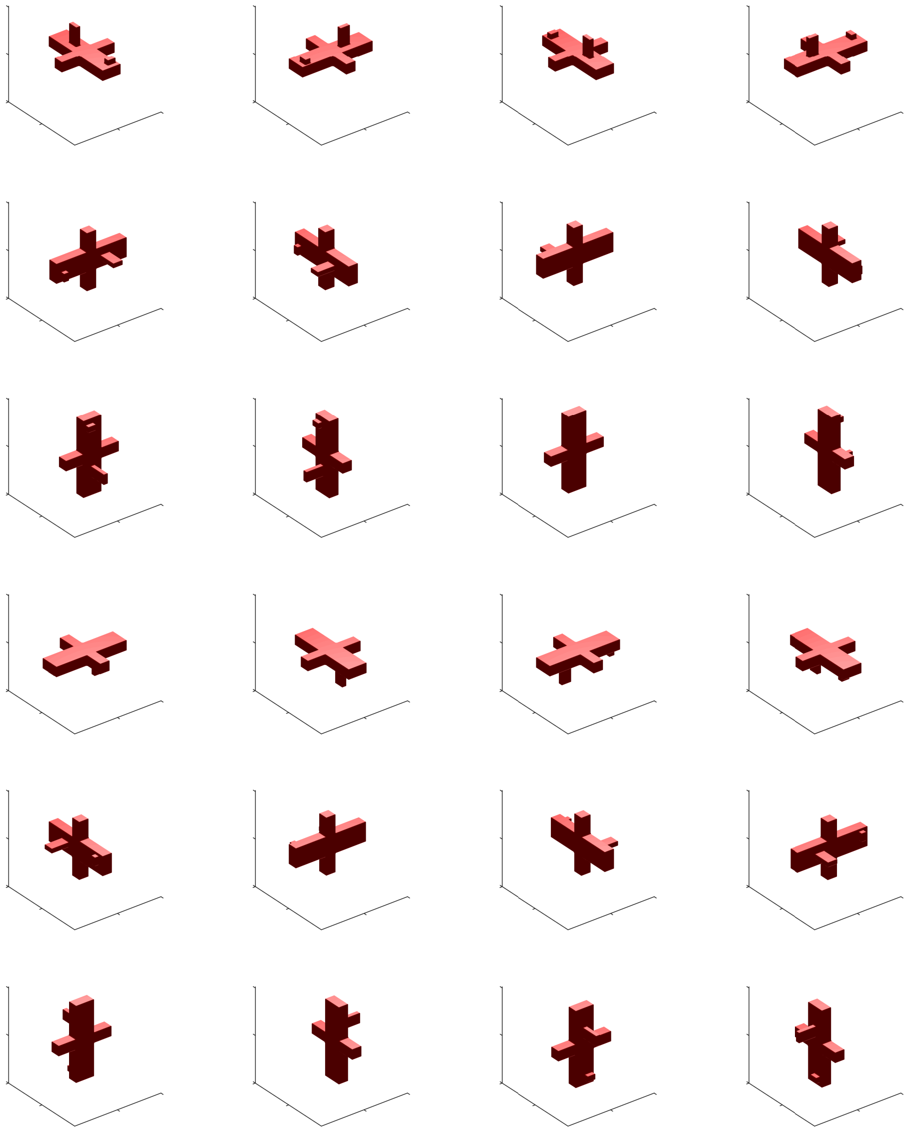


Figure 5.3: Visualisation of all 24 unique global axial rotations of a 3D image volume. This Figure was made using the rotational functionality built into SPAARC radiomics. A binary image was built to resemble a humanoid shape inside a volume. The binary image was structured such that it is not symmetric across any plane, so that each rotation is distinct. As a binary image, each rotation was then visualised using a meshgrid.

Depeursinge *et. al.* [2] refer to this desired property of filters as *equivariant*, which implies that if the input image is translated or rotated, then the response map would follow the same transformation. Common filters such as *separable wavelets* and *Laws* (introduced in Section 5.4) do not have this property inherently due to their dependence on orientation. As demonstrated here, this has a consequence on radiomic features. This limitation has not been considered by many previous radiomics studies that have utilised these filters (e.g. wavelets in [16]) to extract potential imaging biomarkers, and in the aim here is to determine benchmarks that include equivariant representations of these filters.

A strategy for achieving rotational invariance with *separable* filters displaying directional sensitivity is to generate a collection of response maps from all global image orientations. These response maps for each rotation can then be pooled, voxel-wise, by taking either the maximum or average voxel intensity [86]. Note that the rotated image volumes must be re-rotated to the original orientation prior to pooling. In this work, these pooled response maps are referred to as h_{max} or h_{avg}, respectively. For a 2D image there are 4 global right-angle rotations to consider, and for 3D there are 24, which are visualised with an example in Figure 5.3. Instead of rotating the images, one can one can also consider rotating the kernels. For separable filters, this can be achieved via convolution of a collection of permuted and flipped versions of the filters [86].

5.4 Selected Convolutional Filters for Radiomics

This section discusses some of the more common filtering techniques used in radiomics in the search for potential imaging biomarkers. The filters introduced here are as such part of the collection being considered for standardisation in the IBSI [86]. Filter types are discussed alongside implementation decisions that may lead to discrepancy where appropriate.

5.4.1 Mean

The *mean* (or *average*) filter is perhaps the simplest to consider. Each voxel intensity in an image becomes an average of its neighbourhood over the size of the filter kernel. The resulting response map after applying a mean filter is a smoothed representation of the input image. As such, mean filters are often utilised to reduce the influence of potential noise. In practice, it is intuitive to select only odd sized kernels so there is a clear central voxel. To construct a D dimensional mean filter kernel of size M^D, each intensity within the filter is assigned an intensity value of $1/M^D$. For example, a $2D$ mean filter of $M = 5$ would be a 5×5 grid containing voxels of intensity $1/25$. Note that odd M kernels can also be reported in terms of a δ variable, where $M = 2\delta + 1$.

5.4.2 Laws Filtering

An influential approach to quantifying texture in imaging was first presented in 1980 by Laws [123]. Laws' kernels are a set of 5 types of 1D filters designed to emphasise different structures within an image, namely: *ripples*, *edges*, *spots*, *waves*, alongside *Level* (a low pass for grey level averaging). Table 5.1 introduces these filter kernels at both a scale of 3 and 5 voxels where appropriate. These filters can be combined in numerous ways to obtain 2D and 3D filters, as demonstrated in Figure 5.4, though are separable (see Section 5.2.2.1) by design. Response maps are obtained by applying a selected kernel in each direction in the image. For example, h_{E5S5E5} would be acquired with convolution of g_{E5} in the k_1 direction, then g_{S5} in the k_2, and finally g_{E5} in the k_3 direction.

5.4.2.1 Laws Texture Energy Images

The response maps from applying Laws' kernels are transformed into so-called *texture energy* images, h_{energy}, by summarising the amount of variation within a filter window [123]. This is

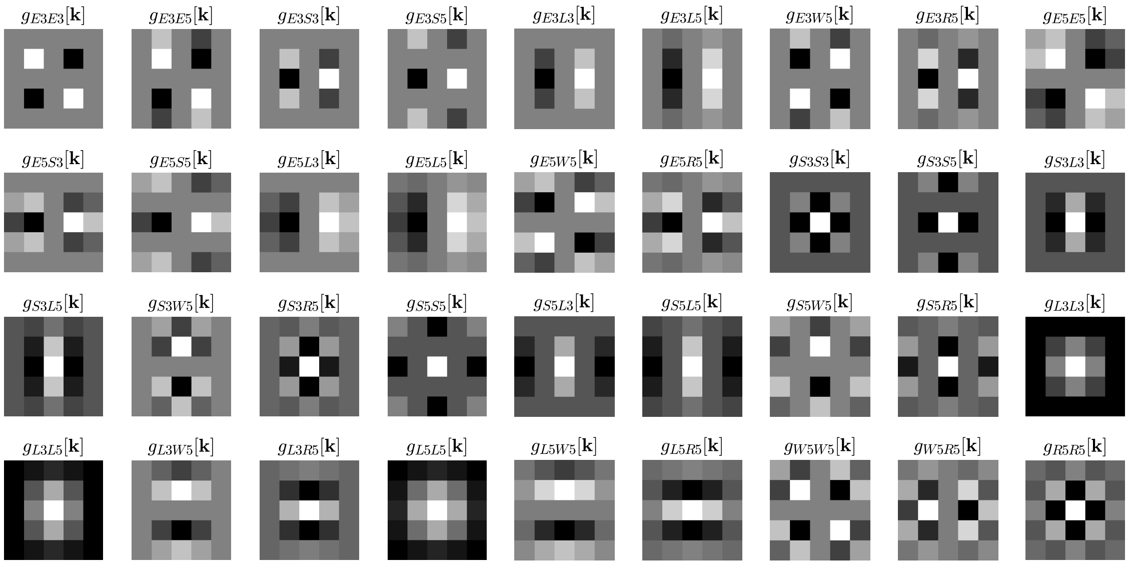


Figure 5.4: Visualisations of a selection of 2D Laws Kernels generated by separable convolution with a 5×5 impulse and zero padding.

achieved by taking the absolute values from the response map and applying a mean filter (Section 5.4.1) of a chosen size. In the original work on 2D imaging, Laws suggested a 15×15 moving window to calculate the texture energy image [123], though for medical imaging purposes this can be varied depending on the physical resolution of the image and task at hand. The size of the mean filter used to calculate the texture energy images is an important parameter to report for reproducibility, alongside the kernels used.

A response map from convolution with a selected set of Laws kernels is not rotational invariant by default (see Figure 5.5). As discussed in Section 5.3.2, rotational invariance can be achieved by calculating response maps for all image orientations, then re-rotating these response maps and pooling. A single rotationally invariant h_{energy} can then be calculated from these pooled response maps. Note that this is not the same as calculating a h_{energy} for each orientation and *then* pooling this set of texture energy images. The author found that these two approaches are

Table 5.1: Normalised representations of Laws 1D kernels [86, 123].

Laws Filter	**1D kernel**
Level	$g_{L3}[k] = \frac{1}{\sqrt{6}} \cdot [1, 2, 1]$ $g_{L5}[k] = \frac{1}{\sqrt{70}} \cdot [1, 4, 6, 4, 1]$
Edges	$g_{E3}[k] = \frac{1}{\sqrt{2}} \cdot [-1, 0, 1]$ $g_{E5}[k] = \frac{1}{\sqrt{10}} \cdot [-1, -2, 0, 2, 1]$
Spots	$g_{S3}[k] = \frac{1}{\sqrt{6}} \cdot [-1, 2, -1]$ $g_{S5}[k] = \frac{1}{\sqrt{10}} \cdot [-1, 0, 2, 0, -1]$
Ripples	$g_{R5}[k] = \frac{1}{\sqrt{70}} \cdot [1, -4, 6, -4, 1]$
Waves	$g_{W5}[k] = \frac{1}{\sqrt{10}} \cdot [-1, 2, 0, -2, 1]$

not be equivalent, so to alleviate a potential source of discrepancy it is recommended to calculate the h_{energy} once, after orientation pooling of the response maps. Additionally, for simplicity when reporting it makes intuitive sense to use the same padding technique for both the application of the Laws kernels and the subsequent computation of h_{energy}.

5.4.3 Laplacian of Gaussian

A Laplacian is a second order derivative operator and as such emphasises sharp intensity transitions (such as edges). A Gaussian smooths intensity values and reduces noise. A popular combination of these, known as the Laplacian of Gaussian (LoG) filter - which corresponds to a second spatial derivative of a Gaussian - is defined in D dimensions as [86]:

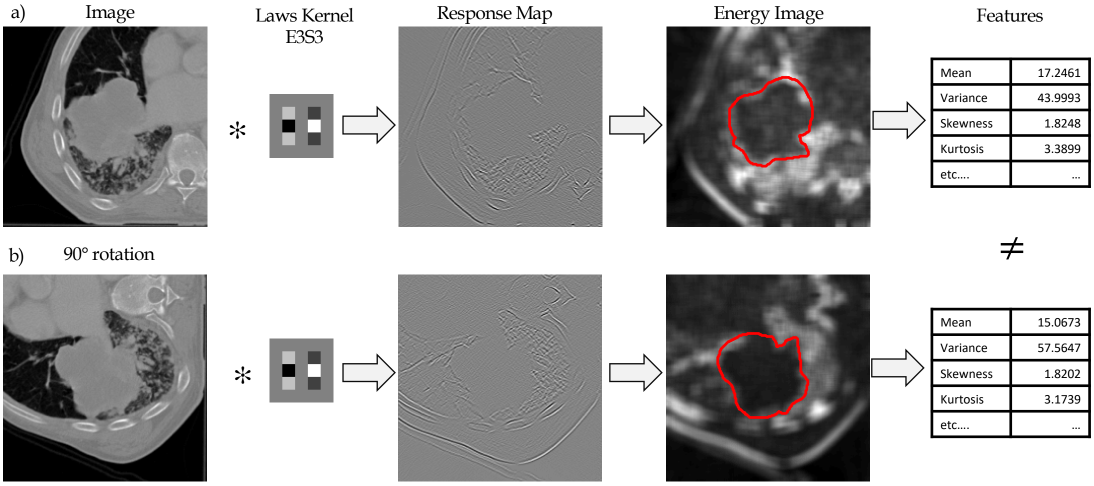
$$g_\sigma[\boldsymbol{k}] = -\frac{1}{\sigma^2}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^D\left(D - \frac{||\boldsymbol{k}||^2}{\sigma^2}\right)e^{-\frac{||\boldsymbol{k}||^2}{2\sigma^2}}. \tag{5.11}$$

As before, $\boldsymbol{k} = (k_1, k_2, ...k_D) \in \mathbb{Z}^D$ are the discrete spatial coordinates, and σ is the Gaussian standard deviation term, which effectively controls the scale of the filter. As a single global σ is utilised here, i.e. the same for each spatial direction, the LoG filter has circular symmetry as shown in Figure 5.6.

For medical imaging it is desirable to define σ in terms of physical measurement, which must be translated into a "voxel spacing" equivalent σ_v via:

$$\sigma_v = \frac{\sigma}{\text{voxel spacing}}. \tag{5.12}$$

Filtering packages often assume σ_v (e.g. MATLAB's *fspecial3* [52]), which must be handled carefully during implementation with parameters reported in mm. This equation assumes that the



Mean	17.2461
Variance	43.9993
Skewness	1.8248
Kurtosis	3.3899
etc....	...

Mean	15.0673
Variance	57.5647
Skewness	1.8202
Kurtosis	3.1739
etc....	...

Figure 5.5: Demonstration that some Laws filter kernels are not rotationally invariant. Example uses a single slice from `Dataset-2` introduced in Chapter 3. The image is convolved with a g_{E3S3} kernel (achieved through separable convolution) that results in a response map h_{E3S3}. The absolute values from this response map are then averaged with a mean filter of size $\delta = 5$ to calculate h_{energy}. Features are then aggregated from the red contour. The only difference between row a) and b) is that the input image has been rotated 90 degrees. This results in different extracted radiomic features (show are the first 4 statistical values). To achieve rotational invariance, the response maps from all orientations should be pooled prior to the calculation of the h_{energy}.

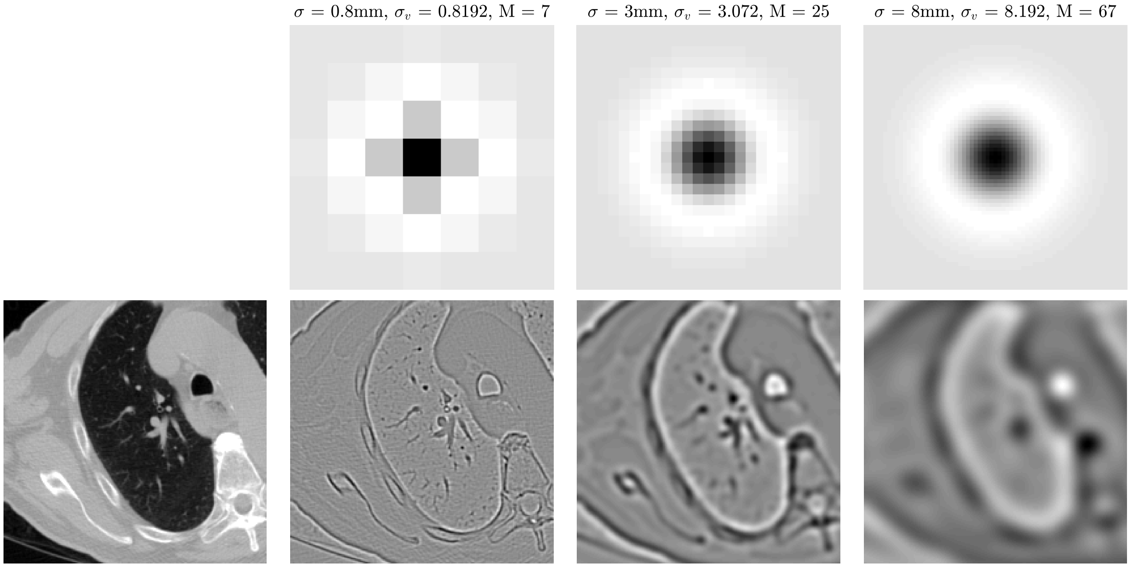

σ = 0.8mm, σ_v = 0.8192, M = 7
σ = 3mm, σ_v = 3.072, M = 25
σ = 8mm, σ_v = 8.192, M = 67

Figure 5.6: Filtering examples on CT imaging using 2D LoG kernels generated at various scales. CT pixel resolution of the original image (left) is 0.976 by 0.976 mm. The LoG kernel for each convolution is visualised above the corresponding response map. Scales shown: σ = 0.8 mm, 3mm, and 8mm. Filter cut-off was set to 4σ, (i.e. $d = 4$ in Eq. 5.13.) Imaging taken from STAGE dataset.

image has isotropic voxels/pixels.

Practically, a LoG filter kernel also needs to be cropped as the spatial support technically extends between $-\infty$ and ∞. The IBSI thus suggest a cut-off to be decided based on a multiple (d) of σ_v. Formally, the recommendation is to set the size of the kernel in each direction (e.g. $M \times M$ in 2D) with

$$M = 1 + 2\lfloor d\,\sigma_v + 0.5 \rfloor, \tag{5.13}$$

so that the kernel will always be odd and cannot be less than 1 voxel.

5.4.4 Separable Wavelets

Wavelet analysis filtering techniques enable the assessment of image frequency content at a range of scales [124]. A set of low- and high- pass filters are convolved with the image in combination to generate band-pass response maps, also referred to as the *wavelet coefficients*. Effectively, multi-scale analysis is achieved through iterative decimation (down-sampling) of the response map by a factor of 2, or up-sampling of the filters by a factor of 2. As such, there are two distinct approaches to wavelet transforms: *decimated* and *undecimated*, both of which are discussed in the following sections. Both techniques have been utilised in radiomics analysis [111]. The undecimated approach is likely more appropriate for the subsequent extraction of imaging biomarkers from a VOI as the response map remains a consistent size, which is easier to implement into a radiomics pipeline as the masks do not need to be decimated to match the new image size. The IBSI core-group recommend the undecimated approach.

With separable wavelets, this *decomposition* analysis is performed starting with a *mother wavelet*, which is the high-pass filter $g_H[k]$, and low-pass filter $g_L[k]$ that is called the *scaling function* [86]. Common types of wavelets include *Haar*, *Daubechies*, and *Coifflet* . There is a substantial theoretical

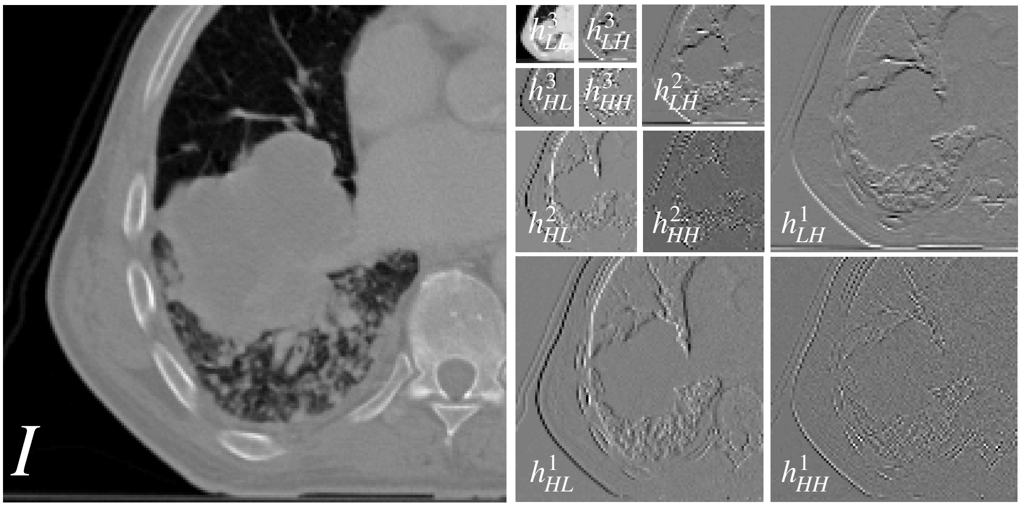


Figure 5.7: An example of a level 3 decimated wavelet transform on a CT imaging slice in 2D, using the Haar wavelet. Note that in this compact visualisation only the final low pass response h^3_{LL} is shown, as the others become the input for each subsequent decomposition. At each decomposition level the image is reduced in size by a factor of 2.

basis behind wavelet analysis, and to keep the scope manageable, this chapter discusses only the key concepts of how these different decomposition response maps are generated for subsequent radiomic analysis.

5.4.4.1 Decimated Transform

The decimated wavelet transform efficiently decomposes the input image into non-redundant wavelet coefficients. With the decimated transform, as you move through the levels of decomposition, the imaging is decimated by a factor of 2. As such, the starting image dimensions must be a multiple of 2^N and during padding this criteria should be met if this is not the case.

To perform a decimated wavelet transform: first, the selected high and low pass filters ($g_L[k]$ and $g_H[k]$) are convolved along each image direction for all unique combinations. In 2D, this results in 4 response maps, h_{LL}, h_{LH}, h_{HL}, and h_{HH}, and in 3D this gives 8, h_{LLL}, h_{LLH}, h_{LHL}, h_{LHH}, h_{HLL}, h_{HHL}, h_{HLH}, and h_{HHH}. At this point, these response maps are then down sampled by a factor of two. This is the first level of the decomposition, e.g. h^1_{LH}.

This process can then be repeated to the desired decomposition level j, where the low-pass response map h^j_{LL} is used as the input for the next $(J+1)$ decomposition level. Figure 5.7 highlights this process with a 2D example using a slice from CT imaging.

5.4.4.2 Undecimated Transform

In contrast, the undecimated or *stationary* wavelet transform does not require the downsampling used by the decimated method. As the name implies, the response map remains stationary in size for each decomposition, which introduces redundancy. However, this work is not concerned with other wavelet uses such as efficient image compression. As such, this approach is actually advantageous for radiomic analysis as the VOI does not need to be down sampled to match the decimated response maps. Rather than downsampling the image, instead the low and high pass

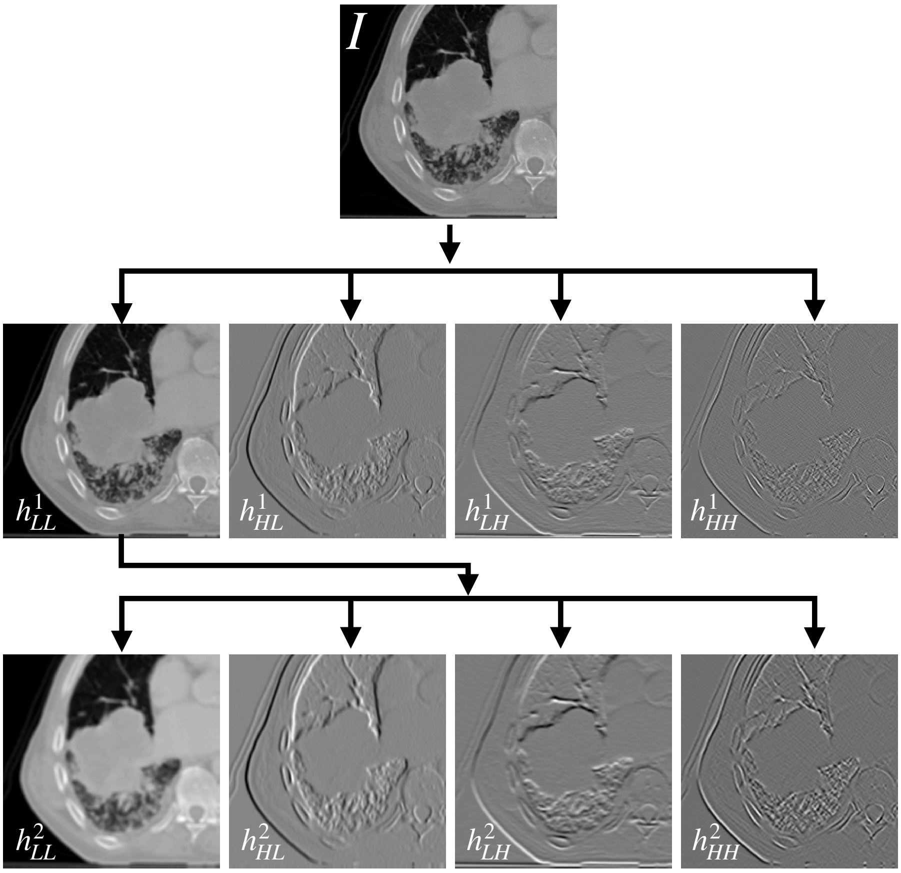


Figure 5.8: An example of an undecimated (stationary) wavelet transform up to decomposition level 2 on a CT imaging slice in 2D.

filters are upsampled for each decomposition level by inserting zeros between coefficients. For example, in practice a *Haar* high pass filter

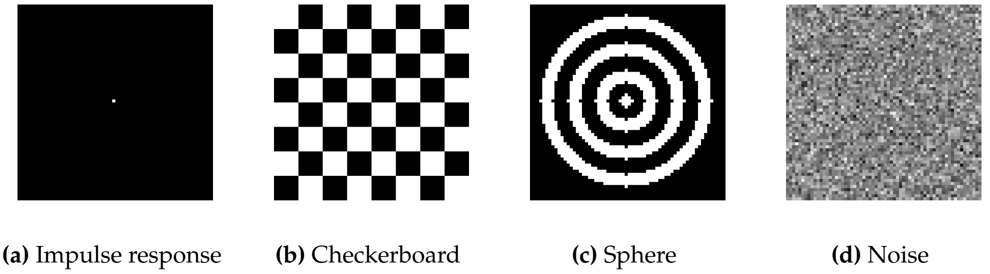
$$g^1_H = [-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$$

is up-sampled to:

$$g^2_H = [-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0]$$

for the next decomposition level. As highlighted in 5.8, as with the decimated transform, h^J_{LL} becomes the input for the $J+1$ decomposition.

This and previous sections have described filters currently implemented in the SPAARC software. Other filters targeted for standardisation include the *Gabor* transform and non-separable *Riez* aligned directional wavelet filters [86]. In the next sections, an overview is given of the aims and standardisation phases that will be utilised to determine a consensus on these filtering techniques within the IBSI. This will improve reproducibility of studies, which is critical when discussing and evaluating the clinical potential of imaging biomarkers derived from these techniques.



(a) Impulse response **(b)** Checkerboard **(c)** Sphere **(d)** Noise

Figure 5.9: Visualisation of the central 2D slice for a selection of the digital phantoms generated by the IBSI to be used for the different filter tests [125].

5.5 Overview of IBSI Filter Standardisation Phases

To standardise the image filtering processes for radiomics, the IBSI aim to utilise the same general three-phase study structure as the first instalment discussed in Section 3.2.2.1 of Chapter 3. In this instance, the goal is to provide reference values for these filter-based features as before. In addition, the aim is to generate valid consensus-based response maps for a range of tests for key filtering techniques. The planned three phases are summarised below.

Phase 1 will evaluate the application of different filters on a set of digital phantoms without any other image processing. These digital phantoms are introduced in Section 5.5.1 and the filter tests designed for this work are summarised in Table 5.2. Rather than assessing aggregated features values, the goal is to compare the entire response maps submitted by teams for each filter test. The aim is to produce a set of validated consensus based response maps for each filter test to act as benchmarks for software calibration.

Phase 2 will then establish a set of reference feature values obtained after applying these convolutional filters to `Dataset-2`, the lung CT utilised in the previous study (see Section 3.2.2.1). The statistical and intensity histogram features will be examined using the same criteria as the first study. Here, other image processing steps such as interpolation and discretisation are stacked on top of the filtering process to represent a typical extraction scenario in radiomics.

Phase 3 will act as a final validation for the reproducibility of filter-based features that were able to be standardised in the first two phases. As before, reproducibility of features will be assessed using `Dataset-3`, which contains PET MRI and CT imaging of 51 patients with STS.

This chapter discusses the key preliminary work required to conduct **Phase 1**. Section 5.6 introduces the methods to evaluate the response maps received from each filter test. The presented techniques can determine if there is a valid *consensus response map* (CRM) and agreement between different software. In Section 5.7, this methodology is used to analyse the initial submissions from contributing teams.

5.5.1 Phase 1 Digital Phantoms

For comparison of different filter implementations, the IBSI has produced a set of 3D digital phantoms to be used as the input images for a variety of filter tests (see Table 5.2)) [86]. These phantoms are available online [125]. All phantoms used for the filter tests have the same form: a dimension of $64 \times 64 \times 64$ voxels, a physical voxel size of $2 \times 2 \times 2$ mm, intensity precision of 8-bit, and voxels ranging from [0, 255]. Prior to filtering, the phantom images should be converted to *at least* 32 bit precision. The four key phantoms utilised for the filter tests discussed in this chapter are introduced in the box below.

The central slice of these phantoms have been visualised in Figure 5.9. When filters are applied to the phantoms, they yield response maps of the same size (i.e. $64 \times 64 \times 64$). By analysing the results of different implementations, a consensus will ideally be reached for each filter test, producing a valid CRM. To compare software implementations, the next section introduces the methodology designed to perform the consensus analysis.

Phantom Overview

`Impulse response`: Designed to visualise the filter kernel. Contains a single voxel with intensity of 255 with the rest equal to 0.

`Sphere`: A spherically symmetric phantom consisting of 4 spherical shells of increasing radius from the centre.

`Checkerboard`: Cubic regions of 8 × 8 × 8 voxels, alternating between minimum of 0 and maximum intensity of 255.

`Noise`: Unstructured integer intensities generated with Gaussian noise centred on intensity 127 with a standard deviation of 48.

5.6 Developing a Methodology for Response Map Comparisons

In pursuit of a *valid* CRM for each filter test, response maps are compared from different software to identify if there are significant discrepancies. Intuitively, if there is no discrepancy between teams there would be zero variance between the submissions.

Presented here is the methodology designed by the author to measure the consensus for each filter test, and a process to iterate towards a meaningful CRM in the presence of discrepancy. The response map data is analysed in two main ways: (1) assessment of variation between all response maps simultaneously, and (2) pairwise comparisons between each response map and the *average* result. Other approaches the author considered are left to the discussion.

The first technique utilises *Principle Component Analysis* (PCA) (Section 5.6.1), the second uses *difference imaging* to evaluate a voxel-wise passing rate based on a set tolerance (Section 5.6.2). These techniques are combined in an effort to iterate towards an optimised CRM by identifying and removing the contributions of outlier submissions if appropriate (Section 5.6.3).

5.6.1 Evaluating Consensus with PCA

From all submissions of a given filter test, a *preliminary* CRM can be considered by simply calculating the average response map in a voxel-wise manner.

Principle Component Analysis (PCA) [126] is used as a way to evaluate a preliminary CRM and visualise variances between all teams simultaneously. The key concept is that similar observations will cluster together on a PCA plot, offering an excellent way to evaluate differences, with the mean result (CRM) corresponding to the centroid of this plot. The euclidean distance to the centroid can give a measure of *similarity* between any one response map and the mean result. The key use of PCA by the author is to identify outliers.

5.6.1.1 PCA technique overview

PCA reduces the dimensionality of data by projecting it into dimensions referred to as principal components (PCs). The PCs are uncorrelated (orthogonal) to one another and retain key trends of the variation found in the data. PCs are ordered by their explained variance, where the first PC represents the maximum variance direction in the data.

Figure 5.10 illustrates finding the principle components for some simulated data consisting of 10 observations of 2 variables (i.e. each observation is 2-dimensional). The concept remains the same when considering much higher dimensions.

The first step in PCA is to organise the data into an n-by-v data matrix, where the rows (n) correspond to observations and the columns correspond to the variables (v). In essence, the PCA approach can then performed in the following way: (1) centre the data for each column by subtracting the mean of each column, (2) calculate a covariance matrix from the centred data matrix, (3) calculate the eigenvectors and corresponding eigenvalues with an eigendecomposition of the covariance matrix, (4) sort eigenvectors from largest to smallest based on the eigenvalues, (5) transform the centred data matrix into the principle components sub-space by multiplying with the matrix of eigenvectors. If required, the reader is encouraged to consult referenced material for further details concerning eigendecomposition [127]. In practice, for this work *pca* functionality from Matlab [128] was utilised once the data was organised into the required matrix, and an example is provided in the following section for clarification of this approach.

5.6.1.2 PCA for Response Map Comparison in Practice

A response map can be thought of as high dimensional data. All filter tests in this work are applied to phantoms with $64 \times 64 \times 64 = 262144$ voxels, and thus the response maps submitted by each team are the same size. For PCA, each voxel is considered a separate dimension (or variable), and each response map is a single observation point in this $\mathbb{R}^{64\times64\times64}$ (or $\mathbb{R}^{262144}$) space. In other words, a response map from these filter tests can be represented as a single point in a 262144-dimensional variable space.

The key variation and distribution of many response maps within the $\mathbb{R}^{64\times64\times64}$ space can thus be visualised by plotting the first two principle components. As mentioned, the average result corresponds to the centroid of these plots.

As a simplified demonstration, let's consider mock results for response maps that are 3 by 3 pixels and thus represented in $\mathbb{R}^{3\times3}$ space instead.

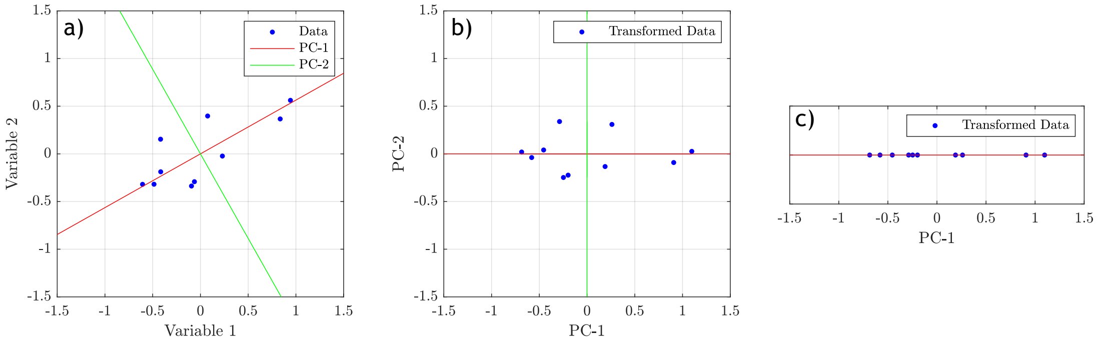


Figure 5.10: Illustration of *Principle Component Analysis* (PCA) with a low dimensional example of 2 variables. Here 10 observations of 2 variables (randomly generated data) are transformed on to the corresponding principle components. The first principle component (PC) is an axis through the data that represents the direction of maximum variation (shown in red). All PCs are orthogonal to one another. Here, **a)** shows the plot of the variables against each other for the 10 observations. For this example the variables are already centred. Also shown is the corresponding principle component axes. In **b)**, the example data is transformed onto the principle components. In **c)**, the example data is project onto just the first PC.

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 2 \\ 3 & 2 & 5 \\ 1 & 5 & 2 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 5 & 1 & 2 \\ 3 & 2 & 5 \\ 1 & 5 & 2 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 4 & 1 & 2 \\ 3 & 8 & 4 \\ 1 & 4 & 2 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 5 & 1 & 2 \\ 3 & 1 & 5 \\ 1 & 5 & 3 \end{bmatrix}$$

In this example, Team **C** has deliberately been made to be the outlier and vary more than the others. Team **A** and **B** are identical, and **D** has 2 pixels that vary from these two teams by 1 intensity value.

To prepare these response maps for PCA, each one is collapsed into a single row and combined into a data matrix, where each row is a single observation (a response map) and each column is the intensity value associated with a pixel (or voxel) at a set position (i.e. the variables). For this example, this would be a 4-by-9 matrix. If Teams are ordered **A** to **D** for each row, the data matrix for this example is as follows:

$$\text{Data Matrix} = \begin{bmatrix} 5 & 3 & 1 & 1 & 2 & 5 & 2 & 5 & 2 \\ 5 & 3 & 1 & 1 & 2 & 5 & 2 & 5 & 2 \\ 4 & 3 & 1 & 1 & 8 & 4 & 2 & 4 & 2 \\ 5 & 3 & 1 & 1 & 1 & 5 & 2 & 5 & 3 \end{bmatrix}$$

Once data is prepared in this way, it is passed to the *pca* function [128] (with the eigendecomposition method selected) that performs the PCA as outlined in Section 5.6.1.1. The returned result represents a centred version of the data matrix transformed into principal component space (referred to as the *score* [128]).

$$\text{PC representation} = \begin{bmatrix} -1.2986 & -0.3553 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1.2986 & -0.3553 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4.9297 & 0.1012 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2.3326 & 0.6094 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

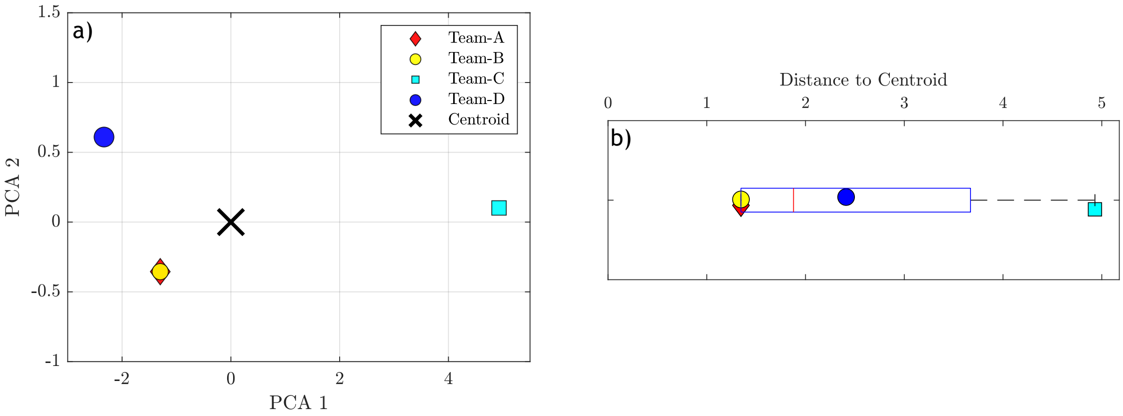


Figure 5.11: Example of *Principle Component Analysis* (PCA) technique used to compare response maps. Shown are the results of PCA analysis for a simplified example of 4 mock response maps discussed in Section 5.6.1.2. Team **A** and **B** are identical, and as such occupy the same point in PCA space. **C** has the most variation, and as such is the greatest euclidean distance from the centroid. **a)** Is a plot of the position of each response map for the first 2 principle components, **b)** plots the euclidean distance of each team to the centroid and includes a box plot to summarise the distribution.

Here each column corresponds to a principle component (in order) and each row is an observation, which is the same order as the data matrix. Values here are shown to 4 decimal places. Intuitively, the first two columns are used for the PCA plots. The PCA plot and distance to centroid for this example are shown in Figure 5.11.

As expected from the example data, Team A and B lie on the same point in the PCA plot as they are identical, and the outlier Team C is furthest away from the centroid. For this example, all of the variation is captured with the first two principle components. In general, the euclidean distance to the centroid is calculated using all non zero components not just the first two. As an example, the euclidean distance to the centroid (which is located at the origin) for Team C is given via

$$\sqrt{4.9297^2 + 0.1012^2} = 4.9307.$$

PCA For Response Map Comparison Key Points

- A response map can be represented as a single *observation* in high dimensional space, where each voxel intensity is a separate *variable*.
- PCA can be used to assess the variation between many observations in high dimensional space.
- Data is projected onto a subspace (the principle components) whilst retaining the essence of variation. Similar observations will cluster together on a PCA plot.
- The euclidean distance to the centroid (corresponding to the average result) can be used to identify outliers.

This PCA technique is found to scale well with increasing numbers of observations, making it well suited for consensus testing. To demonstrate this, Figure 5.12 shows the result of PCA for 50 simulated response maps (of size 64 × 64 × 64). The intensity values of the simulated response

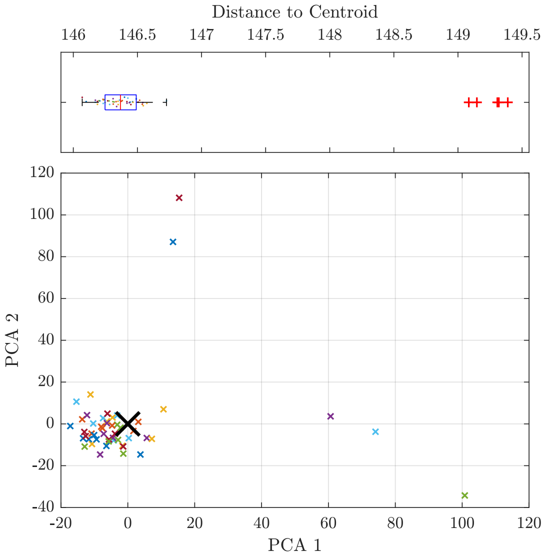


Figure 5.12: Illustration that the *Principle Component Analysis* (PCA) technique scales well when comparing a high number of response maps. Here, 50 simulated response maps (each of size 64 × 64 × 64) randomly generated from two distinct uniform distributions (split into N=45 and N=5) are compared. Correspondingly, 5 response maps can be seen as outliers with a larger distance to centroid (the large cross) measured.

maps were randomly generated from two distinct uniform distributions (split into N=45 and N=5). Correspondingly, in Figure 5.12 there are 5 response maps that have a much larger distance to centroid and show up as outliers compared to consensus (marked as a larger cross).

5.6.1.3 Limitations of PCA

The PCA technique neatly summarises overall discrepancy when comparing response maps. It can efficiently indicate when there is no variation between submissions, or obvious outliers. Through the visualisation, response maps that are alike are closely grouped, and outlier submissions that require further review can easily be identified. Perfect agreement between all teams lead to zero distance to the centroid in all cases, though it should be noted in practice that subtle errors due to machine precision will lead to extremely small non-zero distances. The goal for standardisation is to minimise this distance to centroid as much as possible.

However, when using just PCA on its own it may be difficult to assess the validity of some generated consensus response maps. Subtle disagreements between many submitted response maps may significantly affect a CRM, yet produce no obvious outliers when measuring the distance to the centroid with PCA. The scale variability of the principle components make it challenging to determine tolerances for an *acceptable* distance to consensus (i.e. how far away from the centroid is passable?). This distance can be hard to interpret as a single reported measure. As a result, the next section introduces a further analysis technique using pairwise assessment, where a generated CRM is also compared separately with each response map.

5.6.2 Pairwise Assessment of Response Maps

A CRM can be further evaluated using pairwise techniques. Naturally, a strong CRM should show close agreement with every team used to generate it. Moreover, pairwise assessment allows for definitive tolerances to be reported. If pairwise analysis of all teams with the CRM are within a small tolerance, one can be confident the generated CRM is a standardised representation of a particular filter test. If not, iterative approaches can be considered to remove dissident contributions, as long as an absolute majority of the submissions remain.

5.6.2.1 Difference Image

To compare any two image matrices of the same size, an intuitive approach is to look at voxel-wise differences. This technique results in a single output matrix where each voxel intensity now quantifies changes between the two input images, referred to as a *difference* or *error* image. By visualising this output, one can quickly pinpoint the locations and level of disagreement. This is demonstrated in Figure 5.13.

In general, for each voxel at position $\mathbf{k}$, the difference, $D[\mathbf{k}]$, of two images, $A[\mathbf{k}]$ and $B[\mathbf{k}]$, is calculated via

$$D[\mathbf{k}] = \frac{|A[\mathbf{k}] - B[\mathbf{k}]|}{C}, \tag{5.14}$$

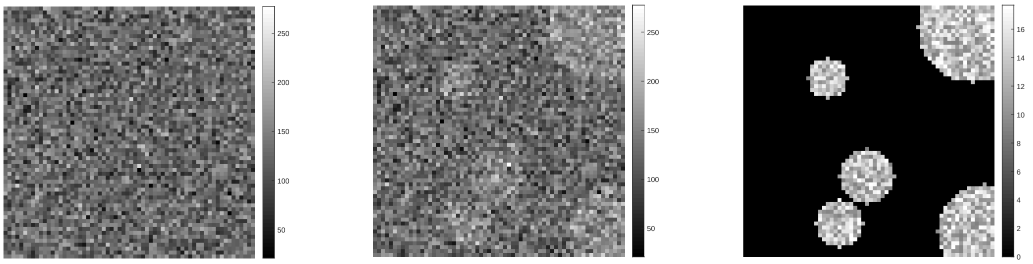
where C is an optional function to scale the absolute difference ($|A[\mathbf{k}] - B[\mathbf{k}]|$). Examples of C could be the combined *max*, *median*, or *range* of the input images. Different filter tests have varying ranges. In the following work, $D[\mathbf{k}]$ is scaled by the range of the response maps being tested, i.e.

$$D_r[\mathbf{k}] = 100 \times \frac{|A[\mathbf{k}] - B[\mathbf{k}]|}{\max(A, B) - \min(A, B)}. \tag{5.15}$$

In this case, the greatest value $D_r[\mathbf{k}]$ can reach is 100%, which would occur if a maximum intensity is compared to a minimum.

Image difference can be summarised further by reducing the dimensionality of $D[\mathbf{k}]$ into a single value. Indeed, one can report a variety of statistical measures of $D[\mathbf{k}]$, such as the *mean*, *median*, *variance*, *minimum*, *maximum*, *kurtosis*, or *skewness*. However, reducing $D[\mathbf{k}]$ into just a single value loses information which can have consequences for measuring discrepancy. As an example, different types of variation can result in the same statistical value. Two different $D[\mathbf{k}]$ could have the a similar mean, yet one is from a slight discrepancy amongst many voxels, and the other is caused by a larger discrepancy in only a few voxels.

What is considered an acceptable pairwise variation will also be based on variation below a set *tolerance*. Instead of single statistical measures of $D[\mathbf{k}]$, this can be evaluated on a voxel-wise level using what the author defines here as *passing rate plots*.



(a) Image A

(b) Image B

(c) Difference image

Figure 5.13: Example of *difference image* analysis. Image **a)** is the central slice from the IBSI2 *noise phantom*, **b)** is the noise phantom after injected with 5 randomly placed spherical patches of variation up to 20%. Image **c)** is the difference between **a)** and **b)**, scale by the range (see Eq. 5.15).

5.6.2.2 Tolerance & Passing Rate Plots

Here a *Passing mask* $P_\gamma[\mathbf{k}]$ is defined by setting a maximum tolerance γ for $D[\mathbf{k}]$. In other words, if γ is the limit of acceptable variation between two voxels, then:

$$P_\gamma[\mathbf{k}] = \begin{cases} 1 & \text{when } D[\mathbf{k}] \leq \gamma \\ 0 & \text{otherwise.} \end{cases} \tag{5.16}$$

Each voxel in $P_\gamma[\mathbf{k}]$ is assigned a value of 1 if below or equal to the tolerance, and 0 if above. The percentage of voxels that pass, R_γ, can be calculated for a given γ using $P_\gamma[\mathbf{k}]$ via

$$R_\gamma = 100 \times \frac{1}{N} \sum_{\mathbf{k}} P_\gamma[\mathbf{k}], \tag{5.17}$$

where N is the total number of voxels. Intuitively, if all voxels are below the tolerance, then $\sum_{\mathbf{k}} P_\gamma[\mathbf{k}] = N$, and $R_\gamma = 100\%$. A passing rate plot is simply a measure of R_γ for a range of γ. These plots are similar in concept to those found in gamma map analysis.

If $D_r[\mathbf{k}]$ is used as defined in Eq. 5.15, then this difference image is expressed as a percentage. As such, the set tolerance γ would also be a percentage. To be clear, these are two distinct percentage measures: R_γ is simply the percentage of voxels that are within the tolerance, and γ is the tolerance used to determine if each voxel in $D_r[\mathbf{k}]$ has acceptable variation or not. When comparing a given response map to the CRM, the desired result is $R_\gamma = 100\%$ for as low a tolerance as possible (i.e. the curve rises rapidly to 100%).

5.6.3 Optimising to Valid Consensus

In the presence of discrepancy, both methods described above can be used to potentially produce a more optimal CRM. A CRM is determined to be *valid* if the pairwise assessment between it and each response map are all within a chosen tolerance. If a preliminary CRM is not valid, an iterative approach is used to remove the submissions with the largest discrepancy. This methodology is outlined in Figure 5.14. In each cycle, if the pairwise assessments is not successful, the distance to consensus from PCA determines the submission to be removed. A more optimal CRM can then

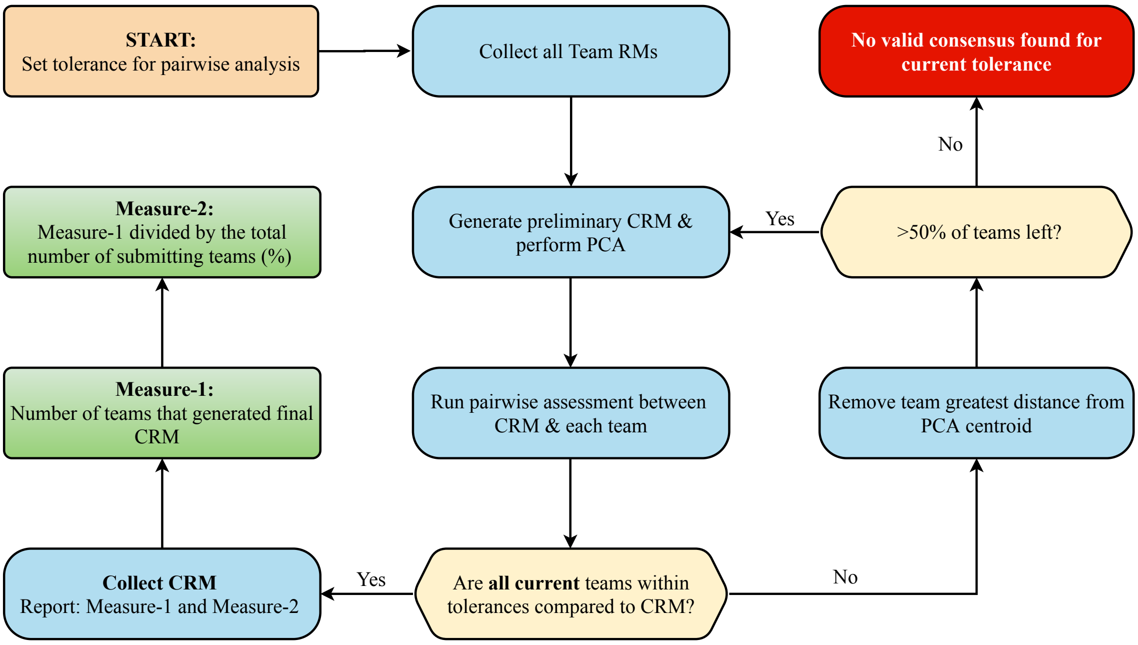

START:
Set tolerance for pairwise analysis
Collect all Team RMs
No valid consensus found for current tolerance
No
Measure-2:
Measure-1 divided by the total number of submitting teams (%)
Generate preliminary CRM & perform PCA
Yes
>50% of teams left?
Measure-1:
Number of teams that generated final CRM
Run pairwise assessment between CRM & each team
Remove team greatest distance from PCA centroid
Collect CRM
Report: Measure-1 and Measure-2
Yes
Are all current teams within tolerances compared to CRM?
No

Figure 5.14: Diagram outlining a designed method for identifying a valid *consensus response map* (CRM) when there are potential discrepancies between submitted *response maps* (RMs). Outliers are identified and iteratively removed. At each iteration a preliminary CRM is generated and the euclidean distance with respect to each RM quantified using PCA (distance to centroid). Then, each RM is compared pair-wise with the CRM. To be accepted, the CRM must be within a given pairwise tolerance of all teams that contributed. If not, the team with the greatest distance to centroid is removed. This is repeated until all remaining teams are with tolerance or no consensus is found.

be generated from the remaining submissions and assessed again. This cycle is repeated until a CRM is reached that is within tolerance of all remaining submissions at that iteration, or $\leq 50\%$ of submissions remain, in which case there is no *valid* identified CRM for the selected tolerance.

As with the standardisation approach discussed in Chapter 3, this work uses the two measures defined previously (Section 3.2.2.3) to report consensus: `Measure-1` is the number of teams that generated the final CRM, and `Measure-2` is the first measure divided by the original number of submitting teams expressed as a percentage. As before, to become a valid consensus-based benchmark an absolute majority of teams need to be used to generate the CRM, i.e. `Measure-2`>50%. The following section demonstrates this iterative methodology with an example.

5.6.3.1 Simulated Example

This section demonstrates the iterative method described in Figure 5.14 to arrive at a valid consensus with a simulated example. For this test, the pairwise assessment uses a maximum voxel-wise tolerance of $\gamma = 1\%$, i.e. the voxel-wise difference between each response map and the CRM cannot exceed an absolute difference of 1% of their range.

By considering the *noise* phantom as simply a *ground truth* (GT) response map, 20 variants were generated by injecting additional noise and variation, as summarised in Figure 5.15. Twelve simulated response maps were allowed to diverge voxel-wise a random amount within 0.5 % of the range of the phantom. These were simulated to represent acceptable levels of difference between response maps. In addition to this baseline variation, another 6 were modified up to 7% of the range in a randomly generated sphere within the response map (similar to Figure 5.13). Likewise,

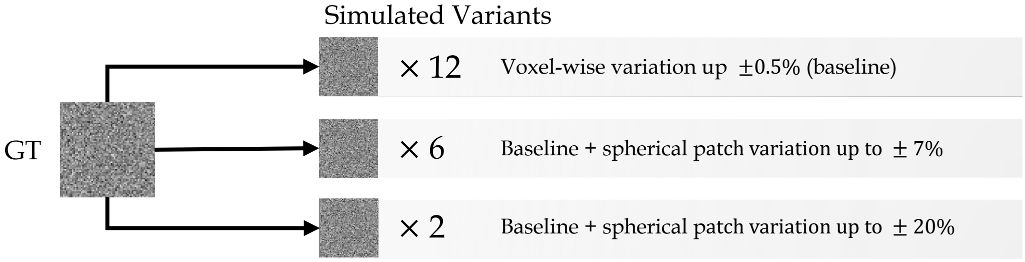

Simulated Variants
GT
× 12
Voxel-wise variation up ±0.5% (baseline)
× 6
Baseline + spherical patch variation up to ± 7%
× 2
Baseline + spherical patch variation up to ± 20%

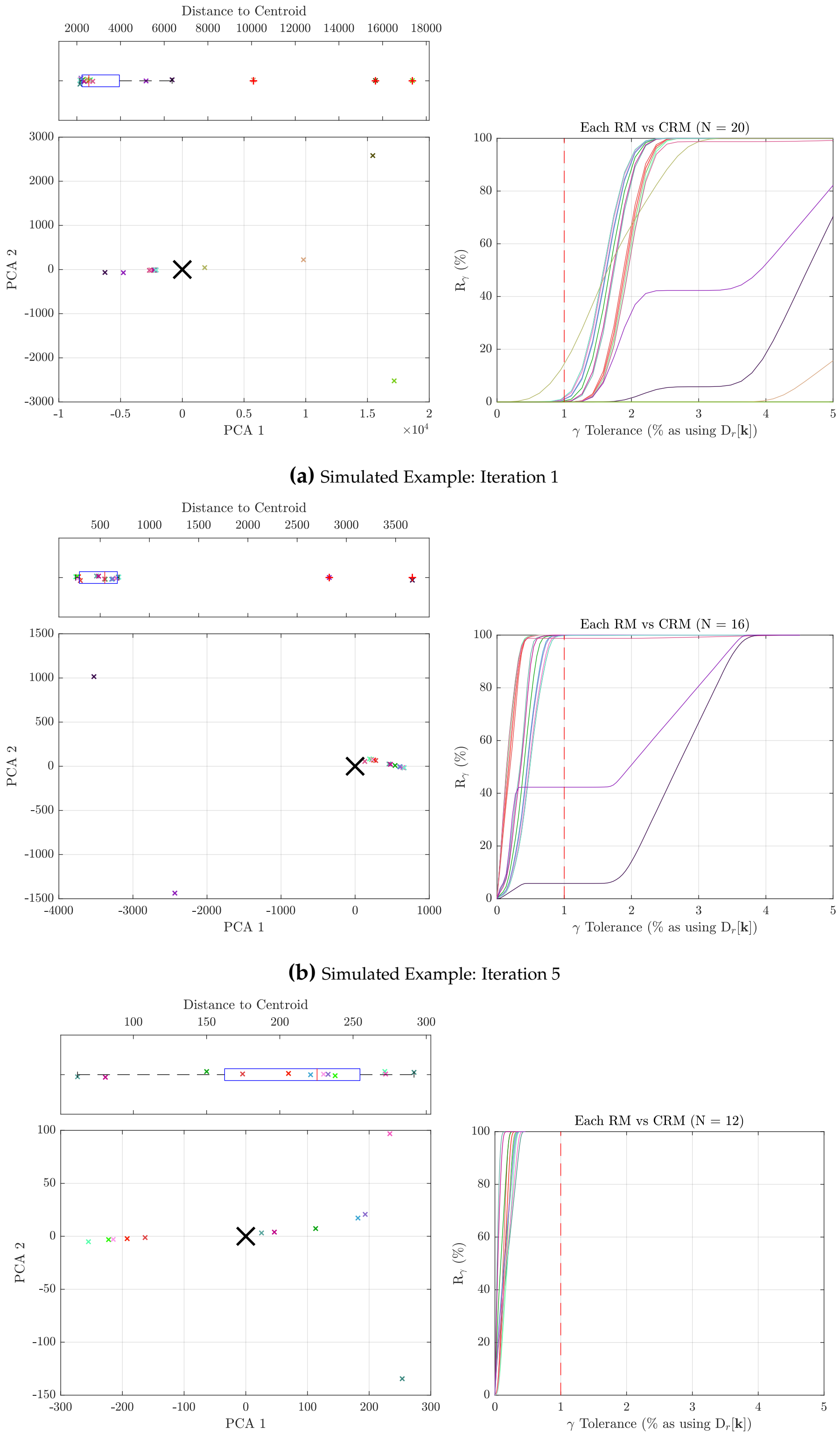
Figure 5.15: Overview of data generated for the simulated example in Section 5.6.3.1. This data was used to demonstrate the iterative method developed in this work to reach a valid consensus response map. The *noise phantom* from the IBSI was considered as a ground truth response map and 20 variants produced from it. For all maps, a baseline randomly generated additional noise was added such that each voxel could vary up to 0.5% of the range of the image. For 6 maps, an additional 7% variation was allowed within a randomly generated spherical patch (see Figure 5.13). This was repeated for a final 2 maps but allowing up to 20% variation. From this dataset one would expect 8 response maps to be removed from contributing towards a CRM if the tolerance is set to $\gamma = 1\%$.

the final 2 simulated response map were injected with a spherical patch of variation to 20% of the range. These 8 simulated response maps with additional patches of variation should be identified and removed from contributing to the CRM.

For this example there is a maximum set tolerance $\gamma = 1\%$, though R_γ is plotted for a range of γ values ($0\% \rightarrow 5\%$). Starting from these 20 response maps, the methodology is followed to iterate towards a CRM that is within tolerance as described in Figure 5.14. Intuitively, 12 response maps are expected to be selected to generate the valid CRM based on the set tolerance. Figure 5.16 shows the PCA and R_γ passing rate plots for selected steps in the iteration towards a consensus. The results are shown at iteration 1 (n=20), 5 (n=16), and 9 (n=12), which is the final iteration where all simulated response maps are within pairwise tolerance of the CRM. With the removal of outlier response maps at each cycle, the distance to centroid is shown to decrease, and R_γ curves improve until all are within the set tolerance, as expected. In this example, the final result is: `Measure-1` = 12 (very strong consensus), and `Measure-2` = 60%. As a sanity check, this example was re-ran 100 times, with the same consensus measures achieved in every case. Note that this simulation was to confirm that, given a set of response maps and a tolerance, the iterative methodology will correctly identify and weed out outliers. Of course, for standardisation the lower the tolerance value selected the better. In the following work, $\gamma = 1\%$ is used as the initial tolerance, then once/if a consensus is reached, the *optimal* lowest $\gamma\%$ that would have achieved the same result is calculated and reported.

5.7 Initial Filter Test Submissions for the IBSI

This section presents the set of filter tests and the first results for Phase 1 of the second instalment of the IBSI. This preliminary analysis was performed using the methodology designed above. Submissions were opened in June 2020, with these baseline results presented here collected in October 2020 from team uploads to the initiative website. As with the first study, it is strongly anticipated that the number of submissions and unique teams will increase as the initiative continues. Additional filtering techniques and filter tests will likely also be included.



(a) Simulated Example: Iteration 1

(b) Simulated Example: Iteration 5

(c) Simulated Example: Iteration 9

Figure 5.16: Results from simulated example to demonstrate iterating towards a consensus using the methodology outlined in Figure 5.14. Starting with 20 variant response maps generated using the *noise* phantom as a ground truth (see Figure 5.15). Three iterations are shown (1,5,9). For each iteration, the PCA (left), distance to centroid (top) and R_γ curves are plotted. At each iteration, with the removal of outlier response maps, the distance to centroid decreases and the R_γ passing rate curves begin to fall within tolerance. For pairwise assessment $D_r[\mathbf{k}]$ was used (see Eq. 5.15), and the tolerance for acceptable discrepancy was set to $\gamma = 1\%$.

Table 5.2: A subset of the filter tests for Phase 1 of the IBSI [86]. Included for this prelimary analysis are the first 25 filter tests, where there has been at least 2 independent team submissions as of October 2020.

ID	Filter Type	Phantom	Padding	2D or 3D	Filter Parameters / Settings
1.a.1	mean	checkerboard	zero	3D	size: $M = 15$
1.a.2			nearest	3D	size: $M = 15$
1.a.3			periodic	3D	size: $M = 15$
1.a.4			mirror	3D	size: $M = 15$
2.a	LoG	impulse	zero	3D	scale: $\sigma = 3$mm, cutoff: $d = 4$
2.b		checkerboard	mirror	3D	scale: $\sigma = 5$mm, cutoff: $d = 4$
3.a.1	Laws	impulse	zero	3D	E5L5S5 response map
3.a.2			zero	3D	E5L5S5 response map, 3D rotation invariance, max pooling
3.a.3			zero	3D	E5L5S5 energy image, $\delta = 7$, 3D rotation invariance, max pooling
3.b.1		checkerboard	mirror	3D	E3W5R5 response map
3.b.2			mirror	3D	E3W5R5 response map, 3D rotation invariance, max pooling
3.b.3			mirror	3D	E3W5R5 energy image, $\delta = 7$, 3D rotation invariance, max pooling
4.a.1	Gabor	impulse	zero	2D	modulus, $\sigma = 10$mm, $\lambda = 4$mm, $\gamma = 1/2$, in-plane orientation: $\theta = \pi/3$
4.a.2			zero	2D	modulus, $\sigma = 10$mm, $\lambda = 4$mm, $\gamma = 1/2$, 2D rotation invariance: $\Delta\theta = \pi/4$, average 2D response over orthogonal planes
4.b.1		sphere	mirror	2D	modulus, $\sigma = 20$mm, $\lambda = 8$mm, $\gamma = 5/2$, in-plane orientation: $\theta = 5\pi/4$
4.b.2			mirror	2D	modulus, $\sigma = 20$mm, $\lambda = 8$mm, $\gamma = 5/2$, 2D rotation invariance: $\Delta\theta = \pi/8$, average 2D response over orthogonal planes
5.a.1	Daubechies 2	impulse	zero	3D	Undecimated LHL map, level: 1
5.a.2			zero	3D	Undecimated LHL map, level: 1, 3D rotation invariance, average pooling
6.a.1	Coifflet 1	sphere	periodic	3D	Undecimated HHL map, level: 1
6.a.2			periodic	3D	Undecimated HHL map, level: 1, 3D rotation invariance, average pooling
7.a.1	Haar	checkerboard	mirror	3D	Undecimated LLL map, level: 2
7.a.2			mirror	3D	Undecimated HHH map, level: 2, 3D rotation invariance, average pooling
8.a.1	Simoncelli	checkerboard	periodic	3D	B-map level: 1
8.a.2			periodic	3D	B-map level: 2
8.a.3			periodic	3D	B-map level: 3

Table 5.3: Initial consensus results (October 2020) from the first 25 filter tests (detailed in Table 5.2). From the intial submissions, consensus is evaluated using the methodolgy described in Figure 5.14 using a tolerance of $\gamma = 1\%$. If consensus was found, the optimal γ for that test is shown, which is the lowest γ that would achieve the same consensus measures.

I.D.	Initial Submissions	Consensus Reached?	Submissions Within Consensus	Consensus Strength (measure-1)	Consensus Stability (measure-2) (%)	Optimal γ (%)
1.a.1	9	yes	7	strong	77.78	1.14E-03
1.a.2	8	yes	6	strong	75.00	2.61E-03
1.a.3	9	yes	6	strong	66.67	0.00E+00
1.a.4	9	yes	7	strong	77.78	2.68E-03
2.a	8	yes	5	moderate	62.50	1.85E-03
2.b	8	yes	5	moderate	62.50	4.94E-02
3.a.1	6	yes	4	moderate	66.67	1.51E-06
3.a.2	5	yes	4	moderate	80.00	5.05E-06
3.a.3	5	yes	4	moderate	80.00	1.17E-05
3.b.1	6	yes	4	moderate	66.67	7.00E-06
3.b.2	4	yes	3	moderate	75.00	1.24E-05
3.b.3	5	yes	3	moderate	60.00	9.11E-05
4.a.1	3	yes	2	weak	66.67	5.60E-01
4.a.2	3	no	-	none	-	-
4.b.1	3	yes	2	weak	66.67	1.13E-04
4.b.2	3	yes	2	weak	66.67	3.73E-05
5.a.1	7	yes	5	moderate	71.43	2.59E-06
5.a.2	5	yes	3	moderate	60.00	1.15E-06
6.a.1	7	yes	5	moderate	71.43	2.75E-06
6.a.2	5	yes	3	moderate	60.00	3.20E-06
7.a.1	5	yes	3	moderate	60.00	0.00E+00
7.a.2	5	yes	3	moderate	60.00	1.62E-15
8.a.1	2	yes	2	weak	100.00	5.14E-01
8.a.2	2	no	-	none	-	-
8.a.3	2	no	-	none	-	-

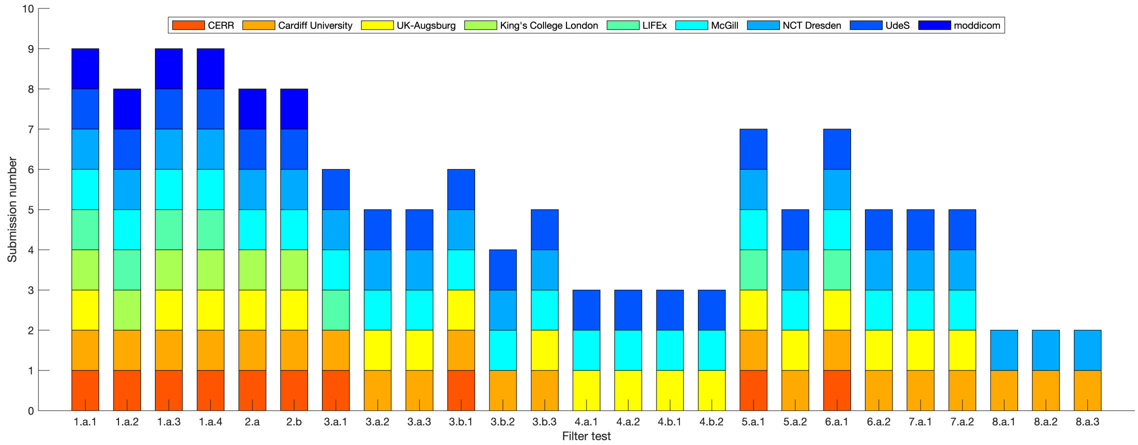


Figure 5.17: Overview of the number of unique submissions for the first 25 filter tests proposed by the IBSI [86], where at least 2 submissions were received (submissions for October 2020). Nine unique teams participated in this initial time point. Note that the color assigned to each team is consistent throughout all plots in this Chapter.

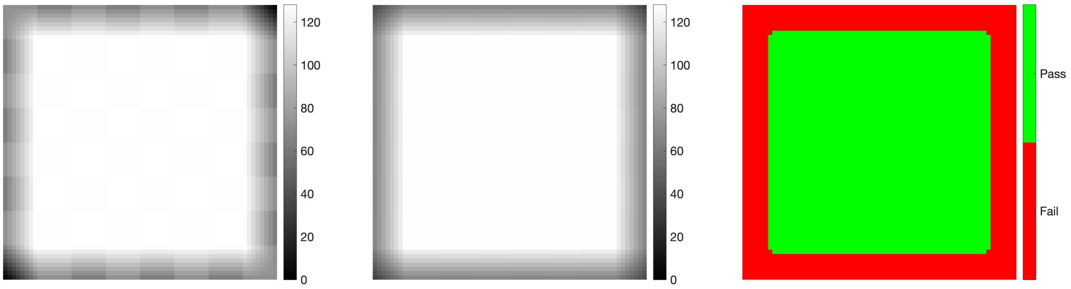
5.7.1 Results Overview

In total, 9 unique teams contributed a first set of response maps following the initial filtering workflow recommendations developed by the IBSI core members [86]. Figure 5.17 summarises the number of submissions for 25 of these filter tests detailed in Table 5.2, where in each case at least 2 submissions were received. At this time point, no individual team provided a response map for every test configuration.

Using the methodology described in Figure 5.14, for each test the validity and strength of the CRM was evaluated. Table 5.3 presents the consensus measures obtained for these initial submissions. All teams provided some results for the mean filter, which is intuitively the simplest to implement. As shown in Figure 5.17, contributions then dropped off with the increasing implementation complexity of the given filters.

As with the simulated example above, the cut off for acceptable pairwise variation between each team and the average response map was initially set to $\gamma = 1\%$, though the final maximum variation was far below this tolerance for most of the valid CRMs obtained. Many of the tests achieved $R_\gamma = 100\%$ for all pairwise assessments at a much lower γ value after the iterative approach was used to identify and remove outlier contributions. In other words, once the outliers were removed, comparing the response maps with the CRM yielded extremely small variation, often at the order of machine precision error. The lowest γ at which $R_\gamma = 100\%$ for all pairwise comparisons (after removing outliers) is recorded in Table 5.3 as the *optimal* γ, which if used as the tolerance would achieve the same consensus results. This is also illustrated by visualising the passing rate plots with specific filter tests in the following sections, as the optimal γ is the threshold at which all curves on the passing rate plots have reached 100%: in most cases this is an extremely small value.

Through outlier removal, a valid CRM was achieved for 18/25 of these initial filter tests, though the strength of the consensus remains moderate in most cases simply due to the number of current submissions. As shown in Table 5.3, nearly every filter test contained outliers that were removed. This explicitly emphasises the need for benchmarks as even for simpler filter tests some software



(a) Outlier submission **(b)** Valid CRM **(c)** $P_\gamma[\mathbf{k}]$

Figure 5.18: Example of outlier discrepancy for test 1.a.1 of the mean filter. A 2D slice of the 3D volumes are visualised of: **(a)** an outlier submission, **(b)** the final valid CRM found, and **(c)** the passing mask $P_\gamma[\mathbf{k}]$ (see E.q. 5.16) with a set tolerance of $\gamma = 1\%$ generated from a comparison of **(a)** and **(b)**. Voxels that were not within tolerance are shown as red. There is a clear discrepancy at the boarder of the response map.

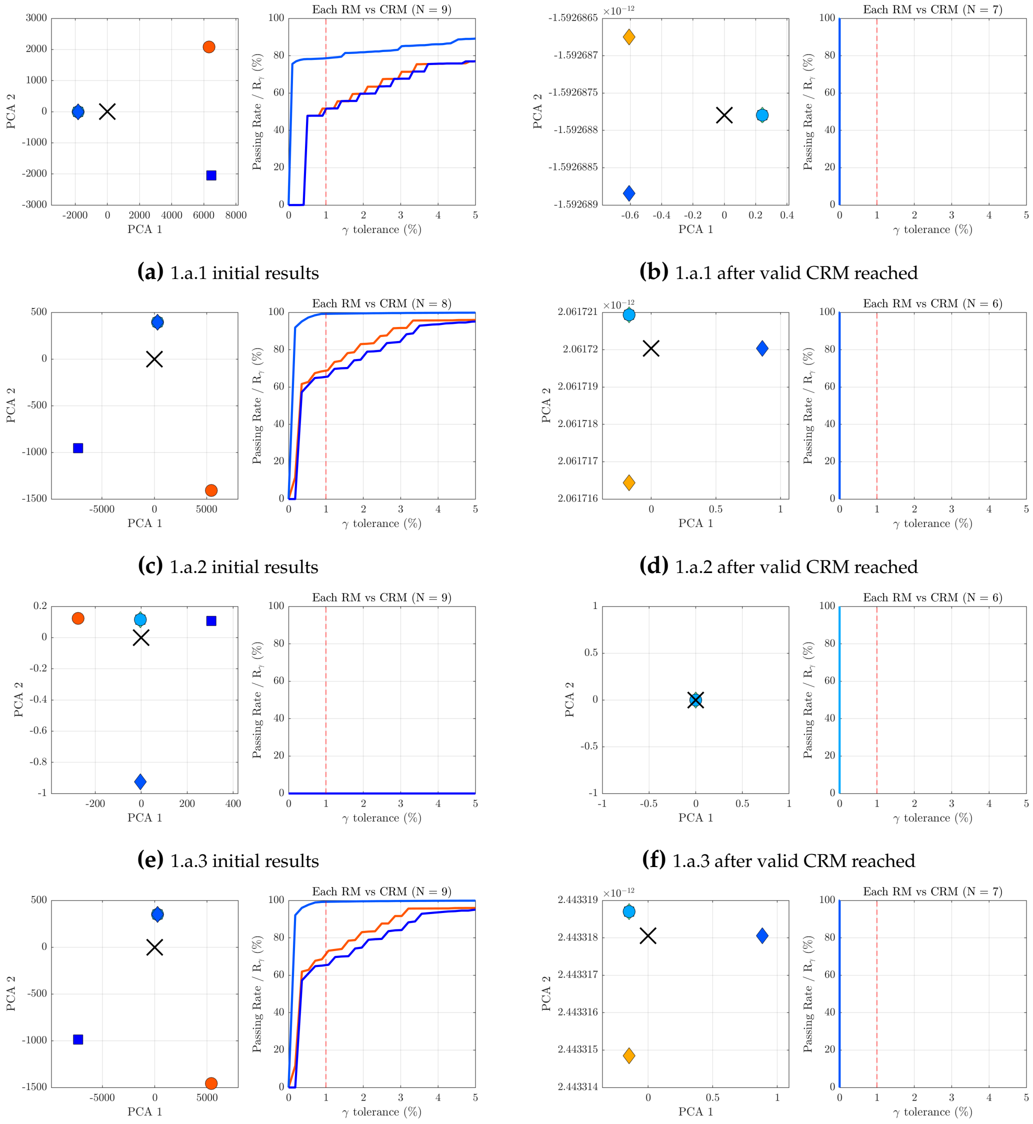
deviated substantially, and as such requires revision to comply. To become a benchmark, as before the IBSI necessitate at least a moderate consensus (3 or more teams). For 4/25 tests there was weak agreement found, with potential to build on these CRM as tentative benchmarks. However, for 3/25 tests the iterative process was not able to determine any valid consensus at all, as it led to $\leq 50\%$ of teams contributing to the CRM. Despite the need for more team submissions to achieve a stronger consensus, these preliminary results can already identify sources of discrepancy and evaluate the consensus methodology. The following sections present more details of results for each filter family.

5.7.1.1 Mean Filter Results

As shown in Table 5.3, though there were outlier submissions identified in all cases, each mean filter test still achieved a valid CRM with a strong consensus. As the mean filter is quite simplistic, these tests were an opportunity to check different padding applications of software (as discussed in Section 5.2.2.2) as each of the 4 tests are the same apart from the selected padding type. Despite apparent simplicity of the mean filter, this work discovered significant variation suggesting software discrepancy in either the filter kernel generated or a deviating padding technique.

Figure 5.18 illustrates a type of discrepancy found using filter test 1.a.1. In this example, a 2D slice of the passing mask ($P_\gamma[\mathbf{k}]$, see E.q. 5.16) is visualised that compares a selected outlier team to the valid CRM. The area of discrepancy is shown to be at the boarders, indicating for this particular submission that the padding technique is likely not set properly or is not correctly implemented.

The PCA and R_γ plots for each of the mean filter tests are presented in Figure 5.19. The validity of the CRM greatly improves after removing the outlier submissions. This is highlighted by the dramatic decrease in the size of the PCA components and the γ at which $R_\gamma = 100\%$. The same two teams are found to be amongst the outliers in all 4 tests, suggesting a systematic error or difference in implementation approach here. Results for filter test 1.a.2 and 1.a.4 appear almost identical from the initial submissions, though this is in part due to the mirror and nearest padding producing a similar extension at the boarders in the case of the checkerboard phantom. For filter test 1.a.3, outliers influenced the initial CRM so much that the passing rate plot is initially flat until beyond $\gamma = 5\%$ for all pairwise comparisons.



(a) 1.a.1 initial results

(b) 1.a.1 after valid CRM reached

(c) 1.a.2 initial results

(d) 1.a.2 after valid CRM reached

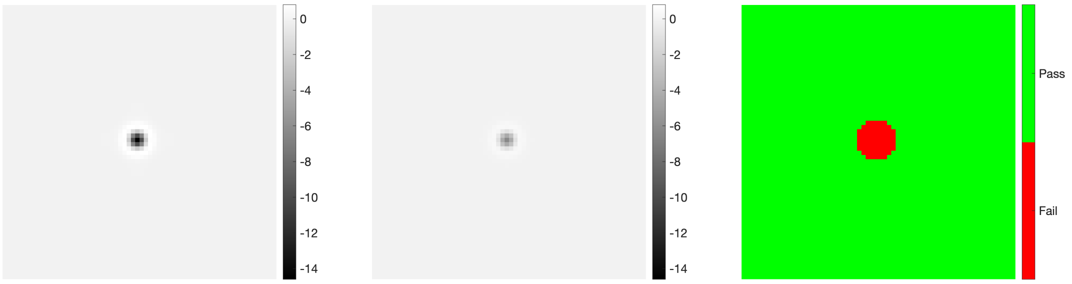
(e) 1.a.3 initial results

(f) 1.a.3 after valid CRM reached

(g) 1.a.4 initial results

(h) 1.a.4 after valid CRM reached

Figure 5.19: Initial results from the mean filter tests (October 2020) to determine valid consensus response maps.. Each subfigure visualises the PCA and R_γ passing rate plots respectively. For each filter test: all submissions are evaluated (**(a)**,**(c)**,**(e)**,**(g)**), and then just the remaining submissions after removing outliers (**(b)**,**(d)**,**(f)**,**(h)**) following the method described in Figure 5.14. Note that after removal of outliers the PCA components are extremely small, and for the passing rate plots $R_\gamma = 100\%$ is reached almost immediately. Also note that many points and curves overlap (which is expected if submissions are identical). In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: 0% $\rightarrow$ 5%, though the curves stop as soon as they hit $R_\gamma = 100\%$.

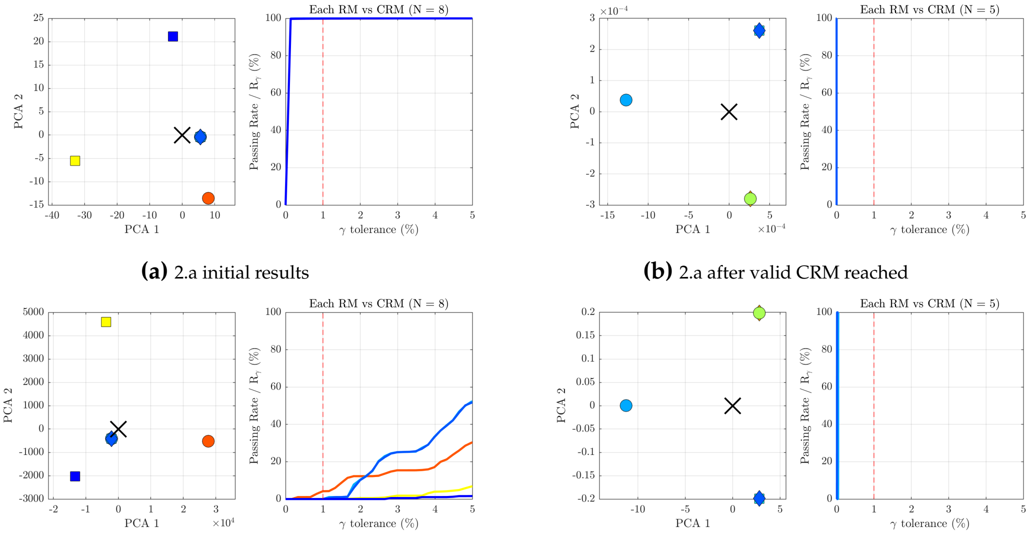


(a) Outlier submission **(b)** Valid CRM **(c)** $P_\gamma[\mathbf{k}]$

Figure 5.20: Example of outlier discrepancy for test 2.a of LoG filter. A 2D slice of the 3D response map is visualised for: **(a)** an outlier submission, **(b)** the final valid CRM found, and **(c)** the passing mask $P_\gamma[\mathbf{k}]$ (see E.q. 5.16) with a set tolerance of $\gamma = 1\%$ generated from a comparison of these two response maps. Voxels that were not within tolerance are shown as red. **(a)** and **(b)** are shown on the same scale.

5.7.1.2 LoG Filter Results

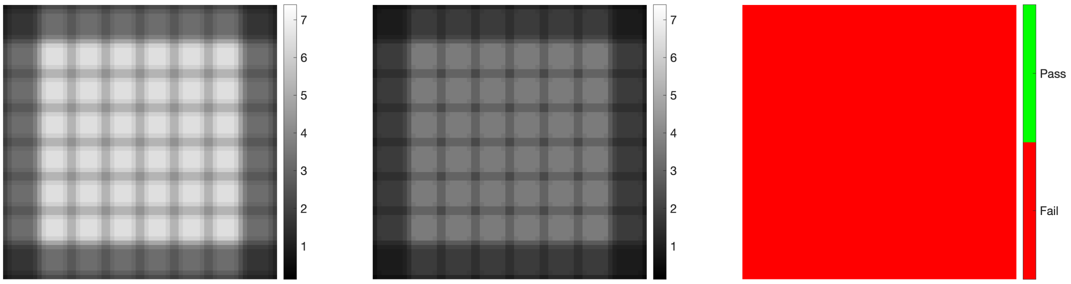
As shown in Table 5.3, though outlier submissions were identified, the two LoG filter tests also reached a valid CRM with a moderate consensus strength. Filter test 2.a utilises the impulse phantom, and by design the resulting response map produces the filter kernel being applied. This filter test uses a kernel that is much smaller than the impulse phantom, so most of the response map remains zero. As an example, Figure 5.20 visualises the discrepancy between one of the outlier submissions and the final valid CRM. Even with clear discrepancy, as most of the voxels remain zero they are within tolerance when compared by default. This illustrates why the key



(a) 2.a initial results **(b)** 2.a after valid CRM reached

(c) 2.b initial results **(d)** 2.b after valid CRM reached

Figure 5.21: Initial results from the LoG filter tests (October 2020). Each subfigure visualises the PCA and R_γ passing rate plots respectively. For each filter test: all submissions are evaluated (**(a)**,**(c)**), and then just the remaining submissions after removing outliers (**(b)**,**(d)**) following the method described in Figure 5.14. Note that after removal of outliers the PCA components are extremely small, and for the passing rate plots $R_\gamma = 100\%$ is reached almost immediately. Also note that many points and curves overlap (which is expected if submissions are identical). In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: $0\% \rightarrow 5\%$, though the curves stop as soon as they hit 100%.



(a) Outlier submission **(b)** Valid CRM **(c)** $P_\gamma[\mathbf{k}]$

Figure 5.22: Example of outlier discrepancy Laws Filtering test 3.b.3. A 2D slice of the 3D response map is visualised for: **(a)** an outlier submission, **(b)** the final valid CRM found, and **(c)** the passing mask $P_\gamma[\mathbf{k}]$ (see E.q. 5.16) with a set tolerance of $\gamma = 1\%$ generated from a comparison of these two response maps. Voxels that were not within tolerance are shown as red. For this example significant discrepancy is measured over the entire response map. **(a)** and **(b)** are shown on the same scale.

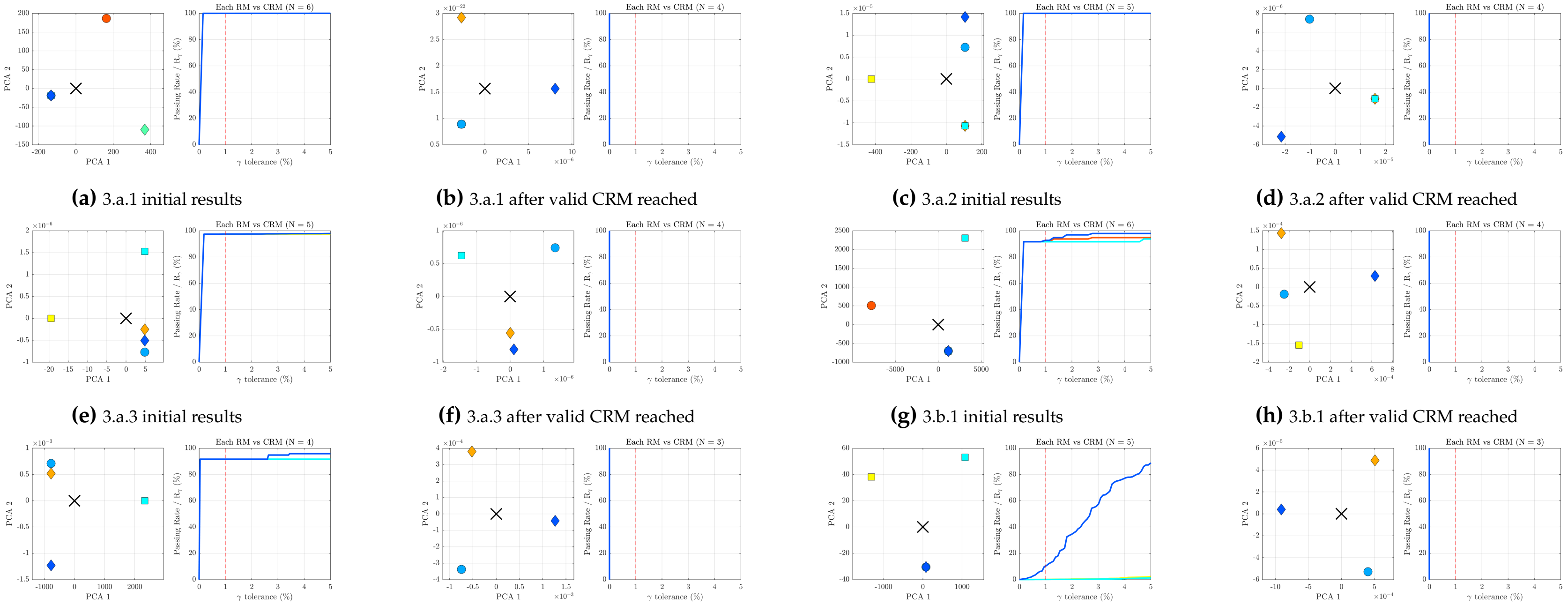
measure is the point at which $R_\gamma = 100\%$.

The PCA and R_γ plots for the two LoG filter tests are shown in Figure 5.21. The same 3 teams were identified as outliers for both filter test 2.a and 2.b. The remaining 5 teams show extremely close agreement as demonstrated by the low optimal γ of both tests. As discussed in Section 5.4.3, the LoG filter kernel size is determined with a cut-off and discrepancy may arise if a different approach was used to that recommended via Eq. 5.13.

5.7.1.3 Laws Filter Results

As outlined in Table 5.3, all Laws filter tests reached at least a moderate consensus strength, producing a valid CRM in each case. As before, some submissions were identified as significant outliers. Tests were designed to examine both response maps generated from applying the Laws filter kernels (3.a.1, 3.a.2, 3.b.1, 3.b.2), as well as averaging into Laws texture energy images (3.a.3, 3.b.3). Also tested was rotational invariance techniques (Section 5.3.2), by specifying max pooling of response maps corresponding to all unique right angle orientations prior to the calculation of energy images. Again, as an example Figure 5.22 visualises the central slice from one of the identified outlier response maps compared to the valid CRM for filter test 3.b.3. In this case, every voxel has variation above the tolerance, hence the completely red passing mask $P_\gamma[\mathbf{k}]$. The PCA and R_γ plots for the 6 Laws filter tests are shown in Figure 5.23 from the initial and post iterative approach to remove outliers. Very little variation is found between remaining teams and the CRM, with small PCA components and optimal γ measured at the final iteration.

The Laws filtering approach is prone to several implementation decisions that can lead to discrepancy. As an example, on a preliminary implementation within the Cardiff's SPAARC software, energy images were generated for each orientation and pooled, instead of pooling the response maps first and generating one energy image. These two approaches are not equivalent. Overall, the rotational invariance adds significant complexity to the pipeline and proved challenging to implement for several participating. As before with the first study (Chapter 3), it is anticipated by improving the guidelines, more submissions will be collected, leading to an increase in consensus strength as the initiative continues.



(a) 3.a.1 initial results

(b) 3.a.1 after valid CRM reached

(c) 3.a.2 initial results

(d) 3.a.2 after valid CRM reached

(e) 3.a.3 initial results

(f) 3.a.3 after valid CRM reached

(g) 3.b.1 initial results

(h) 3.b.1 after valid CRM reached

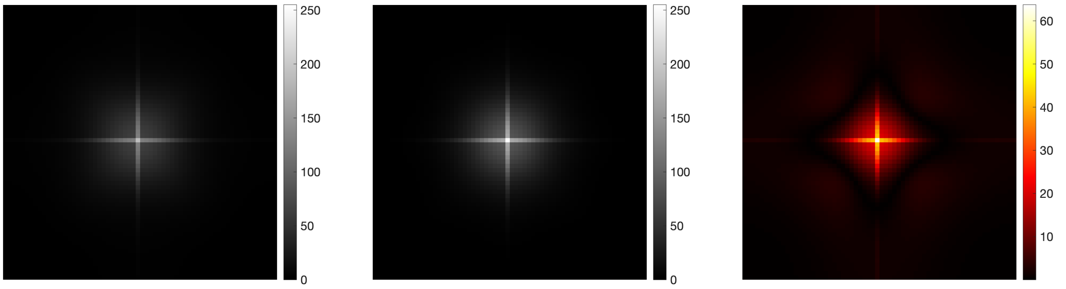
(i) 3.b.2 initial results

(j) 3.b.2 after valid CRM reached

(k) 3.b.3 initial results

(l) 3.b.3 after valid CRM reached

Figure 5.23: Initial results from the Laws filter tests (October 2020) to determine valid consensus response maps. Each subfigure visualises the PCA and R_γ passing rate plots respectively. For each filter test: all submissions are evaluated (**(a)**,**(c)**,**(e)**,**(g)**,**(i)**,**(k)**), and then just the remaining submissions after removing outliers (**(b)**,**(d)**,**(f)**,**(h)**,**(j)**,**(l)**) following the method described in Figure 5.14. Note that after removal of outliers the PCA components are extremely small, and for the passing rate plots $R_\gamma = 100\%$ is reached almost immediately. Also note that many points and curves overlap (which is expected if submissions are identical). In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: $0\% \rightarrow 5\%$, though the curves stop as soon as they hit $R_\gamma = 100\%$.

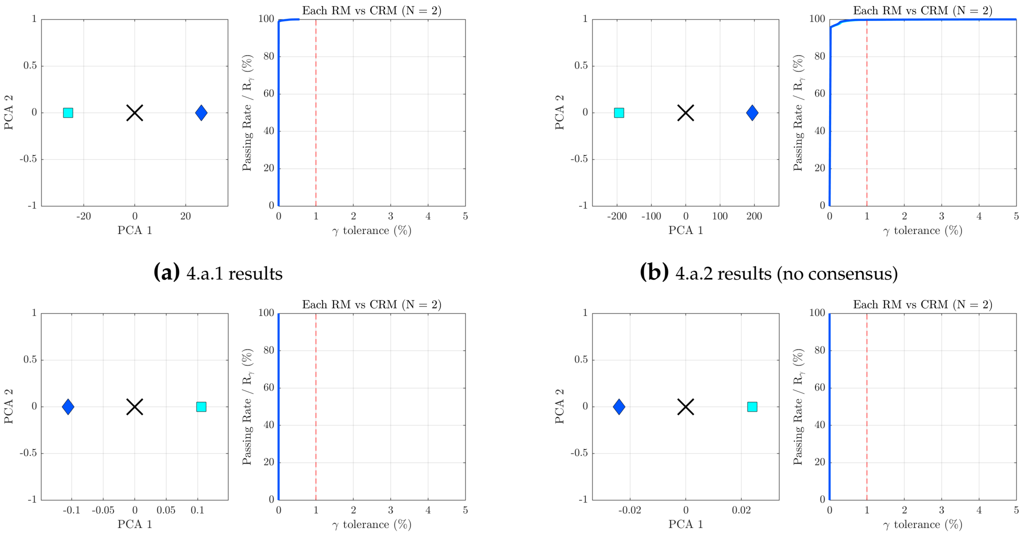


(a) Submission 1 **(b)** Submission 2 **(c)** Absolute difference

Figure 5.24: Visualising discrepancy in Gabor test 4.a.2. A 2D slice of the 3D response map is visualised for two submitting teams (**(a)** and **(b)**), and **(c)** the absolute difference between them. **(a)** and **(b)** are shown on the same scale.

5.7.1.4 Gabor Filter Results

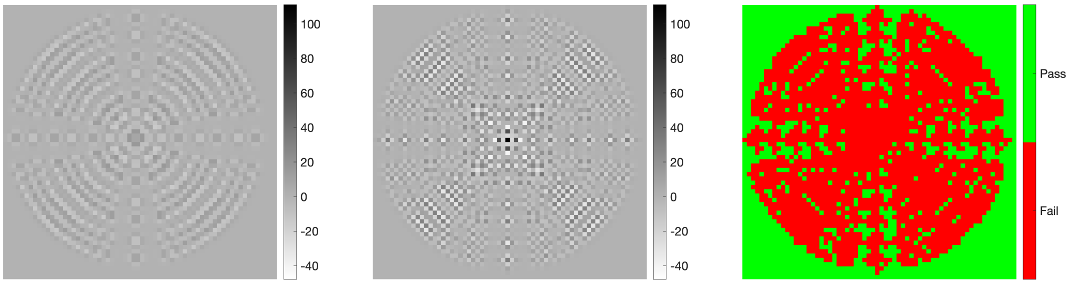
As summarised in Table 5.3, there were a low number of contributions to the Gabor filter tests in this initial set with only 3 teams submitting results. From those, only a weak or no consensus was found in every case. More over, one of the team's filter tests were determined to be a submission error. This teams response maps were unsurprisingly picked up as clear outliers with the analysis methods discussed, however they could be excluded prior based on solely a qualitatively assessment. For 3 of the 4 filter tests, a weak consensus was achieved with agreement found between the 2 remaining teams. Interestingly, for test 4.a.2 there was no consensus found. A 2D slice for this test is visualised for each team alongside the difference image for this slice, which highlights a significant scaling change. However, note that only a small percentage of voxels actually have a significant difference when considering the full 3D response map as this is an impulse response



(a) 4.a.1 results **(b)** 4.a.2 results (no consensus)

(c) 4.b.1 results **(d)** 4.b.2 results

Figure 5.25: Initial results for the Gabor filter tests (October 2020) to determine valid consensus response maps. Each subfigure visualises the PCA and R_γ passing rate plots respectively. Due to a submission error of 1 of the teams, only results from the remaining 2 teams are shown. As such, there is only one set of plots shown for each test. Filter test 4.a.2,**(b)** does not fall within tolerance. In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: 0% → 5%, though the curves stop as soon as they hit 100%.



(a) Outlier submission **(b)** Valid CRM **(c)** $P_\gamma[\mathbf{k}]$

Figure 5.26: Example of outlier discrepancy for separable wavelets filter test 6.a.2. A 2D slice of the 3D response map is visualised for: **(a)** an outlier submission, **(b)** the final valid CRM found, and **(c)** the passing mask $P_\gamma[\mathbf{k}]$ (see E.q. 5.16) with a set tolerance of $\gamma = 1\%$ generated from a comparison of these two response maps. Voxels that were not within tolerance are shown as red. Here, **(a)** and **(b)** are shown on the same scale with an inverted gray scale to better visualise the discrepancy.

test so many voxels have zero intensity. Figure 5.25 shows the PCA and R_γ plots for the Gabor tests for the two teams.

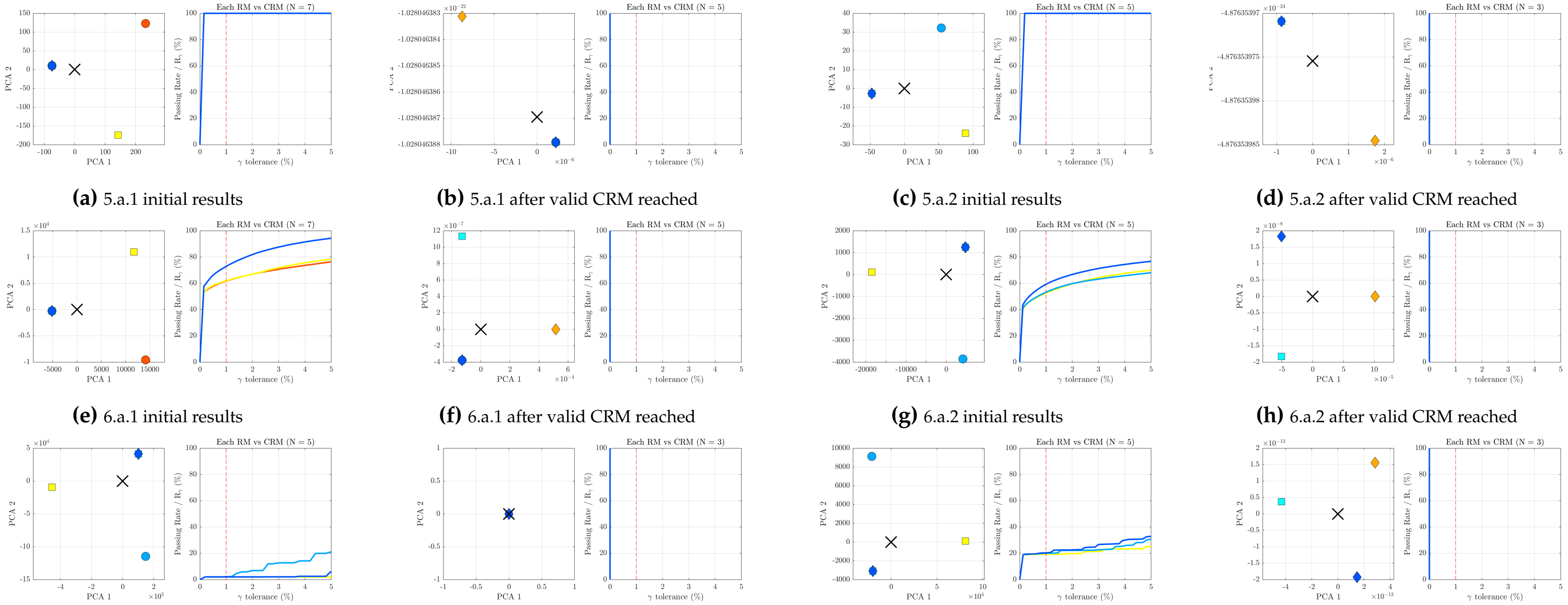
The Gabor filter was the only one not to be implemented in Cardiff's SPAARC pipeline sufficiently to contribute to this set of filter tests at the collection of these initial submissions, though it will be included in a future update. As such the technical detail behind Gabor filters have not been included in this chapter, though this can be found in the corresponding IBSI reference manual [86]. The initial submissions for these tests have yet to produce at least a moderate consensus, thus, improving here will be a considerable focus as the initiative continues and a greater number of teams contribute.

5.7.1.5 Separable Wavelets Filter Results

Filter tests 5-7 are a set of separable wavelet tests using commonly selected *mother* wavelets (*Daubechies 2*, *Coifflet 1*, and *Haar*). As outlined in Table 5.3, all of these tests reached a moderate consensus strength, producing a valid CRM in each case. Again, in every test some number of submissions were identified as significant outliers based on the PCA and R_γ analysis.

As an example, Figure 5.26 visualises the central slice from one of the identified outlier response maps compared to the valid CRM for filter test 6.a.2. Interestingly, the requirement for rotational invariance led to highly diverging response maps. Teams that were in agreement for filter test 6.a.1 were not for 6.a.2, indicating a fundamental cause of discrepancy in the implementations that needs to be addressed when trying to satisfy this criteria of rotational invariance. This is discussed more in Section 5.8.

Figure 5.27 visualises the PCA and R_γ plots for the all of the separable wavelet tests (5.a.1 up to 7.a.2). For each test the initial plots are shown including all teams, alongside the subsequent result after iteratively removing identified outlier submissions and reaching a valid moderate consensus. The remaining submissions show extremely strong agreement as highlighted in these plots and by the optimal γ in Table 5.3.



(a) 5.a.1 initial results

(b) 5.a.1 after valid CRM reached

(c) 5.a.2 initial results

(d) 5.a.2 after valid CRM reached

(e) 6.a.1 initial results

(f) 6.a.1 after valid CRM reached

(g) 6.a.2 initial results

(h) 6.a.2 after valid CRM reached

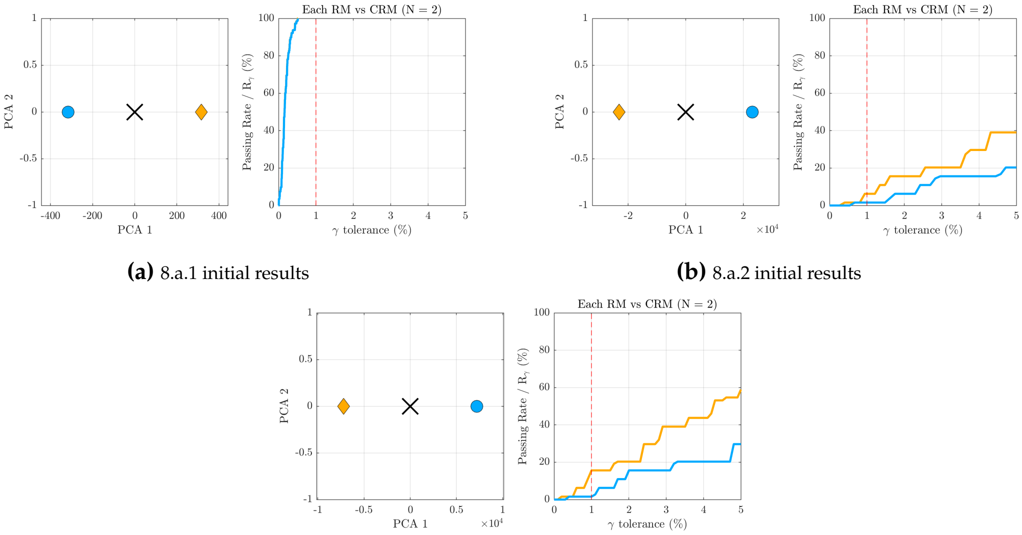
(i) 7.a.1 initial results

(j) 7.a.1 after valid CRM reached

(k) 7.a.2 initial results

(l) 7.a.2 after valid CRM reached

Figure 5.27: Initial results from the separable wavelet filter tests (October 2020) to determine valid consensus response maps. Each subfigure visualises the PCA and R_γ passing rate plots respectively. For each filter test: all submissions are evaluated (**(a)**,**(c)**,**(e)**,**(g)**,**(i)**,**(k)**), and then just the remaining submissions after removing outliers (**(b)**,**(d)**,**(f)**,**(h)**,**(j)**,**(l)**) following the method described in Figure 5.14. Note that after removal of outliers the PCA components are extremely small, and for the passing rate plots $R_\gamma = 100\%$ is reached almost immediately. Also note that many points and curves overlap (which is expected if submissions are identical). In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: $0\% \rightarrow 5\%$, though the curves stop as soon as they hit $R_\gamma = 100\%$.



(a) 8.a.1 initial results

(b) 8.a.2 initial results

(c) 8.a.3 initial results

Figure 5.28: Initial results from the simoncelli filter tests (October 2020) to determine valid consensus response maps. Each subfigure visualises the PCA and R_γ passing rate plots respectively. Only two teams had submitted results for this filter at the intial time evaluation presented here. In the PCA plots the large cross represents the CRM. For consistent comparison the x axis of the R_γ plots are limited to γ: 0% → 5%, though the curves stop as soon as they hit 100%. Significant differences are found between the two implementations for the higher decomposition levels.

5.7.1.6 Non-Separable Wavelets Filter Results

The final set of results where at least 2 initial submissions were received were the Simoncelli non-separable wavelet tests. As with the Gabor tests, there were not enough submissions to reach at least a moderate consensus and more implementations are needed to produce a valid benchmark reference value for this particular filter. However, for completeness the PCA and R_γ plots were included here in Figure 5.28. There was a weak consensus found for 1/3 of the filter tests, with increasing divergence at higher decomposition levels. As this work is limited by participation, this needs to be further explored before a valid, standardised CRM can be offered as a benchmark in this case.

5.8 Discussion

Convolutional approaches offer a further paradigm for image biomarker extraction. The ability to boost heterogeneous characteristics with different filtering techniques may prove valuable in the search for biomarkers in oncology - such as those that predict tumour aggression - by enhancing biologically relevant patterns within the imaging. Filtering is an attractive proposition for radiomic studies as a result. As argued throughout this thesis, radiomic techniques require standards for effective, *reproducible*, translational studies. Radiomic features aggregated from filtered imaging are no different.

The work in this chapter discusses the first of a multi-phase effort to produce an extended set of standardised reference response maps for a subset of common filtering techniques that have seen some use in radiomics research. Through a consensus based approach, by examining independent software implemented with different languages across different platforms, the aim was to determine if significant discrepancy was present. And if so, discuss potential causes and provide recommendations for implementation and feature use in future studies.

To determine discrepancy in response maps generated from filter tests, a method for comparison was developed and evaluated by the author. The key objective was to design a criteria to assess submissions for consensus. To efficiently identify significant outliers, this work presented a novel combination of PCA and pairwise assessment using so-called *passing rate plots* that evaluated a γ measure of difference. The main output was a preliminary set of consensus based response maps (CRMs) for a subset of filter tests. A valid CRM was achieved for 18/25 of the tests presented in Table 5.2. By following guidelines discussed here and in the developing IBSI documentation [86], at least a moderate majority (3+) of teams reproduced a result that was within an extremely small tolerance for these 18 filter tests, providing a benchmark for other software to match.

However, these preliminary results have also highlighted a significant level of discrepancy present between some software. This has clear consequences for reproducibility of radiomics studies, and offers further evidence for the need for guidelines for extracting potential filter based biomarkers.

5.8.1 Reflection on Methodology Developed for Response Map Comparison

The author introduced and tested, to the best of their knowledge, a unique approach to analysing volumetric images for similarity through a combination of techniques. The first technique was to use PCA [126], a classic dimensionality reduction method. In the context of this work, PCA enables an overall evaluation of any number of response maps by comparing their positions in high dimensional space. This technique is only possible if all response maps are the same size (e.g. $\mathbb{R}^{64\times64\times64}$), as is the case with this set of tests. The mean submission is represented as a centroid within PCA space and any voxel-wise variation is captured by deviation in the principle components. This makes it extremely useful for judging consensus. The euclidean distance to this centroid is used to rank contributions, with those furthest from the centroid having shown the most variation compared to those clustered closer together. Obvious outliers are identified immediately with this technique, as exactly identical response maps occupy the same coordinate point within PCA space. As shown throughout the results plots within this chapter, many of the submissions overlap when plotting the first two PCA components. Take Figure 5.19 **(a)**, the initial PCA plot contains the evaluation of 9 teams, though there are 3 clear regions of submissions. One is the cluster of 7 teams on top of one another and the other 2 represent the submissions that are later identified as outliers.

Experimentally, this work finds that in cases where evaluated submissions are extremely close (effectively to within machine precision), their corresponding points within PCA space are still technically different, though extremely small and on the order of $10^{-12}/10^{-13}$ in some cases (e.g. Figure 5.19). As expected, the relative scale of PCA components decreases dramatically after the removal of outliers. When there is a very prominent outlier, this will dominate the visualisation and scale of the first principle component. It is important to remember that only the first 2 principle components are visualised (which capture the most significant variation), though all are used to calculate the euclidean distance to the centroid. In the tests where only 2 submissions are analysed, there is only 1 non zero principle component (e.g. as shown in Figure 5.28).

In cases where there are not obvious outliers and only subtle divergence, the distance to centroid from PCA alone is a difficult measure to use to determine if a consensus was in fact reached. As such, the author developed an additional analysis technique using a pairwise assessment of the average result with each submission. This yields a set of difference images, which are then evaluated voxel by voxel to give a pass/fail rate using a range of γ thresholds. This pairwise technique gives a more intuitive understanding of the level of agreement. Passing masks can also be easily visualised to identify the areas of discrepancy, which is beneficial for troubleshooting over using the PCA approach alone. The resulting technique is loosely related to the concept of gamma maps, that were developed as a way to evaluate agreement between treatment plan and dose measurement [129]. However, the idea has been heavily modified to the specific needs of this standardisation effort, restricted to a voxel by voxel comparison.

The passing rate plots R_γ display the percentage of voxels that are within a given γ threshold. As with the PCA plots, there is a curve drawn for each individual team which represents the pairwise comparison between that team's submission and the mean result at that iteration. However, in many cases the curves overlap as the submissions are extremely similar, as mentioned previously with the PCA plots.

In this work, a maximum threshold of $\gamma = 1\%$ was selected as the input value for the iteration process described in Section 5.6.3 and summarised in Figure 5.14. All submissions had to be within this tolerance for a valid CRM. This starting value is selected as a reasonable first estimate of acceptable error, yet it should be stated clearly that this is an arbitrary choice to start the iteration process. The optimal γ is thus reported, which is subsequently calculated as the minimum γ threshold that would have yielded the same level of consensus. The optimal γ evidently is found to be much lower than $\gamma = 1\%$, often by many orders of magnitude. As such, this suggests a much stricter tolerance could be selected as a starting point. Encouragingly, close agreement is clearly demonstrated through these plots, and this initial analysis shows that $R_\gamma = 100\%$ is reached when γ is much higher than 1% when there are outliers, and then far below it when they are subsequently removed. Here γ is expressed as a percentage, as the difference image intensities are converted into a percentage of the intensity range of the compared response maps (Eq. 5.15). The filter tests produce response maps with a variety of ranges, so it follows that the set tolerance is effectively more lenient in terms of absolute difference for the filters with a greater range.

5.8.2 Alternative Approaches to Measuring Response Map Differences

Other analysis metrics based on image comparison and quality assessments were considered and tested for this work. One of the simplest pairwise quality metrics is *mean-squared error* (MSE). This computes the average squared intensity differences between pixels/voxels in one image compared to another. For standardisation and reproducibility, one would then aim to minimise all pairwise MSE. However, as this MSE is a scalar global measure it will lose information via averaging the incidence of error. For example, if there is only a small region within a response map has discrepancy, as is the case with several of the filter tests using the impulse phantom, then a very low MSE can still be achieved, despite a qualitatively clear difference in response maps. This issue with scalar measurements is summarised nicely in an early study by Eskicioglu and Fisher [130], who assessed performance of a number of pairwise image quality measures including MSE, and concluded no single scalar measure could reliably describe a variety of discrepancies.

Another prominent pairwise image quality technique put forward by Wang *et. al.* [131] is the *Structural Similarity* (SSIM) index. The SSIM index was developed to evaluate the perceptual difference between two similar input images, inspired by the human visual system, which is highly adapted to structural information [131]. This metric is based on locally measuring 3 image characteristics: luminance, contrast and structure [131]. The SSIM index can be summarised as a value where 1 represents the highest quality match between the two compared images and usually falls between 0 and 1. This can be reported as a single scalar global measure, yet the technique results in a SSIM *map*, that is the same size as the compared images, containing locally calculated SSIM index values. The SSIM technique offers a viable alternative measure to the pairwise analysis presented here. An implementation for SSIM is available within the MATLAB image processing toolbox [132] and as such was added to the analysis pipeline here. Much like a *passing mask* (e.g. see Figure 5.26), visualising the SSIM map identifies local areas of discrepancy in the submitted response maps. However, for the purposes of these filter tests where the concerned is voxel-wise agreement, the SSIM adds a level of complexity to the analysis without apparent additional benefit. An arbitrary choice still must be made for an acceptable local SSIM index, and there are also several hyper-parameters for SSIM that can be tuned [132]. As such, a passing rate plot is utilised as an intuitive and acceptable pairwise method for consensus measurement for this study.

5.8.3 3D vs 2D Filtering and Voxel Size

In most tests a 3D convolution is evaluated. In general, a 3D assessment of structure and texture is thought to best leverage the available information within a volumetric medical image such as those obtained from PET, CT and MRI [96]. Correspondingly, this adds to computational complexity, particularly when considering right-angle rotational invariance (with 24 orientations to consider compared to 4). Optionally, one can of course apply the filters slice by slice in 2D, though this yields fundamentally different response maps to the 3D approach. For completeness, continuing future phases of this work aim to obtain benchmark consensus response maps for more 2D variants as well, where filters are generated and applied in this slice by slice manner. However, in principle the 3D filters are more relevant and challenging to implement for radiomics and thus where the focus of the preliminary set of filter tests discussed in this work.

As discussed in previous chapters, isotropic voxel sizes are recommended for 3D analysis. Voxel spacing can often vary in studies merging datasets from multiple clinics that have different acquisition protocols. It is important to note that the frequency responses of the same filter kernel applied to images with different physical voxel dimensions are not directly comparable. However, some filter kernels can be constructed incorporating the physical distance of the voxels (as discussed for LoG Filter), so the kernel can dynamically adapt to the input image, though in practice many filters are computed using “voxel space” irrespective of the physical dimensions of the scan (such as Laws or separable wavelets). Interpolation is thus recommended before the application of filtering in the radiomics pipeline so that all images in the dataset have the same physical voxel size, though interpolation also inherently changes the frequency content of any image by definition [86].

5.8.4 Considerations For Separable Wavelets

In Section 5.4.4, both the *decimated* and *undecimated* approaches to the separable wavelet transform are discussed. A study by Bogowicz *et*. *al*. [111] looked at the influence of two different implementations on the reproducibility of local control tumour models in head and neck cancer using PET imaging. The two implementations were from *University Hospital Zurich* and the *MAASTRO* clinic. They found the biggest discrepancies by far were observed in filter-based *wavelet* features (using a *coiflet* mother wavelet). Although supposedly based on the same mathematical definitions, overall, Bogowicz *et*. *al*. [111] found 88% of the 649 feature variants tested (568 of which were wavelet based) were not reproducible between the two implementations (ICC$\leq$ 0.8), with nearly all wavelet based features being non-reproducible. In particular, there were fundamental differences in the method and workflow when applying the wavelet filters. On review of the supplementary material [111], one software utilised a *decimated* wavelet transform with re-sampling of the contour to the lower resolution grid, where as the other used the *undecimated* approach. There also appears to be no consideration of rotational invariance techniques to account for the directional dependency of this type of filtering approach for either software. This study is a clear example of the need for standardisation and reference material for these features.

For the separable wavelet filters examined in this work, tests were designed for the *undecimated* (stationary) technique. Although both methods have been implemented by the author as part of SPAARC's internal development at Cardiff, the undecimated approach is more appropriate for image analysis with the aggregation of features from subsequent response maps, as it avoids having to down-sample the VOI contour alongside the image. As a result, although both methods have seen use in radiomic studies, the undecimated technique was the preferred and recommended approach to be benchmarked after discussion with participating teams.

5.8.5 Finding Discrepancy

The initial results obtained from the simplest mean filter tests (Section 5.7.1.1) highlight that the tasks of loading in phantoms correctly and padding were enough to generate tangible discrepancy. Padding choice is probably unlikely to have a significant effect on radiomic studies in the majority of cases as the tumour is often located far from the image boundary compared to the size of the filter kernel. Generally, most languages in which radiomics software is developed should have these common padding options as a standard function. Clearly, padding should not be ignored entirely from reporting for the benefit of reproducibility. More over, padding needs to be applied precisely for certain filtering techniques such as Simoncelli wavelets that use a cubic array to ensure the corresponding filters are circularly/spherically symmetric [86].

Notably, significant differences were found in filter tests requiring rotational invariance. On careful review, this is thought to have occurred in some cases depending on which approach was se-

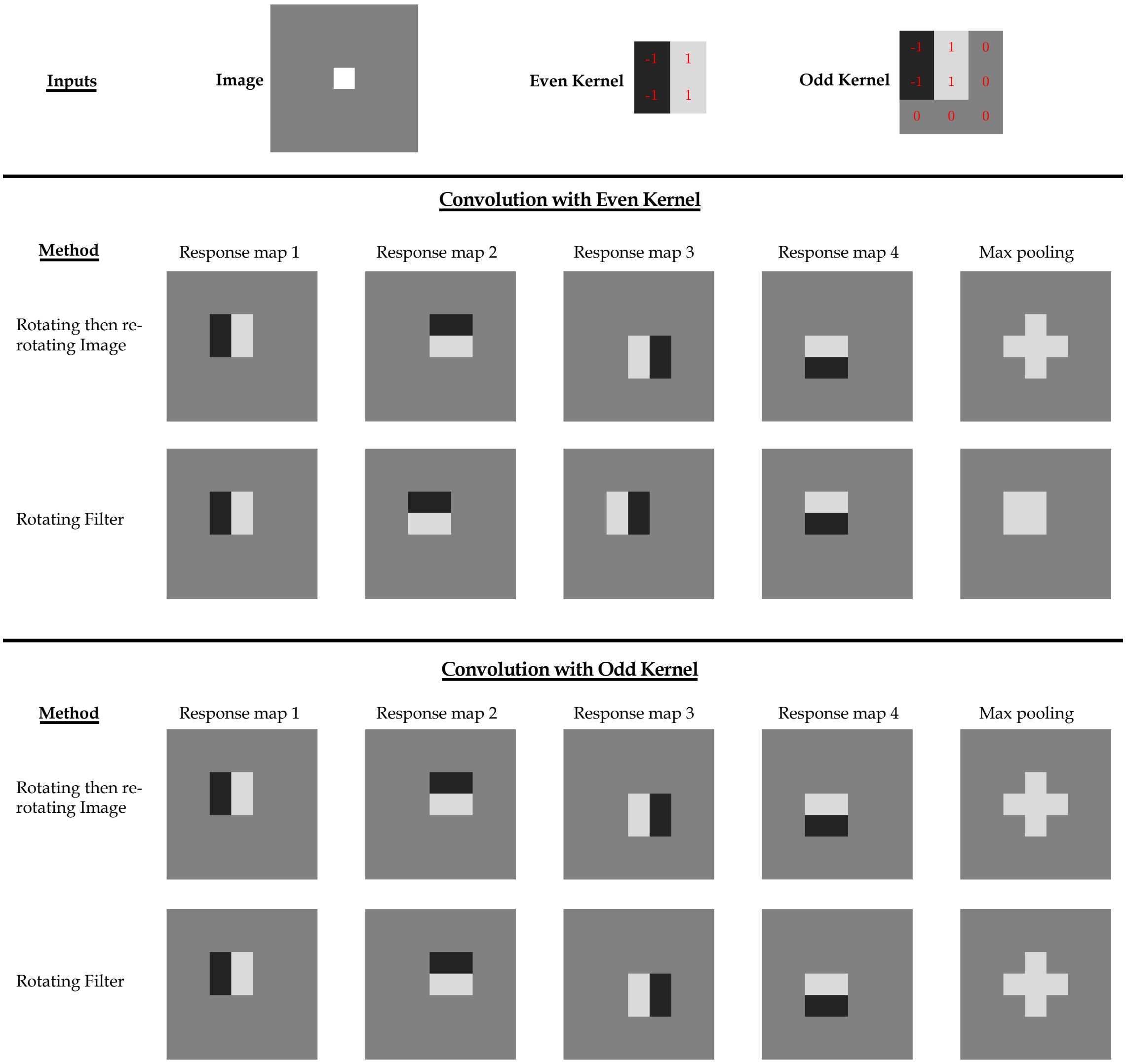


Figure 5.29: A visual comparison of two methods used to achieve rotational invariance, demonstrated with a simple impulse image and both an even and odd filter kernel. The two methods are: 1) rotate the image, convolve the rotated image with the filter kernel, re-rotate the response map, repeat from all rotations and pool results, or 2), rotate the filter kernel, convolve with the image, repeat for all filter rotations and pool results. As this is a 2D example there are 4 right angle rotations and so 4 response maps (in this case they are then max pooled). A discrepancy between the two methods arises when an even sized kernel is used. In this example, zeros are appended to the even kernel to create an odd kernel to demonstrate that this results in a match for both methods. This observation generalises for all even and odd 3D kernels.

lected to fulfil this criteria. The two approaches were either: 1) rotate the image, convolve the filter kernel with the rotated image, re-rotate the response map back to original orientation and then pool the response maps from each rotation, or 2), rotate the filter kernel, convolve with the rotated kernel image and pool the resulting response maps from each rotation. When the filter kernel is odd (and there is a clear central voxel in the kernel) these method are found to be equivalent. However, when the filter kernel is even (and there is *not* a clear central voxel) these two methods evidently produce slightly shifted results that subsequently do not pool in the same manner. Figure 5.29 provides a visual demonstration of this issue with a very basic 2D impulse image and kernels. In essence, this discrepancy occurs as a position is effectively chosen in the even kernel as the "centre" by convention, and this centre is not the same when the filter is rotated. Also shown in Figure 5.29, this effect can be corrected by extending the kernel to be odd by appropriately appending zeros. As such, odd kernels are recommended whenever possible if pursing rotational invariance, and in particular to append zeros to create an odd kernel if technique 2) is used in the software. Intuitively, rotating filter kernels is computationally more efficient than rotating large image volumes, though the latter method may be necessary if a team is incorporating an *off the shelf* function in their pipeline from a standard toolbox (for example, making use of MATLAB or pywavelets stationary wavelet transform *swt* function) instead of coding from scratch. As became clear from the set of submissions discussed in this chapter, this finding is a significant potential discrepancy point for software that must be addressed to achieve a standardised result.

5.8.6 Future Work

This chapter contains only the initial set of results for Phase 1 of this consensus-based research. As with the first study discussed in Chapter 3, the goal will be to further iterate to improve the strength of consensus for the filter test results presented here. Any consensus study is clearly limited by participation. In total, 9 teams contributed to this initial assessment of filter tests, and no one team contributed to all filter tests for this first round. Especially for the Gabor filter and Simoncelli wavelets, where only 2 teams provided results, no valid consensus was achievable by default. However, greater participation is expected as the study continues, similarly to the progression in the previous study (see Figure 3.4 and Figure 3.5 in Chapter 3). Several of the top performing teams and widely used open source radiomics implementations from the first study have not yet contributed results, though are expected to in upcoming iterations. Despite this, the results presented in this chapter reveal necessary insight into the challenge of reproducible filtering for radiomics and develops the methodological basis to underpin the continuation of the study.

An inherent limitation of this work is the necessarily narrow scope. It is not feasible to provide exhaustive references and benchmarks for the abundance of designed imaging filters one could deploy for radiomics studies. However, to give a reasonable representation, filtering techniques selected here were identified through availability in several prominent radiomics toolboxes and published results, including PyRadiomics [133], IBEX [134], LIFEx [135], QuantImage [136] and CGITA [137]. Filter tests for advanced techniques such as steerable wavelets will also be included as the initiative continues [86]. Although not exhaustive, the filter selection represents a number of common techniques with enough diversity to reveal clear causes of discrepancy in prominent radiomics software that can also be instructive and generalisable beyond any one specific filter technique. As feature algorithms themselves have already been standardised in the previous study, the aggregation of the features from the filtered image that will be tested in the next phases should rapidly cohere between teams, once there is a strong consensus in the critical Phase 1 tests explored in this chapter.

5.9 Conclusion

This chapter discusses the challenges of implementing a set of convolutional image filters for radiomic analysis. A methodology was developed by the author to determine consensus-based *response maps* as part of further benchmarking to feed into the *Image Biomarker Standardisation Initiative*. These CRMs can be used as references to aid reproducibility of future radiomic studies that utilise these filters in the search for clinically relevant image biomarkers. Here, a set of benchmark tests were compared for a collection of prevalent filtering techniques, and the initial results of 9 unique software implementations were evaluated using the designed evaluation method. At least a moderate (3+) or better majority consensus was achieved in 18/25 of the filter tests presented. Nearly every filter test had submissions that were detected as outliers that varied beyond an acceptable threshold. From this analysis, the author identified several key potential causes of discrepancy that significantly affected the reproducibility of filter tests between different software. This study forms the preliminary baseline of consensus that will be iteratively improved on in the pursuit of stronger benchmarks for filter-based radiomics.

Take home message

1. Application of image filters prior to radiomic feature aggregation must be carefully considered.
2. A methodology was developed to evaluate response map discrepancy between software for a set of filter tests.
3. Initial results from software contributing to the IBSI achieved at least a moderate consensus for 18/25 filter tests.
4. Significant discrepancy and outlier submissions were identified in comparison of radiomics software.
5. Continuation of this standardisation study will benefit radiomics reproducibility.

6

Further Discussion, Future Work and Conclusions

"The beautiful thing about learning is no one can take it away from you."

— B.B. King

6.1 Summary of Significant Contributions

This project resulted in a comprehensive, standardised, reproducible pipeline for radiomic analysis in oncology that can be used to analyse 3D regions of interest from medical image volumes obtained from modalities such as CT, PET and MRI. These radiomic features are designed to quantify tumour heterogeneity. In building this pipeline, several key causes of feature extraction discrepancy were identified and evaluated, such as interpolation grid alignment and thresholding technique. Prior to this work, there was no definitive guidelines for best practices in radiomics for pre-processing techniques, which severely undermined study reproducibility and thus chances of clinical translation.

This project work provided a core and value contribution in an international collaborative effort with the IBSI to establish feature definitions and define a clear image processing scheme to produce a set of standardised radiomic features that quantify tumour morphology, statistics and texture. The key novel aspect of the IBSI was to achieve collaboratively determined benchmarks through exhaustive implementation, an approach that proved instrumental to identifying causes of discrepancy that effect reproducibility of radiomics research. Open datasets were made available to ensure one can check if they are compliant to the achieved standard. Over 160 baseline feature algorithms were benchmarked using a variety of extraction settings, providing well over 1500 reference values in total. These reference values and this work have already become highly used and cited in the field.

Furthermore, in exploration of feature robustness, the author identified a gap in knowledge concerning the effect of 3D isotropic interpolation on radiomics features. Through a study utilising a large set of staging PET images of patients with oesophageal cancer, the stability of features extracted from the primary tumour were assessed after resampling the imaging to a range of voxel sizes. Features were categorised based on their response to interpolation. For apparently systematically varying features, this work was able to model how many of these features changed when extracted at different voxel sizes for this dataset. Two standardised interpolation methods were tested (linear and spline). This study showed that feature values often varied depending on the interpolation method, but the stability of a feature to voxel size resampling remained consistent. From this study, it is recommended to test for and use stable features in modelling, particularly if re-sampling datasets containing imaging from different protocols with different reconstructed voxel sizes. This study also confirmed that extracting features at multiple voxel dimensions by interpolating the imaging will lead to a large amount of feature redundancy. Optimisation strategies can therefore prune a large amount of features before modelling.

Moreover, this project has further contributed to radiomics standardisation in the context of image filtering prior to feature extraction. To continue the IBSI work, the author developed and tested a consensus methodology to analyse response maps produced from a number of image filter tests on digital phantoms. In essence, these filter tests were designed to evaluate the reproducibility of image filtering techniques across different radiomics software. For each filter test, the analysis methods described were used to measure variation between outputs from different software to identify and remove outliers from contributing to a majority consensus result (a CRM benchmark). Analysis of initial submissions from a subset of teams identified very close agreement for some software and produced valid initial references maps for a variety of filter tests. However, this initial analysis also identified significant discrepancies in software, which has important consequences for radiomic studies using additional filtering for analysis. This work presented several potential causes of discrepancy and provides the ground work for future iteration towards another set of radiomics consensus-based reference values and recommendations.

In line with the aims laid out in Chapter 1, this thesis primarily has focused on the task of standardisation through benchmarking and interpolation robustness in the radiomics pipeline. Alongside this, there has been a significant technological transfer aspect to this project with the repackaging of the SPAARC code into an extension that interfaces with MIM (MIM Software Inc.), a leading commercial provider of medical imaging software (as discussed more in Section 6.4). The SPAARC pipeline has been used in a number of radiomics studies. The following section briefly discusses some of the published work that has utilised SPAARC tools.

6.2 Additional Research Output using the SPAARC Pipeline

Development of the SPAARC radiomics software has facilitated exploration of a variety of radiomics research questions during the course of this project. This section summarises the themes and challenges explored in several of these other co-authored published studies.

Theme: Challenge posed to radiomics by segmentation method

Parkinson *et. al.* [76] evaluated how a selection of classic PET semi-automated segmentation methods affected development of prognostic models built incorporating a small set of standardised radiomic features. This study was performed using the large STAGE dataset (PET imaging of patients with biopsy-proven oesophageal cancer) that was also used in the research described in Chapter 4. Separate prognostic models were developed with the same clinical data and selection of standardised features extracted (with the same settings) from the metabolic tumour volume, segmented with *different* semi-automatic methods.

This work has shown that variability in segmentation from the different semi-automated methods led to variability in extracted radiomic features. As a result, the prognostic models were also dependent on the segmentation. Importantly, the prognostic value of incorporating radiomic features into a model was found to change based on the segmentation used prior to model development. These different models led to different patient risk stratification. In other words, patients could change risk group (e.g. from low to intermediate) based soley on the selected segmentation method, holding all other factors in the pipeline constant. This study emphasised a known key challenge for radiomic studies, reiterating that the segmentation method must also be standardised and reproducible alongside the other radiomic extraction techniques. The detriment if models like these saw clinical use is clear, one could imagine a scenario where a subset of patients who were assigned to a lower risk group (based on one model) are denied effective therapeutic treatment that they would have received if only another segmentation method had been used to generate the model and placed them in a higher risk group. In reverse, a different subset of patients could be assigned to a higher risk group and thus received treatment that is more aggressive than necessary.

Theme: Challenges in development and external validation of prognostic models incorporating image features in PET

Foley *et. al.* [87] explored an external validation of a prognostic model developed on patients with oesophageal cancer (OC) incorporating both clinical variables and PET imaging features. In essence, generalisability of a model originally developed on the UK-based STAGE dataset (again, as used in Chapter 4) was tested against an independent cohort of patients treated with the CROSS regimen in The Netherlands.

Although the original model only incorporated simple first-order image features, to ensure reproducibility the tools developed in this work facilitated extraction of the features in both cohorts of this study. The key finding in this research was that a prognostic model developed with the STAGE dataset combining clinical variables and simple image features was not able to significantly discriminate between risk groups in the CROSS validation cohort. Inherent limitations of the methodology may have contributed to the lack of external validity for the model developed in the STAGE dataset. For example, a key limitation was the variation in PET/CT scanners and acquisition protocol in the CROSS dataset, compared to the STAGE cohort. To address this, *combat harmonisation* [138, 139], a method proposed to limit potential batch effects and harmonise features extracted from centres, was also explored. However, combat harmonised features showed no significant improvement to the model validity between the two cohorts in this case.

This study confirmed the difficult realities of identifying truly generalisable and universal image biomarkers for prognostic assessment to aid staging in OC. As such it presents a negative result, which are lacking in radiomics literature.

> `Theme:` Stability and prognostic value of radiomic features from contrast and non-contrast enhanced CT (Oesophageal Cancer)

Piazzese *et. al.* [88] utilised the SPAARC radiomics pipeline to investigate feature stability, prognostic value, dimensionality effects, and contrast agent dependency, using planning CT images of patients with OC. In line with other published studies (eg. [140]), results from this study suggested radiomic features were more stable (in terms of feature distributions) if assessing cohorts of contrast and non-contrast enhanced CT images separately, instead of in a combined cohort. This study also assessed both 2D and 3D aggregation methods for the texture features, with slightly more 2D features found to be stable over their 3D counter parts. This study also identified a potential imaging biomarker (GLDZM zone distance variance) for OC CT imaging that was both statistically correlated with overall survival and was independent to contrast administration. However, a major limitation of this study was that the survival model was developed and predictive power tested on the same dataset. There was not a split into training and validation to attempt to keep the power of the study high, as such, in terms of the *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD) evaluation scheme [141] (see Section 6.3.2), this study would be a type 1 prognostic model, and needs further external validation on a separate dataset to confirm or refute the validity of this finding.

> `Theme:` Radiomics-based biomarkers in advanced pancreatic cancer.

Mori *et. al.* [89] deployed the SPAARC software in the training and validation of a radiomic-based model for locally advanced pancreatic cancer (LAPC) in an effort to predict *distant relapse free survival* (DRFS). Two robust PET based radiomic features combined in a prognostic index, derived from Cox analysis, showed superior ability to identify risk of metastatic relapse when compared to currently used clinical variables/biomarkers. Notably, the two image features were morphological and statistical based and not textural. Importantly, reliable models to predict DRFS in this area could guide treatment personalisation, for example, in tailoring radiotherapy dose to be more palliative for patients at a high risk of rapidly developing metastatic disease, or increasing local therapies for those at lower risk. This study presented a successful training and validation of a radiomics model in pancreatic cancer though splitting the cohort. However, further validation with independent cohorts are needed to confirm the potential of such a radiomics index.

In review, a particular key challenge of radiomics in PET imaging of pancreatic cancer is the scan resolution compared to the small tumour volumes usually involved. If a tumour is only represented by a small number of voxels within the scan, then it will be more unstable to slight variations in segmentation. As previously discussed, segmentation is already a major factor for radiomic studies, and for small tumours, this problem is amplified. As well as this, there is limited textural information one can extract from limited voxels. In fact this most often results in highly correlated features or static features values across the whole cohort that would not be useful for patient stratification.

The purpose of this section was to briefly highlight some of the additional impact of this projects work through the use of this standardised software in published research by the Cancer Imaging and Data Analytics (CIDA) team at Cardiff University School of Engineering and external collaborators. A collection of other studies are still ongoing that are utilising the SPAARC pipeline. The following section reflects further on the radiomics field in general and the lessons learnt from undertaking this research.

6.3 Critical Reflection of Radiomics Research

6.3.1 Engineered Features vs Deep Learning

As briefly mentioned in Section 1.5, this project set out to address key challenges within the domain of the *traditional* radiomics approach, which uses feature algorithms with pre-defined mathematical formula that can be interpreted precisely. However, there is a parallel track that has gained widespread attention: so-called *deep-learning* (DL) based radiomics. Under the headline grabbing banner of *artificial intelligence* (AI), the field of DL has seen monumental advances and uses in many areas of science and engineering over the last few years, such as: speech recognition, object detection in imaging, self driving cars and drug discovery [142]. Particularly, digital imaging has become a successful frontier for DL techniques.

The interest and accessibility of AI tools [143, 144] has enabled researchers from many domains to try to harness its potential, and it is no surprise that medical applications represent a broad frontier for AI research as a result. A comprehensive review by Asfhar *et. al*. [145] provided excellent discussion of advancements and advantages between these two approaches to radiomics. This section reflects on the engineered feature approach of this thesis compared to DL and the potential future of radiomics in this context.

At its core, DL replaces the mathematically defined features aggregated from a segmented region of the image, with features that are iteratively learnt by a complex neural network. This *end-to-end* process requires large amounts of training data and careful design of the network architecture. There are an abundance of excellent introductions to the DL process to which the reader is referred [142]. In DL radiomics, large datasets of labelled images are used to train the neural network to independently learn to recognise disease from normal tissue. In effect, to automatically develop a set of relevant features, without human intervention, that quantify tumour heterogeneity. Evidently, DL techniques replace much of the processing pipeline that was standardised through the work in this project (refer to Figure 2.2).

In comparison of these two techniques, Asfhar *et. al*. [145] highlighted some key advantages and disadvantages between engineered features and DL. One key advantage for DL is that it does not require (in principle) a contour defining the tumour. The idea is that the network will learn which regions are important, which then negates the challenges of inter and intra-reader variability in the generation of accurate contours that is a core component of traditional image biomarker research. As discussed, a key limitation of radiomics is the clear dependency on segmentation. In reality of course, DL requires many of its own processing steps to prepare images to be fed into the network. For example, network architectures may need to be constrained to a set image matrix size, which would require re-sampling or the use of bounding boxes, introducing necessary image processing that can undermine reproducibility as demonstrated through the standardisation of the hand engineered approach. DL is not an automatic safe haven for reproducibility.

Unlike DL, the traditional approach to radiomics of course requires design and selection of features. This is both an advantage and disadvantage. In one sense, you are limited by the imagination of the scientists that engineered said features. We may not have defined the right combination that can precisely capture and partition patients based on a signature of disease. On the other hand, as they are pre-designed, they are tangible.

A key disadvantage of DL is that there is a loss of intuition, and the decisions that are reached are effectively the result of a *black box* [146, 147]. In this context, a black box refers to a model that is so complex that it is fundamentally not understandable by humans. This is especially problematic for DL medical applications as clinical decisions could rely on black boxes that may have unrealised and unexplainable bias. To address this, there have been many attempts to ”explain” how DL techniques arrive at a given decision: this has itself received significant push back as the ”explanation” often still does not provide enough detail to evaluate what the black box model is actually doing. In fact, it could be dangerously misleading.

A perspective by Rudin [147] argues that in the case of high-stakes decisions (such as in healthcare), we should *”stop trying to explain the black box and use interpretable models instead”*. One of the examples they explore criticises the use of so-called *saliency maps* that are often touted as offering an explanation of DL image processing classifiers. These saliency maps are used to determine which parts of the image are being omitted by the classifier as not relevant, highlighting regions within an image that are significant in the decision made. However, Rudin [147] argues that knowing where the network is ”looking” does not inform what it is actually doing, and these saliency maps can appear basically the same for completely different classes and give an illusion of understandability that isn’t accurate to the workings of the underlying model. Beyond imaging, in light of the recent surge of general interest in AI in medicine, guidelines such as Challen *et. al*. [146] summarise many of the major short to long term clinical safety aspects that must be considered in this rapidly evolving domain.

The traditional radiomics approach has a significant advantage in that models will be generally more interpretable as the identified imaging biomarkers are tangible. As they are engineered, they can be understood. Even if a remarkably accurate prediction model were to be developed with DL methods, clinicians would likely and rightly resist blind trust in the machines recommendation. This alone explains why the techniques that were standardised in this work remain prevalent within the literature, alongside the wave of impressive DL results.

As well as this, inherently in DL one cannot select a predefined set of features that are known to be robust to various parameters, such as acquisition protocol or voxel resampling. One must evaluate DL robustness at the level of the model (which requires careful collection and curation of outcome data by design), instead of at the level of individual features (which can be done without outcome data) [148]. The former is again a far more difficult task. Many image processing issues that affect the reproducibility of engineered features will generalise to DL models. For example, a DL model could overfit to particular aspects of an acquisition protocol if the dataset is not sufficiently diverse. This could be avoided in a traditional radiomics approach by using features that are known to be robust to this issue, hence the importance of robustness and repeatablilty testing. As the features are independent from the data, these models can be developed on smaller datasets [145]. Practically, for rare cancers there may only be a small cohort of relevant patients which would disqualify a DL approach from scratch by default, unless federated learning approaches can be adopted (as discussed later in Section 6.3.3).

Importantly, interesting opportunities exist with hybrid solutions involving both traditional and DL techniques. For example, DL based segmentation has shown increasing potential within the medical image space, with the caveat that there is still much ongoing search into the reproducibility of many DL contouring methods [149]. As such, a clear hybrid solution would be a radiomics model that is built with a DL segmentation algorithm to outline the tumour, such as the popular U-Net [150], with the contour subsequently fed into the traditional radiomics pipeline to extract features, as discussed throughout this thesis.

Another interesting hybrid approach is to use both techniques to train two classifiers separately, one with engineered features and the other using a selected DL neural network, and then combine the decision of each. For example, a study by Antropova *et. al.* [151] in breast imaging demonstrated increased predictive performance in lesion diagnosis after fusing features pooled from a pre-trained convolutional neural network (CNN) with handcrafted radiomic features. Similarly, Diamant *et. al.* [152] found that a model combining results from a traditional radiomics approach with a deep learning model produced the best outcome prediction for head and neck cancer.

Whether engineered features, DL, or a hybrid combination of both will prevail remains under rapid investigation within the radiomics community. Certainly, as the radiomics field has matured over the last 5 years the expectations of feature based techniques has become more grounded with the realities and difficulties of reproducibility [153, 154]. Either way, the standardisation of the traditional radiomic approach was a clear necessity, and the work underwent in this project has provided important contributions in this regard.

6.3.2 Avoiding Fortuitous Features

A well known aspect of the engineered feature approach that has become self evident to the author throughout implementation is the sheer number of tweaks to parameters that can be made, and how this can lead to an almost limitless number of potential feature variants for tumour analysis. This is a major consideration in machine learning in general, often referenced throughout the literature with common phrases such as the *"curse of dimensionality"* (e.g. [63, 64, 145]). High dimensional data can lead to over-fitting to a dataset, which limits generalisability of any model and causes a high rate of false positive results. This is particularly prevalent in medical data, where the cohorts often contain a relatively small number of patients compared to the potential number of features that could be extracted. In the context of radiomics, the more image features one has to consider, the higher the chance of finding *at least one* that correlates with a clinical endpoint. Without careful navigation, radiomic studies can be subject to severe limitations and methodological pitfalls when this is not carefully handled. Comprehensive textbook resources (e.g. [155, 156]) are available that discuss this challenge of modelling with high dimensional data, along with published recommendations to avoid common pitfalls in the context of radiomics [148].

As initiated with the work in Chapter 5, adding different filtering techniques only compounds this challenge of high dimensionality. The potential combinations of filters and features on top of other extraction settings rapidly increases the number of quantitative metrics one can obtain from a single image. A small baseline set of features can easily reach many thousands after incorporating only a limited selection of filters, if the full gamut of texture algorithms are utilised on each response map. It is not clear if extracting texture features on top of complex filtered imaging is optimal. The aggregation of only first-order features from response maps remains a sensible self-imposed limitation, with an added advantage that it also helps to preserve some interpretability of the features. However, best practices in feature aggregation of filtered imaging remain open to further study.

Data reduction and *feature selection* techniques are used to reduce chances of fortuitous feature identification in radiomics modelling [148, 155]. Robustness testing, as this work has explored, facilitates this need. A non-robust feature may show a fortuitous statistical correlation with the clinical endpoint and prior removal is thus pivotal to obtaining a generalisable model. As well as removing non-robust features, those remaining should not be strongly correlated to other clinical metrics or each other. Following the principle of *Occam's razor* [157], selecting the *"simplest"* image biomarker is likely the correct choice if it is tightly correlated with something more complex. A complex texture feature that ends up highly correlated with volume is likely of little additional benefit [77]. As touched on in the previous section, this notion applies to any prognostic or predictive modelling technique as well: a simpler model is favourable if performance is comparable to something more complex. Often in pursuit of novelty in radiomics studies there is a failure to compare to simpler baseline clinical models and practices [148]. On this point, as discussed by Kazmierska *et. al*. [154], greater collaboration between data scientists and clinicians at all stages of research can guarantee that a proposed model could in fact have a clinical role by answering a previously unmet clinical need. Radiomic models should always attempt to demonstrate their value in addition to other existing clinical decision tools, such as normal staging or RECIST (Response Evaluation Criteria in Solid Tumors) [158].

As a form of feature reduction, families of features can be *clustered* into a single representative value (through techniques such as PCA, that was utilised in another context in Chapter 5). However, this again loses interpretability and could potentially keep a collection of redundant features in the extraction and modelling pipeline that simply amount to noise. To reduce features, many different selection algorithms have been explored in radiomics to varying success. For example, a comprehensive study by Leger *et. al*. [159] assessed the ability of 11 machine learning approaches combined with 12 feature selection techniques to predict loco-regional tumour control and overall survival for patients with head and neck squamous cell carcinoma. By performing a systematic evaluation they found that no one combination of methods noticeably outperformed all others. A subset of different feature selection and learning methods achieved similar results, suggesting that a number of approaches are viable for radiomic analysis. They also suggested the feature selection method was more important to development of a useful model than the learning algorithms that were tested. Although one of the simplest, Spearman rank correlation proved to be one of the most effective selection methods in this study [159].

To avoid fortuitous feature finding, stringent validation strategies will also need to become the norm in published research [158]. This requires the partition of cohorts, and optimally the utilisation of both internal and external (large) datasets to really test a model's viability. Naturally, to build a prognostic model that is suitable for clinical prospective use requires much rigour. As such, a number of prominent published guidelines have become recommended in the radiomics community to systematically judge study quality [7, 141, 160]. The combination of the following resources can facilitate effective, reproducible results in radiomics:

- **Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis** (TRIPOD). The TRIPOD statement is a landmark reference providing guidelines for reporting prognostic models [141]. This article does not specifically mention radiomics. Rather, it is applicable broadly to any study investigating diagnostic or prognostic modelling. This statement defined different types of prediction and prognostic modelling studies using four main groups. Briefly, in studies of *type 1* a single dataset is used for model development and reporting results. In *type 2* studies, one dataset is still used but partitioned into separate development and validation groups (this is effectively an *internal* validation). In *type 3* studies, a different dataset is used to evaluate model performance (e.g. acquired from another study or obtained from another institution). Finally, in *type 4* studies no development is used, rather it is just the evaluation of an already published model independently applied to a new dataset. As such, type 3 and 4 studies are *external* validations. Intuitively, a model performing well in a type 3 study is proven to be more generalisable than type 1 or 2. Type 3 have been demonstrated to be more reliable in the context of radiomics [39], and will likely become the minimum standard for publication as the field matures and moves beyond the *"hypothesis generating"* discovery phase [161]. Radiomic studies should report their TRIPOD study type as standard.

- **Radiomics Quality Score** (RQS). Building on the example of TRIPOD, the RQS proposed by Lambin *et. al.* [7] provides a specific radiomics orientated assessment checklist. This allows one to rapidly establish whether a study is complaint with proposed best practices. The RQS evaluates sixteen key components relevant to radiomic studies with a point based system that relates to importance of a given criteria, with a maxiumum of 36 points obtainable. Notably, penalty in the form of negative points are given for studies that do *not* consider feature reduction when the number of features exceeds the number of samples. The RQS is presented as a comprehensive checklist (available online [162]). Importantly, reviews utilising the RQS - such as by Valdora *et. al.* [14] on radiomics and breast cancer - have found substantial quality limitations in published research, highlighting the need for higher quality prospective and reproducible results to further evaluate the potential of radiomics. The RQS score was also used in a review of reporting quality in radiomics research in nerooncology by Park *et*. *al*. [163], who from 51 articles identified a median RQS of 11/36, with 29.4% of studies performing some external validation. They also concluded in many cases the quality of reporting across studies was insufficient.

- **Checklist for Artificial Intelligence in Medical Imaging** (CLAIM). Another study assessment criteria, CLAIM, has also been recently proposed by Mongan *et*. *al*. [160] in light of the ever growing interested in AI based techniques within medicine. This checklist of best practices for AI in medicine acts as a resource to guide authors (and reviewers) to present research in a clear and scientifically rigorous manner. It should be noted that CLAIM guidelines contain deep learning specific considerations for studies, though are still broadly applicable to the engineered feature extraction methods as well.

Adherence to these resources - alongside the standardisation of feature extraction guidelines of the IBSI to which this work has contributed - will improve the research quality and output in radiomics moving forward.

Optimal investigations that will produce clinically relevant and usable results will have been developed and validated on very large multi centre cohorts, the best studies of which will be prospectively designed. Image biomarkers identified in this way are unlikely to be fortuitous. Naturally, this requires access to data, which is a significant challenge in the context of medicine when factoring in patient privacy and safety. Although open access to data is desired, securing patient privacy is essential [154]. A challenge encountered in this project was access to abundant curated clinical data. The path for radiomics to succeed is through data availability, and the lack of multi centric datasets is a clear barrier.

6.3.3 Data Sharing and Availability

A way to avoid fortuitous biomarkers is external validation through access to various independent datasets. More over - once an imaging biomarker has been identified - as imaging acquisition, protocols and treatment strategies evolve, there will be a constant pressure to re-evaluate and update models to account for potential *distribution shifts* [146, 164], where a mismatch develops between the older training data and the environment in which the model is in operation. This mismatch would render a model less effective. Challenges such as distribution shift can be mitigated by multi-centric continuous collection and curation of new data that is readily assessable. There are three key avenues to address this validation data bottle neck: (1) publicly available datasets, (2) centralised data-centre infrastructures, and (3) distributed / federated learning infrastructures.

Public online sharing of oncology imaging files has been facilitated through initiatives such as *The Cancer Imaging Achieve* (TCIA) [165]. This is an open access resource that host millions of imaging files and has been leveraged by many radiomics studies, including this project, as a feasible way to enable sharing of clinical images and other metadata across multiple research sites [166]. However, the sensitive nature of medical imaging requires it to be appropriately anonymised for the public domain. This is not necessarily just a simple task of removing identifying metadata from imaging files (e.g. stripping information from the DICOM header). For example, with some 3D medical data, even after metadata anonymisation, the patients face can still be reconstructed and potential identified with facial recognition software [167]. Gaining ethical approval to open-access imaging data necessarily has stringent ethical and legal hurdles as a result. However, if overcome, this availability is clearly a valuable asset to the research community.

Rather than making data public, another approach is to have an entity maintain one centralised repository where instead of open to all, only trusted users are granted access [154]. This type of infrastructure houses scan collections in one central location (sometimes referred to as a *Data Lake* [164]) from many clinics and institutions that are in collaboration. To add to security, a computational environment can be constructed around the central repository, where users do not directly access the data, but package and send their algorithms to the central sever and then collect back the results. Of course, this approach necessarily still requires data to leave individual clinics, which can have many of the same administrative and political barriers as open access. To circumvent this issue, distributed and federated learning approaches have thus gained interest as a way to enable collaborative learning on multiple datasets without any need for data exchange [164, 168, 169].

With federated learning, rather than sending patient data across institutional firewalls, only the models themselves are distributed. Learning occurs on the data locally, and models are aggregated either by a central server or exchanged and updated in a peer to peer process: full federated learning work flows and their potential future in digital healthcare are discussed in detail by Rieke *et*. *al*. [164]. This approach has clear benefit to radiomics modelling, offering a mechanism to address the lacking validation of models in the literature identified in recent reviews [158, 163]. In particular, the potential for large scale distributed learning and data analysis was demonstrated in a study by Deist *et*. *al*. [169] on over 20,000 lung cancer patients: they connected lung cancer databases containing tumour staging and post treatment survival information of 8 healthcare institutes in 5 countries to train and validate a logistic regression prediction model for post-treatment two-year survival. Although this model did not involve image analysis, they demonstrated the successful adoption of a federated IT infrastructure. Our group (Cardiff) were one of the eight centres in this research and this experience with the infrastructure could help explore more avenues for radiomics incorporated federated learning.

Finally, to facilitate federated learning - or in essence any validation of a radiomics model - there is a substantial push to ensure data used in research is FAIR (findable, accessible, interoperable, and reusable) [170]. Particularly for federated systems, data needs to be curated such that it is interoperable, e.g. following a standardised medical ontology and semantically organised so that it can be consistently queried and interpreted. For multi-source data - from imaging to electronic health records to outcome data - FAIR guiding principles are argued by Kazmierska *et*. *al*. [154] as vital to the translation of multi-source models into used decision support systems in the clinic. To this end, part of the IBSI output has been to consolidate an ontology from the set of common radiomic features that were standardised.

6.3.4 Positive Study Bias

In reflection, it should also be emphasised that despite the excitement and flurry of activity within the domain of radiomics (and prior simpler CAD systems) since 2012, to the best of the author's knowledge, there has still not been a case where a radiomic model has made the transition to use for personalised patient management in oncology. Notably, this slight undermining of the initial high expectations is in conflict with the continuing strong bias in positive results within the literature. The recent excellent review by Dercle *et. al.* [158] highlighted the overwhelming majority of the 165 radiomics articles they evaluated were positive. As such, it is likely that many negative results go unpublished due to their perceived lack of novelty or after facing many of the challenges discussed in this work (such as different imaging protocols or not enough available validation data). These negative results however are important to guiding future research.

For example, van Timmeren *et. al.* [171] attempted validation of a radiomics signature utilising cone-beam CT (CBCT) taken of NSCLC patients treated with (chemo)radiotherapy. The initial signature trained on 141 patients could not be validated with 3 external data sets, suggesting longitudinal CBCT radiomics could be of little value in outcome prediction in this domain. Moreover, a study by Welch *et. al.* [172] highlighted the vulnerabilities of potential radiomic signatures by performing an external validation of a model, then deliberately randomising signal intensity in the VOIs. They found a similar performance from the model after randomisation of intensity values, suggesting the intensity and textural distribution was not the relevant prognostic driver, rather, it was the volume. Similarly, in a retrospective study by Ger *et. al.* [173] utilising 726 CT and 686 PET images from head and neck cancer patients, they found that radiomic features were not consistently associated with survival and failed to improve prediction compared to just using volume. Even when sub-setting the datasets to examine only imaging with the same protocols, radiomics features were not found to be beneficial. They concluded that for head and neck cancer, radiomic signatures may not be reproducible even across similar cohorts, which contridicts other promising studies such as Vallières *et. al.* [118].

In summary, despite the promising applications and pilot studies of radiomics in oncology, whether it will ultimately provide reliable decision support in many domains remains to be demonstrated. The fact there is still no prominent subset of features that are clear imaging biomarker with definitive clinical value in the literature highlights the intricate challenges of each cancer cite and situation.

6.4 Future Work

Many aspects discussed in this thesis will form the basis for future work. Most pressingly is the continuation of filter-based image biomarker standardisation with the next phases of the second IBSI study. Chapter 5 discussed and developed initial ground work for the second standardisation effort. Tentative benchmarks were determined for 18/25 filter tests with at least a moderate consensus (3+). This work will be continued and extended both in the number of filter tests conducted, and ideally the number of contributing teams. The results presented here will be iterated on in pursuit of a strong (6+) consensus in the response maps produced for each of the filter tests. Subsequently, filter-based features will be benchmarked and further validated using the same datasets as the original study discussed in Chapter 3. As also mentioned in this discussion, the addition of many filters can lead to extremely high dimensional data. Once standardised, the most appropriate aggregation algorithms to use on filter imaging remains a clear area of investigation.

The primary research aims within this thesis have been the standardisation of image processing and feature algorithms. The author explored a hypothesis that implementation discrepancy plays a major role in radiomics reproducibility issues. As discussed, the resulting SPAARC software was utilised in focused studies that developed and evaluated radiomics models for particular cancer sites, such as the oesophagus [76, 87, 88] and pancreas [89]. However, the bulk of the work presented in this thesis did not explore model development or identify any potential imaging biomarkers directly. In one aspect of future work, there is potential to utilise the standardised algorithms in SPAARC, along with the recently implement filter-based methods, to analyse tumour heterogeneity in a collected MRI dataset of patients with glioblastoma, in light of the principles discussed in the critical reflection above. In particular, the variety in MRI acquisition protocol will require careful consideration to identify the appropriate radiomic extraction settings.

The study in Chapter 4 provided insight into the robustness of radiomic features to isotropic interpolation. This is just one of a number of extraction settings that requires careful consideration by a researcher when conducting a radiomics study. Although this thesis explored the standardisation efforts of a large set of algorithms, it is important to note that the features only remain standardised if the extraction settings are consistent. Standardisation offers no recommendation for the "correct" settings to use, only that the same result is achieved for the same settings. For example, Section 3.3.3 discussed how the ordering of thresholding techniques caused discrepancy and thus required consistent application between software. However, the optimal threshold to select, or whether to use one in the first place, remains an open question that is likely highly specific to a given cancer site and clinical question.

Recent work by Fornacon-Wood *et. al*. [174] comparing now standardised radiomics software emphasised the point of setting consideration by reporting that default settings between a small selection of open source implementations that had been standardised were still not the same. They found different default settings led to varying reliability in the prognostic value of certain features, in some cases reversing the relationship between the feature and survival from protective to harmful. They showed that harmonisation of the calculation settings in the software dramatically improved this result. As such, many avenues for future work remain to identify the optimal use and impact of extraction settings. Any published radiomics model should be scrutinised by its dependence on user selected settings.

If clinical utility of a set of standardised image biomarkers is proven, a clear remaining challenge is the frictionless integration of radiomics tools into a clinical workflow. SPAARC has been developed primarily as a scripting based software package in Matlab, and as such is mainly a tool for the researcher. In anticipation of any potential clinical adoption of radiomic methods, there is ongoing work to develop an extension that utilises SPAARC's standardised algorithms in a readily assessable workflow that is compatible with MIM (MIM Software Inc.) a clinically utilised commercial imaging software. There is a signed user agreement to evaluate this extension in the coming year. In the future, an extension like this could benefit both researcher and clinician. Of course, extraction will need to be made suitably customisable to any radiomics signature of interest. It will also be intrinsically linked with the segmentation method. In an imaged future workflow, important radiomic imaging biomarkers and prognostic information could be displayed directly within the software used by clinicans, alongside the imaging, and collected as a report. For this reason, the SPAARC implementation has currently remained in-house as this commercial application is pursued. In the interest of reproducibility, many software packages are published as open-source (e.g. [133, 134]). This remains an option to increase the impact and reach of the tools developed in this project, subject to institution approval to release the code. SPAARC automatically saves extraction settings, though future work could expand this to automatically generate reporting tables for publication.

Future work in the CIDA team has also begun to extend the goal of standardisation of radiomic algorithms, to that of image acquisition by using physical calibration phantoms. As highlighted in several studies, acquisition protocols play a crucial role in the reproducibility of given features [66–69], to the detriment of model generalisability. Identifying and promoting a *radiomics protocol* for each modality would be particularly important for prospective clinical studies in this topic [154]. This can be facilitated by developing new physical phantoms tailored to specifically emphasise texture patterns [154]. Scanners could be calibrated such that analysis of these phantoms produce a set of expected reference values. These would be used in combination with the software-based digital phantom calibrations developed as part of this work. Outside of clinical modelling, radiomic metrics could also play a role in assessing scan quality in this regard. Furthermore, as radiomics features change with segmentation, this could be turned on its head, and the features could become a quality metric for segmentations as well. For example, automatic rejection of segmentations with sharp corners (as likely biologically unnatural tumour shapes) based on morphology features.

6.5 Final Conclusion

The main aims of this project were to: (1) produce benchmarks and recommendations for radiomic feature extraction that will support clinical adoption of radiomics techniques, (2) develop analysis methods that enable evaluation of multiple radiomics software to reach standardised benchmarks through consensus, and (3), identify features that are robust to prominent image processing steps to facilitate more optimal and generalisable radiomics modelling. The developed SPAARC radiomics package was utilised in participation with an internationally reconsigned initiative to determine a set of consensus-base benchmarks for over 160 baseline features and a variety of image processing configurations through exhaustive implementation. The author's work directly led to the clarification of key standardisation recommendations, such as the interpolation grid generation, that enabled these comprehensive benchmarks to be produced. Furthermore, this work specifically explored the role of image interpolation on standardised features in PET imaging and identified many that were stable or potentially systematically varying. The author found that extraction at isotropically decreasing voxel sizes for many features resulted in large amounts of redundancy. Comparison of two standardised interpolation methods found that feature stability assessment remained consistent despite changes to the numeric value of the extracted features. Finally, the author developed analysis methods to evaluate the application of filters to medical imaging in radiomics software in continuation of the consensus-based standardisation effort of the IBSI. This analysis method was used to identify significant discrepancy between software for a number of filter tests, though at least a moderate majority consensus was found in 18/25 filter tests discussed in this work. Efforts are ongoing to improve these initial results to provide additional filter-based benchmarks for the wider research community.

Bibliography

[1] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. Parkin, M. Piñeros, A. Znaor, and F. Bray, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International Journal of Cancer*, vol. 144, p. ijc.31937, dec 2018.

[2] A. Depeursinge, O. Al-Kadi, and J. Mitchell, *Biomedical Texture Analysis*. London, United Kingdom: Academic Press, 1st ed., oct 1 2017.

[3] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, pp. 563–577, feb 2016.

[4] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, pp. 441–446, mar 2012.

[5] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, pp. 1234–1248, nov 2012.

[6] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2, pp. 115–152, 2008.

[7] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. de Jong, J. van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, pp. 749–762, dec 2017.

[8] W. Phillip Law and K. A. Miles, "Incorporating prognostic imaging biomarkers into clinical practice," *Cancer Imaging*, vol. 13, no. 3, pp. 332–341, 2013.

[9] S. S. F. Yip and H. J. W. L. Aerts, "Applications and limitations of radiomics," *Physics in Medicine and Biology*, vol. 61, pp. R150–R166, jul 2016.

[10] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis, "Characterization of PET/CT images using texture analysis: the past, the present… any future?," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, pp. 151–165, jan 2017.

[11] M. Scrivener, E. E. C. de Jong, J. E. van Timmeren, T. Pieters, B. Ghaye, and X. Geets, "Radiomics applied to lung cancer: a review," *Translational Cancer Research*, vol. 5, pp. 398–409, aug 2016.

[12] A. Chaddad, M. J. Kucharczyk, P. Daniel, S. Sabri, B. J. Jean-Claude, T. Niazi, and B. Abdulkarim, "Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation," *Frontiers in Oncology*, vol. 9, pp. 1–9, may 2019.

[13] A. Jethanandani, T. A. Lin, S. Volpe, H. Elhalawani, A. S. Mohamed, P. Yang, and C. D. Fuller, "Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: A systematic review," *Frontiers in Oncology*, vol. 8, no. MAY, 2018.

[14] F. Valdora, N. Houssami, F. Rossi, M. Calabrese, and A. S. Tagliafico, "Rapid review: radiomics and breast cancer," *Breast Cancer Research and Treatment*, vol. 169, pp. 217–229, jun 2018.

[15] P. S. van Rossum, C. Xu, D. V. Fried, L. Goense, L. E. Court, and S. H. Lin, "The emerging field of radiomics in esophageal cancer: current evidence and future potential," *Translational Cancer Research*, vol. 5, no. 4, pp. 410–423, 2016.

[16] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, dec 2014.

[17] K. G. Foley, R. K. Hills, B. Berthon, C. Marshall, C. Parkinson, W. G. Lewis, T. D. Crosby, E. Spezi, and S. A. Roberts, "Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer," *European Radiology*, vol. 28, pp. 428–436, jan 2018.

[18] "What is cancer?." National Cancer Insitute, February 9, 2015, Avaliable online: `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`, [Accessed: 20/08/2020].

[19] R. Hesketh, "What is a tumour?," in *Introduction to Cancer Biology*, ch. 5, pp. 102–149, Cambridge University Press, 2013.

[20] H. Dillekås, M. S. Rogers, and O. Straume, "Are 90% of deaths from cancer caused by metastases?," *Cancer Medicine*, vol. 8, pp. 5574–5576, sep 2019.

[21] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[22] S. B. Edge and C. C. Compton, "The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM," *Annals of Surgical Oncology*, vol. 17, pp. 1471–1474, jun 2010.

[23] "Overview: radiotherapy." National Health Service, March 29, 2017, Avaliable online: `https://www.nhs.uk/conditions/radiotherapy/`, [Accessed: 25/08/2019].

[24] T. M. Pawlik and K. Keyomarsi, "Role of cell cycle in mediating sensitivity to radiotherapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 59, pp. 928–942, jul 2004.

[25] S. Webb, "The physical basis of IMRT and inverse planning," *The British Journal of Radiology*, vol. 76, pp. 678–689, oct 2003.

[26] J. Bhosle and G. Hall, "Principles of cancer treatment by chemotherapy," *Surgery (Oxford)*, vol. 27, pp. 173–177, apr 2009.

[27] "Hormone therapy for cancer." Cancer Research UK, August 13, 2019, Avaliable online: `https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/hormone-therapy`, [Accessed: 09/09/2019].

[28] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature Reviews Clinical Oncology*, vol. 15, pp. 81–94, feb 2018.

[29] "Improving outcomes through personalised medicine." National Health Service England, September 7, 2016, Avaliable online: `https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf`, [Accessed: 09/09/2019].

[30] "Precision medicine in cancer treatment." National Cancer Insitute, October 3, 2017, Avaliable online: `https://www.cancer.gov/about-cancer/treatment/types/precision-medicine`, [Accessed: 09/09/2019].

[31] J. J. Berman, "Chapter 1 - introduction: Seriously, what is precision medicine?," in *Precision Medicine and the Reinvention of Human Disease* (J. J. Berman, ed.), pp. 1 – 15, Academic Press, 2018.

[32] K.-H. Yu and M. Snyder, "Omics Profiling in Precision Oncology," *Molecular & Cellular Proteomics*, vol. 15, pp. 2525–2536, aug 2016.

[33] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi,

J. Downward, P. A. Futreal, and C. Swanton, “Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing,” *New England Journal of Medicine*, vol. 366, pp. 883–892, mar 2012.

[34] M. Gerlinger and C. Swanton, “How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine,” *British Journal of Cancer*, vol. 103, pp. 1139–1143, oct 2010.

[35] M. Jamal-Hanjani, G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. Watkins, S. Veeriah, S. Shafi, D. H. Johnson, R. Mitter, R. Rosenthal, M. Salm, S. Horswell, M. Escudero, N. Matthews, A. Rowan, T. Chambers, D. A. Moore, S. Turajlic, H. Xu, S.-M. Lee, M. D. Forster, T. Ahmad, C. T. Hiley, C. Abbosh, M. Falzon, E. Borg, T. Marafioti, D. Lawrence, M. Hayward, S. Kolvekar, N. Panagiotopoulos, S. M. Janes, R. Thakrar, A. Ahmed, F. Blackhall, Y. Summers, R. Shah, L. Joseph, A. M. Quinn, P. A. Crosbie, B. Naidu, G. Middleton, G. Langman, S. Trotter, M. Nicolson, H. Remmen, K. Kerr, M. Chetty, L. Gomersall, D. A. Fennell, A. Nakas, S. Rathinam, G. Anand, S. Khan, P. Russell, V. Ezhil, B. Ismail, M. Irvin-Sellers, V. Prakash, J. F. Lester, M. Kornaszewska, R. Attanoos, H. Adams, H. Davies, S. Dentro, P. Taniere, B. O’Sullivan, H. L. Lowe, J. A. Hartley, N. Iles, H. Bell, Y. Ngai, J. A. Shaw, J. Herrero, Z. Szallasi, R. F. Schwarz, A. Stewart, S. A. Quezada, J. Le Quesne, P. Van Loo, C. Dive, A. Hackshaw, and C. Swanton, “Tracking the Evolution of Non–Small-Cell Lung Cancer,” *New England Journal of Medicine*, vol. 376, no. 22, pp. 2109–2121, 2017.

[36] A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. L. Zeger, “Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework,” *Clinical Pharmacology and Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[37] D. R. Hodgson, R. D. Whittaker, A. Herath, D. Amakye, and G. Clack, “Biomarkers in oncology drug development,” *Molecular Oncology*, vol. 3, pp. 24–32, feb 2009.

[38] J. A. Wagner, “Biomarkers: Principles, policies, and practice,” *Clinical Pharmacology and Therapeutics*, vol. 86, no. 1, pp. 3–7, 2009.

[39] A. Zwanenburg and S. Löck, “Why validation of prognostic models matters?,” *Radiotherapy and Oncology*, vol. 127, pp. 370–373, jun 2018.

[40] I. El Naqa, S. L. Kerns, J. Coates, Y. Luo, C. Speers, C. M. L. West, B. S. Rosenstein, and R. K. Ten Haken, “Radiogenomics and radiotherapy response modeling,” *Physics in Medicine & Biology*, vol. 62, pp. R179–R206, aug 2017.

[41] T. E. Yankeelov, R. G. Abramson, and C. C. Quarles, “Quantitative multimodality imaging in cancer research and therapy,” *Nature Reviews Clinical Oncology*, vol. 11, pp. 670–680, nov 2014.

[42] T. Beyer, D. W. Townsend, T. Brun, P. E. Kinahan, M. Charron, R. Roddy, J. Jerin, J. Young, L. Byars, and R. Nutt, “A combined PET/CT scanner for clinical oncology.,” *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 41, pp. 1369–79, aug 2000.

[43] “Diagnostic imaging dataset statistical release.” National Health Service, *NHS England*, Avaliable online: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2019/07/Provisional-Monthly-Diagnostic-Imaging-Dataset-Statistics-2019-07-18.pdf, [Accessed: 10/06/2020].

[44] W. Birkfellner, “CT Reconstruction,” in *Applied Medical Image Processing: A Basic Course*, ch. 10, pp. 340–341, CRC Press, 2nd ed., 2014.

[45] P. Allisy-Roberts and J. Williams, “Computed tomography,” in *Farr’s Physics for Medical Imaging*, ch. 7, pp. 103–119, Saunders, Elsevier, 2nd ed., 2008.

[46] L. W. Goldman, “Principles of CT and CT technology,” *Journal of Nuclear Medicine Technology*, vol. 35, no. 3, pp. 115–128, 2007.

[47] W. Birkfellner, “A Few Basics of Medical Image Sources,” in *Applied Medical Image Processing: A Basic Course*, ch. 1, pp. 1–43, CRC Press, 2nd ed., 2014.

[48] P. Allisy-Roberts and J. Williams, “Gamma imaging,” in *Farr’s Physics for Medical Imaging*, ch. 7, pp. 121–144, Saunders, Elsevier, 2nd ed., 2008.

[49] P. E. Kinahan and J. W. Fletcher, “Positron Emission Tomography-Computed Tomography Standardized Uptake Values in Clinical Practice and Assessing Response to Therapy,” *Seminars in Ultrasound, CT and MRI*, vol. 31, pp. 496–505, dec 2010.

[50] P. Allisy-Roberts and J. Williams, “Magentic resonance imaging,” in *Farr’s Physics for Medical Imaging*, ch. 10, pp. 169–195, Saunders, Elsevier, 2nd ed., 2008.

[51] M. L. Lipton, *Totally Accessible MRI*. New York, NY: Springer New York, 2008.

[52] “Matlab programming language.” MATLAB, *Mathworks*, Avaliable online: https://uk.mathworks.com, [Accessed: 10/01/2020].

[53] M. R. P. Donald W. McRobbie, Elizabeth A. Moore, Martin J. Graves, *MRI from Picture to Proton*. Cambridge University Press, 3rd ed., 2017.

[54] S. Currie, N. Hoggard, I. J. Craven, M. Hadjivassiliou, and I. D. Wilkinson, “Understanding MRI: basic MR physics for physicians,” *Postgraduate Medical Journal*, vol. 89, pp. 209–223, apr 2013.

[55] “Scopus document search.” *Elsevier*, Avaliable online: https://www.scopus.com, [Accessed: 20/02/2020].

[56] R. Haralick, K. Shanmugan, and I. Dinstein, “Textural features for image classification,” 1973.

[57] M. M. Galloway, “Texture analysis using gray level run lengths,” *Computer Graphics and Image Processing*, vol. 4, pp. 172–179, jun 1975.

[58] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, “Shape and Texture Indexes Application to Cell Nuclei Classification,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, p. 1357002, feb 2013.

[59] G. Thibault, J. Angulo, and F. Meyer, “Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification,” *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 630–637, mar 2014.

[60] M. Amadasun and R. King, “Texural Features Corresponding to Texural Properties,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, 1989.

[61] C. Sun and W. G. Wee, “Neighboring gray level dependence matrix for texture classification,” *Computer Vision, Graphics, and Image Processing*, vol. 23, pp. 341–352, sep 1983.

[62] M. L. Giger, H.-P. Chan, and J. Boone, “Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM,” *Medical Physics*, vol. 35, pp. 5799–5820, nov 2008.

[63] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, “Machine Learning methods for Quantitative Radiomic Biomarkers,” *Scientific Reports*, vol. 5, p. 13087, oct 2015.

[64] M. Hatt, C. C. Le Rest, F. Tixier, B. Badic, U. Schick, and D. Visvikis, “Radiomics: Data Are Also Images,” *Journal of Nuclear Medicine*, vol. 60, pp. 38S–44S, sep 2019.

[65] L. Beaton, S. Bandula, M. N. Gaze, and R. A. Sharma, “How rapid advances in imaging are defining the future of precision radiation oncology,” *British Journal of Cancer*, vol. 120, pp. 779–790, apr 2019.

[66] Y. Balagurunathan, V. Kumar, Y. Gu, J. Kim, H. Wang, Y. Liu, D. B. Goldgof, L. O. Hall, R. Korn, B. Zhao, L. H. Schwartz, S. Basu, S. Eschrich, R. A. Gatenby, and R. J. Gillies, “Test–Retest Reproducibility Analysis of Lung CT Image Features,” *Journal of Digital Imaging*, vol. 27, pp. 805–823, dec 2014.

[67] J. E. van Timmeren, R. T. H. Leijenaar, W. J. C. van Elmpt, J. Wang, Z. Zhang, A. Dekker, and P. Lambin, “Test-retest data for radiomics feature stability analysis: generalizable or study specific?,” *Tomography*, vol. 2, no. 4, pp. 361–365, 2016.

[68] R. T. H. Leijenaar, S. Carvalho, E. R. Velazquez, W. J. C. van Elmpt, C. Parmar, O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, A. L. A. J. Dekker, R. J. Gillies, H. J. W. L. Aerts, and P. Lambin, “Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability,” *Acta Oncologica*, vol. 52, pp. 1391–1397, oct 2013.

[69] R. Berenguer, M. d. R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. Mansilla Legorburo, and S. Sabater, "Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters," *Radiology*, vol. 288, pp. 407–415, aug 2018.

[70] C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. U. Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin, and H. J. W. L. Aerts, "Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation," *PLoS ONE*, vol. 9, p. e102107, jul 2014.

[71] G. Doumou, M. Siddique, C. Tsoumpas, V. Goh, and G. J. Cook, "The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer," *European Radiology*, vol. 25, no. 9, pp. 2805–2812, 2015.

[72] F. H. van Velden, G. M. Kramer, V. Frings, I. A. Nissen, E. R. Mulder, A. J. de Langen, O. S. Hoekstra, E. F. Smit, and R. Boellaard, "Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation," *Molecular Imaging and Biology*, vol. 18, no. 5, pp. 788–795, 2016.

[73] M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky, M. Friess, P. Veit-Haibach, M. Huellner, I. Opitz, W. Weder, T. Frauenfelder, M. Guckenberger, and S. Tanadini-Lang, "Influence of inter-observer delineation variability on radiomics stability in different tumor sites," *Acta Oncologica*, vol. 57, pp. 1070–1074, aug 2018.

[74] M. Hatt, J. A. Lee, C. R. Schmidtlein, I. El Naqa, C. Caldwell, E. De Bernardi, W. Lu, S. Das, X. Geets, V. Gregoire, R. Jeraj, M. P. MacManus, O. R. Mawlawi, U. Nestle, A. B. Pugachev, H. Schöder, T. Shepherd, E. Spezi, D. Visvikis, H. Zaidi, and A. S. Kirov, "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211," *Medical Physics*, vol. 44, no. 6, pp. e1–e42, 2017.

[75] B. Berthon, C. Marshall, M. Evans, and E. Spezi, "ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography," *Physics in Medicine and Biology*, vol. 61, no. 13, pp. 4855–4869, 2016.

[76] C. Parkinson, K. Foley, P. Whybra, R. Hills, A. Roberts, C. Marshall, J. Staffurth, and E. Spezi, "Evaluation of prognostic models developed using standardised image features from different PET automated segmentation methods," *EJNMMI Research*, vol. 8, p. 29, apr 2018.

[77] M. Vallières, D. Visvikis, and M. Hatt, "Dependency of a validated radiomics signature on tumor volume and potential corrections," *The Journal of Nuclear Medicine*, vol. 59, pp. 640–640, 2018.

[78] R. A. Bundschuh, J. Dinges, L. Neumann, M. Seyfried, N. Zsótér, L. Papp, R. Rosenberg, K. Becker, S. T. Astner, M. Henninger, K. Herrmann, S. I. Ziegler, M. Schwaiger, and M. Essler, "Textural parameters of tumor heterogeneity in18F-FDG PET/CT for therapy response assessment and prognosis in patients with locally advanced rectal cancer," *Journal of Nuclear Medicine*, vol. 55, no. 6, pp. 891–897, 2014.

[79] J.-h. Kim, E. S. Ko, Y. Lim, K. S. Lee, B.-k. Han, E. Y. Ko, S. Y. Hahn, and S. J. Nam, "Breast Cancer Heterogeneity: MR Imaging Texture Analysis and Survival Outcomes," *Radiology*, vol. 282, pp. 665–675, mar 2017.

[80] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," *International Journal of Radiation Oncology Biology Physics*, vol. 102, no. 4, pp. 1143–1158, 2018.

[81] L. Lu, W. Lv, J. Jiang, J. Ma, Q. Feng, A. Rahmim, and W. Chen, "Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization," *Molecular Imaging and Biology*, vol. 18, no. 6, pp. 935–945, 2016.

[82] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, "Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters," *Acta Oncologica*, vol. 49, pp. 1012–1016, oct 2010.

[83] J. Yan, J. L. Chu-Shern, H. Y. Loi, L. K. Khor, A. K. Sinha, S. T. Quek, I. W. K. Tham, and D. Townsend, "Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET," *Journal of Nuclear Medicine*, vol. 56, pp. 1667–1673, nov 2015.

[84] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. Leijenaar, J. Lenkowicz,

F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. Pfaehler, A. Rahmim, A. U. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, E. Spezi, R. J. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E. G. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. van Velden, P. Whybra, C. Richter, and S. Löck, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, p. 191145, mar 2020.

[85] P. Whybra, C. Parkinson, K. Foley, J. Staffurth, and E. Spezi, "Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[86] A. Depeursinge, V. Andrearczyk, P. Whybra, J. van Griethuysen, H. Müller, R. Schaer, M. Vallières, and A. Zwanenburg, "Standardised convolutional filtering for radiomics," jun 2020.

[87] K. G. Foley, Z. Shi, P. Whybra, P. Kalendralis, R. Larue, M. Berbee, M. N. Sosef, C. Parkinson, J. Staffurth, T. D. Crosby, S. A. Roberts, A. Dekker, L. Wee, and E. Spezi, "External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer," *Radiotherapy and Oncology*, vol. 133, pp. 205–212, apr 2019.

[88] C. Piazzese, K. Foley, P. Whybra, C. Hurt, T. Crosby, and E. Spezi, "Discovery of stable and prognostic CT-based radiomic features independent of contrast administration and dimensionality in oesophageal cancer," *PLoS ONE*, vol. 14, no. 11, pp. 1–13, 2019.

[89] M. Mori, P. Passoni, E. Incerti, V. Bettinardi, S. Broggi, M. Reni, P. Whybra, E. Spezi, E. G. Vanoli, L. Gianolli, M. Picchio, N. G. Di Muzio, and C. Fiorino, "Training and validation of a robust PET radiomic-based index to predict distant-relapse-free-survival after radio-chemotherapy for locally advanced pancreatic cancer," *Radiotherapy and Oncology*, vol. 153, pp. 258–264, dec 2020.

[90] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv:1612.07003.*, dec 2016.

[91] "Cerr github repository." Avaliable online: https://github.com/cerr/CERR, [Accessed: 09/09/2019].

[92] "Digital imaging and commuications in medicine." The DICOM Standard, Avaliable online: https://www.dicomstandard.org/current/, [Accessed: 10/09/2019].

[93] "Nifti data format." Avaliable online: https://nifti.nimh.nih.gov/nifti-1/, [Accessed: 03/10/2020].

[94] C. Mayo id Fuller Ellen D. Yorke Jatinder R. Palta Peter, J. Moran, W. Bosch, Y. Xiao, T. McNutt, R. Popple, J. Michalski, M. Feng, L. Marks, C. D. Fuller, E. Yorke, J. Palta, P. Gabriel, A. Molineu, M. Matuszak, E. Covington, K. Masi, S. Richardson, T. Ritter, T. Morgas, S. Flampouri, L. Santanam, J. Moore, T. Purdie, R. C. Miller, C. Hurkmans, J. Adams, Q.-R. J. Wu, C. Fox, R. A. Siochi, N. L. Brown, W. Verbakel, Y. Archambault, S. Chmura, D. Eagle, T. Fitzgerald, A. Dekker, T. Hong, R. Kapoor, B. Lansing, S. Jolly, M. Napolitano, J. Percy, M. Rose, S. Siddiqui, C. Schadt, W. Simon, W. Straube, S. St. James, K. Ulin, S. Yom, and T. Yock, "Standardizing Nomenclatures in Radiation Oncology," tech. rep., American Association of Physicists in Medicine, jan 2018.

[95] R. T. Larue, L. Van De Voorde, J. E. van Timmeren, R. T. Leijenaar, M. Berbée, M. N. Sosef, W. M. Schreurs, W. van Elmpt, and P. Lambin, "4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers," *Radiotherapy and Oncology*, vol. 125, no. 1, pp. 147–153, 2017.

[96] A. Depeursinge, A. Foncubierta-Rodriguez, D. Van De Ville, and H. Müller, "Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities," *Medical Image Analysis*, vol. 18, pp. 176–196, jan 2014.

[97] "Matlab documentation: interp3." MATLAB, *Mathworks*, Avaliable online: https://uk.mathworks.com/help/matlab/ref/interp3.html#bt2rbzl-1-method, [Accessed: 03/01/2020].

[98] R. G. McClarren, "Interpolation," in *Computational Nuclear Engineering and Radiological Science Using Python*, ch. 10, pp. 173–192, Elsevier, 2018.

[99] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, and K. Miles, "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: Preliminary evidence of an association with tumour metabolism, stage, and survival," *Clinical Radiology*, vol. 67, no. 2, pp. 157–164, 2012.

[100] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine and Biology*, vol. 60, pp. 5471–5496, jul 2015.

[101] "Matlab: Infinity and nan." MATLAB, *Mathworks*, Avaliable online: `https://www.mathworks.com/help/matlab/matlab_prog/infinity-and-nan.html`, [Accessed: 15/01/2020].

[102] R. T. Leijenaar, G. Nalbantov, S. Carvalho, W. J. van Elmpt, E. G. Troost, R. Boellaard, H. J. Aerts, R. J. Gillies, and P. Lambin, "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis," *Scientific Reports*, vol. 5, p. 11075, sep 2015.

[103] "Matlab file exchange: Marching cubes." MATLAB, *Mathworks*, Avaliable online: `https://www.mathworks.com/matlabcentral/fileexchange/32506-marching-cubes`, [Accessed: 03/03/2020].

[104] "Matlab documentation: isosurface." MATLAB, *Mathworks*, Avaliable online: `https://www.mathworks.com/help/matlab/ref/isosurface.html`, [Accessed: 03/03/2020].

[105] M. Hatt, F. Tixier, C. Cheze Le Rest, O. Pradier, and D. Visvikis, "Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 40, no. 11, pp. 1662–1671, 2013.

[106] H. Y. Wang, E. M. Donovan, A. Nisbet, C. P. South, S. Alobaidli, V. Ezhil, I. Phillips, V. Prakash, M. Ferreira, P. Webster, and P. M. Evans, "The stability of imaging biomarkers in radiomics: A framework for evaluation," *Physics in Medicine and Biology*, vol. 64, no. 16, 2019.

[107] D. Giavarina, "Understanding Bland Altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[108] M.-C. Desseroit, F. Tixier, W. A. Weber, B. A. Siegel, C. Cheze Le Rest, D. Visvikis, and M. Hatt, "Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non–Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort," *Journal of Nuclear Medicine*, vol. 58, no. 3, pp. 406–411, 2017.

[109] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, pp. 155–163, jun 2016.

[110] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients.," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[111] M. Bogowicz, R. T. Leijenaar, S. Tanadini-Lang, O. Riesterer, M. Pruschy, G. Studer, J. Unkelbach, M. Guckenberger, E. Konukoglu, and P. Lambin, "Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models," *Radiotherapy and Oncology*, vol. 125, no. 3, pp. 385–391, 2017.

[112] "Data sets used by the ibsi for benchmarking and standardisation." *GitHub*, Avaliable online: `https://github.com/theibsi/data_sets`, [Accessed: 20/03/2020].

[113] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E. G. Troost, C. Richter, and S. Löck, "Assessing robustness of radiomic features by image perturbation," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[114] M. Shafiq-ul Hassan, G. G. Zhang, K. Latifi, G. Ullah, D. C. Hunt, Y. Balagurunathan, M. A. Abdalah, M. B. Schabath, D. G. Goldgof, D. Mackin, L. E. Court, R. J. Gillies, and E. G. Moros, "Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels," *Medical Physics*, vol. 44, pp. 1050–1062, mar 2017.

[115] M. Shafiq-Ul-Hassan, K. Latifi, G. Zhang, G. Ullah, R. Gillies, and E. Moros, "Voxel size and gray level normalization of CT radiomic features in lung cancer," *Scientific Reports*, vol. 8, no. 1, pp. 1–9, 2018.

[116] D. Mackin, X. Fave, L. Zhang, J. Yang, A. K. Jones, C. S. Ng, and L. Court, "Harmonizing the pixel size in retrospective computed tomography radiomics studies," *PLoS ONE*, vol. 12, no. 9, pp. 1–17, 2017.

[117] S. S. Yip, C. Parmar, J. Kim, E. Huynh, R. H. Mak, and H. J. Aerts, "Impact of experimental design on PET radiomics in predicting somatic mutation status," *European Journal of Radiology*, vol. 97, pp. 8–15, dec 2017.

[118] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–14, 2017.

[119] "R: A language and environment for statistical computing." Avaliable online: https://www.r-project.org.

[120] R. B. Ger, S. Zhou, P.-C. M. Chi, H. J. Lee, R. R. Layman, A. K. Jones, D. L. Goff, C. D. Fuller, R. M. Howell, H. Li, R. J. Stafford, L. E. Court, and D. S. Mackin, "Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies," *Scientific Reports*, vol. 8, p. 13047, dec 2018.

[121] S. Brunton, *Data-driven science and engineering : machine learning, dynamical systems, and control*. Cambridge, United Kingdom New York, NY: Cambridge University Press, 2019.

[122] R. C. Gonzalez, *Digital Image Processing 4th Edition*. Pearson, 2018.

[123] K. I. Laws, "Rapid Texture Identification," in *Proceedings SPIE* (T. F. Wiener, ed.), vol. D, pp. 376–381, dec 1980.

[124] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, "Texture analysis of medical images," *Clinical radiology*, vol. 59, no. 12, pp. 1061–9, 2004.

[125] "Image biomarker standardisation intiative datasets." Avaliable online: https://theibsi.github.io/datasets/, [Accessed: 30/09/2020].

[126] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, pp. 641–642, jul 2017.

[127] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20150202, apr 2016.

[128] "Matlab documentation: Pca." MATLAB, *Mathworks*, Avaliable online: https://www.mathworks.com/help/stats/pca.html, [Accessed: 01/09/2020].

[129] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," *Medical Physics*, vol. 25, pp. 656–661, may 1998.

[130] A. Eskicioglu and P. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, vol. 43, no. 12, pp. 2959–2965, 1995.

[131] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, apr 2004.

[132] "Matlab documentation: Strctural similarity index (ssim)." MATLAB, *Mathworks*, Avaliable online: https://uk.mathworks.com/help/images/ref/ssim.html, [Accessed: 03/10/2020].

[133] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, nov 2017.

[134] L. Zhang, D. V. Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court, "<scp>ibex</scp> : An open infrastructure software platform to facilitate collaborative work in radiomics," *Medical Physics*, vol. 42, pp. 1341–1353, feb 2015.

[135] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, "LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity," *Cancer Research*, vol. 78, pp. 4786–4789, aug 2018.

[136] Y. Dicente Cid, J. Castelli, R. Schaer, N. Scher, A. Pomoni, J. O. Prior, and A. Depeursinge, "QuantImage: An Online Tool for High-Throughput 3D Radiomics Feature Extraction in PET-CT," in *Biomedical Texture Analysis*, pp. 349–377, Elsevier, 1 ed., 2017.

[137] Y.-H. D. Fang, C.-Y. Lin, M.-J. Shih, H.-M. Wang, T.-Y. Ho, C.-T. Liao, and T.-C. Yen, "Development and Evaluation of an Open-Source Software Package "CGITA" for Quantifying Tumor Heterogeneity with Molecular Images," *BioMed Research International*, vol. 2014, pp. 1–9, 2014.

[138] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, pp. 118–127, jan 2007.

[139] R. N. Mahon, M. Ghita, G. D. Hugo, and E. Weiss, "ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets," *Physics in Medicine and Biology*, vol. 65, no. 1, 2020.

[140] B. Badic, M. C. Desseroit, M. Hatt, and D. Visvikis, "Potential Complementary Value of Noncontrast and Contrast Enhanced CT Radiomics in Colorectal Cancers," *Academic Radiology*, vol. 26, no. 4, pp. 469–479, 2019.

[141] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *BMJ (Online)*, vol. 350, no. January, pp. 1–9, 2015.

[142] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, may 2015.

[143] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[144] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[145] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities," aug 2018.

[146] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, pp. 231–237, mar 2019.

[147] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, nov 2018.

[148] A. Zwanenburg, "Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis," *European Journal of Nuclear Medicine and Molecular Imaging*, 2019.

[149] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020.

[150] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[151] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Medical Physics*, vol. 44, no. 10, pp. 5162–5171, 2017.

[152] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[153] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into Imaging*, vol. 11, no. 1, 2020.

[154] J. Kazmierska, A. Hope, E. Spezi, S. Beddar, W. H. Nailon, B. Osong, A. Ankolekar, A. Choudhury, A. Dekker, K. R. Redalen, and A. Traverso, "From multisource data to clinical decision aids in radiation oncology: The need for a clinical data science community," *Radiotherapy and Oncology*, no. xxxx, 2020.

[155] P. Kubben, M. Dumontier, and A. Dekker, eds., *Fundamentals of Clinical Data Science*. Cham: Springer International Publishing, 2019.

[156] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer New York, 2009.

[157] "Occam's razor." Avaliable online: https://www.britannica.com/topic/Occams-razor, [Accessed: 20/12/2020].

[158] L. Dercle, T. Henry, A. Carré, N. Paragios, E. Deutsch, and C. Robert, "Reinventing radiation therapy with machine learning and imaging bio-markers (radiomics): State-of-the-art, challenges and perspectives," *Methods*, no. May, pp. 0–1, 2020.

[159] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, A. Sak, M. Stuschke, P. Balermpas, C. Rödel, U. Ganswindt, C. Belka, S. Pigorsch, S. E. Combs, D. Mönnich, D. Zips, M. Krause, M. Baumann, E. G. C. Troost, S. Löck, and C. Richter, "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling," *Scientific Reports*, vol. 7, p. 13206, dec 2017.

[160] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers," *Radiology: Artificial Intelligence*, vol. 2, p. e200029, mar 2020.

[161] J. P. B. O'Connor, E. O. Aboagye, J. E. Adams, H. J. W. L. Aerts, S. F. Barrington, A. J. Beer, R. Boellaard, S. E. Bohndiek, M. Brady, G. Brown, D. L. Buckley, T. L. Chenevert, L. P. Clarke, S. Collette, G. J. Cook, N. M. DeSouza, J. C. Dickson, C. Dive, J. L. Evelhoch, C. Faivre-Finn, F. A. Gallagher, F. J. Gilbert, R. J. Gillies, V. Goh, J. R. Griffiths, A. M. Groves, S. Halligan, A. L. Harris, D. J. Hawkes, O. S. Hoekstra, E. P. Huang, B. F. Hutton, E. F. Jackson, G. C. Jayson, A. Jones, D.-M. Koh, D. Lacombe, P. Lambin, N. Lassau, M. O. Leach, T.-Y. Lee, E. L. Leen, J. S. Lewis, Y. Liu, M. F. Lythgoe, P. Manoharan, R. J. Maxwell, K. A. Miles, B. Morgan, S. Morris, T. Ng, A. R. Padhani, G. J. M. Parker, M. Partridge, A. P. Pathak, A. C. Peet, S. Punwani, A. R. Reynolds, S. P. Robinson, L. K. Shankar, R. A. Sharma, D. Soloviev, S. Stroobants, D. C. Sullivan, S. A. Taylor, P. S. Tofts, G. M. Tozer, M. van Herk, S. Walker-Samuel, J. Wason, K. J. Williams, P. Workman, T. E. Yankeelov, K. M. Brindle, L. M. McShane, A. Jackson, and J. C. Waterton, "Imaging biomarker roadmap for cancer studies," *Nature Reviews Clinical Oncology*, vol. 14, pp. 169–186, mar 2017.

[162] "Radiomics quality score - rqs." Avaliable online: https://www.radiomics.world/rqs, [Accessed: 10/12/2020].

[163] J. E. Park, H. S. Kim, D. Kim, S. Y. Park, J. Y. Kim, S. J. Cho, and J. H. Kim, "A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features," *BMC Cancer*, vol. 20, p. 29, dec 2020.

[164] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, p. 119, dec 2020.

[165] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, dec 2013.

[166] J. Kalpathy-Cramer, J. B. Freymann, J. S. Kirby, P. E. Kinahan, and F. W. Prior, "Quantitative Imaging Network: Data Sharing and Competitive AlgorithmValidation Leveraging The Cancer Imaging Archive," *Translational Oncology*, vol. 7, pp. 147–152, feb 2014.

[167] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, R. C. Petersen, and C. R. Jack, "Identification of Anonymous MRI Research Participants with Face-Recognition Software," *New England Journal of Medicine*, vol. 381, pp. 1684–1686, oct 2019.

[168] A. Jochems, T. M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin, and A. Dekker, "Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept," *Radiotherapy and Oncology*, vol. 121, pp. 459–467, dec 2016.

[169] T. M. Deist, F. J. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, and A. Dekker, "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train," *Radiotherapy and Oncology*, vol. 144, pp. 189–200, mar 2020.

[170] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, dec 2016.

[171] J. E. van Timmeren, W. van Elmpt, R. T. Leijenaar, B. Reymen, R. Monshouwer, J. Bussink, L. Paelinck, E. Bogaert, C. De Wagter, E. Elhaseen, Y. Lievens, O. Hansen, C. Brink, and P. Lambin, "Longitudinal radiomics of cone-beam CT images from non-small cell lung cancer patients: Evaluation of the added prognostic value for overall survival and locoregional recurrence," *Radiotherapy and Oncology*, vol. 136, pp. 78–85, jul 2019.

[172] M. L. Welch, C. McIntosh, B. Haibe-Kains, M. F. Milosevic, L. Wee, A. Dekker, S. H. Huang, T. G. Purdie, B. O'Sullivan, H. J. Aerts, and D. A. Jaffray, "Vulnerabilities of radiomic signature development: The need for safeguards," *Radiotherapy and Oncology*, vol. 130, pp. 2–9, jan 2019.

[173] R. B. Ger, S. Zhou, B. Elgohari, H. Elhalawani, D. M. Mackin, J. G. Meier, C. M. Nguyen, B. M. Anderson, C. Gay, J. Ning, C. D. Fuller, H. Li, R. M. Howell, R. R. Layman, O. Mawlawi, R. J. Stafford, H. Aerts, and L. E. Court, "Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- And PET-imaged head and neck cancer patients," *PLoS ONE*, vol. 14, no. 9, pp. 1–13, 2019.

[174] I. Fornacon-Wood, H. Mistry, C. J. Ackermann, F. Blackhall, A. McPartlin, C. Faivre-Finn, G. J. Price, and J. P. B. O'Connor, "Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform," *European Radiology*, vol. 30, pp. 6241–6250, nov 2020.

Appendix

A

Table A1: List of teams participating in the first IBSI study. Shown are team names used in this document to match the published study and the institution where the corresponding software was developed, alongside the main programming language used. More details for each team can be found in Zwanenburg *et*. *al*. [84]. Note that some institutions had more than one team. These were independent software developed in the same institution. This is discussed in Section 3.4.

Team Name	**Institution**	**Language**
Brest (BCOM)	INSERM Brest	C++
Brest (MaCha)	INSERM Brest	C++
CaPTk	University of Pennsylvania	C++
Cardiff	Cardiff University	Matlab
CERR	Memorial Sloan-Kettering Cancer Center	Matlab
Gemelli	Fondazione Policlinico Universitario Agostino Gemelli	R
KCL	King's College London	Matlab
LIFEx	Universite Paris Saclay	Java
LUMC	Leiden University Medical Center (LUMC), VU University Medical Center	IDL
MAASTRO	Maastricht University Medical Centre	Matlab
McGill	McGill University	Matlab
MIRP	OncoRay–National Center for Radiation Research in Oncology	Python
MITK	German Cancer Research Center (DKFZ)	C++
Moffitt	Moffitt Cancer Center	C++
NKI	Netherlands Cancer Institute (NKI)	C++
Pyradiomics	Netherlands Cancer Institute (NKI), Maastricht University, Dana-Farber Cancer Institute	Python
QIFE	Stanford University	Matlab
QuantImage	University of Applied Sciences Western Switzerland (HES-SO)	Matlab
RaCaT	University Medical Center Groningen (UMCG)	C++
SERA	Johns Hopkins University	Matlab
Tuebingen	University of Tubingen	Python
UCSF	University of California, San Francisco (UCSF)	Python
UMCG (Beukinga)	University Medical Center Groningen (UMCG)	Matlab
UMCG (van Dijk)	University Medical Center Groningen (UMCG)	Matlab
USZ	University of Zurich	Python

B

Morphology Features

Introduced in Section 2.4.1 of main text. These follow the standardised definitions [90].

Table B1: Definitions for 23 Morphology features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Mesh-based volume	From a triangle mesh of the ROI, each face k in the mesh can be defined by 3 coordinates: $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. The signed volume for the tetrahedron V_k formed by each face and the origin is given via $V_k = \frac{\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})}{6}$. If the orientation of all face normals is kept consistently defined (either pointing into the ROI or outward), the mesh volume can be found by summing over all (N_f) volumes: $F_1 = V = \left\lvert \sum_{k=1}^{N_f} V_k \right\rvert$
Voxel counting volume	$F_2 = \sum_{k=1}^{N_v} V_k$ when V_k is the volume of one voxel.
Surface area (mesh)	Each face in the mesh has an area: $A_k = \frac{\lvert \mathbf{ab} \times \mathbf{ac} \rvert}{2}$ Sum over all faces: $F_3 = A = \sum_{k=1}^{N_f} A_k$
Surface to volume ratio	$F_4 = \frac{A}{V}$
Compactness 1	$F_5 = \frac{V}{\pi^{1/2} A^{3/2}}$
Compactness 2	$F_6 = 36\pi \frac{V^2}{A^3}$
Spherical disproportion	$F_7 = \frac{A}{4\pi R^2} = \frac{A}{(36\pi V^2)^{1/3}}$
Sphericity	$F_8 = \frac{(36\pi V^2)^{1/3}}{A}$
Asphericity	$F_9 = \left(\frac{1}{36\pi}\frac{A^3}{V^2}\right)^{1/3} - 1$

Table B1: Morphology features continued.

Name	**Definition**
Centre of mass shift	Using set of coordinate points for morphological mask $\mathbf{X}_c$ with $N_{v,m}$ voxels. The geometric centre of mass is: $\overrightarrow{CoM}_{geom} = \frac{1}{N_{v,m}} \sum_{k=1}^{N_{v,m}} \vec{X}_{c,k}$ The position of each voxel is then weighted by its corresponding intensity. (Note, this is now using the intensity mask, where $\mathbf{X}_{c,gl}$ are the coordinate points for intensity mask with intensities $\mathbf{X}_{gl}$. Number of voxels in intensity mask is $N_{v,gl}$) $\overrightarrow{CoM}_{gl} = \frac{\sum_{k=1}^{N_{v,gl}} X_{gl,k} \vec{X}_{c,gl,k}}{\sum_{k=1}^{N_{v,gl}} X_{gl,k}}$ the shift is the distance between these two measures: $F_{10} = \Vert\overrightarrow{CoM}_{geom} - \overrightarrow{CoM}_{gl}\Vert_2$
Maximum 3D diameter	The largest pairwise distance between all of the surface mesh vertices.
Major axis length	The largest axial length of an ellipsoid enclosing the ROI. This can be found with PCA on the physical coordinates of the voxel centres that define the ROI. There are three eigenvalues corresponding to the major, minor and leas axes ($\lambda_{major} \geq \lambda_{minor} \geq \lambda_{least}$). $F_{12} = 2a = 4\sqrt{\lambda_{major}}$
Minor axis length	$F_{13} = 2b = 4\sqrt{\lambda_{minor}}$
Least axis length	$F_{14} = 2c = 4\sqrt{\lambda_{least}}$
Elongation	$F_{15} = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$
Flatness	$F_{16} = \sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$
Volume density (axis-aligned bounding box)	Compare the ROI volume to the volume of the smallest axis-aligned bounding box ($V_a abb$) that encompasses it. $F_{17} = \frac{V}{V_{aabb}}$
Area density (axis-aligned bounding box)	Compare the ROI are to the are of the smallest axis-aligned bounding box ($A_a abb$) that encompasses it. $F_{18} = \frac{A}{A_{aabb}}$
Volume density (approximate enclosing ellipsoid)	$F_{19} = \frac{V}{V_{aee}}$ where V_{aee} is the volume of the ellipsoid enclosing the ROI (V_{aee}) (as described above).

Table B1: Morphology features continued.

Name	**Definition**
Area density (approximate enclosing ellipsoid)	$F_{20} = \frac{A}{A_{aee}}$ where A_{aee} is the approximated surface area of the ellipsoid enclosing the ROI (A_{aee}) (see [90]).
Volume density (convex hull)	Compare the volume of the convex hull that encloses the ROI mesh (V_{con} to the ROI volume. $F_{21} = \frac{V}{V_{convex}}$
Area density (convex hull)	Compare the area of the convex hull that encloses the ROI mesh (A_{con} to the ROI area. $F_{22} = \frac{A}{A_{convex}}$
Integrated intensity	$F_{23} = V \; \frac{1}{N_{v,gl}} \sum_{k=1}^{N_{v,gl}} X_{gl,k}$

Intensity-based Statistical Features

Introduced in Section 2.4.2 of main text. Following the standardised definitions [90], for these features the set of intensities inside the ROI intensity mask is defined as $\mathbf{X}_{gl} = \{X_{gl,1}, X_{gl,2}, \ldots, X_{gl,N_v}\}$, where N_v is the total number of voxels being assessed.

Table B2: Definitions for 18 intensity-based statistics features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	**Definition**
Mean intensity	$F_1 = \frac{1}{N_v} \sum_{k=1}^{N_v} X_{gl,k}$
Intensity variance	$F_2 = \frac{1}{N_v} \sum_{k=1}^{N_v} \left(X_{gl,k} - \mu\right)^2$ where μ is the mean intensity feature.
Intensity skewness	$F_3 = \frac{\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^3}{\left(\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2\right)^{3/2}}$
Intensity kurtosis	$F_4 = \frac{\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^4}{\left(\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2\right)^{2}} - 3$
Median intensity	$F_5 = \text{midpoint}(\mathbf{X}_{gl})$
Minimum intensity	$F6 = \min(\mathbf{X}_{gl})$
10th intensity percentile	10th percentile of $\mathbf{X}_{gl}$
90th intensity percentile	90th percentile of $\mathbf{X}_{gl}$

Table B2: Intensity-based statistics features continued.

Name	Definition
Maximum intensity	$F_9 = \max(\mathbf{X}_{gl})$
Interquartile range	$F_{10} = P_{75} - P_{25}$ where P_x is the x^{th} percentile of $\mathbf{X}_{gl}$.
Intensity range	$F_{11} = \max(\mathbf{X}_{gl}) - \min(\mathbf{X}_{gl})$
Intensity-based Mean absolute deviation	$F_{12} = \frac{1}{N_v}\sum_{k=1}^{N_v} \lvert X_{gl,k} - \mu \rvert$ where μ is the mean intensity.
Intensity-based robust mean absolute deviation	This is the same as above, but on a subset of X_{gl} that are in-between (or equal to) the 10th and 90th percentile.
Intensity-based median absolute deviation	$F_{14} = \frac{1}{N_v}\sum_{k=1}^{N_v} \lvert X_{gl,k} - M \rvert$ where M is the median intensity.
Intensity-based coefficient of variation	$F_{15} = \frac{\sigma}{\mu}$ where μ is the mean intensity and σ is the square root of the variance (standard deviation).
Intensity-based quartile coefficient of dispersion	$F_{16} = \frac{P_{75} - P_{25}}{P_{75} + P_{25}}$
Intensity-based energy	$F_{17} = \sum_{k=1}^{N_v} X_{gl,k}^2$
Root mean square intensity	$F_{18} = \sqrt{\frac{\sum_{k=1}^{N_v} X_{gl,k}^2}{N_v}}$

Intensity Histogram Features

Introduced in Section 2.4.3 of main text. Following the standardised definitions [90], for these features the discretised intensities inside the ROI intensity mask are a set $\mathbf{X}_d = \{X_{d,1}, X_{d,2}, \ldots, X_{d,N_v}\}$, where N_v is the total number of voxels in the ROI intensity mask.

Table B3: Definitions for 23 Intensity Histogram features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Mean discretised intensity	$F_1 = \frac{1}{N_v}\sum_{k=1}^{N_v} X_{d,k}$
Discretised intensity variance	$F_2 = \frac{1}{N_v}\sum_{k=1}^{N_v} \left(X_{d,k} - \mu\right)^2$ where μ is the mean discretised intensity

Table B3: Intensity Histogram features continued.

Name	**Definition**
Discretised intensity skewness	$F_3 = \frac{\frac{1}{N_v}\sum_{k=1}^{N_v}(X_{d,k}-\mu)^3}{\left(\frac{1}{N_v}\sum_{k=1}^{N_v}(X_{d,k}-\mu)^2\right)^{3/2}}$
Excess discretised intensity kurtosis	$F_4 = \frac{\frac{1}{N_v}\sum_{k=1}^{N_v}(X_{d,k}-\mu)^4}{\left(\frac{1}{N_v}\sum_{k=1}^{N_v}(X_{d,k}-\mu)^2\right)^{2}} - 3$
Median discretised intensity	$F_5 = \text{midpoint}(\mathbf{X}_d)$
Minimum discretised intensity	$F_6 = \min(\mathbf{X}_d)$
10^{th} discretised intensity percentile	10^{th} percentile of $\mathbf{X}_d$
90^{th} discretised intensity percentile	90^{th} percentile of $\mathbf{X}_d$
Maximum discretised intensity	$F_9 = \max(\mathbf{X}_d)$
Intensity histogram mode	Most common value in $\mathbf{X}_d$. If there is not a unique value, select the one closest to mean. If equidistant values, select the lower value.
Discretised intensity interquartile range	$F_{11} = P_{75} - P_{25}$
Discretised intensity range	$F_{12} = \max(\mathbf{X}_d) - \min(\mathbf{X}_d)$
Intensity histogram mean absolute deviation	$F_{13} = \frac{1}{N_v}\sum_{i=1}^{N_v}\lvert X_{d,i} - \mu\rvert$
Intensity histogram robust mean absolute deviation	This is the same as above, but on a subset of X_d that are in-between (or equal to) the 10th and 90th percentile.
Intensity histogram median absolute deviation	$F_{15} = \frac{1}{N_v}\sum_{k=1}^{N_v}\lvert X_{d,k} - M\rvert$ where M is the median value.
Intensity histogram coefficient of variation	$F_{16} = \frac{\sigma}{\mu}$
Intensity histogram quartile coefficient of dispersion	$F_{17} = \frac{P_{75}-P_{25}}{P_{75}+P_{25}}$
Discretised intensity entropy	$F_{18} = -\sum_{i=1}^{N_g} p_i \log_2 p_i$
Discretised intensity uniformity	$F_{19} = \sum_{i=1}^{N_g} p_i^2$ where $p_i = n_i/N_v$
Maximum histogram gradient	Let $\mathbf{H}'$ be the numerical gradient of $\mathbf{H}$. $F_{20} = \max(\mathbf{H}')$
Maximum histogram gradient intensity	The discretised intensity value corresponding to $\max(\mathbf{H}')$.

Table B3: Intensity Histogram features continued.

Name	Definition
Minimum histogram gradient	$F_{22} = \min(\mathbf{H}')$
Minimum histogram gradient intensity	The discretised intensity value corresponding to $\min(\mathbf{H}')$.

IVH features

Introduced in Section 2.4.4 of main text. Features calculated using the intensity fraction γ and fractional volume v_i described in that section only.

Table B4: Definitions for 7 IVH features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Volume fraction at 10% intensity	The volume fraction v_i that has intensity fraction γ of at least 10% (F_1)
Volume fraction at 90% intensity	The volume fraction v_i that has intensity fraction grey γ of at least 90% (F_2)
Intensity at 10% volume	The minimum intensity present in at most 10 % of the volume v_i. (F_3)
Intensity at 90% volume	The minimum intensity present in at most 90 % of the volume v_i. (F_4)
Volume fraction difference between 10% and 90% intensity	Difference between first two features: F_5 = F_1 -F_2
Intensity difference between 10% and 90% volume	Difference between second two features: F_6 = F_3 -F_4
Area under the IVH curve	Approximated using the trapezoidal rule [90].

GLCM features

Introduced in Section 2.4.6 of main text.

Table B5: Definitions for 25 GLCM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Joint Maximum	$F_1 = \max(p_{ij})$

Table B5: GLCM features continued.

Name	**Definition**
Joint Average	$F_2 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i\, p_{ij}$
Joint Variance	$F_3 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2\, p_{ij}$ Here μ corresponds to the Joint Average.
Joint Entropy	$F_4 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} \log_2(p_{ij})$
Difference Average	$F_5 = \sum_{k=0}^{N_g-1} k\, p_{i-j,k}$
Difference Variance	$F_6 = \sum_{k=0}^{N_g-1} (k-\mu)^2 p_{i-j,k}$ Here μ corresponds to Difference Average.
Difference Entropy	$F_7 = -\sum_{k=0}^{N_g-1} p_{i-j,k} \log_2(p_{i-j,k})$
Sum Average	$F_8 = \sum_{k=2}^{2N_g} k\, p_{i+j,k}$
Sum Variance	$F_9 = \sum_{k=2}^{2N_g} (k-\mu)^2 p_{i+j,k}$ Here μ corresponds to Sum Average.
Sum Entropy	$F_{10} = -\sum_{k=2}^{2N_g} p_{i+j,k} \log_2 p_{i+j,k}$
Angular Second Moment	$F_{11} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij}^2$
Contrast	$F_{12} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-j)^2\, p_{ij}$
Dissimilarity	$F_{13} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \lvert i-j \rvert\, p_{ij}$
Inverse Difference	$F_{14} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1+\lvert i-j \rvert}$
Normalised Inverse Difference	$F_{15} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1+\lvert i-j \rvert/N_g}$
Inverse Difference Moment	$F_{16} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1+(i-j)^2}$
Normalised Inverse Difference moment	$F_{17} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1+(i-j)^2/N_g^2}$
Inverse Variance	$F_{18} = 2\sum_{i=1}^{N_g} \sum_{j>i}^{N_g} \frac{p_{ij}}{(i-j)^2}$
Correlation	$F_{19} = \frac{1}{\sigma_{i.}^2} \left(-\mu_{i.}^2 + \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i\, j\, p_{ij}\right)$ Here, $\mu_{i.} = \sum_{i=1}^{N_g} i\, p_{i.}$ and $\sigma_{i.} = \left(\sum_{i=1}^{N_g} (i-\mu_{i.})^2 p_{i.}\right)^{1/2}$ give respectively the mean and standardised deviation of $p_{i.}$.

Table B5: GLCM features continued.

Name	**Definition**
Autocorrelation	$F_{20} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i\, j\, p_{ij}$
Cluster Tendency	$F_{21} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - 2\mu_{i.})^2\, p_{ij}$
Cluster Shade	$F_{22} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - 2\mu_{i.})^3\, p_{ij}$
Cluster Prominence	$F_{23} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - 2\mu_{i.})^4\, p_{ij}$
Information correlation 1	$F_{24} = \frac{HXY - HXY_1}{HX}$ where: $HXY = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} \log_2 p_{ij}$ $HX = -\sum_{i=1}^{N_g} p_{i.} \log_2 p_{i.}$ $HXY_1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} \log_2 (p_{i.} p_{.j})$
Information correlation 2	$F_{25} = \sqrt{1 - \exp(-2(HXY_2 - HXY))}$ where: HXY same as above. $HXY_2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{i.} p_{.j} \log_2 (p_{i.} p_{.j})$

GLRLM features

Introduced in Section 2.4.7 of main text.

Table B6: Definitions for 16 GLRLM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	**Definition**
Short Runs Emphasis	$F_1 = \frac{1}{N_s} \sum_{j=1}^{N_r} \frac{r_{.j}}{j^2}$ Here $N_s = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} r_{ij}$
Long Runs Emphasis	$F_2 = \frac{1}{N_s} \sum_{j=1}^{N_r} j^2 r_{.j}$
Low Grey Level Run Emphasis	$F_3 = \frac{1}{N_s} \sum_{i=1}^{N_g} \frac{r_{i.}}{i^2}$
High Grey Level Run Emphasis	$F_4 = \frac{1}{N_s} \sum_{i=1}^{N_g} i^2 r_{i.}$
Short Run Low Grey Level Emphasis	$F_5 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{r_{ij}}{i^2 j^2}$
Short Run High Grey Level Emphasis	$F_6 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{i^2 r_{ij}}{j^2}$
Long Run Low Grey Level Emphasis	$F_7 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{j^2 r_{ij}}{i^2}$

Table B6: GLRLM features continued.

Name	**Definition**
Long Run High Grey Level Emphasis	$F_8 = \frac{1}{N_s}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 j^2 r_{ij}$
Grey Level Non-uniformity	$F_9 = \frac{1}{N_s}\sum_{i=1}^{N_g} r_{i.}^2$
Normalised Grey Level Non-uniformity	$F_{10} = \frac{1}{N_s^2}\sum_{i=1}^{N_g} r_{i.}^2$
Run Length Non-uniformity	$F_{11} = \frac{1}{N_s}\sum_{j=1}^{N_r} r_{.j}^2$
Normalised Run Length Non-uniformity	$F_{12} = \frac{1}{N_s^2}\sum_{j=1}^{N_r} r_{.j}^2$
Run Percentage	$F_{13} = \frac{N_s}{N_v}$
Grey Level Variance	$F_{14} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}(i-\mu)^2 p_{ij}$ where $p_{ij} = r_{ij}/N_s$ and $\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i\, p_{ij}$.
Run Length Variance	$F_{15} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}(j-\mu)^2 p_{ij}$ where $\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j\, p_{ij}$.
Run Entropy	$F_{16} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p_{ij}\log_2 p_{ij}$ where $p_{ij} = r_{ij}/N_s$

GLSZM features

Introduced in Section 2.4.8 of main text.

Table B7: Definitions for 16 GLSZM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	**Definition**
Small zone emphasis	$F_1 = \frac{1}{N_s}\sum_{j=1}^{N_z}\frac{s_{.j}}{j^2}$
Large zone emphasis	$F_2 = \frac{1}{N_s}\sum_{j=1}^{N_z} j^2 s_{.j}$
Low grey level zone emphasis	$F_3 = \frac{1}{N_s}\sum_{i=1}^{N_g}\frac{s_{i.}}{i^2}$
High grey level zone emphasis	$F_4 = \frac{1}{N_s}\sum_{i=1}^{N_g} i^2 s_{i.}$
Small zone low grey level emphasis	$F_5 = \frac{1}{N_s}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\frac{s_{ij}}{i^2 j^2}$
Small zone high grey level emphasis	$F_6 = \frac{1}{N_s}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\frac{i^2 s_{ij}}{j^2}$

Table B7: GLSZM features continued.

Name	Definition
Large zone low grey level emphasis	$F_7 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} \frac{j^2 s_{ij}}{i^2}$
Large zone high grey level emphasis	$F_8 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} i^2 j^2 s_{ij}$
Grey level non-uniformity	$F_9 = \frac{1}{N_s} \sum_{i=1}^{N_g} s_{i.}^2$
Normalised grey level non-uniformity	$F_{10} = \frac{1}{N_s^2} \sum_{i=1}^{N_g} s_{i.}^2$
Zone size non-uniformity	$F_{11} = \frac{1}{N_s} \sum_{j=1}^{N_z} s_{.j}^2$
Normalised zone size non-uniformity	$F_{12} = \frac{1}{N_s^2} \sum_{i=1}^{N_z} s_{.j}^2$
Zone percentage	$F_{13} = \frac{N_s}{N_v}$
Grey level variance	$F_{14} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} (i-\mu)^2 p_{ij}$ where $p_{ij} = s_{ij}/N_s$
Zone size variance	$F_{15} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} (j-\mu)^2 p_{ij}$
Zone size entropy	$F_{16} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} p_{ij} \log_2 p_{ij}$

GLDZM features

Introduced in Section 2.4.9 of main text.

Table B8: Definitions for 16 GLDZM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Small Distance Emphasis	$F_1 = \frac{1}{N_s} \sum_{j=1}^{N_d} \frac{d_{.j}}{j^2}$
Large Distance Emphasis	$F_2 = \frac{1}{N_s} \sum_{j=1}^{N_d} j^2 d_{.j}$
Low Grey Level Zone Emphasis	$F_3 = \frac{1}{N_s} \sum_{i=1}^{N_g} \frac{d_{i.}}{i^2}$
High Grey Level Zone Emphasis	$F_4 = \frac{1}{N_s} \sum_{i=1}^{N_g} i^2 d_{i.}$
Small Distance Low Grey Level Emphasis	$F_5 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{d_{ij}}{i^2 j^2}$
Small Distance High Grey Level Emphasis	$F_6 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{i^2 d_{ij}}{j^2}$

Table B8: GLDZM features continued.

Name	Definition
Large Distance Low Grey Level Emphasis	$F_7 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{j^2 d_{ij}}{i^2}$
Large Distance High Grey Level Emphasis	$F_8 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} i^2 j^2 d_{ij}$
Grey Level Non-uniformity	$F_9 = \frac{1}{N_s} \sum_{i=1}^{N_g} d_{i.}^2$
Normalised Grey Level Non-uniformity	$F_{10} = \frac{1}{N_s^2} \sum_{i=1}^{N_g} d_{i.}^2$
Zone Distance Non-uniformity	$F_{11} = \frac{1}{N_s} \sum_{j=1}^{N_d} d_{.j}^2$
Normalised Zone Distance Non-uniformity	$F_{12} = \frac{1}{N_s^2} \sum_{i=1}^{N_d} d_{.j}^2$
Zone Percentage	$F_{13} = \frac{N_s}{N_v}$
Grey Level Variance	$F_{14} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} (i - \mu)^2 p_{ij}$ where $p_{ij} = d_{ij}/N_s$ and $\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} i\, p_{ij}$.
Zone Distance Variance	$F_{15} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} (j - \mu)^2 p_{ij}$ where $\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} j\, p_{ij}$ (mean zone size).
Zone Distance Entropy	$F_{16} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} p_{ij} \log_2 p_{ij}$

NGTDM features

Introduced in Section 2.4.10 of main text.

Table B9: Definitions for 5 NGTDM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Coarseness	$F_1 = \frac{1}{\sum_{i=1}^{N_g} p_i\, s_i}$
Contrast	$N_{g,p}$ = number of grey levels $p_i > 0$ $a = \frac{1}{N_{g,p}(N_{g,p}-1)}$ $b = \left(\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} p_{i_1} p_{i_2} \, (i_1 - i_2)^2 \right)$ $c = \left(\frac{1}{N_{v,c}} \sum_{i=1}^{N_g} s_i \right)$ $F_2 = a \times b \times c.$

Table B9: NGTDM features continued.

Name	Definition
Busyness	$F_3 = \frac{\sum_{i=1}^{N_g} p_i \, s_i}{\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} \lvert i_1 \, p_{i_1} - i_2 \, p_{i_2} \rvert}$ when $p_{i_1} \neq 0$ and $p_{i_2} \neq 0$
Complexity	$F_4 = \frac{1}{N_{v,c}} \sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} \lvert i_1 - i_2 \rvert \, \frac{p_{i_1} \, s_{i_1} + p_{i_2} \, s_{i_2}}{p_{i_1} + p_{i_2}}$ when $p_{i_1} \neq 0$ and $p_{i_2} \neq 0$
Strength	$F_5 = \frac{\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} (p_{i_1} + p_{i_2})(i_1 - i_2)^2}{\sum_{i=1}^{N_g} s_i},$ when $p_{i_1} \neq 0$ and $p_{i_2} \neq 0$

NGLDM features

Introduced in Section 2.4.11 of main text.

Table B10: Definitions for 15 NGLDM features implemented in the SPAARC radiomics package. Definitions here are those collected and documented by the IBSI [90].

Name	Definition
Low Dependence Emphasis	$F_1 = \frac{1}{N_s} \sum_{j=1}^{N_n} \frac{s_{.j}}{j^2}$
High dependence emphasis	$F_2 = \frac{1}{N_s} \sum_{j=1}^{N_n} j^2 s_{.j}$
Low grey level count emphasis	$F_3 = \frac{1}{N_s} \sum_{i=1}^{N_g} \frac{s_{i.}}{i^2}$
High grey level count emphasis	$F_4 = \frac{1}{N_s} \sum_{i=1}^{N_g} i^2 s_{i.}$
Low dependence low grey level emphasis	$F_5 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} \frac{s_{ij}}{i^2 j^2}$
Low dependence high grey level emphasis	$F_6 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} \frac{i^2 s_{ij}}{j^2}$
High dependence low grey level emphasis	$F_7 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} \frac{j^2 s_{ij}}{i^2}$
High dependence high grey level emphasis	$F_8 = \frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} i^2 j^2 s_{ij}$
Grey level non-uniformity	$F_9 = \frac{1}{N_s} \sum_{i=1}^{N_g} s_{i.}^2$
Normalised grey level non-uniformity	$F_{10} = \frac{1}{N_s^2} \sum_{i=1}^{N_g} s_{i.}^2$
Dependence count non-uniformity	$F_{11} = \frac{1}{N_s} \sum_{j=1}^{N_n} s_{.j}^2$

Table B10: NGLDM features continued.

Name	**Definition**
Grey level variance	$F_{12} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} (i - \mu)^2 p_{ij}$ where $\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} i\, p_{ij}$.
Dependence count variance	$F_{13} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} (j - \mu)^2 p_{ij}$ μ same as feature above.
Dependence count entropy	$F_{14} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_n} p_{ij} \log_2 p_{ij}$
Dependence count energy	$F_{15} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} p_{ij}^2$

C

Table C1: Cardiff results for every configuration tested in the IBSI consensus study compared to final benchmarks (3.s.f.).

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	Morphology	Volume (mesh)	556	4	556.3333	0	match
digital phantom	Morphology	Volume (voxel counting)	592	4	592	0	match
digital phantom	Morphology	Surface area (mesh)	388	3	388.0706	0	match
digital phantom	Morphology	Surface to volume ratio	0.698	0.004	0.69755	0	match
digital phantom	Morphology	Compactness 1	0.0411	0.0003	0.041058	0	match
digital phantom	Morphology	Compactness 2	0.599	0.004	0.59895	0	match
digital phantom	Morphology	Spherical disproportion	1.19	0.01	1.1863	0	match
digital phantom	Morphology	Sphericity	0.843	0.005	0.84294	0	match
digital phantom	Morphology	Asphericity	0.186	0.001	0.18632	0	match
digital phantom	Morphology	Centre of mass shift	0.672	0.004	0.67154	0	match
digital phantom	Morphology	Maximum 3D diameter	13.1	0.1	13.1149	0	match
digital phantom	Morphology	Major axis length	11.4	0.1	11.4024	0	match
digital phantom	Morphology	Minor axis length	9.31	0.06	9.308	0	match
digital phantom	Morphology	Least axis length	8.54	0.05	8.536	0	match
digital phantom	Morphology	Elongation	0.816	0.005	0.81632	0	match
digital phantom	Morphology	Flatness	0.749	0.005	0.74861	0	match
digital phantom	Morphology	Volume density (AABB)	0.869	0.005	0.86927	0	match
digital phantom	Morphology	Area density (AABB)	0.866	0.005	0.86623	0	match
digital phantom	Morphology	Volume density (OMBB)	0.869	0.005			
digital phantom	Morphology	Area density (OMBB)	0.866	0.005			
digital phantom	Morphology	Volume density (AEE)	1.17	0.01	1.1728	0	match
digital phantom	Morphology	Area density (AEE)	1.36	0.01	1.3551	0	match
digital phantom	Morphology	Volume density (MVEE)					
digital phantom	Morphology	Area density (MVEE)					
digital phantom	Morphology	Volume density (convex hull)	0.961	0.006	0.96085	0	match
digital phantom	Morphology	Area density (convex hull)	1.03	0.01	1.0329	0	match
digital phantom	Morphology	Integrated intensity	1200	10	1195.3649	0	match
digital phantom	Morphology	Moran’s I index	0.0397	0.0003			
digital phantom	Morphology	Geary’s C measure	0.974	0.006			
digital phantom	Local intensity	Local intensity peak	2.6	0			
digital phantom	Local intensity	Global intensity peak	3.1	0			
digital phantom	Statistics	Mean	2.15	0	2.1486	0	match
digital phantom	Statistics	Variance	3.05	0	3.0455	0	match
digital phantom	Statistics	Skewness	1.08	0	1.0838	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	Statistics	(Excess) kurtosis	-0.355	0	-0.35462	0	match
digital phantom	Statistics	Median	1	0	1	0	match
digital phantom	Statistics	Minimum	1	0	1	0	match
digital phantom	Statistics	10th percentile	1	0	1	0	match
digital phantom	Statistics	90th percentile	4	0	4.2	0.2	no match
digital phantom	Statistics	Maximum	6	0	6	0	match
digital phantom	Statistics	Interquartile range	3	0	3	0	match
digital phantom	Statistics	Range	5	0	5	0	match
digital phantom	Statistics	Mean absolute deviation	1.55	0	1.5522	0	match
digital phantom	Statistics	Robust mean absolute deviation	1.11	0	1.1138	0	match
digital phantom	Statistics	Median absolute deviation	1.15	0	1.1486	0	match
digital phantom	Statistics	Coefficient of variation	0.812	0	0.8122	0	match
digital phantom	Statistics	Quartile coefficient of dispersion	0.6	0	0.6	0	match
digital phantom	Statistics	Energy	567	0	567	0	match
digital phantom	Statistics	Root mean square	2.77	0	2.7681	0	match
digital phantom	Intensity histogram	Mean	2.15	0	2.1486	0	match
digital phantom	Intensity histogram	Variance	3.05	0	3.0455	0	match
digital phantom	Intensity histogram	Skewness	1.08	0	1.0838	0	match
digital phantom	Intensity histogram	(Excess) kurtosis	-0.355	0	-0.35462	0	match
digital phantom	Intensity histogram	Median	1	0	1	0	match
digital phantom	Intensity histogram	Minimum	1	0	1	0	match
digital phantom	Intensity histogram	10th percentile	1	0	1	0	match
digital phantom	Intensity histogram	90th percentile	4	0	4.2	0.2	no match
digital phantom	Intensity histogram	Maximum	6	0	6	0	match
digital phantom	Intensity histogram	Mode	1	0	1	0	match
digital phantom	Intensity histogram	Interquartile range	3	0	3	0	match
digital phantom	Intensity histogram	Range	5	0	5	0	match
digital phantom	Intensity histogram	Mean absolute deviation	1.55	0	1.5522	0	match
digital phantom	Intensity histogram	Robust mean absolute deviation	1.11	0	1.1138	0	match
digital phantom	Intensity histogram	Median absolute deviation	1.15	0	1.1486	0	match
digital phantom	Intensity histogram	Coefficient of variation	0.812	0	0.8122	0	match
digital phantom	Intensity histogram	Quartile coefficient of dispersion	0.6	0	0.6	0	match
digital phantom	Intensity histogram	Entropy	1.27	0	1.2656	0	match
digital phantom	Intensity histogram	Uniformity	0.512	0	0.51242	0	match
digital phantom	Intensity histogram	Maximum histogram gradient	8	0	8	0	match
digital phantom	Intensity histogram	Maximum histogram gradient intensity	3	0	3	0	match
digital phantom	Intensity histogram	Minimum histogram gradient	-50	0	-50	0	match
digital phantom	Intensity histogram	Minimum histogram gradient intensity	1	0	1	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	Intensity volume histogram	Volume fraction at 10% intensity	0.324	0	0.32432	0	match
digital phantom	Intensity volume histogram	Volume fraction at 90% intensity	0.0946	0	0.094595	0	match
digital phantom	Intensity volume histogram	Intensity at 10% volume	5	0	5	0	match
digital phantom	Intensity volume histogram	Intensity at 90% volume	2	0	2	0	match
digital phantom	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.23	0	0.22973	0	match
digital phantom	Intensity volume histogram	Intensity difference between 10% and 90% volume	3	0	3	0	match
digital phantom	Intensity volume histogram	Area under the IVH curve	0.32	0	0.32027	0	match
digital phantom	GLCM (2D averaged)	Joint maximum	0.519	0	0.5188	0	match
digital phantom	GLCM (2D averaged)	Joint average	2.14	0	2.1424	0	match
digital phantom	GLCM (2D averaged)	Joint variance	2.69	0	2.6877	0	match
digital phantom	GLCM (2D averaged)	Joint entropy	2.05	0	2.0497	0	match
digital phantom	GLCM (2D averaged)	Difference average	1.42	0	1.4225	0	match
digital phantom	GLCM (2D averaged)	Difference variance	2.9	0	2.9016	0	match
digital phantom	GLCM (2D averaged)	Difference entropy	1.4	0	1.3961	0	match
digital phantom	GLCM (2D averaged)	Sum average	4.28	0	4.2848	0	match
digital phantom	GLCM (2D averaged)	Sum variance	5.47	0	5.4729	0	match
digital phantom	GLCM (2D averaged)	Sum entropy	1.6	0	1.6032	0	match
digital phantom	GLCM (2D averaged)	Angular second moment	0.368	0	0.36753	0	match
digital phantom	GLCM (2D averaged)	Contrast	5.28	0	5.2779	0	match
digital phantom	GLCM (2D averaged)	Dissimilarity	1.42	0	1.4225	0	match
digital phantom	GLCM (2D averaged)	Inverse difference	0.678	0	0.67795	0	match
digital phantom	GLCM (2D averaged)	Normalised inverse difference	0.851	0	0.8514	0	match
digital phantom	GLCM (2D averaged)	Inverse difference moment	0.619	0	0.61874	0	match
digital phantom	GLCM (2D averaged)	Normalised inverse difference moment	0.899	0	0.89922	0	match
digital phantom	GLCM (2D averaged)	Inverse variance	0.0567	0	0.056698	0	match
digital phantom	GLCM (2D averaged)	Correlation	-0.0121	0	-0.012107	0	match
digital phantom	GLCM (2D averaged)	Autocorrelation	5.09	0	5.0944	0	match
digital phantom	GLCM (2D averaged)	Cluster tendency	5.47	0	5.4729	0	match
digital phantom	GLCM (2D averaged)	Cluster shade	7	0	6.9978	0	match
digital phantom	GLCM (2D averaged)	Cluster prominence	79.1	0	79.1126	0	match
digital phantom	GLCM (2D averaged)	Information correlation 1	-0.155	0	-0.15512	0	match
digital phantom	GLCM (2D averaged)	Information correlation 2	0.487	0	0.48746	0	match
digital phantom	GLCM (2D slice-merged)	Joint maximum	0.512	0	0.51229	0	match
digital phantom	GLCM (2D slice-merged)	Joint average	2.14	0	2.1434	0	match
digital phantom	GLCM (2D slice-merged)	Joint variance	2.71	0	2.7116	0	match
digital phantom	GLCM (2D slice-merged)	Joint entropy	2.24	0	2.2384	0	match
digital phantom	GLCM (2D slice-merged)	Difference average	1.4	0	1.399	0	match
digital phantom	GLCM (2D slice-merged)	Difference variance	3.06	0	3.0643	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLCM (2D slice-merged)	Difference entropy	1.49	0	1.4926	0	match
digital phantom	GLCM (2D slice-merged)	Sum average	4.29	0	4.2869	0	match
digital phantom	GLCM (2D slice-merged)	Sum variance	5.66	0	5.6561	0	match
digital phantom	GLCM (2D slice-merged)	Sum entropy	1.79	0	1.7949	0	match
digital phantom	GLCM (2D slice-merged)	Angular second moment	0.352	0	0.35168	0	match
digital phantom	GLCM (2D slice-merged)	Contrast	5.19	0	5.1902	0	match
digital phantom	GLCM (2D slice-merged)	Dissimilarity	1.4	0	1.399	0	match
digital phantom	GLCM (2D slice-merged)	Inverse difference	0.683	0	0.68329	0	match
digital phantom	GLCM (2D slice-merged)	Normalised inverse difference	0.854	0	0.85385	0	match
digital phantom	GLCM (2D slice-merged)	Inverse difference moment	0.625	0	0.625	0	match
digital phantom	GLCM (2D slice-merged)	Normalised inverse difference moment	0.901	0	0.90088	0	match
digital phantom	GLCM (2D slice-merged)	Inverse variance	0.0553	0	0.055286	0	match
digital phantom	GLCM (2D slice-merged)	Correlation	0.0173	0	0.017307	0	match
digital phantom	GLCM (2D slice-merged)	Autocorrelation	5.14	0	5.1395	0	match
digital phantom	GLCM (2D slice-merged)	Cluster tendency	5.66	0	5.6561	0	match
digital phantom	GLCM (2D slice-merged)	Cluster shade	6.98	0	6.9766	0	match
digital phantom	GLCM (2D slice-merged)	Cluster prominence	80.4	0	80.3855	0	match
digital phantom	GLCM (2D slice-merged)	Information correlation 1	-0.0341	0	-0.034089	0	match
digital phantom	GLCM (2D slice-merged)	Information correlation 2	0.263	0	0.26251	0	match
digital phantom	GLCM (2.5D direction-merged)	Joint maximum	0.489	0	0.48896	0	match
digital phantom	GLCM (2.5D direction-merged)	Joint average	2.2	0	2.2046	0	match
digital phantom	GLCM (2.5D direction-merged)	Joint variance	3.22	0	3.2181	0	match
digital phantom	GLCM (2.5D direction-merged)	Joint entropy	2.48	0	2.484	0	match
digital phantom	GLCM (2.5D direction-merged)	Difference average	1.46	0	1.4608	0	match
digital phantom	GLCM (2.5D direction-merged)	Difference variance	3.11	0	3.1077	0	match
digital phantom	GLCM (2.5D direction-merged)	Difference entropy	1.61	0	1.6143	0	match
digital phantom	GLCM (2.5D direction-merged)	Sum average	4.41	0	4.4091	0	match
digital phantom	GLCM (2.5D direction-merged)	Sum variance	7.48	0	7.4795	0	match
digital phantom	GLCM (2.5D direction-merged)	Sum entropy	2.01	0	2.0137	0	match
digital phantom	GLCM (2.5D direction-merged)	Angular second moment	0.286	0	0.28628	0	match
digital phantom	GLCM (2.5D direction-merged)	Contrast	5.39	0	5.3928	0	match
digital phantom	GLCM (2.5D direction-merged)	Dissimilarity	1.46	0	1.4608	0	match
digital phantom	GLCM (2.5D direction-merged)	Inverse difference	0.668	0	0.66762	0	match
digital phantom	GLCM (2.5D direction-merged)	Normalised inverse difference	0.847	0	0.84708	0	match
digital phantom	GLCM (2.5D direction-merged)	Inverse difference moment	0.606	0	0.60645	0	match
digital phantom	GLCM (2.5D direction-merged)	Normalised inverse difference moment	0.897	0	0.89682	0	match
digital phantom	GLCM (2.5D direction-merged)	Inverse variance	0.0597	0	0.059714	0	match
digital phantom	GLCM (2.5D direction-merged)	Correlation	0.178	0	0.17755	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLCM (2.5D direction-merged)	Autocorrelation	5.4	0	5.4031	0	match
digital phantom	GLCM (2.5D direction-merged)	Cluster tendency	7.48	0	7.4795	0	match
digital phantom	GLCM (2.5D direction-merged)	Cluster shade	16.6	0	16.6057	0	match
digital phantom	GLCM (2.5D direction-merged)	Cluster prominence	147	0	147.2016	0	match
digital phantom	GLCM (2.5D direction-merged)	Information correlation 1	-0.124	0	-0.12406	0	match
digital phantom	GLCM (2.5D direction-merged)	Information correlation 2	0.487	0	0.4871	0	match
digital phantom	GLCM (2.5D merged)	Joint maximum	0.492	0	0.49223	0	match
digital phantom	GLCM (2.5D merged)	Joint average	2.2	0	2.2047	0	match
digital phantom	GLCM (2.5D merged)	Joint variance	3.24	0	3.2353	0	match
digital phantom	GLCM (2.5D merged)	Joint entropy	2.61	0	2.6068	0	match
digital phantom	GLCM (2.5D merged)	Difference average	1.44	0	1.4352	0	match
digital phantom	GLCM (2.5D merged)	Difference variance	3.23	0	3.2303	0	match
digital phantom	GLCM (2.5D merged)	Difference entropy	1.67	0	1.6746	0	match
digital phantom	GLCM (2.5D merged)	Sum average	4.41	0	4.4093	0	match
digital phantom	GLCM (2.5D merged)	Sum variance	7.65	0	7.6511	0	match
digital phantom	GLCM (2.5D merged)	Sum entropy	2.14	0	2.1404	0	match
digital phantom	GLCM (2.5D merged)	Angular second moment	0.277	0	0.27681	0	match
digital phantom	GLCM (2.5D merged)	Contrast	5.29	0	5.2902	0	match
digital phantom	GLCM (2.5D merged)	Dissimilarity	1.44	0	1.4352	0	match
digital phantom	GLCM (2.5D merged)	Inverse difference	0.673	0	0.67314	0	match
digital phantom	GLCM (2.5D merged)	Normalised inverse difference	0.85	0	0.84967	0	match
digital phantom	GLCM (2.5D merged)	Inverse difference moment	0.613	0	0.61287	0	match
digital phantom	GLCM (2.5D merged)	Normalised inverse difference moment	0.899	0	0.89869	0	match
digital phantom	GLCM (2.5D merged)	Inverse variance	0.0582	0	0.058164	0	match
digital phantom	GLCM (2.5D merged)	Correlation	0.182	0	0.18244	0	match
digital phantom	GLCM (2.5D merged)	Autocorrelation	5.45	0	5.4508	0	match
digital phantom	GLCM (2.5D merged)	Cluster tendency	7.65	0	7.6511	0	match
digital phantom	GLCM (2.5D merged)	Cluster shade	16.4	0	16.4065	0	match
digital phantom	GLCM (2.5D merged)	Cluster prominence	142	0	142.4179	0	match
digital phantom	GLCM (2.5D merged)	Information correlation 1	-0.0334	0	-0.033417	0	match
digital phantom	GLCM (2.5D merged)	Information correlation 2	0.291	0	0.29117	0	match
digital phantom	GLCM (3D averaged)	Joint maximum	0.503	0	0.50281	0	match
digital phantom	GLCM (3D averaged)	Joint average	2.14	0	2.143	0	match
digital phantom	GLCM (3D averaged)	Joint variance	3.1	0	3.0993	0	match
digital phantom	GLCM (3D averaged)	Joint entropy	2.4	0	2.3997	0	match
digital phantom	GLCM (3D averaged)	Difference average	1.43	0	1.431	0	match
digital phantom	GLCM (3D averaged)	Difference variance	3.06	0	3.0563	0	match
digital phantom	GLCM (3D averaged)	Difference entropy	1.56	0	1.5627	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLCM (3D averaged)	Sum average	4.29	0	4.286	0	match
digital phantom	GLCM (3D averaged)	Sum variance	7.07	0	7.0728	0	match
digital phantom	GLCM (3D averaged)	Sum entropy	1.92	0	1.9226	0	match
digital phantom	GLCM (3D averaged)	Angular second moment	0.303	0	0.30298	0	match
digital phantom	GLCM (3D averaged)	Contrast	5.32	0	5.3245	0	match
digital phantom	GLCM (3D averaged)	Dissimilarity	1.43	0	1.431	0	match
digital phantom	GLCM (3D averaged)	Inverse difference	0.677	0	0.67662	0	match
digital phantom	GLCM (3D averaged)	Normalised inverse difference	0.851	0	0.85068	0	match
digital phantom	GLCM (3D averaged)	Inverse difference moment	0.618	0	0.61774	0	match
digital phantom	GLCM (3D averaged)	Normalised inverse difference moment	0.898	0	0.89844	0	match
digital phantom	GLCM (3D averaged)	Inverse variance	0.0604	0	0.060416	0	match
digital phantom	GLCM (3D averaged)	Correlation	0.157	0	0.15735	0	match
digital phantom	GLCM (3D averaged)	Autocorrelation	5.06	0	5.0554	0	match
digital phantom	GLCM (3D averaged)	Cluster tendency	7.07	0	7.0728	0	match
digital phantom	GLCM (3D averaged)	Cluster shade	16.6	0	16.6441	0	match
digital phantom	GLCM (3D averaged)	Cluster prominence	145	0	144.7034	0	match
digital phantom	GLCM (3D averaged)	Information correlation 1	-0.157	0	-0.15685	0	match
digital phantom	GLCM (3D averaged)	Information correlation 2	0.52	0	0.51959	0	match
digital phantom	GLCM (3D merged)	Joint maximum	0.509	0	0.50854	0	match
digital phantom	GLCM (3D merged)	Joint average	2.15	0	2.149	0	match
digital phantom	GLCM (3D merged)	Joint variance	3.13	0	3.1325	0	match
digital phantom	GLCM (3D merged)	Joint entropy	2.57	0	2.5739	0	match
digital phantom	GLCM (3D merged)	Difference average	1.38	0	1.3795	0	match
digital phantom	GLCM (3D merged)	Difference variance	3.21	0	3.2146	0	match
digital phantom	GLCM (3D merged)	Difference entropy	1.64	0	1.6409	0	match
digital phantom	GLCM (3D merged)	Sum average	4.3	0	4.2979	0	match
digital phantom	GLCM (3D merged)	Sum variance	7.41	0	7.4122	0	match
digital phantom	GLCM (3D merged)	Sum entropy	2.11	0	2.1099	0	match
digital phantom	GLCM (3D merged)	Angular second moment	0.291	0	0.29095	0	match
digital phantom	GLCM (3D merged)	Contrast	5.12	0	5.1176	0	match
digital phantom	GLCM (3D merged)	Dissimilarity	1.38	0	1.3795	0	match
digital phantom	GLCM (3D merged)	Inverse difference	0.688	0	0.6877	0	match
digital phantom	GLCM (3D merged)	Normalised inverse difference	0.856	0	0.8559	0	match
digital phantom	GLCM (3D merged)	Inverse difference moment	0.631	0	0.63064	0	match
digital phantom	GLCM (3D merged)	Normalised inverse difference moment	0.902	0	0.90221	0	match
digital phantom	GLCM (3D merged)	Inverse variance	0.0574	0	0.057445	0	match
digital phantom	GLCM (3D merged)	Correlation	0.183	0	0.18313	0	match
digital phantom	GLCM (3D merged)	Autocorrelation	5.19	0	5.1917	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLCM (3D merged)	Cluster tendency	7.41	0	7.4122	0	match
digital phantom	GLCM (3D merged)	Cluster shade	17.4	0	17.4192	0	match
digital phantom	GLCM (3D merged)	Cluster prominence	147	0	147.4639	0	match
digital phantom	GLCM (3D merged)	Information correlation 1	-0.0288	0	-0.0288	0	match
digital phantom	GLCM (3D merged)	Information correlation 2	0.269	0	0.26917	0	match
digital phantom	GLRLM (2D averaged)	Short runs emphasis	0.641	0	0.64062	0	match
digital phantom	GLRLM (2D averaged)	Long runs emphasis	3.78	0	3.7784	0	match
digital phantom	GLRLM (2D averaged)	Low grey level run emphasis	0.604	0	0.60436	0	match
digital phantom	GLRLM (2D averaged)	High grey level run emphasis	9.82	0	9.8243	0	match
digital phantom	GLRLM (2D averaged)	Short run low grey level emphasis	0.294	0	0.29397	0	match
digital phantom	GLRLM (2D averaged)	Short run high grey level emphasis	8.57	0	8.5731	0	match
digital phantom	GLRLM (2D averaged)	Long run low grey level emphasis	3.14	0	3.1445	0	match
digital phantom	GLRLM (2D averaged)	Long run high grey level emphasis	17.4	0	17.387	0	match
digital phantom	GLRLM (2D averaged)	Grey level non-uniformity	5.2	0	5.1971	0	match
digital phantom	GLRLM (2D averaged)	Normalised grey level non-uniformity	0.46	0	0.45973	0	match
digital phantom	GLRLM (2D averaged)	Run length non-uniformity	6.12	0	6.1229	0	match
digital phantom	GLRLM (2D averaged)	Normalised run length non-uniformity	0.492	0	0.49174	0	match
digital phantom	GLRLM (2D averaged)	Run percentage	0.627	0	0.6271	0	match
digital phantom	GLRLM (2D averaged)	Grey level variance	3.35	0	3.353	0	match
digital phantom	GLRLM (2D averaged)	Run length variance	0.761	0	0.76148	0	match
digital phantom	GLRLM (2D averaged)	Run entropy	2.17	0	2.1696	0	match
digital phantom	GLRLM (2D slice-merged)	Short runs emphasis	0.661	0	0.6612	0	match
digital phantom	GLRLM (2D slice-merged)	Long runs emphasis	3.51	0	3.5119	0	match
digital phantom	GLRLM (2D slice-merged)	Low grey level run emphasis	0.609	0	0.60852	0	match
digital phantom	GLRLM (2D slice-merged)	High grey level run emphasis	9.74	0	9.7426	0	match
digital phantom	GLRLM (2D slice-merged)	Short run low grey level emphasis	0.311	0	0.31081	0	match
digital phantom	GLRLM (2D slice-merged)	Short run high grey level emphasis	8.67	0	8.6731	0	match
digital phantom	GLRLM (2D slice-merged)	Long run low grey level emphasis	2.92	0	2.9201	0	match
digital phantom	GLRLM (2D slice-merged)	Long run high grey level emphasis	16.1	0	16.119	0	match
digital phantom	GLRLM (2D slice-merged)	Grey level non-uniformity	20.5	0	20.4873	0	match
digital phantom	GLRLM (2D slice-merged)	Normalised grey level non-uniformity	0.456	0	0.45553	0	match
digital phantom	GLRLM (2D slice-merged)	Run length non-uniformity	21.6	0	21.5992	0	match
digital phantom	GLRLM (2D slice-merged)	Normalised run length non-uniformity	0.441	0	0.4411	0	match
digital phantom	GLRLM (2D slice-merged)	Run percentage	0.627	0	0.6271	0	match
digital phantom	GLRLM (2D slice-merged)	Grey level variance	3.37	0	3.3742	0	match
digital phantom	GLRLM (2D slice-merged)	Run length variance	0.778	0	0.77818	0	match
digital phantom	GLRLM (2D slice-merged)	Run entropy	2.57	0	2.5701	0	match
digital phantom	GLRLM (2.5D direction-merged)	Short runs emphasis	0.665	0	0.66466	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLRLM (2.5D direction-merged)	Long runs emphasis	3.46	0	3.4613	0	match
digital phantom	GLRLM (2.5D direction-merged)	Low grey level run emphasis	0.58	0	0.57996	0	match
digital phantom	GLRLM (2.5D direction-merged)	High grey level run emphasis	10.3	0	10.3072	0	match
digital phantom	GLRLM (2.5D direction-merged)	Short run low grey level emphasis	0.296	0	0.29641	0	match
digital phantom	GLRLM (2.5D direction-merged)	Short run high grey level emphasis	9.03	0	9.0265	0	match
digital phantom	GLRLM (2.5D direction-merged)	Long run low grey level emphasis	2.79	0	2.7864	0	match
digital phantom	GLRLM (2.5D direction-merged)	Long run high grey level emphasis	17.9	0	17.8986	0	match
digital phantom	GLRLM (2.5D direction-merged)	Grey level non-uniformity	19.5	0	19.4627	0	match
digital phantom	GLRLM (2.5D direction-merged)	Normalised grey level non-uniformity	0.413	0	0.41312	0	match
digital phantom	GLRLM (2.5D direction-merged)	Run length non-uniformity	22.3	0	22.2921	0	match
digital phantom	GLRLM (2.5D direction-merged)	Normalised run length non-uniformity	0.461	0	0.46101	0	match
digital phantom	GLRLM (2.5D direction-merged)	Run percentage	0.632	0	0.63176	0	match
digital phantom	GLRLM (2.5D direction-merged)	Grey level variance	3.58	0	3.5807	0	match
digital phantom	GLRLM (2.5D direction-merged)	Run length variance	0.758	0	0.75806	0	match
digital phantom	GLRLM (2.5D direction-merged)	Run entropy	2.52	0	2.5176	0	match
digital phantom	GLRLM (2.5D merged)	Short runs emphasis	0.68	0	0.68016	0	match
digital phantom	GLRLM (2.5D merged)	Long runs emphasis	3.27	0	3.2727	0	match
digital phantom	GLRLM (2.5D merged)	Low grey level run emphasis	0.585	0	0.58512	0	match
digital phantom	GLRLM (2.5D merged)	High grey level run emphasis	10.2	0	10.2086	0	match
digital phantom	GLRLM (2.5D merged)	Short run low grey level emphasis	0.312	0	0.31195	0	match
digital phantom	GLRLM (2.5D merged)	Short run high grey level emphasis	9.05	0	9.0494	0	match
digital phantom	GLRLM (2.5D merged)	Long run low grey level emphasis	2.63	0	2.6257	0	match
digital phantom	GLRLM (2.5D merged)	Long run high grey level emphasis	17	0	17.0267	0	match
digital phantom	GLRLM (2.5D merged)	Grey level non-uniformity	77.1	0	77.1176	0	match
digital phantom	GLRLM (2.5D merged)	Normalised grey level non-uniformity	0.412	0	0.41239	0	match
digital phantom	GLRLM (2.5D merged)	Run length non-uniformity	83.2	0	83.246	0	match
digital phantom	GLRLM (2.5D merged)	Normalised run length non-uniformity	0.445	0	0.44517	0	match
digital phantom	GLRLM (2.5D merged)	Run percentage	0.632	0	0.63176	0	match
digital phantom	GLRLM (2.5D merged)	Grey level variance	3.59	0	3.5924	0	match
digital phantom	GLRLM (2.5D merged)	Run length variance	0.767	0	0.76719	0	match
digital phantom	GLRLM (2.5D merged)	Run entropy	2.76	0	2.7603	0	match
digital phantom	GLRLM (3D averaged)	Short runs emphasis	0.705	0	0.70523	0	match
digital phantom	GLRLM (3D averaged)	Long runs emphasis	3.06	0	3.0611	0	match
digital phantom	GLRLM (3D averaged)	Low grey level run emphasis	0.603	0	0.60298	0	match
digital phantom	GLRLM (3D averaged)	High grey level run emphasis	9.7	0	9.6976	0	match
digital phantom	GLRLM (3D averaged)	Short run low grey level emphasis	0.352	0	0.35158	0	match
digital phantom	GLRLM (3D averaged)	Short run high grey level emphasis	8.54	0	8.5397	0	match
digital phantom	GLRLM (3D averaged)	Long run low grey level emphasis	2.39	0	2.391	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLRLM (3D averaged)	Long run high grey level emphasis	17.6	0	17.5662	0	match
digital phantom	GLRLM (3D averaged)	Grey level non-uniformity	21.8	0	21.7762	0	match
digital phantom	GLRLM (3D averaged)	Normalised grey level non-uniformity	0.43	0	0.43018	0	match
digital phantom	GLRLM (3D averaged)	Run length non-uniformity	26.9	0	26.8534	0	match
digital phantom	GLRLM (3D averaged)	Normalised run length non-uniformity	0.513	0	0.51277	0	match
digital phantom	GLRLM (3D averaged)	Run percentage	0.68	0	0.67983	0	match
digital phantom	GLRLM (3D averaged)	Grey level variance	3.46	0	3.465	0	match
digital phantom	GLRLM (3D averaged)	Run length variance	0.574	0	0.57354	0	match
digital phantom	GLRLM (3D averaged)	Run entropy	2.43	0	2.4321	0	match
digital phantom	GLRLM (3D merged)	Short runs emphasis	0.729	0	0.72913	0	match
digital phantom	GLRLM (3D merged)	Long runs emphasis	2.76	0	2.7615	0	match
digital phantom	GLRLM (3D merged)	Low grey level run emphasis	0.607	0	0.60665	0	match
digital phantom	GLRLM (3D merged)	High grey level run emphasis	9.64	0	9.6376	0	match
digital phantom	GLRLM (3D merged)	Short run low grey level emphasis	0.372	0	0.3716	0	match
digital phantom	GLRLM (3D merged)	Short run high grey level emphasis	8.67	0	8.6724	0	match
digital phantom	GLRLM (3D merged)	Long run low grey level emphasis	2.16	0	2.1629	0	match
digital phantom	GLRLM (3D merged)	Long run high grey level emphasis	15.6	0	15.6346	0	match
digital phantom	GLRLM (3D merged)	Grey level non-uniformity	281	0	281.2813	0	match
digital phantom	GLRLM (3D merged)	Normalised grey level non-uniformity	0.43	0	0.43009	0	match
digital phantom	GLRLM (3D merged)	Run length non-uniformity	328	0	327.7187	0	match
digital phantom	GLRLM (3D merged)	Normalised run length non-uniformity	0.501	0	0.5011	0	match
digital phantom	GLRLM (3D merged)	Run percentage	0.68	0	0.67983	0	match
digital phantom	GLRLM (3D merged)	Grey level variance	3.48	0	3.479	0	match
digital phantom	GLRLM (3D merged)	Run length variance	0.598	0	0.59778	0	match
digital phantom	GLRLM (3D merged)	Run entropy	2.62	0	2.6244	0	match
digital phantom	GLSZM (2D)	Small zone emphasis	0.363	0	0.36331	0	match
digital phantom	GLSZM (2D)	Large zone emphasis	43.9	0	43.8667	0	match
digital phantom	GLSZM (2D)	Low grey level emphasis	0.371	0	0.3712	0	match
digital phantom	GLSZM (2D)	High grey level emphasis	16.4	0	16.4405	0	match
digital phantom	GLSZM (2D)	Small zone low grey level emphasis	0.0259	0	0.025855	0	match
digital phantom	GLSZM (2D)	Small zone high grey level emphasis	10.3	0	10.278	0	match
digital phantom	GLSZM (2D)	Large zone low grey level emphasis	40.4	0	40.3981	0	match
digital phantom	GLSZM (2D)	Large zone high grey level emphasis	113	0	112.5214	0	match
digital phantom	GLSZM (2D)	Grey level non-uniformity	1.41	0	1.4143	0	match
digital phantom	GLSZM (2D)	Normalised grey level non-uniformity	0.323	0	0.32299	0	match
digital phantom	GLSZM (2D)	Zone size non-uniformity	1.49	0	1.4857	0	match
digital phantom	GLSZM (2D)	Normalised zone size non-uniformity	0.333	0	0.3332	0	match
digital phantom	GLSZM (2D)	Zone percentage	0.24	0	0.24039	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLSZM (2D)	Grey level variance	3.97	0	3.9695	0	match
digital phantom	GLSZM (2D)	Zone size variance	21	0	20.9971	0	match
digital phantom	GLSZM (2D)	Zone size entropy	1.93	0	1.9319	0	match
digital phantom	GLSZM (2.5D)	Small zone emphasis	0.368	0	0.36752	0	match
digital phantom	GLSZM (2.5D)	Large zone emphasis	34.2	0	34.2222	0	match
digital phantom	GLSZM (2.5D)	Low grey level emphasis	0.368	0	0.36806	0	match
digital phantom	GLSZM (2.5D)	High grey level emphasis	16.2	0	16.1667	0	match
digital phantom	GLSZM (2.5D)	Small zone low grey level emphasis	0.0295	0	0.029541	0	match
digital phantom	GLSZM (2.5D)	Small zone high grey level emphasis	9.87	0	9.8706	0	match
digital phantom	GLSZM (2.5D)	Large zone low grey level emphasis	30.6	0	30.554	0	match
digital phantom	GLSZM (2.5D)	Large zone high grey level emphasis	107	0	106.6111	0	match
digital phantom	GLSZM (2.5D)	Grey level non-uniformity	5.44	0	5.4444	0	match
digital phantom	GLSZM (2.5D)	Normalised grey level non-uniformity	0.302	0	0.30247	0	match
digital phantom	GLSZM (2.5D)	Zone size non-uniformity	3.44	0	3.4444	0	match
digital phantom	GLSZM (2.5D)	Normalised zone size non-uniformity	0.191	0	0.19136	0	match
digital phantom	GLSZM (2.5D)	Zone percentage	0.243	0	0.24324	0	match
digital phantom	GLSZM (2.5D)	Grey level variance	3.92	0	3.9167	0	match
digital phantom	GLSZM (2.5D)	Zone size variance	17.3	0	17.321	0	match
digital phantom	GLSZM (2.5D)	Zone size entropy	3.08	0	3.0805	0	match
digital phantom	GLSZM (3D)	Small zone emphasis	0.255	0	0.25518	0	match
digital phantom	GLSZM (3D)	Large zone emphasis	550	0	550	0	match
digital phantom	GLSZM (3D)	Low grey level emphasis	0.253	0	0.25278	0	match
digital phantom	GLSZM (3D)	High grey level emphasis	15.6	0	15.6	0	match
digital phantom	GLSZM (3D)	Small zone low grey level emphasis	0.0256	0	0.025604	0	match
digital phantom	GLSZM (3D)	Small zone high grey level emphasis	2.76	0	2.7633	0	match
digital phantom	GLSZM (3D)	Large zone low grey level emphasis	503	0	502.7944	0	match
digital phantom	GLSZM (3D)	Large zone high grey level emphasis	1490	0	1494.6	0	match
digital phantom	GLSZM (3D)	Grey level non-uniformity	1.4	0	1.4	0	match
digital phantom	GLSZM (3D)	Normalised grey level non-uniformity	0.28	0	0.28	0	match
digital phantom	GLSZM (3D)	Zone size non-uniformity	1	0	1	0	match
digital phantom	GLSZM (3D)	Normalised zone size non-uniformity	0.2	0	0.2	0	match
digital phantom	GLSZM (3D)	Zone percentage	0.0676	0	0.067568	0	match
digital phantom	GLSZM (3D)	Grey level variance	2.64	0	2.64	0	match
digital phantom	GLSZM (3D)	Zone size variance	331	0	330.96	0	match
digital phantom	GLSZM (3D)	Zone size entropy	2.32	0	2.3219	0	match
digital phantom	GLDZM (2D)	Small distance emphasis	0.946	0	0.94643	0	match
digital phantom	GLDZM (2D)	Large distance emphasis	1.21	0	1.2143	0	match
digital phantom	GLDZM (2D)	Low grey level emphasis	0.371	0	0.3712	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLDZM (2D)	High grey level emphasis	16.4	0	16.4405	0	match
digital phantom	GLDZM (2D)	Small distance low grey level emphasis	0.367	0	0.36748	0	match
digital phantom	GLDZM (2D)	Small distance high grey level emphasis	15.2	0	15.2351	0	match
digital phantom	GLDZM (2D)	Large distance low grey level emphasis	0.386	0	0.38608	0	match
digital phantom	GLDZM (2D)	Large distance high grey level emphasis	21.3	0	21.2619	0	match
digital phantom	GLDZM (2D)	Grey level non-uniformity	1.41	0	1.4143	0	match
digital phantom	GLDZM (2D)	Normalised grey level non-uniformity	0.323	0	0.32299	0	match
digital phantom	GLDZM (2D)	Zone distance non-uniformity	3.79	0	3.7857	0	match
digital phantom	GLDZM (2D)	Normalised zone distance non-uniformity	0.898	0	0.89796	0	match
digital phantom	GLDZM (2D)	Zone percentage	0.24	0	0.24039	0	match
digital phantom	GLDZM (2D)	Grey level variance	3.97	0	3.9695	0	match
digital phantom	GLDZM (2D)	Zone distance variance	0.051	0	0.05102	0	match
digital phantom	GLDZM (2D)	Zone distance entropy	1.73	0	1.7319	0	match
digital phantom	GLDZM (2.5D)	Small distance emphasis	0.917	0	0.91667	0	match
digital phantom	GLDZM (2.5D)	Large distance emphasis	1.33	0	1.3333	0	match
digital phantom	GLDZM (2.5D)	Low grey level emphasis	0.368	0	0.36806	0	match
digital phantom	GLDZM (2.5D)	High grey level emphasis	16.2	0	16.1667	0	match
digital phantom	GLDZM (2.5D)	Small distance low grey level emphasis	0.362	0	0.36227	0	match
digital phantom	GLDZM (2.5D)	Small distance high grey level emphasis	14.3	0	14.2917	0	match
digital phantom	GLDZM (2.5D)	Large distance low grey level emphasis	0.391	0	0.3912	0	match
digital phantom	GLDZM (2.5D)	Large distance high grey level emphasis	23.7	0	23.6667	0	match
digital phantom	GLDZM (2.5D)	Grey level non-uniformity	5.44	0	5.4444	0	match
digital phantom	GLDZM (2.5D)	Normalised grey level non-uniformity	0.302	0	0.30247	0	match
digital phantom	GLDZM (2.5D)	Zone distance non-uniformity	14.4	0	14.4444	0	match
digital phantom	GLDZM (2.5D)	Normalised zone distance non-uniformity	0.802	0	0.80247	0	match
digital phantom	GLDZM (2.5D)	Zone percentage	0.243	0	0.24324	0	match
digital phantom	GLDZM (2.5D)	Grey level variance	3.92	0	3.9167	0	match
digital phantom	GLDZM (2.5D)	Zone distance variance	0.0988	0	0.098765	0	match
digital phantom	GLDZM (2.5D)	Zone distance entropy	2	0	2.0022	0	match
digital phantom	GLDZM (3D)	Small distance emphasis	1	0	1	0	match
digital phantom	GLDZM (3D)	Large distance emphasis	1	0	1	0	match
digital phantom	GLDZM (3D)	Low grey level emphasis	0.253	0	0.25278	0	match
digital phantom	GLDZM (3D)	High grey level emphasis	15.6	0	15.6	0	match
digital phantom	GLDZM (3D)	Small distance low grey level emphasis	0.253	0	0.25278	0	match
digital phantom	GLDZM (3D)	Small distance high grey level emphasis	15.6	0	15.6	0	match
digital phantom	GLDZM (3D)	Large distance low grey level emphasis	0.253	0	0.25278	0	match
digital phantom	GLDZM (3D)	Large distance high grey level emphasis	15.6	0	15.6	0	match
digital phantom	GLDZM (3D)	Grey level non-uniformity	1.4	0	1.4	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	GLDZM (3D)	Normalised grey level non-uniformity	0.28	0	0.28	0	match
digital phantom	GLDZM (3D)	Zone distance non-uniformity	5	0	5	0	match
digital phantom	GLDZM (3D)	Normalised zone distance non-uniformity	1	0	1	0	match
digital phantom	GLDZM (3D)	Zone percentage	0.0676	0	0.067568	0	match
digital phantom	GLDZM (3D)	Grey level variance	2.64	0	2.64	0	match
digital phantom	GLDZM (3D)	Zone distance variance	0	0	0	0	match
digital phantom	GLDZM (3D)	Zone distance entropy	1.92	0	1.9219	0	match
digital phantom	NGTDM (2D)	Coarseness	0.121	0	0.12051	0	match
digital phantom	NGTDM (2D)	Contrast	0.925	0	0.92526	0	match
digital phantom	NGTDM (2D)	Busyness	2.99	0	2.9888	0	match
digital phantom	NGTDM (2D)	Complexity	10.4	0	10.4001	0	match
digital phantom	NGTDM (2D)	Strength	2.88	0	2.8764	0	match
digital phantom	NGTDM (2.5D)	Coarseness	0.0285	0	0.02847	0	match
digital phantom	NGTDM (2.5D)	Contrast	0.601	0	0.60118	0	match
digital phantom	NGTDM (2.5D)	Busyness	6.8	0	6.8042	0	match
digital phantom	NGTDM (2.5D)	Complexity	14.1	0	14.0829	0	match
digital phantom	NGTDM (2.5D)	Strength	0.741	0	0.74131	0	match
digital phantom	NGTDM (3D)	Coarseness	0.0296	0	0.029604	0	match
digital phantom	NGTDM (3D)	Contrast	0.584	0	0.58371	0	match
digital phantom	NGTDM (3D)	Busyness	6.54	0	6.5436	0	match
digital phantom	NGTDM (3D)	Complexity	13.5	0	13.5398	0	match
digital phantom	NGTDM (3D)	Strength	0.763	0	0.7635	0	match
digital phantom	NGLDM (2D)	Low dependence emphasis	0.158	0	0.15807	0	match
digital phantom	NGLDM (2D)	High dependence emphasis	19.2	0	19.1738	0	match
digital phantom	NGLDM (2D)	Low grey level count emphasis	0.702	0	0.70175	0	match
digital phantom	NGLDM (2D)	High grey level count emphasis	7.49	0	7.4869	0	match
digital phantom	NGLDM (2D)	Low dependence low grey level emphasis	0.0473	0	0.04729	0	match
digital phantom	NGLDM (2D)	Low dependence high grey level emphasis	3.06	0	3.0649	0	match
digital phantom	NGLDM (2D)	High dependence low grey level emphasis	17.6	0	17.5997	0	match
digital phantom	NGLDM (2D)	High dependence high grey level emphasis	49.5	0	49.4777	0	match
digital phantom	NGLDM (2D)	Grey level non-uniformity	10.2	0	10.2464	0	match
digital phantom	NGLDM (2D)	Normalised grey level non-uniformity	0.562	0	0.56186	0	match
digital phantom	NGLDM (2D)	Dependence count non-uniformity	3.96	0	3.9646	0	match
digital phantom	NGLDM (2D)	Normalised dependence count non-uniformity	0.212	0	0.21177	0	match
digital phantom	NGLDM (2D)	Dependence count percentage	1	0			
digital phantom	NGLDM (2D)	Grey level variance	2.7	0	2.7037	0	match
digital phantom	NGLDM (2D)	Dependence count variance	2.73	0	2.7295	0	match
digital phantom	NGLDM (2D)	Dependence count entropy	2.71	0	2.7143	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
digital phantom	NGLDM (2D)	Dependence count energy	0.17	0	0.17025	0	match
digital phantom	NGLDM (2.5D)	Low dependence emphasis	0.159	0	0.15922	0	match
digital phantom	NGLDM (2.5D)	High dependence emphasis	18.8	0	18.8378	0	match
digital phantom	NGLDM (2.5D)	Low grey level count emphasis	0.693	0	0.69332	0	match
digital phantom	NGLDM (2.5D)	High grey level count emphasis	7.66	0	7.6622	0	match
digital phantom	NGLDM (2.5D)	Low dependence low grey level emphasis	0.0477	0	0.04774	0	match
digital phantom	NGLDM (2.5D)	Low dependence high grey level emphasis	3.07	0	3.0704	0	match
digital phantom	NGLDM (2.5D)	High dependence low grey level emphasis	17.2	0	17.1817	0	match
digital phantom	NGLDM (2.5D)	High dependence high grey level emphasis	50.8	0	50.7703	0	match
digital phantom	NGLDM (2.5D)	Grey level non-uniformity	37.9	0	37.9189	0	match
digital phantom	NGLDM (2.5D)	Normalised grey level non-uniformity	0.512	0	0.51242	0	match
digital phantom	NGLDM (2.5D)	Dependence count non-uniformity	12.4	0	12.3514	0	match
digital phantom	NGLDM (2.5D)	Normalised dependence count non-uniformity	0.167	0	0.16691	0	match
digital phantom	NGLDM (2.5D)	Dependence count percentage	1	0			
digital phantom	NGLDM (2.5D)	Grey level variance	3.05	0	3.0455	0	match
digital phantom	NGLDM (2.5D)	Dependence count variance	3.27	0	3.2673	0	match
digital phantom	NGLDM (2.5D)	Dependence count entropy	3.36	0	3.363	0	match
digital phantom	NGLDM (2.5D)	Dependence count energy	0.122	0	0.12199	0	match
digital phantom	NGLDM (3D)	Low dependence emphasis	0.045	0	0.044996	0	match
digital phantom	NGLDM (3D)	High dependence emphasis	109	0	109	0	match
digital phantom	NGLDM (3D)	Low grey level count emphasis	0.693	0	0.69332	0	match
digital phantom	NGLDM (3D)	High grey level count emphasis	7.66	0	7.6622	0	match
digital phantom	NGLDM (3D)	Low dependence low grey level emphasis	0.00963	0	0.0096306	0	match
digital phantom	NGLDM (3D)	Low dependence high grey level emphasis	0.736	0	0.73617	0	match
digital phantom	NGLDM (3D)	High dependence low grey level emphasis	102	0	102.4508	0	match
digital phantom	NGLDM (3D)	High dependence high grey level emphasis	235	0	234.9865	0	match
digital phantom	NGLDM (3D)	Grey level non-uniformity	37.9	0	37.9189	0	match
digital phantom	NGLDM (3D)	Normalised grey level non-uniformity	0.512	0	0.51242	0	match
digital phantom	NGLDM (3D)	Dependence count non-uniformity	4.86	0	4.8649	0	match
digital phantom	NGLDM (3D)	Normalised dependence count non-uniformity	0.0657	0	0.065741	0	match
digital phantom	NGLDM (3D)	Dependence count percentage	1	0			
digital phantom	NGLDM (3D)	Grey level variance	3.05	0	3.0455	0	match
digital phantom	NGLDM (3D)	Dependence count variance	22.1	0	22.057	0	match
digital phantom	NGLDM (3D)	Dependence count entropy	4.4	0	4.4037	0	match
digital phantom	NGLDM (3D)	Dependence count energy	0.0533	0	0.053324	0	match
configuration A	Diagnostics-initial image	Image dimension x	204	0	204	0	match
configuration A	Diagnostics-initial image	Image dimension y	201	0	201	0	match
configuration A	Diagnostics-initial image	Image dimension z	60	0	60	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	Diagnostics-initial image	Voxel dimension x	0.977	0	0.977	0	match
configuration A	Diagnostics-initial image	Voxel dimension y	0.977	0	0.977	0	match
configuration A	Diagnostics-initial image	Voxel dimension z	3	0	3	0	match
configuration A	Diagnostics-initial image	Mean intensity	-266	0	-266.4704	0	match
configuration A	Diagnostics-initial image	Minimum intensity	-1000	0	-1000	0	match
configuration A	Diagnostics-initial image	Maximum intensity	3065	0	3065	0	match
configuration A	Diagnostics-interpolated image	Image dimension x	204	1	204	0	match
configuration A	Diagnostics-interpolated image	Image dimension y	201	1	201	0	match
configuration A	Diagnostics-interpolated image	Image dimension z	60	0	60	0	match
configuration A	Diagnostics-interpolated image	Voxel dimension x	0.977	0	0.977	0	match
configuration A	Diagnostics-interpolated image	Voxel dimension y	0.977	0	0.977	0	match
configuration A	Diagnostics-interpolated image	Voxel dimension z	3	0	3	0	match
configuration A	Diagnostics-interpolated image	Mean intensity	-266	3	-266.4704	0	match
configuration A	Diagnostics-interpolated image	Minimum intensity	-1000	0	-1000	0	match
configuration A	Diagnostics-interpolated image	Maximum intensity	3065	40	3065	0	match
configuration A	Diagnostics-initial ROI	Int. mask dimension x	204	0	204	0	match
configuration A	Diagnostics-initial ROI	Int. mask dimension y	201	0	201	0	match
configuration A	Diagnostics-initial ROI	Int. mask dimension z	60	0	60	0	match
configuration A	Diagnostics-initial ROI	Int. mask bounding box dimension x	100	0	100	0	match
configuration A	Diagnostics-initial ROI	Int. mask bounding box dimension y	99	0	99	0	match
configuration A	Diagnostics-initial ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration A	Diagnostics-initial ROI	Morph. mask bounding box dimension x	100	0	100	0	match
configuration A	Diagnostics-initial ROI	Morph. mask bounding box dimension y	99	0	99	0	match
configuration A	Diagnostics-initial ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration A	Diagnostics-initial ROI	Int. mask voxel count	125256	0	125256	0	match
configuration A	Diagnostics-initial ROI	Morph. mask voxel count	125256	0	125256	0	match
configuration A	Diagnostics-initial ROI	Int. mask mean intensity	-46.9	0	-46.8827	0	match
configuration A	Diagnostics-initial ROI	Int. mask minimum intensity	-1000	0	-1000	0	match
configuration A	Diagnostics-initial ROI	Int. mask maximum intensity	723	0	723	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask dimension x	204	1	204	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask dimension y	201	1	201	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask dimension z	60	0	60	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask bounding box dimension x	100	1	100	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask bounding box dimension y	99	0.3	99	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration A	Diagnostics-interpolated ROI	Morph. mask bounding box dimension x	100	1	100	0	match
configuration A	Diagnostics-interpolated ROI	Morph. mask bounding box dimension y	99	0.3	99	0	match
configuration A	Diagnostics-interpolated ROI	Morph. mask bounding box dimension z	26	0	26	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	Diagnostics-interpolated ROI	Int. mask voxel count	125256	1000	125256	0	match
configuration A	Diagnostics-interpolated ROI	Morph. mask voxel count	125256	1000	125256	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask mean intensity	-46.9	0.1	-46.8827	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask minimum intensity	-1000	10	-1000	0	match
configuration A	Diagnostics-interpolated ROI	Int. mask maximum intensity	723	7	723	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask dimension x	204	1	204	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask dimension y	201	1	201	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask dimension z	60	0	60	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask bounding box dimension x	100	1	100	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask bounding box dimension y	99	0.3	99	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration A	Diagnostics-resegmented ROI	Morph. mask bounding box dimension x	100	1	100	0	match
configuration A	Diagnostics-resegmented ROI	Morph. mask bounding box dimension y	99	0.3	99	0	match
configuration A	Diagnostics-resegmented ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask voxel count	114596	1000	114596	0	match
configuration A	Diagnostics-resegmented ROI	Morph. mask voxel count	125256	1000	125256	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask mean intensity	13.4	1.1	13.4084	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask minimum intensity	-500	0	-500	0	match
configuration A	Diagnostics-resegmented ROI	Int. mask maximum intensity	377	9	377	0	match
configuration A	Morphology	Volume (mesh)	358000	5000	358407.6214	0	match
configuration A	Morphology	Volume (voxel counting)	359000	5000	358681.4519	0	match
configuration A	Morphology	Surface area (mesh)	35700	300	35702.1565	0	match
configuration A	Morphology	Surface to volume ratio	0.0996	0.0005	0.099613	0	match
configuration A	Morphology	Compactness 1	0.03	0.0001	0.029975	0	match
configuration A	Morphology	Compactness 2	0.319	0.001	0.31924	0	match
configuration A	Morphology	Spherical disproportion	1.46	0.01	1.4632	0	match
configuration A	Morphology	Sphericity	0.683	0.001	0.68345	0	match
configuration A	Morphology	Asphericity	0.463	0.002	0.46316	0	match
configuration A	Morphology	Centre of mass shift	52.9	28.7	52.9152	0	match
configuration A	Morphology	Maximum 3D diameter	125	1	125.2003	0	match
configuration A	Morphology	Major axis length	92.7	0.4	92.7209	0	match
configuration A	Morphology	Minor axis length	81.5	0.4	81.5232	0	match
configuration A	Morphology	Least axis length	70.1	0.3	70.0635	0	match
configuration A	Morphology	Elongation	0.879	0.001	0.87923	0	match
configuration A	Morphology	Flatness	0.756	0.001	0.75564	0	match
configuration A	Morphology	Volume density (AABB)	0.486	0.003	0.48625	0	match
configuration A	Morphology	Area density (AABB)	0.725	0.003	0.72522	0	match
configuration A	Morphology	Volume density (OMBB)					

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	Morphology	Area density (OMBB)					
configuration A	Morphology	Volume density (AEE)	1.29	0.01	1.2925	0	match
configuration A	Morphology	Area density (AEE)	1.71	0.01	1.7052	0	match
configuration A	Morphology	Volume density (MVEE)					
configuration A	Morphology	Area density (MVEE)					
configuration A	Morphology	Volume density (convex hull)	0.827	0.001	0.82676	0	match
configuration A	Morphology	Area density (convex hull)	1.18	0.01	1.1834	0	match
configuration A	Morphology	Integrated intensity	4810000	320000	4805682.113	0	match
configuration A	Morphology	Moran’s I index	0.0322	0.0002			
configuration A	Morphology	Geary’s C measure	0.863	0.001			
configuration A	Local intensity	Local intensity peak	-277	10			
configuration A	Local intensity	Global intensity peak	189	5			
configuration A	Statistics	Mean	13.4	1.1	13.4084	0	match
configuration A	Statistics	Variance	14200	400	14233.4046	0	match
configuration A	Statistics	Skewness	-2.47	0.05	-2.4739	0	match
configuration A	Statistics	(Excess) kurtosis	5.96	0.24	5.9579	0	match
configuration A	Statistics	Median	46	0.3	46	0	match
configuration A	Statistics	Minimum	-500	0	-500	0	match
configuration A	Statistics	10th percentile	-129	8	-129	0	match
configuration A	Statistics	90th percentile	95	0	95	0	match
configuration A	Statistics	Maximum	377	9	377	0	match
configuration A	Statistics	Interquartile range	56	0.5	56	0	match
configuration A	Statistics	Range	877	9	877	0	match
configuration A	Statistics	Mean absolute deviation	73.6	1.4	73.5831	0	match
configuration A	Statistics	Robust mean absolute deviation	27.7	0.8	27.7292	0	match
configuration A	Statistics	Median absolute deviation	64.3	1	64.2906	0	match
configuration A	Statistics	Coefficient of variation	8.9	4.98	8.8977	0	match
configuration A	Statistics	Quartile coefficient of dispersion	0.636	0.008	0.63636	0	match
configuration A	Statistics	Energy	1650000000	20000000	1651693976	0	match
configuration A	Statistics	Root mean square	120	2	120.0549	0	match
configuration A	Intensity histogram	Mean	21.1	0.1	21.0568	0	match
configuration A	Intensity histogram	Variance	22.8	0.6	22.8238	0	match
configuration A	Intensity histogram	Skewness	-2.46	0.05	-2.4579	0	match
configuration A	Intensity histogram	(Excess) kurtosis	5.9	0.24	5.8966	0	match
configuration A	Intensity histogram	Median	22	0	22	0	match
configuration A	Intensity histogram	Minimum	1	0	1	0	match
configuration A	Intensity histogram	10th percentile	15	0.4	15	0	match
configuration A	Intensity histogram	90th percentile	24	0	24	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	Intensity histogram	Maximum	36	0.4	36	0	match
configuration A	Intensity histogram	Mode	23	0	23	0	match
configuration A	Intensity histogram	Interquartile range	2	0	2	0	match
configuration A	Intensity histogram	Range	35	0.4	35	0	match
configuration A	Intensity histogram	Mean absolute deviation	2.94	0.06	2.9402	0	match
configuration A	Intensity histogram	Robust mean absolute deviation	1.18	0.04	1.1818	0	match
configuration A	Intensity histogram	Median absolute deviation	2.58	0.05	2.5805	0	match
configuration A	Intensity histogram	Coefficient of variation	0.227	0.004	0.22688	0	match
configuration A	Intensity histogram	Quartile coefficient of dispersion	0.0455	0	0.045455	0	match
configuration A	Intensity histogram	Entropy	3.36	0.03	3.356	0	match
configuration A	Intensity histogram	Uniformity	0.15	0.002	0.15037	0	match
configuration A	Intensity histogram	Maximum histogram gradient	11000	100	11039.5	0	match
configuration A	Intensity histogram	Maximum histogram gradient intensity	21	0	21	0	match
configuration A	Intensity histogram	Minimum histogram gradient	-10100	100	-10101.5	0	match
configuration A	Intensity histogram	Minimum histogram gradient intensity	24	0	24	0	match
configuration A	Intensity volume histogram	Volume fraction at 10% intensity	0.978	0.001	0.97819	0	match
configuration A	Intensity volume histogram	Volume fraction at 90% intensity	0.0000698	0.0000103	6.98E-05	0	match
configuration A	Intensity volume histogram	Intensity at 10% volume	96	0	96	0	match
configuration A	Intensity volume histogram	Intensity at 90% volume	-128	8	-128	0	match
configuration A	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.978	0.001	0.97812	0	match
configuration A	Intensity volume histogram	Intensity difference between 10% and 90% volume	224	8	224	0	match
configuration A	Intensity volume histogram	Area under the IVH curve			0.58598		
configuration A	GLCM (2D averaged)	Joint maximum	0.109	0.001	0.10889	0	match
configuration A	GLCM (2D averaged)	Joint average	20.6	0.1	20.6482	0	match
configuration A	GLCM (2D averaged)	Joint variance	27	0.4	27.0143	0	match
configuration A	GLCM (2D averaged)	Joint entropy	5.82	0.04	5.82	0	match
configuration A	GLCM (2D averaged)	Difference average	1.58	0.03	1.5768	0	match
configuration A	GLCM (2D averaged)	Difference variance	4.94	0.19	4.9392	0	match
configuration A	GLCM (2D averaged)	Difference entropy	2.27	0.03	2.2707	0	match
configuration A	GLCM (2D averaged)	Sum average	41.3	0.1	41.2964	0	match
configuration A	GLCM (2D averaged)	Sum variance	100	1	100.2058	0	match
configuration A	GLCM (2D averaged)	Sum entropy	4.19	0.03	4.1907	0	match
configuration A	GLCM (2D averaged)	Angular second moment	0.045	0.0008	0.044951	0	match
configuration A	GLCM (2D averaged)	Contrast	7.85	0.26	7.8514	0	match
configuration A	GLCM (2D averaged)	Dissimilarity	1.58	0.03	1.5768	0	match
configuration A	GLCM (2D averaged)	Inverse difference	0.581	0.003	0.58111	0	match
configuration A	GLCM (2D averaged)	Normalised inverse difference	0.961	0.001	0.96112	0	match
configuration A	GLCM (2D averaged)	Inverse difference moment	0.544	0.003	0.54382	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLCM (2D averaged)	Normalised inverse difference moment	0.994	0.001	0.99437	0	match
configuration A	GLCM (2D averaged)	Inverse variance	0.441	0.001	0.44096	0	match
configuration A	GLCM (2D averaged)	Correlation	0.778	0.002	0.77796	0	match
configuration A	GLCM (2D averaged)	Autocorrelation	455	2	455.3685	0	match
configuration A	GLCM (2D averaged)	Cluster tendency	100	1	100.2058	0	match
configuration A	GLCM (2D averaged)	Cluster shade	-1040	20	-1042.9917	0	match
configuration A	GLCM (2D averaged)	Cluster prominence	52700	500	52672.4808	0	match
configuration A	GLCM (2D averaged)	Information correlation 1	-0.236	0.001	-0.2361	0	match
configuration A	GLCM (2D averaged)	Information correlation 2	0.863	0.003	0.86345	0	match
configuration A	GLCM (2D slice-merged)	Joint maximum	0.109	0.001	0.10876	0	match
configuration A	GLCM (2D slice-merged)	Joint average	20.6	0.1	20.6465	0	match
configuration A	GLCM (2D slice-merged)	Joint variance	27	0.4	27.0374	0	match
configuration A	GLCM (2D slice-merged)	Joint entropy	5.9	0.04	5.9028	0	match
configuration A	GLCM (2D slice-merged)	Difference average	1.57	0.03	1.5744	0	match
configuration A	GLCM (2D slice-merged)	Difference variance	4.96	0.19	4.9627	0	match
configuration A	GLCM (2D slice-merged)	Difference entropy	2.28	0.03	2.2836	0	match
configuration A	GLCM (2D slice-merged)	Sum average	41.3	0.1	41.293	0	match
configuration A	GLCM (2D slice-merged)	Sum variance	100	1	100.3268	0	match
configuration A	GLCM (2D slice-merged)	Sum entropy	4.21	0.03	4.2067	0	match
configuration A	GLCM (2D slice-merged)	Angular second moment	0.0446	0.0008	0.044645	0	match
configuration A	GLCM (2D slice-merged)	Contrast	7.82	0.26	7.8226	0	match
configuration A	GLCM (2D slice-merged)	Dissimilarity	1.57	0.03	1.5744	0	match
configuration A	GLCM (2D slice-merged)	Inverse difference	0.581	0.003	0.58128	0	match
configuration A	GLCM (2D slice-merged)	Normalised inverse difference	0.961	0.001	0.96117	0	match
configuration A	GLCM (2D slice-merged)	Inverse difference moment	0.544	0.003	0.54402	0	match
configuration A	GLCM (2D slice-merged)	Normalised inverse difference moment	0.994	0.001	0.99439	0	match
configuration A	GLCM (2D slice-merged)	Inverse variance	0.441	0.001	0.4411	0	match
configuration A	GLCM (2D slice-merged)	Correlation	0.78	0.002	0.78023	0	match
configuration A	GLCM (2D slice-merged)	Autocorrelation	455	2	455.331	0	match
configuration A	GLCM (2D slice-merged)	Cluster tendency	100	1	100.3268	0	match
configuration A	GLCM (2D slice-merged)	Cluster shade	-1050	20	-1045.7191	0	match
configuration A	GLCM (2D slice-merged)	Cluster prominence	52800	500	52772.3395	0	match
configuration A	GLCM (2D slice-merged)	Information correlation 1	-0.214	0.001	-0.21356	0	match
configuration A	GLCM (2D slice-merged)	Information correlation 2	0.851	0.002	0.85102	0	match
configuration A	GLCM (2.5D direction-merged)	Joint maximum	0.0943	0.0008	0.09434	0	match
configuration A	GLCM (2.5D direction-merged)	Joint average	21.3	0.1	21.3403	0	match
configuration A	GLCM (2.5D direction-merged)	Joint variance	18.6	0.5	18.6128	0	match
configuration A	GLCM (2.5D direction-merged)	Joint entropy	5.78	0.04	5.7796	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLCM (2.5D direction-merged)	Difference average	1.35	0.03	1.3493	0	match
configuration A	GLCM (2.5D direction-merged)	Difference variance	4.12	0.2	4.1215	0	match
configuration A	GLCM (2.5D direction-merged)	Difference entropy	2.16	0.03	2.1573	0	match
configuration A	GLCM (2.5D direction-merged)	Sum average	42.7	0.1	42.6807	0	match
configuration A	GLCM (2.5D direction-merged)	Sum variance	68.5	1.3	68.4902	0	match
configuration A	GLCM (2.5D direction-merged)	Sum entropy	4.17	0.03	4.1742	0	match
configuration A	GLCM (2.5D direction-merged)	Angular second moment	0.0429	0.0007	0.042867	0	match
configuration A	GLCM (2.5D direction-merged)	Contrast	5.96	0.27	5.961	0	match
configuration A	GLCM (2.5D direction-merged)	Dissimilarity	1.35	0.03	1.3493	0	match
configuration A	GLCM (2.5D direction-merged)	Inverse difference	0.605	0.003	0.60519	0	match
configuration A	GLCM (2.5D direction-merged)	Normalised inverse difference	0.966	0.001	0.96629	0	match
configuration A	GLCM (2.5D direction-merged)	Inverse difference moment	0.573	0.003	0.57312	0	match
configuration A	GLCM (2.5D direction-merged)	Normalised inverse difference moment	0.996	0.001	0.99571	0	match
configuration A	GLCM (2.5D direction-merged)	Inverse variance	0.461	0.002	0.46117	0	match
configuration A	GLCM (2.5D direction-merged)	Correlation	0.839	0.003	0.839	0	match
configuration A	GLCM (2.5D direction-merged)	Autocorrelation	471	2	471.0435	0	match
configuration A	GLCM (2.5D direction-merged)	Cluster tendency	68.5	1.3	68.4902	0	match
configuration A	GLCM (2.5D direction-merged)	Cluster shade	-1490	30	-1485.2865	0	match
configuration A	GLCM (2.5D direction-merged)	Cluster prominence	47600	700	47642.8398	0	match
configuration A	GLCM (2.5D direction-merged)	Information correlation 1	-0.231	0.001	-0.2307	0	match
configuration A	GLCM (2.5D direction-merged)	Information correlation 2	0.879	0.001	0.8789	0	match
configuration A	GLCM (2.5D merged)	Joint maximum	0.0943	0.0008	0.094274	0	match
configuration A	GLCM (2.5D merged)	Joint average	21.3	0.1	21.3401	0	match
configuration A	GLCM (2.5D merged)	Joint variance	18.6	0.5	18.6165	0	match
configuration A	GLCM (2.5D merged)	Joint entropy	5.79	0.04	5.7894	0	match
configuration A	GLCM (2.5D merged)	Difference average	1.35	0.03	1.3486	0	match
configuration A	GLCM (2.5D merged)	Difference variance	4.14	0.2	4.1354	0	match
configuration A	GLCM (2.5D merged)	Difference entropy	2.16	0.03	2.1618	0	match
configuration A	GLCM (2.5D merged)	Sum average	42.7	0.1	42.6803	0	match
configuration A	GLCM (2.5D merged)	Sum variance	68.5	1.3	68.5117	0	match
configuration A	GLCM (2.5D merged)	Sum entropy	4.18	0.03	4.1752	0	match
configuration A	GLCM (2.5D merged)	Angular second moment	0.0427	0.0007	0.042727	0	match
configuration A	GLCM (2.5D merged)	Contrast	5.95	0.27	5.9542	0	match
configuration A	GLCM (2.5D merged)	Dissimilarity	1.35	0.03	1.3486	0	match
configuration A	GLCM (2.5D merged)	Inverse difference	0.605	0.003	0.60529	0	match
configuration A	GLCM (2.5D merged)	Normalised inverse difference	0.966	0.001	0.9663	0	match
configuration A	GLCM (2.5D merged)	Inverse difference moment	0.573	0.003	0.57323	0	match
configuration A	GLCM (2.5D merged)	Normalised inverse difference moment	0.996	0.001	0.99572	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLCM (2.5D merged)	Inverse variance	0.461	0.002	0.46117	0	match
configuration A	GLCM (2.5D merged)	Correlation	0.84	0.003	0.84008	0	match
configuration A	GLCM (2.5D merged)	Autocorrelation	471	2	471.0411	0	match
configuration A	GLCM (2.5D merged)	Cluster tendency	68.5	1.3	68.5117	0	match
configuration A	GLCM (2.5D merged)	Cluster shade	-1490	30	-1486.4033	0	match
configuration A	GLCM (2.5D merged)	Cluster prominence	47700	700	47686.2721	0	match
configuration A	GLCM (2.5D merged)	Information correlation 1	-0.228	0.001	-0.22791	0	match
configuration A	GLCM (2.5D merged)	Information correlation 2	0.88	0.001	0.88002	0	match
configuration A	GLRLM (2D averaged)	Short runs emphasis	0.785	0.003	0.785	0	match
configuration A	GLRLM (2D averaged)	Long runs emphasis	2.91	0.03	2.9055	0	match
configuration A	GLRLM (2D averaged)	Low grey level run emphasis	0.0264	0.0003	0.026416	0	match
configuration A	GLRLM (2D averaged)	High grey level run emphasis	428	3	428.1961	0	match
configuration A	GLRLM (2D averaged)	Short run low grey level emphasis	0.0243	0.0003	0.024332	0	match
configuration A	GLRLM (2D averaged)	Short run high grey level emphasis	320	1	320.0485	0	match
configuration A	GLRLM (2D averaged)	Long run low grey level emphasis	0.0386	0.0003	0.038556	0	match
configuration A	GLRLM (2D averaged)	Long run high grey level emphasis	1410	20	1405.9383	0	match
configuration A	GLRLM (2D averaged)	Grey level non-uniformity	432	1	432.121	0	match
configuration A	GLRLM (2D averaged)	Normalised grey level non-uniformity	0.128	0.003	0.12842	0	match
configuration A	GLRLM (2D averaged)	Run length non-uniformity	1650	10	1653.9307	0	match
configuration A	GLRLM (2D averaged)	Normalised run length non-uniformity	0.579	0.003	0.57864	0	match
configuration A	GLRLM (2D averaged)	Run percentage	0.704	0.003	0.70415	0	match
configuration A	GLRLM (2D averaged)	Grey level variance	33.7	0.6	33.6905	0	match
configuration A	GLRLM (2D averaged)	Run length variance	0.828	0.008	0.82816	0	match
configuration A	GLRLM (2D averaged)	Run entropy	4.73	0.02	4.7347	0	match
configuration A	GLRLM (2D slice-merged)	Short runs emphasis	0.786	0.003	0.78582	0	match
configuration A	GLRLM (2D slice-merged)	Long runs emphasis	2.89	0.03	2.8934	0	match
configuration A	GLRLM (2D slice-merged)	Low grey level run emphasis	0.0264	0.0003	0.026405	0	match
configuration A	GLRLM (2D slice-merged)	High grey level run emphasis	428	3	428.2514	0	match
configuration A	GLRLM (2D slice-merged)	Short run low grey level emphasis	0.0243	0.0003	0.02433	0	match
configuration A	GLRLM (2D slice-merged)	Short run high grey level emphasis	320	1	320.4963	0	match
configuration A	GLRLM (2D slice-merged)	Long run low grey level emphasis	0.0385	0.0003	0.038469	0	match
configuration A	GLRLM (2D slice-merged)	Long run high grey level emphasis	1400	20	1399.968	0	match
configuration A	GLRLM (2D slice-merged)	Grey level non-uniformity	1730	10	1728.0207	0	match
configuration A	GLRLM (2D slice-merged)	Normalised grey level non-uniformity	0.128	0.003	0.12841	0	match
configuration A	GLRLM (2D slice-merged)	Run length non-uniformity	6600	30	6603.7152	0	match
configuration A	GLRLM (2D slice-merged)	Normalised run length non-uniformity	0.579	0.003	0.57889	0	match
configuration A	GLRLM (2D slice-merged)	Run percentage	0.704	0.003	0.70415	0	match
configuration A	GLRLM (2D slice-merged)	Grey level variance	33.7	0.6	33.6851	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLRLM (2D slice-merged)	Run length variance	0.826	0.008	0.82596	0	match
configuration A	GLRLM (2D slice-merged)	Run entropy	4.76	0.02	4.7568	0	match
configuration A	GLRLM (2.5D direction-merged)	Short runs emphasis	0.768	0.003	0.76798	0	match
configuration A	GLRLM (2.5D direction-merged)	Long runs emphasis	3.09	0.03	3.0899	0	match
configuration A	GLRLM (2.5D direction-merged)	Low grey level run emphasis	0.0148	0.0004	0.014751	0	match
configuration A	GLRLM (2.5D direction-merged)	High grey level run emphasis	449	3	448.551	0	match
configuration A	GLRLM (2.5D direction-merged)	Short run low grey level emphasis	0.0135	0.0004	0.013511	0	match
configuration A	GLRLM (2.5D direction-merged)	Short run high grey level emphasis	332	1	332.4351	0	match
configuration A	GLRLM (2.5D direction-merged)	Long run low grey level emphasis	0.0229	0.0004	0.022863	0	match
configuration A	GLRLM (2.5D direction-merged)	Long run high grey level emphasis	1500	20	1504.3611	0	match
configuration A	GLRLM (2.5D direction-merged)	Grey level non-uniformity	9850	10	9848.4564	0	match
configuration A	GLRLM (2.5D direction-merged)	Normalised grey level non-uniformity	0.126	0.003	0.12637	0	match
configuration A	GLRLM (2.5D direction-merged)	Run length non-uniformity	42700	200	42749.8939	0	match
configuration A	GLRLM (2.5D direction-merged)	Normalised run length non-uniformity	0.548	0.003	0.54755	0	match
configuration A	GLRLM (2.5D direction-merged)	Run percentage	0.68	0.003	0.6798	0	match
configuration A	GLRLM (2.5D direction-merged)	Grey level variance	29.1	0.6	29.1177	0	match
configuration A	GLRLM (2.5D direction-merged)	Run length variance	0.916	0.011	0.91596	0	match
configuration A	GLRLM (2.5D direction-merged)	Run entropy	4.87	0.01	4.8726	0	match
configuration A	GLRLM (2.5D merged)	Short runs emphasis	0.769	0.003	0.76884	0	match
configuration A	GLRLM (2.5D merged)	Long runs emphasis	3.08	0.03	3.0776	0	match
configuration A	GLRLM (2.5D merged)	Low grey level run emphasis	0.0147	0.0004	0.014739	0	match
configuration A	GLRLM (2.5D merged)	High grey level run emphasis	449	3	448.6109	0	match
configuration A	GLRLM (2.5D merged)	Short run low grey level emphasis	0.0135	0.0004	0.013507	0	match
configuration A	GLRLM (2.5D merged)	Short run high grey level emphasis	333	1	332.908	0	match
configuration A	GLRLM (2.5D merged)	Long run low grey level emphasis	0.0228	0.0004	0.022791	0	match
configuration A	GLRLM (2.5D merged)	Long run high grey level emphasis	1500	20	1498.2951	0	match
configuration A	GLRLM (2.5D merged)	Grey level non-uniformity	39400	100	39391.0509	0	match
configuration A	GLRLM (2.5D merged)	Normalised grey level non-uniformity	0.126	0.003	0.12641	0	match
configuration A	GLRLM (2.5D merged)	Run length non-uniformity	171000	1000	170731.0869	0	match
configuration A	GLRLM (2.5D merged)	Normalised run length non-uniformity	0.548	0.003	0.5479	0	match
configuration A	GLRLM (2.5D merged)	Run percentage	0.68	0.003	0.6798	0	match
configuration A	GLRLM (2.5D merged)	Grey level variance	29.1	0.6	29.1021	0	match
configuration A	GLRLM (2.5D merged)	Run length variance	0.914	0.011	0.91376	0	match
configuration A	GLRLM (2.5D merged)	Run entropy	4.87	0.01	4.8731	0	match
configuration A	GLSZM (2D)	Small zone emphasis	0.688	0.003	0.68783	0	match
configuration A	GLSZM (2D)	Large zone emphasis	625	9	625.4341	0	match
configuration A	GLSZM (2D)	Low grey level emphasis	0.0368	0.0005	0.036792	0	match
configuration A	GLSZM (2D)	High grey level emphasis	363	3	363.3366	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLSZM (2D)	Small zone low grey level emphasis	0.0298	0.0005	0.029777	0	match
configuration A	GLSZM (2D)	Small zone high grey level emphasis	226	1	226.2031	0	match
configuration A	GLSZM (2D)	Large zone low grey level emphasis	1.35	0.03	1.3522	0	match
configuration A	GLSZM (2D)	Large zone high grey level emphasis	316000	5000	315526.6344	0	match
configuration A	GLSZM (2D)	Grey level non-uniformity	82.2	0.1	82.2339	0	match
configuration A	GLSZM (2D)	Normalised grey level non-uniformity	0.0728	0.0014	0.072835	0	match
configuration A	GLSZM (2D)	Zone size non-uniformity	479	4	478.7686	0	match
configuration A	GLSZM (2D)	Normalised zone size non-uniformity	0.44	0.004	0.4405	0	match
configuration A	GLSZM (2D)	Zone percentage	0.3	0.003	0.30007	0	match
configuration A	GLSZM (2D)	Grey level variance	42.7	0.7	42.7127	0	match
configuration A	GLSZM (2D)	Zone size variance	609	9	609.4434	0	match
configuration A	GLSZM (2D)	Zone size entropy	5.92	0.02	5.9216	0	match
configuration A	GLSZM (2.5D)	Small zone emphasis	0.68	0.003	0.67988	0	match
configuration A	GLSZM (2.5D)	Large zone emphasis	675	8	675.3305	0	match
configuration A	GLSZM (2.5D)	Low grey level emphasis	0.0291	0.0005	0.029079	0	match
configuration A	GLSZM (2.5D)	High grey level emphasis	370	3	370.2776	0	match
configuration A	GLSZM (2.5D)	Small zone low grey level emphasis	0.0237	0.0005	0.023711	0	match
configuration A	GLSZM (2.5D)	Small zone high grey level emphasis	229	1	229.2384	0	match
configuration A	GLSZM (2.5D)	Large zone low grey level emphasis	1.44	0.02	1.4358	0	match
configuration A	GLSZM (2.5D)	Large zone high grey level emphasis	338000	5000	337783.4809	0	match
configuration A	GLSZM (2.5D)	Grey level non-uniformity	1800	10	1803.0481	0	match
configuration A	GLSZM (2.5D)	Normalised grey level non-uniformity	0.0622	0.0007	0.062159	0	match
configuration A	GLSZM (2.5D)	Zone size non-uniformity	12400	100	12391.9454	0	match
configuration A	GLSZM (2.5D)	Normalised zone size non-uniformity	0.427	0.004	0.42721	0	match
configuration A	GLSZM (2.5D)	Zone percentage	0.253	0.004	0.25312	0	match
configuration A	GLSZM (2.5D)	Grey level variance	47.9	0.4	47.9397	0	match
configuration A	GLSZM (2.5D)	Zone size variance	660	8	659.723	0	match
configuration A	GLSZM (2.5D)	Zone size entropy	6.39	0.01	6.3862	0	match
configuration A	GLDZM (2D)	Small distance emphasis	0.192	0.006	0.19222	0	match
configuration A	GLDZM (2D)	Large distance emphasis	161	1	160.5272	0	match
configuration A	GLDZM (2D)	Low grey level emphasis	0.0368	0.0005	0.036792	0	match
configuration A	GLDZM (2D)	High grey level emphasis	363	3	363.3366	0	match
configuration A	GLDZM (2D)	Small distance low grey level emphasis	0.00913	0.00023	0.009126	0	match
configuration A	GLDZM (2D)	Small distance high grey level emphasis	60.1	3.3	60.0754	0	match
configuration A	GLDZM (2D)	Large distance low grey level emphasis	2.96	0.02	2.9575	0	match
configuration A	GLDZM (2D)	Large distance high grey level emphasis	70100	100	70106.9591	0	match
configuration A	GLDZM (2D)	Grey level non-uniformity	82.2	0.1	82.2339	0	match
configuration A	GLDZM (2D)	Normalised grey level non-uniformity	0.0728	0.0014	0.072835	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	GLDZM (2D)	Zone distance non-uniformity	64	0.4	63.9755	0	match
configuration A	GLDZM (2D)	Normalised zone distance non-uniformity	0.0716	0.0022	0.071634	0	match
configuration A	GLDZM (2D)	Zone percentage	0.3	0.003	0.30007	0	match
configuration A	GLDZM (2D)	Grey level variance	42.7	0.7	42.7127	0	match
configuration A	GLDZM (2D)	Zone distance variance	69.4	0.1	69.3737	0	match
configuration A	GLDZM (2D)	Zone distance entropy	8	0.04	7.9986	0	match
configuration A	GLDZM (2.5D)	Small distance emphasis	0.168	0.005	0.16832	0	match
configuration A	GLDZM (2.5D)	Large distance emphasis	178	1	178.154	0	match
configuration A	GLDZM (2.5D)	Low grey level emphasis	0.0291	0.0005	0.029079	0	match
configuration A	GLDZM (2.5D)	High grey level emphasis	370	3	370.2776	0	match
configuration A	GLDZM (2.5D)	Small distance low grey level emphasis	0.00788	0.00022	0.0078835	0	match
configuration A	GLDZM (2.5D)	Small distance high grey level emphasis	49.5	2.8	49.4974	0	match
configuration A	GLDZM (2.5D)	Large distance low grey level emphasis	2.31	0.01	2.3108	0	match
configuration A	GLDZM (2.5D)	Large distance high grey level emphasis	79500	100	79522.1637	0	match
configuration A	GLDZM (2.5D)	Grey level non-uniformity	1800	10	1803.0481	0	match
configuration A	GLDZM (2.5D)	Normalised grey level non-uniformity	0.0622	0.0007	0.062159	0	match
configuration A	GLDZM (2.5D)	Zone distance non-uniformity	1570	10	1573.8269	0	match
configuration A	GLDZM (2.5D)	Normalised zone distance non-uniformity	0.0543	0.0014	0.054257	0	match
configuration A	GLDZM (2.5D)	Zone percentage	0.253	0.004	0.25312	0	match
configuration A	GLDZM (2.5D)	Grey level variance	47.9	0.4	47.9397	0	match
configuration A	GLDZM (2.5D)	Zone distance variance	78.9	0.1	78.9261	0	match
configuration A	GLDZM (2.5D)	Zone distance entropy	8.87	0.03	8.8706	0	match
configuration A	NGTDM (2D)	Coarseness	0.00629	0.00046	0.0062923	0	match
configuration A	NGTDM (2D)	Contrast	0.107	0.002	0.10745	0	match
configuration A	NGTDM (2D)	Busyness	0.489	0.001	0.48886	0	match
configuration A	NGTDM (2D)	Complexity	438	9	438.2235	0	match
configuration A	NGTDM (2D)	Strength	3.33	0.08	3.3256	0	match
configuration A	NGTDM (2.5D)	Coarseness	0.0000906	0.0000033	9.06E-05	0	match
configuration A	NGTDM (2.5D)	Contrast	0.0345	0.0009	0.034488	0	match
configuration A	NGTDM (2.5D)	Busyness	8.84	0.01	8.8356	0	match
configuration A	NGTDM (2.5D)	Complexity	580	19	580.0524	0	match
configuration A	NGTDM (2.5D)	Strength	0.0904	0.0027	0.090406	0	match
configuration A	NGLDM (2D)	Low dependence emphasis	0.281	0.003	0.28149	0	match
configuration A	NGLDM (2D)	High dependence emphasis	14.8	0.1	14.757	0	match
configuration A	NGLDM (2D)	Low grey level count emphasis	0.0233	0.0003	0.023299	0	match
configuration A	NGLDM (2D)	High grey level count emphasis	446	2	446.1926	0	match
configuration A	NGLDM (2D)	Low dependence low grey level emphasis	0.0137	0.0002	0.013723	0	match
configuration A	NGLDM (2D)	Low dependence high grey level emphasis	94.2	0.4	94.2438	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration A	NGLDM (2D)	High dependence low grey level emphasis	0.116	0.001	0.11585	0	match
configuration A	NGLDM (2D)	High dependence high grey level emphasis	7540	60	7539.7905	0	match
configuration A	NGLDM (2D)	Grey level non-uniformity	757	1	757.2614	0	match
configuration A	NGLDM (2D)	Normalised grey level non-uniformity	0.151	0.003	0.15142	0	match
configuration A	NGLDM (2D)	Dependence count non-uniformity	709	2	708.7854	0	match
configuration A	NGLDM (2D)	Normalised dependence count non-uniformity	0.175	0.001	0.17529	0	match
configuration A	NGLDM (2D)	Dependence count percentage	1	0			
configuration A	NGLDM (2D)	Grey level variance	31.1	0.5	31.1051	0	match
configuration A	NGLDM (2D)	Dependence count variance	3.12	0.02	3.1244	0	match
configuration A	NGLDM (2D)	Dependence count entropy	5.76	0.02	5.7607	0	match
configuration A	NGLDM (2D)	Dependence count energy	0.0268	0.0004	0.026832	0	match
configuration A	NGLDM (2.5D)	Low dependence emphasis	0.243	0.004	0.24265	0	match
configuration A	NGLDM (2.5D)	High dependence emphasis	16.1	0.2	16.0556	0	match
configuration A	NGLDM (2.5D)	Low grey level count emphasis	0.0115	0.0003	0.011454	0	match
configuration A	NGLDM (2.5D)	High grey level count emphasis	466	2	466.2145	0	match
configuration A	NGLDM (2.5D)	Low dependence low grey level emphasis	0.00664	0.0002	0.0066443	0	match
configuration A	NGLDM (2.5D)	Low dependence high grey level emphasis	91.9	0.5	91.8718	0	match
configuration A	NGLDM (2.5D)	High dependence low grey level emphasis	0.0674	0.0004	0.067386	0	match
configuration A	NGLDM (2.5D)	High dependence high grey level emphasis	8100	60	8097.6139	0	match
configuration A	NGLDM (2.5D)	Grey level non-uniformity	17200	100	17232.3631	0	match
configuration A	NGLDM (2.5D)	Normalised grey level non-uniformity	0.15	0.002	0.15037	0	match
configuration A	NGLDM (2.5D)	Dependence count non-uniformity	17500	100	17519.4394	0	match
configuration A	NGLDM (2.5D)	Normalised dependence count non-uniformity	0.153	0.001	0.15288	0	match
configuration A	NGLDM (2.5D)	Dependence count percentage	1	0			
configuration A	NGLDM (2.5D)	Grey level variance	22.8	0.6	22.8238	0	match
configuration A	NGLDM (2.5D)	Dependence count variance	3.37	0.01	3.3708	0	match
configuration A	NGLDM (2.5D)	Dependence count entropy	5.93	0.02	5.9322	0	match
configuration A	NGLDM (2.5D)	Dependence count energy	0.0245	0.0003	0.024451	0	match
configuration B	Diagnostics-initial image	Image dimension x	204	0	204	0	match
configuration B	Diagnostics-initial image	Image dimension y	201	0	201	0	match
configuration B	Diagnostics-initial image	Image dimension z	60	0	60	0	match
configuration B	Diagnostics-initial image	Voxel dimension x	0.977	0	0.977	0	match
configuration B	Diagnostics-initial image	Voxel dimension y	0.977	0	0.977	0	match
configuration B	Diagnostics-initial image	Voxel dimension z	3	0	3	0	match
configuration B	Diagnostics-initial image	Mean intensity	-266	0	-266.4704	0	match
configuration B	Diagnostics-initial image	Minimum intensity	-1000	0	-1000	0	match
configuration B	Diagnostics-initial image	Maximum intensity	3065	0	3065	0	match
configuration B	Diagnostics-interpolated image	Image dimension x	100	1	100	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	Diagnostics-interpolated image	Image dimension y	99	0.8	99	0	match
configuration B	Diagnostics-interpolated image	Image dimension z	60	0	60	0	match
configuration B	Diagnostics-interpolated image	Voxel dimension x	2	0	2	0	match
configuration B	Diagnostics-interpolated image	Voxel dimension y	2	0	2	0	match
configuration B	Diagnostics-interpolated image	Voxel dimension z	3	0	3	0	match
configuration B	Diagnostics-interpolated image	Mean intensity	-270	3	-263.3183	6	partial match
configuration B	Diagnostics-interpolated image	Minimum intensity	-1000	0	-1000	0	match
configuration B	Diagnostics-interpolated image	Maximum intensity	2257	30	2257	0	match
configuration B	Diagnostics-initial ROI	Int. mask dimension x	204	0	204	0	match
configuration B	Diagnostics-initial ROI	Int. mask dimension y	201	0	201	0	match
configuration B	Diagnostics-initial ROI	Int. mask dimension z	60	0	60	0	match
configuration B	Diagnostics-initial ROI	Int. mask bounding box dimension x	100	0	100	0	match
configuration B	Diagnostics-initial ROI	Int. mask bounding box dimension y	99	0	99	0	match
configuration B	Diagnostics-initial ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-initial ROI	Morph. mask bounding box dimension x	100	0	100	0	match
configuration B	Diagnostics-initial ROI	Morph. mask bounding box dimension y	99	0	99	0	match
configuration B	Diagnostics-initial ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-initial ROI	Int. mask voxel count	125256	0	125256	0	match
configuration B	Diagnostics-initial ROI	Morph. mask voxel count	125256	0	125256	0	match
configuration B	Diagnostics-initial ROI	Int. mask mean intensity	-46.9	0	-46.8827	0	match
configuration B	Diagnostics-initial ROI	Int. mask minimum intensity	-1000	0	-1000	0	match
configuration B	Diagnostics-initial ROI	Int. mask maximum intensity	723	0	723	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask dimension x	100	1	100	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask dimension y	99	0.8	99	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask dimension z	60	0	60	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask bounding box dimension x	49	0.2	49	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-interpolated ROI	Morph. mask bounding box dimension x	49	0.2	49	0	match
configuration B	Diagnostics-interpolated ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration B	Diagnostics-interpolated ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask voxel count	29842	100	29842	0	match
configuration B	Diagnostics-interpolated ROI	Morph. mask voxel count	29842	100	29842	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask mean intensity	-47	0.1	-46.9811	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask minimum intensity	-956	1	-956	0	match
configuration B	Diagnostics-interpolated ROI	Int. mask maximum intensity	525	6	525	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask dimension x	100	1	100	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	Diagnostics-resegmented ROI	Int. mask dimension y	99	0.8	99	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask dimension z	60	0	60	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask bounding box dimension x	49	0.3	49	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-resegmented ROI	Morph. mask bounding box dimension x	49	0.3	49	0	match
configuration B	Diagnostics-resegmented ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration B	Diagnostics-resegmented ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask voxel count	27359	300	27359	0	match
configuration B	Diagnostics-resegmented ROI	Morph. mask voxel count	29842	400	29842	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask mean intensity	11.5	1.1	11.5301	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask minimum intensity	-500	0	-500	0	match
configuration B	Diagnostics-resegmented ROI	Int. mask maximum intensity	391	9	391	0	match
configuration B	Morphology	Volume (mesh)	358000	5000	357502.5	0	match
configuration B	Morphology	Volume (voxel counting)	358000	5000	358104	0	match
configuration B	Morphology	Surface area (mesh)	33700	300	33731.1415	0	match
configuration B	Morphology	Surface to volume ratio	0.0944	0.0005	0.094352	0	match
configuration B	Morphology	Compactness 1	0.0326	0.0001	0.032558	0	match
configuration B	Morphology	Compactness 2	0.377	0.001	0.37663	0	match
configuration B	Morphology	Spherical disproportion	1.38	0.01	1.3847	0	match
configuration B	Morphology	Sphericity	0.722	0.001	0.72217	0	match
configuration B	Morphology	Asphericity	0.385	0.001	0.38472	0	match
configuration B	Morphology	Centre of mass shift	63.1	29.6	63.141	0	match
configuration B	Morphology	Maximum 3D diameter	125	1	125.06	0	match
configuration B	Morphology	Major axis length	92.6	0.4	92.6019	0	match
configuration B	Morphology	Minor axis length	81.3	0.4	81.3262	0	match
configuration B	Morphology	Least axis length	70.2	0.3	70.1757	0	match
configuration B	Morphology	Elongation	0.878	0.001	0.87823	0	match
configuration B	Morphology	Flatness	0.758	0.001	0.75782	0	match
configuration B	Morphology	Volume density (AABB)	0.477	0.003	0.47724	0	match
configuration B	Morphology	Area density (AABB)	0.678	0.003	0.67755	0	match
configuration B	Morphology	Volume density (OMBB)					
configuration B	Morphology	Area density (OMBB)					
configuration B	Morphology	Volume density (AEE)	1.29	0.01	1.2919	0	match
configuration B	Morphology	Area density (AEE)	1.62	0.01	1.6164	0	match
configuration B	Morphology	Volume density (MVEE)					
configuration B	Morphology	Area density (MVEE)					
configuration B	Morphology	Volume density (convex hull)	0.829	0.001	0.8294	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	Morphology	Area density (convex hull)	1.12	0.01	1.1242	0	match
configuration B	Morphology	Integrated intensity	4120000	320000	4122026.431	0	match
configuration B	Morphology	Moran’s I index	0.0329	0.0001			
configuration B	Morphology	Geary’s C measure	0.862	0.001			
configuration B	Local intensity	Local intensity peak	178	10			
configuration B	Local intensity	Global intensity peak	178	5			
configuration B	Statistics	Mean	11.5	1.1	11.5301	0	match
configuration B	Statistics	Variance	14400	400	14415.6649	0	match
configuration B	Statistics	Skewness	-2.49	0.05	-2.4897	0	match
configuration B	Statistics	(Excess) kurtosis	5.93	0.24	5.9297	0	match
configuration B	Statistics	Median	45	0.3	45	0	match
configuration B	Statistics	Minimum	-500	0	-500	0	match
configuration B	Statistics	10th percentile	-136	8	-136	0	match
configuration B	Statistics	90th percentile	91	0	91	0	match
configuration B	Statistics	Maximum	391	9	391	0	match
configuration B	Statistics	Interquartile range	52	0.5	52	0	match
configuration B	Statistics	Range	891	9	891	0	match
configuration B	Statistics	Mean absolute deviation	74.4	1.4	74.4142	0	match
configuration B	Statistics	Robust mean absolute deviation	27.3	0.8	27.2727	0	match
configuration B	Statistics	Median absolute deviation	63.8	1	63.7862	0	match
configuration B	Statistics	Coefficient of variation	10.4	5.2	10.4132	0	match
configuration B	Statistics	Quartile coefficient of dispersion	0.591	0.008	0.59091	0	match
configuration B	Statistics	Energy	398000000	11000000	398035345	0	match
configuration B	Statistics	Root mean square	121	2	120.6176	0	match
configuration B	Intensity histogram	Mean	18.9	0.3	18.8728	0	match
configuration B	Intensity histogram	Variance	18.7	0.2	18.6643	0	match
configuration B	Intensity histogram	Skewness	-2.47	0.05	-2.466	0	match
configuration B	Intensity histogram	(Excess) kurtosis	5.84	0.24	5.8394	0	match
configuration B	Intensity histogram	Median	20	0.3	20	0	match
configuration B	Intensity histogram	Minimum	1	0	1	0	match
configuration B	Intensity histogram	10th percentile	14	0.5	14	0	match
configuration B	Intensity histogram	90th percentile	22	0.3	22	0	match
configuration B	Intensity histogram	Maximum	32	0	32	0	match
configuration B	Intensity histogram	Mode	20	0.3	20	0	match
configuration B	Intensity histogram	Interquartile range	2	0	2	0	match
configuration B	Intensity histogram	Range	31	0	31	0	match
configuration B	Intensity histogram	Mean absolute deviation	2.67	0.03	2.6693	0	match
configuration B	Intensity histogram	Robust mean absolute deviation	1.03	0.03	1.0323	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	Intensity histogram	Median absolute deviation	2.28	0.02	2.281	0	match
configuration B	Intensity histogram	Coefficient of variation	0.229	0.004	0.22891	0	match
configuration B	Intensity histogram	Quartile coefficient of dispersion	0.05	0.0005	0.05	0	match
configuration B	Intensity histogram	Entropy	3.16	0.01	3.1562	0	match
configuration B	Intensity histogram	Uniformity	0.174	0.001	0.17408	0	match
configuration B	Intensity histogram	Maximum histogram gradient	3220	50	3224	0	match
configuration B	Intensity histogram	Maximum histogram gradient intensity	19	0.3	19	0	match
configuration B	Intensity histogram	Minimum histogram gradient	-3020	50	-3017.5	0	match
configuration B	Intensity histogram	Minimum histogram gradient intensity	22	0.3	22	0	match
configuration B	Intensity volume histogram	Volume fraction at 10% intensity	0.977	0.001	0.97697	0	match
configuration B	Intensity volume histogram	Volume fraction at 90% intensity	0.0000731	0.0000103	7.31E-05	0	match
configuration B	Intensity volume histogram	Intensity at 10% volume	92	0	92	0	match
configuration B	Intensity volume histogram	Intensity at 90% volume	-135	8	-135	0	match
configuration B	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.977	0.001	0.9769	0	match
configuration B	Intensity volume histogram	Intensity difference between 10% and 90% volume	227	8	227	0	match
configuration B	Intensity volume histogram	Area under the IVH curve			0.57467		
configuration B	GLCM (2D averaged)	Joint maximum	0.156	0.002	0.15602	0	match
configuration B	GLCM (2D averaged)	Joint average	18.7	0.3	18.6827	0	match
configuration B	GLCM (2D averaged)	Joint variance	21	0.3	20.9947	0	match
configuration B	GLCM (2D averaged)	Joint entropy	5.26	0.02	5.2568	0	match
configuration B	GLCM (2D averaged)	Difference average	1.81	0.01	1.8142	0	match
configuration B	GLCM (2D averaged)	Difference variance	7.74	0.05	7.7412	0	match
configuration B	GLCM (2D averaged)	Difference entropy	2.35	0.01	2.354	0	match
configuration B	GLCM (2D averaged)	Sum average	37.4	0.5	37.3653	0	match
configuration B	GLCM (2D averaged)	Sum variance	72.1	1	72.0944	0	match
configuration B	GLCM (2D averaged)	Sum entropy	3.83	0.01	3.8281	0	match
configuration B	GLCM (2D averaged)	Angular second moment	0.0678	0.0006	0.067815	0	match
configuration B	GLCM (2D averaged)	Contrast	11.9	0.1	11.8844	0	match
configuration B	GLCM (2D averaged)	Dissimilarity	1.81	0.01	1.8142	0	match
configuration B	GLCM (2D averaged)	Inverse difference	0.592	0.001	0.59239	0	match
configuration B	GLCM (2D averaged)	Normalised inverse difference	0.952	0.001	0.95214	0	match
configuration B	GLCM (2D averaged)	Inverse difference moment	0.557	0.001	0.55739	0	match
configuration B	GLCM (2D averaged)	Normalised inverse difference moment	0.99	0.001	0.98971	0	match
configuration B	GLCM (2D averaged)	Inverse variance	0.401	0.002	0.40129	0	match
configuration B	GLCM (2D averaged)	Correlation	0.577	0.002	0.57676	0	match
configuration B	GLCM (2D averaged)	Autocorrelation	369	11	368.6717	0	match
configuration B	GLCM (2D averaged)	Cluster tendency	72.1	1	72.0944	0	match
configuration B	GLCM (2D averaged)	Cluster shade	-668	17	-667.6259	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLCM (2D averaged)	Cluster prominence	29400	1400	29358.5221	0	match
configuration B	GLCM (2D averaged)	Information correlation 1	-0.239	0.001	-0.23864	0	match
configuration B	GLCM (2D averaged)	Information correlation 2	0.837	0.001	0.83672	0	match
configuration B	GLCM (2D slice-merged)	Joint maximum	0.156	0.002	0.15566	0	match
configuration B	GLCM (2D slice-merged)	Joint average	18.7	0.3	18.679	0	match
configuration B	GLCM (2D slice-merged)	Joint variance	21	0.3	21.0423	0	match
configuration B	GLCM (2D slice-merged)	Joint entropy	5.45	0.01	5.4522	0	match
configuration B	GLCM (2D slice-merged)	Difference average	1.81	0.01	1.8103	0	match
configuration B	GLCM (2D slice-merged)	Difference variance	7.76	0.05	7.7578	0	match
configuration B	GLCM (2D slice-merged)	Difference entropy	2.38	0.01	2.3832	0	match
configuration B	GLCM (2D slice-merged)	Sum average	37.4	0.5	37.358	0	match
configuration B	GLCM (2D slice-merged)	Sum variance	72.3	1	72.3373	0	match
configuration B	GLCM (2D slice-merged)	Sum entropy	3.89	0.01	3.8913	0	match
configuration B	GLCM (2D slice-merged)	Angular second moment	0.0669	0.0006	0.066865	0	match
configuration B	GLCM (2D slice-merged)	Contrast	11.8	0.1	11.832	0	match
configuration B	GLCM (2D slice-merged)	Dissimilarity	1.81	0.01	1.8103	0	match
configuration B	GLCM (2D slice-merged)	Inverse difference	0.593	0.001	0.59269	0	match
configuration B	GLCM (2D slice-merged)	Normalised inverse difference	0.952	0.001	0.95222	0	match
configuration B	GLCM (2D slice-merged)	Inverse difference moment	0.558	0.001	0.55772	0	match
configuration B	GLCM (2D slice-merged)	Normalised inverse difference moment	0.99	0.001	0.98975	0	match
configuration B	GLCM (2D slice-merged)	Inverse variance	0.401	0.002	0.40137	0	match
configuration B	GLCM (2D slice-merged)	Correlation	0.58	0.002	0.58046	0	match
configuration B	GLCM (2D slice-merged)	Autocorrelation	369	11	368.5971	0	match
configuration B	GLCM (2D slice-merged)	Cluster tendency	72.3	1	72.3373	0	match
configuration B	GLCM (2D slice-merged)	Cluster shade	-673	17	-672.7734	0	match
configuration B	GLCM (2D slice-merged)	Cluster prominence	29500	1400	29522.4793	0	match
configuration B	GLCM (2D slice-merged)	Information correlation 1	-0.181	0.001	-0.1812	0	match
configuration B	GLCM (2D slice-merged)	Information correlation 2	0.792	0.001	0.79224	0	match
configuration B	GLCM (2.5D direction-merged)	Joint maximum	0.126	0.002	0.12607	0	match
configuration B	GLCM (2.5D direction-merged)	Joint average	19.2	0.3	19.2324	0	match
configuration B	GLCM (2.5D direction-merged)	Joint variance	14.2	0.1	14.1787	0	match
configuration B	GLCM (2.5D direction-merged)	Joint entropy	5.45	0.01	5.4538	0	match
configuration B	GLCM (2.5D direction-merged)	Difference average	1.47	0.01	1.4737	0	match
configuration B	GLCM (2.5D direction-merged)	Difference variance	6.48	0.06	6.4753	0	match
configuration B	GLCM (2.5D direction-merged)	Difference entropy	2.24	0.01	2.2351	0	match
configuration B	GLCM (2.5D direction-merged)	Sum average	38.5	0.6	38.4648	0	match
configuration B	GLCM (2.5D direction-merged)	Sum variance	48.1	0.4	48.0556	0	match
configuration B	GLCM (2.5D direction-merged)	Sum entropy	3.91	0.01	3.9123	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLCM (2.5D direction-merged)	Angular second moment	0.0581	0.0006	0.058075	0	match
configuration B	GLCM (2.5D direction-merged)	Contrast	8.66	0.09	8.6591	0	match
configuration B	GLCM (2.5D direction-merged)	Dissimilarity	1.47	0.01	1.4737	0	match
configuration B	GLCM (2.5D direction-merged)	Inverse difference	0.628	0.001	0.62775	0	match
configuration B	GLCM (2.5D direction-merged)	Normalised inverse difference	0.96	0.001	0.96044	0	match
configuration B	GLCM (2.5D direction-merged)	Inverse difference moment	0.6	0.001	0.59972	0	match
configuration B	GLCM (2.5D direction-merged)	Normalised inverse difference moment	0.992	0.001	0.99246	0	match
configuration B	GLCM (2.5D direction-merged)	Inverse variance	0.424	0.003	0.42367	0	match
configuration B	GLCM (2.5D direction-merged)	Correlation	0.693	0.003	0.69321	0	match
configuration B	GLCM (2.5D direction-merged)	Autocorrelation	380	11	379.7348	0	match
configuration B	GLCM (2.5D direction-merged)	Cluster tendency	48.1	0.4	48.0556	0	match
configuration B	GLCM (2.5D direction-merged)	Cluster shade	-905	19	-904.8024	0	match
configuration B	GLCM (2.5D direction-merged)	Cluster prominence	25200	1000	25213.964	0	match
configuration B	GLCM (2.5D direction-merged)	Information correlation 1	-0.188	0.001	-0.18808	0	match
configuration B	GLCM (2.5D direction-merged)	Information correlation 2	0.821	0.001	0.82064	0	match
configuration B	GLCM (2.5D merged)	Joint maximum	0.126	0.002	0.12609	0	match
configuration B	GLCM (2.5D merged)	Joint average	19.2	0.3	19.2321	0	match
configuration B	GLCM (2.5D merged)	Joint variance	14.2	0.1	14.1837	0	match
configuration B	GLCM (2.5D merged)	Joint entropy	5.46	0.01	5.4645	0	match
configuration B	GLCM (2.5D merged)	Difference average	1.47	0.01	1.4726	0	match
configuration B	GLCM (2.5D merged)	Difference variance	6.48	0.06	6.4777	0	match
configuration B	GLCM (2.5D merged)	Difference entropy	2.24	0.01	2.2364	0	match
configuration B	GLCM (2.5D merged)	Sum average	38.5	0.6	38.4641	0	match
configuration B	GLCM (2.5D merged)	Sum variance	48.1	0.4	48.0886	0	match
configuration B	GLCM (2.5D merged)	Sum entropy	3.91	0.01	3.914	0	match
configuration B	GLCM (2.5D merged)	Angular second moment	0.058	0.0006	0.05804	0	match
configuration B	GLCM (2.5D merged)	Contrast	8.65	0.09	8.6461	0	match
configuration B	GLCM (2.5D merged)	Dissimilarity	1.47	0.01	1.4726	0	match
configuration B	GLCM (2.5D merged)	Inverse difference	0.628	0.001	0.62786	0	match
configuration B	GLCM (2.5D merged)	Normalised inverse difference	0.96	0.001	0.96047	0	match
configuration B	GLCM (2.5D merged)	Inverse difference moment	0.6	0.001	0.59984	0	match
configuration B	GLCM (2.5D merged)	Normalised inverse difference moment	0.992	0.001	0.99247	0	match
configuration B	GLCM (2.5D merged)	Inverse variance	0.424	0.003	0.42369	0	match
configuration B	GLCM (2.5D merged)	Correlation	0.695	0.003	0.69521	0	match
configuration B	GLCM (2.5D merged)	Autocorrelation	380	11	379.7324	0	match
configuration B	GLCM (2.5D merged)	Cluster tendency	48.1	0.4	48.0886	0	match
configuration B	GLCM (2.5D merged)	Cluster shade	-906	19	-906.1285	0	match
configuration B	GLCM (2.5D merged)	Cluster prominence	25300	1000	25259.3038	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLCM (2.5D merged)	Information correlation 1	-0.185	0.001	-0.18487	0	match
configuration B	GLCM (2.5D merged)	Information correlation 2	0.819	0.001	0.81943	0	match
configuration B	GLRLM (2D averaged)	Short runs emphasis	0.781	0.001	0.78141	0	match
configuration B	GLRLM (2D averaged)	Long runs emphasis	3.52	0.04	3.5201	0	match
configuration B	GLRLM (2D averaged)	Low grey level run emphasis	0.0331	0.0006	0.033101	0	match
configuration B	GLRLM (2D averaged)	High grey level run emphasis	342	11	342.0131	0	match
configuration B	GLRLM (2D averaged)	Short run low grey level emphasis	0.0314	0.0006	0.031375	0	match
configuration B	GLRLM (2D averaged)	Short run high grey level emphasis	251	8	251.4374	0	match
configuration B	GLRLM (2D averaged)	Long run low grey level emphasis	0.0443	0.0008	0.044337	0	match
configuration B	GLRLM (2D averaged)	Long run high grey level emphasis	1390	30	1391.3359	0	match
configuration B	GLRLM (2D averaged)	Grey level non-uniformity	107	1	106.9443	0	match
configuration B	GLRLM (2D averaged)	Normalised grey level non-uniformity	0.145	0.001	0.14467	0	match
configuration B	GLRLM (2D averaged)	Run length non-uniformity	365	3	365.0723	0	match
configuration B	GLRLM (2D averaged)	Normalised run length non-uniformity	0.578	0.001	0.57837	0	match
configuration B	GLRLM (2D averaged)	Run percentage	0.681	0.002	0.68054	0	match
configuration B	GLRLM (2D averaged)	Grey level variance	28.3	0.3	28.2575	0	match
configuration B	GLRLM (2D averaged)	Run length variance	1.22	0.03	1.2152	0	match
configuration B	GLRLM (2D averaged)	Run entropy	4.53	0.02	4.5272	0	match
configuration B	GLRLM (2D slice-merged)	Short runs emphasis	0.782	0.001	0.78231	0	match
configuration B	GLRLM (2D slice-merged)	Long runs emphasis	3.5	0.04	3.4998	0	match
configuration B	GLRLM (2D slice-merged)	Low grey level run emphasis	0.033	0.0006	0.033028	0	match
configuration B	GLRLM (2D slice-merged)	High grey level run emphasis	342	11	342.1562	0	match
configuration B	GLRLM (2D slice-merged)	Short run low grey level emphasis	0.0313	0.0006	0.031298	0	match
configuration B	GLRLM (2D slice-merged)	Short run high grey level emphasis	252	8	251.9847	0	match
configuration B	GLRLM (2D slice-merged)	Long run low grey level emphasis	0.0442	0.0008	0.044236	0	match
configuration B	GLRLM (2D slice-merged)	Long run high grey level emphasis	1380	30	1381.9563	0	match
configuration B	GLRLM (2D slice-merged)	Grey level non-uniformity	427	1	427.4716	0	match
configuration B	GLRLM (2D slice-merged)	Normalised grey level non-uniformity	0.145	0.001	0.14455	0	match
configuration B	GLRLM (2D slice-merged)	Run length non-uniformity	1460	10	1458.4825	0	match
configuration B	GLRLM (2D slice-merged)	Normalised run length non-uniformity	0.578	0.001	0.57838	0	match
configuration B	GLRLM (2D slice-merged)	Run percentage	0.681	0.002	0.68054	0	match
configuration B	GLRLM (2D slice-merged)	Grey level variance	28.3	0.3	28.2742	0	match
configuration B	GLRLM (2D slice-merged)	Run length variance	1.21	0.03	1.2093	0	match
configuration B	GLRLM (2D slice-merged)	Run entropy	4.58	0.01	4.5762	0	match
configuration B	GLRLM (2.5D direction-merged)	Short runs emphasis	0.759	0.001	0.75883	0	match
configuration B	GLRLM (2.5D direction-merged)	Long runs emphasis	3.82	0.05	3.8177	0	match
configuration B	GLRLM (2.5D direction-merged)	Low grey level run emphasis	0.0194	0.0006	0.019413	0	match
configuration B	GLRLM (2.5D direction-merged)	High grey level run emphasis	356	11	355.9821	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLRLM (2.5D direction-merged)	Short run low grey level emphasis	0.0181	0.0006	0.018142	0	match
configuration B	GLRLM (2.5D direction-merged)	Short run high grey level emphasis	257	9	257.4901	0	match
configuration B	GLRLM (2.5D direction-merged)	Long run low grey level emphasis	0.0293	0.0009	0.029278	0	match
configuration B	GLRLM (2.5D direction-merged)	Long run high grey level emphasis	1500	30	1503.9619	0	match
configuration B	GLRLM (2.5D direction-merged)	Grey level non-uniformity	2400	10	2398.929	0	match
configuration B	GLRLM (2.5D direction-merged)	Normalised grey level non-uniformity	0.137	0.001	0.13652	0	match
configuration B	GLRLM (2.5D direction-merged)	Run length non-uniformity	9380	70	9380.1306	0	match
configuration B	GLRLM (2.5D direction-merged)	Normalised run length non-uniformity	0.533	0.001	0.53349	0	match
configuration B	GLRLM (2.5D direction-merged)	Run percentage	0.642	0.002	0.64217	0	match
configuration B	GLRLM (2.5D direction-merged)	Grey level variance	25.7	0.2	25.6535	0	match
configuration B	GLRLM (2.5D direction-merged)	Run length variance	1.39	0.03	1.3885	0	match
configuration B	GLRLM (2.5D direction-merged)	Run entropy	4.84	0.01	4.8366	0	match
configuration B	GLRLM (2.5D merged)	Short runs emphasis	0.759	0.001	0.75911	0	match
configuration B	GLRLM (2.5D merged)	Long runs emphasis	3.81	0.05	3.8111	0	match
configuration B	GLRLM (2.5D merged)	Low grey level run emphasis	0.0194	0.0006	0.019407	0	match
configuration B	GLRLM (2.5D merged)	High grey level run emphasis	356	11	356.008	0	match
configuration B	GLRLM (2.5D merged)	Short run low grey level emphasis	0.0181	0.0006	0.018139	0	match
configuration B	GLRLM (2.5D merged)	Short run high grey level emphasis	258	9	257.6241	0	match
configuration B	GLRLM (2.5D merged)	Long run low grey level emphasis	0.0292	0.0009	0.029241	0	match
configuration B	GLRLM (2.5D merged)	Long run high grey level emphasis	1500	30	1501.3439	0	match
configuration B	GLRLM (2.5D merged)	Grey level non-uniformity	9600	20	9595.2206	0	match
configuration B	GLRLM (2.5D merged)	Normalised grey level non-uniformity	0.137	0.001	0.13653	0	match
configuration B	GLRLM (2.5D merged)	Run length non-uniformity	37500	300	37503.0795	0	match
configuration B	GLRLM (2.5D merged)	Normalised run length non-uniformity	0.534	0.001	0.53365	0	match
configuration B	GLRLM (2.5D merged)	Run percentage	0.642	0.002	0.64217	0	match
configuration B	GLRLM (2.5D merged)	Grey level variance	25.7	0.2	25.6502	0	match
configuration B	GLRLM (2.5D merged)	Run length variance	1.39	0.03	1.3862	0	match
configuration B	GLRLM (2.5D merged)	Run entropy	4.84	0.01	4.8406	0	match
configuration B	GLSZM (2D)	Small zone emphasis	0.745	0.003	0.74514	0	match
configuration B	GLSZM (2D)	Large zone emphasis	439	8	439.4119	0	match
configuration B	GLSZM (2D)	Low grey level emphasis	0.0475	0.001	0.047462	0	match
configuration B	GLSZM (2D)	High grey level emphasis	284	11	284.3636	0	match
configuration B	GLSZM (2D)	Small zone low grey level emphasis	0.0415	0.0008	0.041528	0	match
configuration B	GLSZM (2D)	Small zone high grey level emphasis	190	7	189.9456	0	match
configuration B	GLSZM (2D)	Large zone low grey level emphasis	1.15	0.04	1.1477	0	match
configuration B	GLSZM (2D)	Large zone high grey level emphasis	181000	3000	181366.4615	0	match
configuration B	GLSZM (2D)	Grey level non-uniformity	20.5	0.1	20.4698	0	match
configuration B	GLSZM (2D)	Normalised grey level non-uniformity	0.0789	0.001	0.078898	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLSZM (2D)	Zone size non-uniformity	140	3	140.4053	0	match
configuration B	GLSZM (2D)	Normalised zone size non-uniformity	0.521	0.004	0.52139	0	match
configuration B	GLSZM (2D)	Zone percentage	0.324	0.001	0.32449	0	match
configuration B	GLSZM (2D)	Grey level variance	36.1	0.3	36.0879	0	match
configuration B	GLSZM (2D)	Zone size variance	423	8	422.881	0	match
configuration B	GLSZM (2D)	Zone size entropy	5.29	0.01	5.294	0	match
configuration B	GLSZM (2.5D)	Small zone emphasis	0.741	0.003	0.74081	0	match
configuration B	GLSZM (2.5D)	Large zone emphasis	444	8	443.6457	0	match
configuration B	GLSZM (2.5D)	Low grey level emphasis	0.0387	0.001	0.038718	0	match
configuration B	GLSZM (2.5D)	High grey level emphasis	284	11	283.8242	0	match
configuration B	GLSZM (2.5D)	Small zone low grey level emphasis	0.0335	0.0009	0.033509	0	match
configuration B	GLSZM (2.5D)	Small zone high grey level emphasis	190	7	190.2202	0	match
configuration B	GLSZM (2.5D)	Large zone low grey level emphasis	1.16	0.04	1.1575	0	match
configuration B	GLSZM (2.5D)	Large zone high grey level emphasis	181000	3000	181212.224	0	match
configuration B	GLSZM (2.5D)	Grey level non-uniformity	437	3	436.8697	0	match
configuration B	GLSZM (2.5D)	Normalised grey level non-uniformity	0.0613	0.0005	0.061341	0	match
configuration B	GLSZM (2.5D)	Zone size non-uniformity	3630	70	3627.0716	0	match
configuration B	GLSZM (2.5D)	Normalised zone size non-uniformity	0.509	0.004	0.50928	0	match
configuration B	GLSZM (2.5D)	Zone percentage	0.26	0.002	0.26032	0	match
configuration B	GLSZM (2.5D)	Grey level variance	41	0.7	41.0299	0	match
configuration B	GLSZM (2.5D)	Zone size variance	429	8	428.8888	0	match
configuration B	GLSZM (2.5D)	Zone size entropy	5.98	0.02	5.9791	0	match
configuration B	GLDZM (2D)	Small distance emphasis	0.36	0.005	0.35975	0	match
configuration B	GLDZM (2D)	Large distance emphasis	31.6	0.2	31.5958	0	match
configuration B	GLDZM (2D)	Low grey level emphasis	0.0475	0.001	0.047462	0	match
configuration B	GLDZM (2D)	High grey level emphasis	284	11	284.3636	0	match
configuration B	GLDZM (2D)	Small distance low grey level emphasis	0.0192	0.0005	0.01915	0	match
configuration B	GLDZM (2D)	Small distance high grey level emphasis	95.7	5.5	95.7344	0	match
configuration B	GLDZM (2D)	Large distance low grey level emphasis	0.934	0.018	0.93353	0	match
configuration B	GLDZM (2D)	Large distance high grey level emphasis	10600	300	10573.8333	0	match
configuration B	GLDZM (2D)	Grey level non-uniformity	20.5	0.1	20.4698	0	match
configuration B	GLDZM (2D)	Normalised grey level non-uniformity	0.0789	0.001	0.078898	0	match
configuration B	GLDZM (2D)	Zone distance non-uniformity	39.8	0.3	39.7715	0	match
configuration B	GLDZM (2D)	Normalised zone distance non-uniformity	0.174	0.003	0.1744	0	match
configuration B	GLDZM (2D)	Zone percentage	0.324	0.001	0.32449	0	match
configuration B	GLDZM (2D)	Grey level variance	36.1	0.3	36.0879	0	match
configuration B	GLDZM (2D)	Zone distance variance	13.5	0.1	13.4623	0	match
configuration B	GLDZM (2D)	Zone distance entropy	6.47	0.03	6.4739	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	GLDZM (2.5D)	Small distance emphasis	0.329	0.004	0.3285	0	match
configuration B	GLDZM (2.5D)	Large distance emphasis	34.3	0.2	34.3385	0	match
configuration B	GLDZM (2.5D)	Low grey level emphasis	0.0387	0.001	0.038718	0	match
configuration B	GLDZM (2.5D)	High grey level emphasis	284	11	283.8242	0	match
configuration B	GLDZM (2.5D)	Small distance low grey level emphasis	0.0168	0.0005	0.016756	0	match
configuration B	GLDZM (2.5D)	Small distance high grey level emphasis	81.4	4.6	81.357	0	match
configuration B	GLDZM (2.5D)	Large distance low grey level emphasis	0.748	0.017	0.74755	0	match
configuration B	GLDZM (2.5D)	Large distance high grey level emphasis	11600	400	11639.3374	0	match
configuration B	GLDZM (2.5D)	Grey level non-uniformity	437	3	436.8697	0	match
configuration B	GLDZM (2.5D)	Normalised grey level non-uniformity	0.0613	0.0005	0.061341	0	match
configuration B	GLDZM (2.5D)	Zone distance non-uniformity	963	6	962.5015	0	match
configuration B	GLDZM (2.5D)	Normalised zone distance non-uniformity	0.135	0.001	0.13514	0	match
configuration B	GLDZM (2.5D)	Zone percentage	0.26	0.002	0.26032	0	match
configuration B	GLDZM (2.5D)	Grey level variance	41	0.7	41.0299	0	match
configuration B	GLDZM (2.5D)	Zone distance variance	15	0.1	15.0252	0	match
configuration B	GLDZM (2.5D)	Zone distance entropy	7.58	0.01	7.584	0	match
configuration B	NGTDM (2D)	Coarseness	0.0168	0.0005	0.016767	0	match
configuration B	NGTDM (2D)	Contrast	0.181	0.001	0.18067	0	match
configuration B	NGTDM (2D)	Busyness	0.2	0.005	0.20002	0	match
configuration B	NGTDM (2D)	Complexity	391	7	390.5958	0	match
configuration B	NGTDM (2D)	Strength	6.02	0.23	6.0195	0	match
configuration B	NGTDM (2.5D)	Coarseness	0.000314	0.000004	0.00031367	0	match
configuration B	NGTDM (2.5D)	Contrast	0.0506	0.0005	0.050645	0	match
configuration B	NGTDM (2.5D)	Busyness	3.45	0.07	3.4549	0	match
configuration B	NGTDM (2.5D)	Complexity	496	5	496.4973	0	match
configuration B	NGTDM (2.5D)	Strength	0.199	0.009	0.19891	0	match
configuration B	NGLDM (2D)	Low dependence emphasis	0.31	0.001	0.30955	0	match
configuration B	NGLDM (2D)	High dependence emphasis	17.3	0.2	17.3467	0	match
configuration B	NGLDM (2D)	Low grey level count emphasis	0.0286	0.0004	0.028578	0	match
configuration B	NGLDM (2D)	High grey level count emphasis	359	10	358.6691	0	match
configuration B	NGLDM (2D)	Low dependence low grey level emphasis	0.0203	0.0003	0.020319	0	match
configuration B	NGLDM (2D)	Low dependence high grey level emphasis	78.9	2.2	78.8506	0	match
configuration B	NGLDM (2D)	High dependence low grey level emphasis	0.108	0.003	0.10829	0	match
configuration B	NGLDM (2D)	High dependence high grey level emphasis	7210	130	7208.9974	0	match
configuration B	NGLDM (2D)	Grey level non-uniformity	216	3	216.1766	0	match
configuration B	NGLDM (2D)	Normalised grey level non-uniformity	0.184	0.001	0.18423	0	match
configuration B	NGLDM (2D)	Dependence count non-uniformity	157	1	157.3368	0	match
configuration B	NGLDM (2D)	Normalised dependence count non-uniformity	0.179	0.001	0.17855	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration B	NGLDM (2D)	Dependence count percentage	1	0			
configuration B	NGLDM (2D)	Grey level variance	25.3	0.4	25.2791	0	match
configuration B	NGLDM (2D)	Dependence count variance	4.02	0.05	4.0194	0	match
configuration B	NGLDM (2D)	Dependence count entropy	5.38	0.01	5.3786	0	match
configuration B	NGLDM (2D)	Dependence count energy	0.0321	0.0002	0.032145	0	match
configuration B	NGLDM (2.5D)	Low dependence emphasis	0.254	0.002	0.25356	0	match
configuration B	NGLDM (2.5D)	High dependence emphasis	19.6	0.2	19.5517	0	match
configuration B	NGLDM (2.5D)	Low grey level count emphasis	0.0139	0.0005	0.013943	0	match
configuration B	NGLDM (2.5D)	High grey level count emphasis	375	11	374.8469	0	match
configuration B	NGLDM (2.5D)	Low dependence low grey level emphasis	0.00929	0.00026	0.009292	0	match
configuration B	NGLDM (2.5D)	Low dependence high grey level emphasis	73.4	2.1	73.4227	0	match
configuration B	NGLDM (2.5D)	High dependence low grey level emphasis	0.077	0.0019	0.076988	0	match
configuration B	NGLDM (2.5D)	High dependence high grey level emphasis	7970	150	7965.9652	0	match
configuration B	NGLDM (2.5D)	Grey level non-uniformity	4760	50	4762.7865	0	match
configuration B	NGLDM (2.5D)	Normalised grey level non-uniformity	0.174	0.001	0.17408	0	match
configuration B	NGLDM (2.5D)	Dependence count non-uniformity	3710	30	3707.4458	0	match
configuration B	NGLDM (2.5D)	Normalised dependence count non-uniformity	0.136	0.001	0.13551	0	match
configuration B	NGLDM (2.5D)	Dependence count percentage	1	0			
configuration B	NGLDM (2.5D)	Grey level variance	18.7	0.2	18.6643	0	match
configuration B	NGLDM (2.5D)	Dependence count variance	4.63	0.06	4.632	0	match
configuration B	NGLDM (2.5D)	Dependence count entropy	5.78	0.01	5.7831	0	match
configuration B	NGLDM (2.5D)	Dependence count energy	0.0253	0.0001	0.025319	0	match
configuration C	Diagnostics-initial image	Image dimension x	204	0	204	0	match
configuration C	Diagnostics-initial image	Image dimension y	201	0	201	0	match
configuration C	Diagnostics-initial image	Image dimension z	60	0	60	0	match
configuration C	Diagnostics-initial image	Voxel dimension x	0.977	0	0.977	0	match
configuration C	Diagnostics-initial image	Voxel dimension y	0.977	0	0.977	0	match
configuration C	Diagnostics-initial image	Voxel dimension z	3	0	3	0	match
configuration C	Diagnostics-initial image	Mean intensity	-266	0	-266.4704	0	match
configuration C	Diagnostics-initial image	Minimum intensity	-1000	0	-1000	0	match
configuration C	Diagnostics-initial image	Maximum intensity	3065	0	3065	0	match
configuration C	Diagnostics-interpolated image	Image dimension x	100	1	100	0	match
configuration C	Diagnostics-interpolated image	Image dimension y	99	0.8	99	0	match
configuration C	Diagnostics-interpolated image	Image dimension z	90	0	90	0	match
configuration C	Diagnostics-interpolated image	Voxel dimension x	2	0	2	0	match
configuration C	Diagnostics-interpolated image	Voxel dimension y	2	0	2	0	match
configuration C	Diagnostics-interpolated image	Voxel dimension z	2	0	2	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	Diagnostics-interpolated image	Mean intensity	-270	3	-263.509	6	partial match
configuration C	Diagnostics-interpolated image	Minimum intensity	-1000	0	-1000	0	match
configuration C	Diagnostics-interpolated image	Maximum intensity	1854	30	1854	0	match
configuration C	Diagnostics-initial ROI	Int. mask dimension x	204	0	204	0	match
configuration C	Diagnostics-initial ROI	Int. mask dimension y	201	0	201	0	match
configuration C	Diagnostics-initial ROI	Int. mask dimension z	60	0	60	0	match
configuration C	Diagnostics-initial ROI	Int. mask bounding box dimension x	100	0	100	0	match
configuration C	Diagnostics-initial ROI	Int. mask bounding box dimension y	99	0	99	0	match
configuration C	Diagnostics-initial ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration C	Diagnostics-initial ROI	Morph. mask bounding box dimension x	100	0	100	0	match
configuration C	Diagnostics-initial ROI	Morph. mask bounding box dimension y	99	0	99	0	match
configuration C	Diagnostics-initial ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration C	Diagnostics-initial ROI	Int. mask voxel count	125256	0	125256	0	match
configuration C	Diagnostics-initial ROI	Morph. mask voxel count	125256	0	125256	0	match
configuration C	Diagnostics-initial ROI	Int. mask mean intensity	-46.9	0	-46.8827	0	match
configuration C	Diagnostics-initial ROI	Int. mask minimum intensity	-1000	0	-1000	0	match
configuration C	Diagnostics-initial ROI	Int. mask maximum intensity	723	0	723	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask dimension x	100	1	100	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask dimension y	99	0.8	99	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask dimension z	90	0	90	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask bounding box dimension x	49	0.2	49	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask bounding box dimension z	40	0.1	40	0	match
configuration C	Diagnostics-interpolated ROI	Morph. mask bounding box dimension x	49	0.2	49	0	match
configuration C	Diagnostics-interpolated ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration C	Diagnostics-interpolated ROI	Morph. mask bounding box dimension z	40	0.1	40	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask voxel count	45985	100	45985	0	match
configuration C	Diagnostics-interpolated ROI	Morph. mask voxel count	45985	100	45985	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask mean intensity	-48.9	0.1	-48.9321	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask minimum intensity	-939	1	-939	0	match
configuration C	Diagnostics-interpolated ROI	Int. mask maximum intensity	521	5	521	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask dimension x	100	1	100	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask dimension y	99	0.8	99	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask dimension z	90	0	90	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask bounding box dimension x	49	0.3	49	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask bounding box dimension z	40	0.3	40	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	Diagnostics-resegmented ROI	Morph. mask bounding box dimension x	49	0.3	49	0	match
configuration C	Diagnostics-resegmented ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration C	Diagnostics-resegmented ROI	Morph. mask bounding box dimension z	40	0.3	40	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask voxel count	45981	700	45981	0	match
configuration C	Diagnostics-resegmented ROI	Morph. mask voxel count	45985	700	45985	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask mean intensity	-49	2.9	-48.9785	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask minimum intensity	-939	4	-939	0	match
configuration C	Diagnostics-resegmented ROI	Int. mask maximum intensity	393	10	393	0	match
configuration C	Morphology	Volume (mesh)	367000	6000	367453.6667	0	match
configuration C	Morphology	Volume (voxel counting)	368000	6000	367880	0	match
configuration C	Morphology	Surface area (mesh)	34300	400	34306.252	0	match
configuration C	Morphology	Surface to volume ratio	0.0934	0.0007	0.093362	0	match
configuration C	Morphology	Compactness 1	0.0326	0.0002	0.032626	0	match
configuration C	Morphology	Compactness 2	0.378	0.004	0.37821	0	match
configuration C	Morphology	Spherical disproportion	1.38	0.01	1.3828	0	match
configuration C	Morphology	Sphericity	0.723	0.003	0.72318	0	match
configuration C	Morphology	Asphericity	0.383	0.004	0.38278	0	match
configuration C	Morphology	Centre of mass shift	45.6	2.8	45.5674	0	match
configuration C	Morphology	Maximum 3D diameter	125	1	125.06	0	match
configuration C	Morphology	Major axis length	93.3	0.5	93.2704	0	match
configuration C	Morphology	Minor axis length	82	0.5	82.0052	0	match
configuration C	Morphology	Least axis length	70.9	0.4	70.9015	0	match
configuration C	Morphology	Elongation	0.879	0.001	0.87922	0	match
configuration C	Morphology	Flatness	0.76	0.001	0.76017	0	match
configuration C	Morphology	Volume density (AABB)	0.478	0.003	0.47826	0	match
configuration C	Morphology	Area density (AABB)	0.678	0.003	0.67842	0	match
configuration C	Morphology	Volume density (OMBB)					
configuration C	Morphology	Area density (OMBB)					
configuration C	Morphology	Volume density (AEE)	1.29	0.01	1.2941	0	match
configuration C	Morphology	Area density (AEE)	1.62	0.01	1.6168	0	match
configuration C	Morphology	Volume density (MVEE)					
configuration C	Morphology	Area density (MVEE)					
configuration C	Morphology	Volume density (convex hull)	0.834	0.002	0.83366	0	match
configuration C	Morphology	Area density (convex hull)	1.13	0.01	1.1301	0	match
configuration C	Morphology	Integrated intensity	-18000000	1400000	-17997326.15	0	match
configuration C	Morphology	Moran’s I index	0.0824	0.0003			
configuration C	Morphology	Geary’s C measure	0.846	0.001			
configuration C	Local intensity	Local intensity peak	169	10			

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	Local intensity	Global intensity peak	180	5			
configuration C	Statistics	Mean	-49	2.9	-48.9785	0	match
configuration C	Statistics	Variance	50600	1400	50639.4315	0	match
configuration C	Statistics	Skewness	-2.14	0.05	-2.1402	0	match
configuration C	Statistics	(Excess) kurtosis	3.53	0.23	3.5251	0	match
configuration C	Statistics	Median	40	0.4	40	0	match
configuration C	Statistics	Minimum	-939	4	-939	0	match
configuration C	Statistics	10th percentile	-424	14	-424	0	match
configuration C	Statistics	90th percentile	86	0.1	86	0	match
configuration C	Statistics	Maximum	393	10	393	0	match
configuration C	Statistics	Interquartile range	67	4.9	67	0	match
configuration C	Statistics	Range	1330	20	1332	0	match
configuration C	Statistics	Mean absolute deviation	158	4	157.9732	0	match
configuration C	Statistics	Robust mean absolute deviation	66.8	3.5	66.7653	0	match
configuration C	Statistics	Median absolute deviation	119	4	119.1267	0	match
configuration C	Statistics	Coefficient of variation	-4.59	0.29	-4.5945	0	match
configuration C	Statistics	Quartile coefficient of dispersion	1.03	0.4	1.0308	0	match
configuration C	Statistics	Energy	2440000000	120000000	2438755180	0	match
configuration C	Statistics	Root mean square	230	4	230.3005	0	match
configuration C	Intensity histogram	Mean	38.6	0.2	38.5583	0	match
configuration C	Intensity histogram	Variance	81.1	2.1	81.1122	0	match
configuration C	Intensity histogram	Skewness	-2.14	0.05	-2.1371	0	match
configuration C	Intensity histogram	(Excess) kurtosis	3.52	0.23	3.519	0	match
configuration C	Intensity histogram	Median	42	0	42	0	match
configuration C	Intensity histogram	Minimum	3	0.16	3	0	match
configuration C	Intensity histogram	10th percentile	24	0.7	24	0	match
configuration C	Intensity histogram	90th percentile	44	0	44	0	match
configuration C	Intensity histogram	Maximum	56	0.5	56	0	match
configuration C	Intensity histogram	Mode	43	0.1	43	0	match
configuration C	Intensity histogram	Interquartile range	3	0.21	3	0	match
configuration C	Intensity histogram	Range	53	0.6	53	0	match
configuration C	Intensity histogram	Mean absolute deviation	6.32	0.15	6.3212	0	match
configuration C	Intensity histogram	Robust mean absolute deviation	2.59	0.14	2.588	0	match
configuration C	Intensity histogram	Median absolute deviation	4.75	0.12	4.7504	0	match
configuration C	Intensity histogram	Coefficient of variation	0.234	0.005	0.23357	0	match
configuration C	Intensity histogram	Quartile coefficient of dispersion	0.0361	0.0027	0.036145	0	match
configuration C	Intensity histogram	Entropy	3.73	0.04	3.7345	0	match
configuration C	Intensity histogram	Uniformity	0.14	0.003	0.13955	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	Intensity histogram	Maximum histogram gradient	4750	30	4745.5	0	match
configuration C	Intensity histogram	Maximum histogram gradient intensity	41	0	41	0	match
configuration C	Intensity histogram	Minimum histogram gradient	-4680	50	-4677	0	match
configuration C	Intensity histogram	Minimum histogram gradient intensity	44	0	44	0	match
configuration C	Intensity volume histogram	Volume fraction at 10% intensity	0.998	0.001	0.99637	0.001	match
configuration C	Intensity volume histogram	Volume fraction at 90% intensity	0.000152	0.00002	0.00015224	0	match
configuration C	Intensity volume histogram	Intensity at 10% volume	88.8	0.2	88.75	0	match
configuration C	Intensity volume histogram	Intensity at 90% volume	-421	14	-411.25	9	match
configuration C	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.997	0.001	0.99622	0	match
configuration C	Intensity volume histogram	Intensity difference between 10% and 90% volume	510	14	500	10	match
configuration C	Intensity volume histogram	Area under the IVH curve	0.681	0.003	0.67993	0.001	match
configuration C	GLCM (3D averaged)	Joint maximum	0.111	0.002	0.11085	0	match
configuration C	GLCM (3D averaged)	Joint average	39	0.2	38.9779	0	match
configuration C	GLCM (3D averaged)	Joint variance	73.7	2	73.745	0	match
configuration C	GLCM (3D averaged)	Joint entropy	6.39	0.06	6.3894	0	match
configuration C	GLCM (3D averaged)	Difference average	2.17	0.05	2.1672	0	match
configuration C	GLCM (3D averaged)	Difference variance	14.4	0.5	14.3781	0	match
configuration C	GLCM (3D averaged)	Difference entropy	2.64	0.03	2.6354	0	match
configuration C	GLCM (3D averaged)	Sum average	78	0.3	77.9559	0	match
configuration C	GLCM (3D averaged)	Sum variance	276	8	275.8026	0	match
configuration C	GLCM (3D averaged)	Sum entropy	4.56	0.04	4.5556	0	match
configuration C	GLCM (3D averaged)	Angular second moment	0.045	0.001	0.045002	0	match
configuration C	GLCM (3D averaged)	Contrast	19.2	0.7	19.1775	0	match
configuration C	GLCM (3D averaged)	Dissimilarity	2.17	0.05	2.1672	0	match
configuration C	GLCM (3D averaged)	Inverse difference	0.582	0.004	0.5824	0	match
configuration C	GLCM (3D averaged)	Normalised inverse difference	0.966	0.001	0.96619	0	match
configuration C	GLCM (3D averaged)	Inverse difference moment	0.547	0.004	0.54749	0	match
configuration C	GLCM (3D averaged)	Normalised inverse difference moment	0.994	0.001	0.99435	0	match
configuration C	GLCM (3D averaged)	Inverse variance	0.39	0.003	0.39048	0	match
configuration C	GLCM (3D averaged)	Correlation	0.869	0.001	0.8693	0	match
configuration C	GLCM (3D averaged)	Autocorrelation	1580	10	1583.4426	0	match
configuration C	GLCM (3D averaged)	Cluster tendency	276	8	275.8026	0	match
configuration C	GLCM (3D averaged)	Cluster shade	-10600	300	-10616.2341	0	match
configuration C	GLCM (3D averaged)	Cluster prominence	569000	11000	568750.038	0	match
configuration C	GLCM (3D averaged)	Information correlation 1	-0.236	0.001	-0.23621	0	match
configuration C	GLCM (3D averaged)	Information correlation 2	0.9	0.001	0.89999	0	match
configuration C	GLCM (3D merged)	Joint maximum	0.111	0.002	0.11093	0	match
configuration C	GLCM (3D merged)	Joint average	39	0.2	38.9765	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	GLCM (3D merged)	Joint variance	73.8	2	73.7815	0	match
configuration C	GLCM (3D merged)	Joint entropy	6.42	0.06	6.4198	0	match
configuration C	GLCM (3D merged)	Difference average	2.16	0.05	2.1627	0	match
configuration C	GLCM (3D merged)	Difference variance	14.4	0.5	14.4289	0	match
configuration C	GLCM (3D merged)	Difference entropy	2.64	0.03	2.6428	0	match
configuration C	GLCM (3D merged)	Sum average	78	0.3	77.9529	0	match
configuration C	GLCM (3D merged)	Sum variance	276	8	276.0199	0	match
configuration C	GLCM (3D merged)	Sum entropy	4.56	0.04	4.5594	0	match
configuration C	GLCM (3D merged)	Angular second moment	0.0447	0.001	0.044715	0	match
configuration C	GLCM (3D merged)	Contrast	19.1	0.7	19.1061	0	match
configuration C	GLCM (3D merged)	Dissimilarity	2.16	0.05	2.1627	0	match
configuration C	GLCM (3D merged)	Inverse difference	0.583	0.004	0.58275	0	match
configuration C	GLCM (3D merged)	Normalised inverse difference	0.966	0.001	0.96626	0	match
configuration C	GLCM (3D merged)	Inverse difference moment	0.548	0.004	0.54788	0	match
configuration C	GLCM (3D merged)	Normalised inverse difference moment	0.994	0.001	0.99438	0	match
configuration C	GLCM (3D merged)	Inverse variance	0.39	0.003	0.39046	0	match
configuration C	GLCM (3D merged)	Correlation	0.871	0.001	0.87052	0	match
configuration C	GLCM (3D merged)	Autocorrelation	1580	10	1583.3939	0	match
configuration C	GLCM (3D merged)	Cluster tendency	276	8	276.0199	0	match
configuration C	GLCM (3D merged)	Cluster shade	-10600	300	-10629.7981	0	match
configuration C	GLCM (3D merged)	Cluster prominence	570000	11000	569596.1785	0	match
configuration C	GLCM (3D merged)	Information correlation 1	-0.228	0.001	-0.22832	0	match
configuration C	GLCM (3D merged)	Information correlation 2	0.899	0.001	0.89936	0	match
configuration C	GLRLM (3D averaged)	Short runs emphasis	0.786	0.003	0.78595	0	match
configuration C	GLRLM (3D averaged)	Long runs emphasis	3.31	0.04	3.3108	0	match
configuration C	GLRLM (3D averaged)	Low grey level run emphasis	0.00155	0.00005	0.0015484	0	match
configuration C	GLRLM (3D averaged)	High grey level run emphasis	1470	10	1471.6445	0	match
configuration C	GLRLM (3D averaged)	Short run low grey level emphasis	0.00136	0.00005	0.0013603	0	match
configuration C	GLRLM (3D averaged)	Short run high grey level emphasis	1100	10	1097.1669	0	match
configuration C	GLRLM (3D averaged)	Long run low grey level emphasis	0.00317	0.00004	0.0031718	0	match
configuration C	GLRLM (3D averaged)	Long run high grey level emphasis	5590	80	5586.6474	0	match
configuration C	GLRLM (3D averaged)	Grey level non-uniformity	3180	10	3179.105	0	match
configuration C	GLRLM (3D averaged)	Normalised grey level non-uniformity	0.102	0.003	0.10157	0	match
configuration C	GLRLM (3D averaged)	Run length non-uniformity	18000	500	17989.8187	0	match
configuration C	GLRLM (3D averaged)	Normalised run length non-uniformity	0.574	0.004	0.57435	0	match
configuration C	GLRLM (3D averaged)	Run percentage	0.679	0.003	0.67913	0	match
configuration C	GLRLM (3D averaged)	Grey level variance	101	3	101.45	0	match
configuration C	GLRLM (3D averaged)	Run length variance	1.12	0.02	1.1209	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	GLRLM (3D averaged)	Run entropy	5.35	0.03	5.3477	0	match
configuration C	GLRLM (3D merged)	Short runs emphasis	0.787	0.003	0.78712	0	match
configuration C	GLRLM (3D merged)	Long runs emphasis	3.28	0.04	3.2759	0	match
configuration C	GLRLM (3D merged)	Low grey level run emphasis	0.00155	0.00005	0.0015468	0	match
configuration C	GLRLM (3D merged)	High grey level run emphasis	1470	10	1472.5002	0	match
configuration C	GLRLM (3D merged)	Short run low grey level emphasis	0.00136	0.00005	0.0013604	0	match
configuration C	GLRLM (3D merged)	Short run high grey level emphasis	1100	10	1099.9533	0	match
configuration C	GLRLM (3D merged)	Long run low grey level emphasis	0.00314	0.00004	0.0031439	0	match
configuration C	GLRLM (3D merged)	Long run high grey level emphasis	5530	80	5525.4544	0	match
configuration C	GLRLM (3D merged)	Grey level non-uniformity	41300	100	41297.6818	0	match
configuration C	GLRLM (3D merged)	Normalised grey level non-uniformity	0.102	0.003	0.10173	0	match
configuration C	GLRLM (3D merged)	Run length non-uniformity	234000	6000	233618.4646	0	match
configuration C	GLRLM (3D merged)	Normalised run length non-uniformity	0.575	0.004	0.57548	0	match
configuration C	GLRLM (3D merged)	Run percentage	0.679	0.003	0.67913	0	match
configuration C	GLRLM (3D merged)	Grey level variance	101	3	101.3845	0	match
configuration C	GLRLM (3D merged)	Run length variance	1.11	0.02	1.1078	0	match
configuration C	GLRLM (3D merged)	Run entropy	5.35	0.03	5.3505	0	match
configuration C	GLSZM (3D)	Small zone emphasis	0.695	0.001	0.69499	0	match
configuration C	GLSZM (3D)	Large zone emphasis	38900	900	38927.2065	0	match
configuration C	GLSZM (3D)	Low grey level emphasis	0.00235	0.00006	0.00235	0	match
configuration C	GLSZM (3D)	High grey level emphasis	971	7	970.711	0	match
configuration C	GLSZM (3D)	Small zone low grey level emphasis	0.0016	0.00004	0.0015955	0	match
configuration C	GLSZM (3D)	Small zone high grey level emphasis	657	4	656.8282	0	match
configuration C	GLSZM (3D)	Large zone low grey level emphasis	21.6	0.5	21.5513	0	match
configuration C	GLSZM (3D)	Large zone high grey level emphasis	70700000	1500000	70710078.21	0	match
configuration C	GLSZM (3D)	Grey level non-uniformity	195	6	195.032	0	match
configuration C	GLSZM (3D)	Normalised grey level non-uniformity	0.0286	0.0003	0.028643	0	match
configuration C	GLSZM (3D)	Zone size non-uniformity	3040	100	3042.8963	0	match
configuration C	GLSZM (3D)	Normalised zone size non-uniformity	0.447	0.001	0.44689	0	match
configuration C	GLSZM (3D)	Zone percentage	0.148	0.003	0.14808	0	match
configuration C	GLSZM (3D)	Grey level variance	106	1	105.9675	0	match
configuration C	GLSZM (3D)	Zone size variance	38900	900	38881.6038	0	match
configuration C	GLSZM (3D)	Zone size entropy	7	0.01	6.9982	0	match
configuration C	GLDZM (3D)	Small distance emphasis	0.531	0.006	0.53112	0	match
configuration C	GLDZM (3D)	Large distance emphasis	11	0.3	11.0209	0	match
configuration C	GLDZM (3D)	Low grey level emphasis	0.00235	0.00006	0.00235	0	match
configuration C	GLDZM (3D)	High grey level emphasis	971	7	970.711	0	match
configuration C	GLDZM (3D)	Small distance low grey level emphasis	0.00149	0.00004	0.0014937	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration C	GLDZM (3D)	Small distance high grey level emphasis	476	11	475.7232	0	match
configuration C	GLDZM (3D)	Large distance low grey level emphasis	0.0154	0.0005	0.015424	0	match
configuration C	GLDZM (3D)	Large distance high grey level emphasis	13400	200	13356.2989	0	match
configuration C	GLDZM (3D)	Grey level non-uniformity	195	6	195.032	0	match
configuration C	GLDZM (3D)	Normalised grey level non-uniformity	0.0286	0.0003	0.028643	0	match
configuration C	GLDZM (3D)	Zone distance non-uniformity	1870	40	1866.2492	0	match
configuration C	GLDZM (3D)	Normalised zone distance non-uniformity	0.274	0.005	0.27409	0	match
configuration C	GLDZM (3D)	Zone percentage	0.148	0.003	0.14808	0	match
configuration C	GLDZM (3D)	Grey level variance	106	1	105.9675	0	match
configuration C	GLDZM (3D)	Zone distance variance	4.6	0.06	4.6004	0	match
configuration C	GLDZM (3D)	Zone distance entropy	7.56	0.03	7.5634	0	match
configuration C	NGTDM (3D)	Coarseness	0.000216	0.000004	0.00021612	0	match
configuration C	NGTDM (3D)	Contrast	0.0873	0.0019	0.087346	0	match
configuration C	NGTDM (3D)	Busyness	1.39	0.01	1.3894	0	match
configuration C	NGTDM (3D)	Complexity	1810	60	1808.7937	0	match
configuration C	NGTDM (3D)	Strength	0.651	0.015	0.65105	0	match
configuration C	NGLDM (3D)	Low dependence emphasis	0.137	0.003	0.1369	0	match
configuration C	NGLDM (3D)	High dependence emphasis	126	2	126.491	0	match
configuration C	NGLDM (3D)	Low grey level count emphasis	0.0013	0.00004	0.0012975	0	match
configuration C	NGLDM (3D)	High grey level count emphasis	1570	10	1567.8577	0	match
configuration C	NGLDM (3D)	Low dependence low grey level emphasis	0.000306	0.000012	0.00030589	0	match
configuration C	NGLDM (3D)	Low dependence high grey level emphasis	141	2	140.5805	0	match
configuration C	NGLDM (3D)	High dependence low grey level emphasis	0.0828	0.0003	0.082808	0	match
configuration C	NGLDM (3D)	High dependence high grey level emphasis	227000	3000	226736.1645	0	match
configuration C	NGLDM (3D)	Grey level non-uniformity	6420	10	6416.5308	0	match
configuration C	NGLDM (3D)	Normalised grey level non-uniformity	0.14	0.003	0.13955	0	match
configuration C	NGLDM (3D)	Dependence count non-uniformity	2450	60	2447.4365	0	match
configuration C	NGLDM (3D)	Normalised dependence count non-uniformity	0.0532	0.0005	0.053227	0	match
configuration C	NGLDM (3D)	Dependence count percentage	1	0			
configuration C	NGLDM (3D)	Grey level variance	81.1	2.1	81.1122	0	match
configuration C	NGLDM (3D)	Dependence count variance	39.2	0.1	39.2081	0	match
configuration C	NGLDM (3D)	Dependence count entropy	7.54	0.03	7.5367	0	match
configuration C	NGLDM (3D)	Dependence count energy	0.00789	0.00011	0.0078911	0	match
configuration D	Diagnostics-initial image	Image dimension x	204	0	204	0	match
configuration D	Diagnostics-initial image	Image dimension y	201	0	201	0	match
configuration D	Diagnostics-initial image	Image dimension z	60	0	60	0	match
configuration D	Diagnostics-initial image	Voxel dimension x	0.977	0	0.977	0	match
configuration D	Diagnostics-initial image	Voxel dimension y	0.977	0	0.977	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	Diagnostics-initial image	Voxel dimension z	3	0	3	0	match
configuration D	Diagnostics-initial image	Mean intensity	-266	0	-266.4704	0	match
configuration D	Diagnostics-initial image	Minimum intensity	-1000	0	-1000	0	match
configuration D	Diagnostics-initial image	Maximum intensity	3065	0	3065	0	match
configuration D	Diagnostics-interpolated image	Image dimension x	100	1	100	0	match
configuration D	Diagnostics-interpolated image	Image dimension y	99	0.8	99	0	match
configuration D	Diagnostics-interpolated image	Image dimension z	90	0	90	0	match
configuration D	Diagnostics-interpolated image	Voxel dimension x	2	0	2	0	match
configuration D	Diagnostics-interpolated image	Voxel dimension y	2	0	2	0	match
configuration D	Diagnostics-interpolated image	Voxel dimension z	2	0	2	0	match
configuration D	Diagnostics-interpolated image	Mean intensity	-270	3	-263.509	6	partial match
configuration D	Diagnostics-interpolated image	Minimum intensity	-1000	0	-1000	0	match
configuration D	Diagnostics-interpolated image	Maximum intensity	1854	30	1854	0	match
configuration D	Diagnostics-initial ROI	Int. mask dimension x	204	0	204	0	match
configuration D	Diagnostics-initial ROI	Int. mask dimension y	201	0	201	0	match
configuration D	Diagnostics-initial ROI	Int. mask dimension z	60	0	60	0	match
configuration D	Diagnostics-initial ROI	Int. mask bounding box dimension x	100	0	100	0	match
configuration D	Diagnostics-initial ROI	Int. mask bounding box dimension y	99	0	99	0	match
configuration D	Diagnostics-initial ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration D	Diagnostics-initial ROI	Morph. mask bounding box dimension x	100	0	100	0	match
configuration D	Diagnostics-initial ROI	Morph. mask bounding box dimension y	99	0	99	0	match
configuration D	Diagnostics-initial ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration D	Diagnostics-initial ROI	Int. mask voxel count	125256	0	125256	0	match
configuration D	Diagnostics-initial ROI	Morph. mask voxel count	125256	0	125256	0	match
configuration D	Diagnostics-initial ROI	Int. mask mean intensity	-46.9	0	-46.8827	0	match
configuration D	Diagnostics-initial ROI	Int. mask minimum intensity	-1000	0	-1000	0	match
configuration D	Diagnostics-initial ROI	Int. mask maximum intensity	723	0	723	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask dimension x	100	1	100	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask dimension y	99	0.8	99	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask dimension z	90	0	90	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask bounding box dimension x	49	0.2	49	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask bounding box dimension z	40	0.1	40	0	match
configuration D	Diagnostics-interpolated ROI	Morph. mask bounding box dimension x	49	0.2	49	0	match
configuration D	Diagnostics-interpolated ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration D	Diagnostics-interpolated ROI	Morph. mask bounding box dimension z	40	0.1	40	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask voxel count	45985	100	45985	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	Diagnostics-interpolated ROI	Morph. mask voxel count	45985	100	45985	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask mean intensity	-48.9	0.1	-48.9321	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask minimum intensity	-939	1	-939	0	match
configuration D	Diagnostics-interpolated ROI	Int. mask maximum intensity	521	5	521	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask dimension x	100	1	100	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask dimension y	99	0.8	99	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask dimension z	90	0	90	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask bounding box dimension x	49	0.3	49	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask bounding box dimension z	40	0.3	40	0	match
configuration D	Diagnostics-resegmented ROI	Morph. mask bounding box dimension x	49	0.3	49	0	match
configuration D	Diagnostics-resegmented ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration D	Diagnostics-resegmented ROI	Morph. mask bounding box dimension z	40	0.3	40	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask voxel count	44465	800	44465	0	match
configuration D	Diagnostics-resegmented ROI	Morph. mask voxel count	45985	700	45985	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask mean intensity	-23.5	3.9	-23.5179	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask minimum intensity	-724	12	-724	0	match
configuration D	Diagnostics-resegmented ROI	Int. mask maximum intensity	521	22	521	0	match
configuration D	Morphology	Volume (mesh)	367000	6000	367453.6667	0	match
configuration D	Morphology	Volume (voxel counting)	368000	6000	367880	0	match
configuration D	Morphology	Surface area (mesh)	34300	400	34306.252	0	match
configuration D	Morphology	Surface to volume ratio	0.0934	0.0007	0.093362	0	match
configuration D	Morphology	Compactness 1	0.0326	0.0002	0.032626	0	match
configuration D	Morphology	Compactness 2	0.378	0.004	0.37821	0	match
configuration D	Morphology	Spherical disproportion	1.38	0.01	1.3828	0	match
configuration D	Morphology	Sphericity	0.723	0.003	0.72318	0	match
configuration D	Morphology	Asphericity	0.383	0.004	0.38278	0	match
configuration D	Morphology	Centre of mass shift	64.9	2.8	64.9261	0	match
configuration D	Morphology	Maximum 3D diameter	125	1	125.06	0	match
configuration D	Morphology	Major axis length	93.3	0.5	93.2704	0	match
configuration D	Morphology	Minor axis length	82	0.5	82.0052	0	match
configuration D	Morphology	Least axis length	70.9	0.4	70.9015	0	match
configuration D	Morphology	Elongation	0.879	0.001	0.87922	0	match
configuration D	Morphology	Flatness	0.76	0.001	0.76017	0	match
configuration D	Morphology	Volume density (AABB)	0.478	0.003	0.47826	0	match
configuration D	Morphology	Area density (AABB)	0.678	0.003	0.67842	0	match
configuration D	Morphology	Volume density (OMBB)					
configuration D	Morphology	Area density (OMBB)					

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	Morphology	Volume density (AEE)	1.29	0.01	1.2941	0	match
configuration D	Morphology	Area density (AEE)	1.62	0.01	1.6168	0	match
configuration D	Morphology	Volume density (MVEE)					
configuration D	Morphology	Area density (MVEE)					
configuration D	Morphology	Volume density (convex hull)	0.834	0.002	0.83366	0	match
configuration D	Morphology	Area density (convex hull)	1.13	0.01	1.1301	0	match
configuration D	Morphology	Integrated intensity	-8640000	1560000	-8641751.615	0	match
configuration D	Morphology	Moran’s I index	0.0622	0.0013			
configuration D	Morphology	Geary’s C measure	0.851	0.001			
configuration D	Local intensity	Local intensity peak	201	10			
configuration D	Local intensity	Global intensity peak	201	5			
configuration D	Statistics	Mean	-23.5	3.9	-23.5179	0	match
configuration D	Statistics	Variance	32800	2100	32786.8476	0	match
configuration D	Statistics	Skewness	-2.28	0.06	-2.2803	0	match
configuration D	Statistics	(Excess) kurtosis	4.35	0.32	4.3511	0	match
configuration D	Statistics	Median	42	0.4	42	0	match
configuration D	Statistics	Minimum	-724	12	-724	0	match
configuration D	Statistics	10th percentile	-304	20	-304	0	match
configuration D	Statistics	90th percentile	86	0.1	86	0	match
configuration D	Statistics	Maximum	521	22	521	0	match
configuration D	Statistics	Interquartile range	57	4.1	57	0	match
configuration D	Statistics	Range	1240	40	1245	0	match
configuration D	Statistics	Mean absolute deviation	123	6	122.5432	0	match
configuration D	Statistics	Robust mean absolute deviation	46.8	3.6	46.8275	0	match
configuration D	Statistics	Median absolute deviation	94.7	3.8	94.73	0	match
configuration D	Statistics	Coefficient of variation	-7.7	1.01	-7.6993	0	match
configuration D	Statistics	Quartile coefficient of dispersion	0.74	0.011	0.74026	0	match
configuration D	Statistics	Energy	1480000000	140000000	1482460471	0	match
configuration D	Statistics	Root mean square	183	7	182.5923	0	match
configuration D	Intensity histogram	Mean	18.5	0.5	18.503	0	match
configuration D	Intensity histogram	Variance	21.7	0.4	21.69	0	match
configuration D	Intensity histogram	Skewness	-2.27	0.06	-2.2678	0	match
configuration D	Intensity histogram	(Excess) kurtosis	4.31	0.32	4.3076	0	match
configuration D	Intensity histogram	Median	20	0.5	20	0	match
configuration D	Intensity histogram	Minimum	1	0	1	0	match
configuration D	Intensity histogram	10th percentile	11	0.7	11	0	match
configuration D	Intensity histogram	90th percentile	21	0.5	21	0	match
configuration D	Intensity histogram	Maximum	32	0	32	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	Intensity histogram	Mode	20	0.4	20	0	match
configuration D	Intensity histogram	Interquartile range	2	0.06	2	0	match
configuration D	Intensity histogram	Range	31	0	31	0	match
configuration D	Intensity histogram	Mean absolute deviation	3.15	0.05	3.1511	0	match
configuration D	Intensity histogram	Robust mean absolute deviation	1.33	0.06	1.3276	0	match
configuration D	Intensity histogram	Median absolute deviation	2.41	0.04	2.4073	0	match
configuration D	Intensity histogram	Coefficient of variation	0.252	0.006	0.2517	0	match
configuration D	Intensity histogram	Quartile coefficient of dispersion	0.05	0.0021	0.05	0	match
configuration D	Intensity histogram	Entropy	2.94	0.01	2.94	0	match
configuration D	Intensity histogram	Uniformity	0.229	0.003	0.22877	0	match
configuration D	Intensity histogram	Maximum histogram gradient	7260	200	7263	0	match
configuration D	Intensity histogram	Maximum histogram gradient intensity	19	0.4	19	0	match
configuration D	Intensity histogram	Minimum histogram gradient	-6670	230	-6674	0	match
configuration D	Intensity histogram	Minimum histogram gradient intensity	22	0.4	22	0	match
configuration D	Intensity volume histogram	Volume fraction at 10% intensity	0.972	0.003	0.97157	0	match
configuration D	Intensity volume histogram	Volume fraction at 90% intensity	0.00009	0.000415	9.00E-05	0	match
configuration D	Intensity volume histogram	Intensity at 10% volume	87	0.1	87	0	match
configuration D	Intensity volume histogram	Intensity at 90% volume	-303	20	-303	0	match
configuration D	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.971	0.001	0.97148	0	match
configuration D	Intensity volume histogram	Intensity difference between 10% and 90% volume	390	20	390	0	match
configuration D	Intensity volume histogram	Area under the IVH curve	0.563	0.012	0.56304	0	match
configuration D	GLCM (3D averaged)	Joint maximum	0.232	0.007	0.23196	0	match
configuration D	GLCM (3D averaged)	Joint average	18.9	0.5	18.8525	0	match
configuration D	GLCM (3D averaged)	Joint variance	17.6	0.4	17.628	0	match
configuration D	GLCM (3D averaged)	Joint entropy	4.95	0.03	4.9473	0	match
configuration D	GLCM (3D averaged)	Difference average	1.29	0.01	1.2926	0	match
configuration D	GLCM (3D averaged)	Difference variance	5.37	0.11	5.369	0	match
configuration D	GLCM (3D averaged)	Difference entropy	2.13	0.01	2.1339	0	match
configuration D	GLCM (3D averaged)	Sum average	37.7	0.8	37.7049	0	match
configuration D	GLCM (3D averaged)	Sum variance	63.4	1.3	63.441	0	match
configuration D	GLCM (3D averaged)	Sum entropy	3.68	0.02	3.6756	0	match
configuration D	GLCM (3D averaged)	Angular second moment	0.11	0.003	0.10965	0	match
configuration D	GLCM (3D averaged)	Contrast	7.07	0.13	7.071	0	match
configuration D	GLCM (3D averaged)	Dissimilarity	1.29	0.01	1.2926	0	match
configuration D	GLCM (3D averaged)	Inverse difference	0.682	0.003	0.68172	0	match
configuration D	GLCM (3D averaged)	Normalised inverse difference	0.965	0.001	0.96507	0	match
configuration D	GLCM (3D averaged)	Inverse difference moment	0.656	0.003	0.65641	0	match
configuration D	GLCM (3D averaged)	Normalised inverse difference moment	0.994	0.001	0.99367	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	GLCM (3D averaged)	Inverse variance	0.341	0.005	0.34059	0	match
configuration D	GLCM (3D averaged)	Correlation	0.798	0.005	0.79815	0	match
configuration D	GLCM (3D averaged)	Autocorrelation	370	16	369.5105	0	match
configuration D	GLCM (3D averaged)	Cluster tendency	63.4	1.3	63.441	0	match
configuration D	GLCM (3D averaged)	Cluster shade	-1270	40	-1272.9298	0	match
configuration D	GLCM (3D averaged)	Cluster prominence	35700	1400	35664.7177	0	match
configuration D	GLCM (3D averaged)	Information correlation 1	-0.231	0.003	-0.23081	0	match
configuration D	GLCM (3D averaged)	Information correlation 2	0.845	0.003	0.84503	0	match
configuration D	GLCM (3D merged)	Joint maximum	0.232	0.007	0.23207	0	match
configuration D	GLCM (3D merged)	Joint average	18.9	0.5	18.8517	0	match
configuration D	GLCM (3D merged)	Joint variance	17.6	0.4	17.6377	0	match
configuration D	GLCM (3D merged)	Joint entropy	4.96	0.03	4.9648	0	match
configuration D	GLCM (3D merged)	Difference average	1.29	0.01	1.29	0	match
configuration D	GLCM (3D merged)	Difference variance	5.38	0.11	5.3813	0	match
configuration D	GLCM (3D merged)	Difference entropy	2.14	0.01	2.1391	0	match
configuration D	GLCM (3D merged)	Sum average	37.7	0.8	37.7035	0	match
configuration D	GLCM (3D merged)	Sum variance	63.5	1.3	63.5056	0	match
configuration D	GLCM (3D merged)	Sum entropy	3.68	0.02	3.6795	0	match
configuration D	GLCM (3D merged)	Angular second moment	0.109	0.003	0.10934	0	match
configuration D	GLCM (3D merged)	Contrast	7.05	0.13	7.0453	0	match
configuration D	GLCM (3D merged)	Dissimilarity	1.29	0.01	1.29	0	match
configuration D	GLCM (3D merged)	Inverse difference	0.682	0.003	0.68204	0	match
configuration D	GLCM (3D merged)	Normalised inverse difference	0.965	0.001	0.96514	0	match
configuration D	GLCM (3D merged)	Inverse difference moment	0.657	0.003	0.65677	0	match
configuration D	GLCM (3D merged)	Normalised inverse difference moment	0.994	0.001	0.9937	0	match
configuration D	GLCM (3D merged)	Inverse variance	0.34	0.005	0.34046	0	match
configuration D	GLCM (3D merged)	Correlation	0.8	0.005	0.80028	0	match
configuration D	GLCM (3D merged)	Autocorrelation	370	16	369.5033	0	match
configuration D	GLCM (3D merged)	Cluster tendency	63.5	1.3	63.5056	0	match
configuration D	GLCM (3D merged)	Cluster shade	-1280	40	-1275.2616	0	match
configuration D	GLCM (3D merged)	Cluster prominence	35700	1500	35742.8426	0	match
configuration D	GLCM (3D merged)	Information correlation 1	-0.225	0.003	-0.22526	0	match
configuration D	GLCM (3D merged)	Information correlation 2	0.846	0.003	0.84643	0	match
configuration D	GLRLM (3D averaged)	Short runs emphasis	0.734	0.001	0.73442	0	match
configuration D	GLRLM (3D averaged)	Long runs emphasis	6.66	0.18	6.6573	0	match
configuration D	GLRLM (3D averaged)	Low grey level run emphasis	0.0257	0.0012	0.025732	0	match
configuration D	GLRLM (3D averaged)	High grey level run emphasis	326	17	325.7413	0	match
configuration D	GLRLM (3D averaged)	Short run low grey level emphasis	0.0232	0.001	0.02325	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	GLRLM (3D averaged)	Short run high grey level emphasis	219	13	218.6218	0	match
configuration D	GLRLM (3D averaged)	Long run low grey level emphasis	0.0484	0.0031	0.048379	0	match
configuration D	GLRLM (3D averaged)	Long run high grey level emphasis	2670	30	2667.0753	0	match
configuration D	GLRLM (3D averaged)	Grey level non-uniformity	3290	10	3293.6495	0	match
configuration D	GLRLM (3D averaged)	Normalised grey level non-uniformity	0.133	0.002	0.13333	0	match
configuration D	GLRLM (3D averaged)	Run length non-uniformity	12400	200	12351.1018	0	match
configuration D	GLRLM (3D averaged)	Normalised run length non-uniformity	0.5	0.001	0.49988	0	match
configuration D	GLRLM (3D averaged)	Run percentage	0.554	0.005	0.55375	0	match
configuration D	GLRLM (3D averaged)	Grey level variance	31.5	0.4	31.4529	0	match
configuration D	GLRLM (3D averaged)	Run length variance	3.35	0.14	3.3485	0	match
configuration D	GLRLM (3D averaged)	Run entropy	5.08	0.02	5.0806	0	match
configuration D	GLRLM (3D merged)	Short runs emphasis	0.736	0.001	0.73571	0	match
configuration D	GLRLM (3D merged)	Long runs emphasis	6.56	0.18	6.5562	0	match
configuration D	GLRLM (3D merged)	Low grey level run emphasis	0.0257	0.0012	0.025672	0	match
configuration D	GLRLM (3D merged)	High grey level run emphasis	326	17	326.0725	0	match
configuration D	GLRLM (3D merged)	Short run low grey level emphasis	0.0232	0.001	0.023229	0	match
configuration D	GLRLM (3D merged)	Short run high grey level emphasis	219	13	219.4018	0	match
configuration D	GLRLM (3D merged)	Long run low grey level emphasis	0.0478	0.0031	0.047847	0	match
configuration D	GLRLM (3D merged)	Long run high grey level emphasis	2630	30	2625.5931	0	match
configuration D	GLRLM (3D merged)	Grey level non-uniformity	42800	200	42767.9687	0	match
configuration D	GLRLM (3D merged)	Normalised grey level non-uniformity	0.134	0.002	0.13361	0	match
configuration D	GLRLM (3D merged)	Run length non-uniformity	160000	3000	160418.4996	0	match
configuration D	GLRLM (3D merged)	Normalised run length non-uniformity	0.501	0.001	0.50117	0	match
configuration D	GLRLM (3D merged)	Run percentage	0.554	0.005	0.55375	0	match
configuration D	GLRLM (3D merged)	Grey level variance	31.4	0.4	31.4254	0	match
configuration D	GLRLM (3D merged)	Run length variance	3.29	0.13	3.295	0	match
configuration D	GLRLM (3D merged)	Run entropy	5.08	0.02	5.0832	0	match
configuration D	GLSZM (3D)	Small zone emphasis	0.637	0.005	0.6365	0	match
configuration D	GLSZM (3D)	Large zone emphasis	99100	2800	99078.5164	0	match
configuration D	GLSZM (3D)	Low grey level emphasis	0.0409	0.0005	0.040946	0	match
configuration D	GLSZM (3D)	High grey level emphasis	188	10	188.1832	0	match
configuration D	GLSZM (3D)	Small zone low grey level emphasis	0.0248	0.0004	0.024761	0	match
configuration D	GLSZM (3D)	Small zone high grey level emphasis	117	7	116.5533	0	match
configuration D	GLSZM (3D)	Large zone low grey level emphasis	241	14	240.7782	0	match
configuration D	GLSZM (3D)	Large zone high grey level emphasis	41400000	300000	41404349.39	0	match
configuration D	GLSZM (3D)	Grey level non-uniformity	212	6	212.1341	0	match
configuration D	GLSZM (3D)	Normalised grey level non-uniformity	0.0491	0.0008	0.04906	0	match
configuration D	GLSZM (3D)	Zone size non-uniformity	1630	10	1629.1129	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	GLSZM (3D)	Normalised zone size non-uniformity	0.377	0.006	0.37676	0	match
configuration D	GLSZM (3D)	Zone percentage	0.0972	0.0007	0.097245	0	match
configuration D	GLSZM (3D)	Grey level variance	32.7	1.6	32.7185	0	match
configuration D	GLSZM (3D)	Zone size variance	99000	2800	98972.7701	0	match
configuration D	GLSZM (3D)	Zone size entropy	6.52	0.01	6.5153	0	match
configuration D	GLDZM (3D)	Small distance emphasis	0.579	0.004	0.57934	0	match
configuration D	GLDZM (3D)	Large distance emphasis	10.3	0.1	10.2583	0	match
configuration D	GLDZM (3D)	Low grey level emphasis	0.0409	0.0005	0.040946	0	match
configuration D	GLDZM (3D)	High grey level emphasis	188	10	188.1832	0	match
configuration D	GLDZM (3D)	Small distance low grey level emphasis	0.0302	0.0006	0.030208	0	match
configuration D	GLDZM (3D)	Small distance high grey level emphasis	99.3	5.1	99.3004	0	match
configuration D	GLDZM (3D)	Large distance low grey level emphasis	0.183	0.004	0.18285	0	match
configuration D	GLDZM (3D)	Large distance high grey level emphasis	2620	110	2619.1681	0	match
configuration D	GLDZM (3D)	Grey level non-uniformity	212	6	212.1341	0	match
configuration D	GLDZM (3D)	Normalised grey level non-uniformity	0.0491	0.0008	0.04906	0	match
configuration D	GLDZM (3D)	Zone distance non-uniformity	1370	20	1369.4454	0	match
configuration D	GLDZM (3D)	Normalised zone distance non-uniformity	0.317	0.004	0.31671	0	match
configuration D	GLDZM (3D)	Zone percentage	0.0972	0.0007	0.097245	0	match
configuration D	GLDZM (3D)	Grey level variance	32.7	1.6	32.7185	0	match
configuration D	GLDZM (3D)	Zone distance variance	4.61	0.04	4.6139	0	match
configuration D	GLDZM (3D)	Zone distance entropy	6.61	0.03	6.6141	0	match
configuration D	NGTDM (3D)	Coarseness	0.000208	0.000004	0.00020847	0	match
configuration D	NGTDM (3D)	Contrast	0.046	0.0005	0.046022	0	match
configuration D	NGTDM (3D)	Busyness	5.14	0.14	5.1437	0	match
configuration D	NGTDM (3D)	Complexity	400	5	399.6936	0	match
configuration D	NGTDM (3D)	Strength	0.162	0.008	0.16173	0	match
configuration D	NGLDM (3D)	Low dependence emphasis	0.0912	0.0007	0.091222	0	match
configuration D	NGLDM (3D)	High dependence emphasis	223	5	222.7484	0	match
configuration D	NGLDM (3D)	Low grey level count emphasis	0.0168	0.0009	0.016771	0	match
configuration D	NGLDM (3D)	High grey level count emphasis	364	16	364.0495	0	match
configuration D	NGLDM (3D)	Low dependence low grey level emphasis	0.00357	0.00004	0.0035687	0	match
configuration D	NGLDM (3D)	Low dependence high grey level emphasis	18.9	1.1	18.945	0	match
configuration D	NGLDM (3D)	High dependence low grey level emphasis	0.798	0.072	0.79765	0	match
configuration D	NGLDM (3D)	High dependence high grey level emphasis	92800	1300	92761.6252	0	match
configuration D	NGLDM (3D)	Grey level non-uniformity	10200	300	10172.0488	0	match
configuration D	NGLDM (3D)	Normalised grey level non-uniformity	0.229	0.003	0.22877	0	match
configuration D	NGLDM (3D)	Dependence count non-uniformity	1840	30	1836.8652	0	match
configuration D	NGLDM (3D)	Normalised dependence count non-uniformity	0.0413	0.0003	0.04131	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration D	NGLDM (3D)	Dependence count percentage	1	0			
configuration D	NGLDM (3D)	Grey level variance	21.7	0.4	21.69	0	match
configuration D	NGLDM (3D)	Dependence count variance	63.9	1.3	63.9226	0	match
configuration D	NGLDM (3D)	Dependence count entropy	6.98	0.01	6.9811	0	match
configuration D	NGLDM (3D)	Dependence count energy	0.0113	0.0002	0.011291	0	match
configuration E	Diagnostics-initial image	Image dimension x	204	0	204	0	match
configuration E	Diagnostics-initial image	Image dimension y	201	0	201	0	match
configuration E	Diagnostics-initial image	Image dimension z	60	0	60	0	match
configuration E	Diagnostics-initial image	Voxel dimension x	0.977	0	0.977	0	match
configuration E	Diagnostics-initial image	Voxel dimension y	0.977	0	0.977	0	match
configuration E	Diagnostics-initial image	Voxel dimension z	3	0	3	0	match
configuration E	Diagnostics-initial image	Mean intensity	-266	0	-266.4704	0	match
configuration E	Diagnostics-initial image	Minimum intensity	-1000	0	-1000	0	match
configuration E	Diagnostics-initial image	Maximum intensity	3065	0	3065	0	match
configuration E	Diagnostics-interpolated image	Image dimension x	100	1	100	0	match
configuration E	Diagnostics-interpolated image	Image dimension y	99	0.8	99	0	match
configuration E	Diagnostics-interpolated image	Image dimension z	90	0	90	0	match
configuration E	Diagnostics-interpolated image	Voxel dimension x	2	0	2	0	match
configuration E	Diagnostics-interpolated image	Voxel dimension y	2	0	2	0	match
configuration E	Diagnostics-interpolated image	Voxel dimension z	2	0	2	0	match
configuration E	Diagnostics-interpolated image	Mean intensity	-270	3	-271.2408	1	match
configuration E	Diagnostics-interpolated image	Minimum intensity	-1111	10	-1269	150	no match
configuration E	Diagnostics-interpolated image	Maximum intensity	2637	30	2637	0	match
configuration E	Diagnostics-initial ROI	Int. mask dimension x	204	0	204	0	match
configuration E	Diagnostics-initial ROI	Int. mask dimension y	201	0	201	0	match
configuration E	Diagnostics-initial ROI	Int. mask dimension z	60	0	60	0	match
configuration E	Diagnostics-initial ROI	Int. mask bounding box dimension x	100	0	100	0	match
configuration E	Diagnostics-initial ROI	Int. mask bounding box dimension y	99	0	99	0	match
configuration E	Diagnostics-initial ROI	Int. mask bounding box dimension z	26	0	26	0	match
configuration E	Diagnostics-initial ROI	Morph. mask bounding box dimension x	100	0	100	0	match
configuration E	Diagnostics-initial ROI	Morph. mask bounding box dimension y	99	0	99	0	match
configuration E	Diagnostics-initial ROI	Morph. mask bounding box dimension z	26	0	26	0	match
configuration E	Diagnostics-initial ROI	Int. mask voxel count	125256	0	125256	0	match
configuration E	Diagnostics-initial ROI	Morph. mask voxel count	125256	0	125256	0	match
configuration E	Diagnostics-initial ROI	Int. mask mean intensity	-46.9	0	-46.8827	0	match
configuration E	Diagnostics-initial ROI	Int. mask minimum intensity	-1000	0	-1000	0	match
configuration E	Diagnostics-initial ROI	Int. mask maximum intensity	723	0	723	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask dimension x	100	1	100	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	Diagnostics-interpolated ROI	Int. mask dimension y	99	0.8	99	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask dimension z	90	0	90	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask bounding box dimension x	49	0.2	49	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask bounding box dimension z	40	0.1	40	0	match
configuration E	Diagnostics-interpolated ROI	Morph. mask bounding box dimension x	49	0.2	49	0	match
configuration E	Diagnostics-interpolated ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration E	Diagnostics-interpolated ROI	Morph. mask bounding box dimension z	40	0.1	40	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask voxel count	45985	100	45985	0	match
configuration E	Diagnostics-interpolated ROI	Morph. mask voxel count	45985	100	45985	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask mean intensity	-48.3	0.1	-48.2783	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask minimum intensity	-966	1	-966	0	match
configuration E	Diagnostics-interpolated ROI	Int. mask maximum intensity	627	5	627	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask dimension x	100	1	100	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask dimension y	99	0.8	99	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask dimension z	90	0	90	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask bounding box dimension x	49	0.3	49	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask bounding box dimension y	49	0.3	49	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask bounding box dimension z	40	0.3	40	0	match
configuration E	Diagnostics-resegmented ROI	Morph. mask bounding box dimension x	49	0.3	49	0	match
configuration E	Diagnostics-resegmented ROI	Morph. mask bounding box dimension y	49	0.3	49	0	match
configuration E	Diagnostics-resegmented ROI	Morph. mask bounding box dimension z	40	0.3	40	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask voxel count	44484	800	44484	0	match
configuration E	Diagnostics-resegmented ROI	Morph. mask voxel count	45985	700	45985	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask mean intensity	-22.6	4.1	-22.6256	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask minimum intensity	-743	13	-743	0	match
configuration E	Diagnostics-resegmented ROI	Int. mask maximum intensity	345	9	345	0	match
configuration E	Morphology	Volume (mesh)	367000	6000	367453.6667	0	match
configuration E	Morphology	Volume (voxel counting)	368000	6000	367880	0	match
configuration E	Morphology	Surface area (mesh)	34300	400	34306.252	0	match
configuration E	Morphology	Surface to volume ratio	0.0934	0.0007	0.093362	0	match
configuration E	Morphology	Compactness 1	0.0326	0.0002	0.032626	0	match
configuration E	Morphology	Compactness 2	0.378	0.004	0.37821	0	match
configuration E	Morphology	Spherical disproportion	1.38	0.01	1.3828	0	match
configuration E	Morphology	Sphericity	0.723	0.003	0.72318	0	match
configuration E	Morphology	Asphericity	0.383	0.004	0.38278	0	match
configuration E	Morphology	Centre of mass shift	68.5	2.1	68.5402	0	match
configuration E	Morphology	Maximum 3D diameter	125	1	125.06	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	Morphology	Major axis length	93.3	0.5	93.2704	0	match
configuration E	Morphology	Minor axis length	82	0.5	82.0052	0	match
configuration E	Morphology	Least axis length	70.9	0.4	70.9015	0	match
configuration E	Morphology	Elongation	0.879	0.001	0.87922	0	match
configuration E	Morphology	Flatness	0.76	0.001	0.76017	0	match
configuration E	Morphology	Volume density (AABB)	0.478	0.003	0.47826	0	match
configuration E	Morphology	Area density (AABB)	0.678	0.003	0.67842	0	match
configuration E	Morphology	Volume density (OMBB)					
configuration E	Morphology	Area density (OMBB)	0.69	0.002			
configuration E	Morphology	Volume density (AEE)	1.29	0.01	1.2941	0	match
configuration E	Morphology	Area density (AEE)	1.62	0.01	1.6168	0	match
configuration E	Morphology	Volume density (MVEE)					
configuration E	Morphology	Area density (MVEE)					
configuration E	Morphology	Volume density (convex hull)	0.834	0.002	0.83366	0	match
configuration E	Morphology	Area density (convex hull)	1.13	0.01	1.1301	0	match
configuration E	Morphology	Integrated intensity	-8310000	1600000	-8313866.368	0	match
configuration E	Morphology	Moran’s I index	0.0596	0.0014			
configuration E	Morphology	Geary’s C measure	0.853	0.001			
configuration E	Local intensity	Local intensity peak	181	13			
configuration E	Local intensity	Global intensity peak	181	5			
configuration E	Statistics	Mean	-22.6	4.1	-22.6256	0	match
configuration E	Statistics	Variance	35100	2200	35098.3231	0	match
configuration E	Statistics	Skewness	-2.3	0.07	-2.3005	0	match
configuration E	Statistics	(Excess) kurtosis	4.44	0.33	4.441	0	match
configuration E	Statistics	Median	43	0.5	43	0	match
configuration E	Statistics	Minimum	-743	13	-743	0	match
configuration E	Statistics	10th percentile	-310	21	-310	0	match
configuration E	Statistics	90th percentile	93	0.2	93	0	match
configuration E	Statistics	Maximum	345	9	345	0	match
configuration E	Statistics	Interquartile range	62	3.5	62	0	match
configuration E	Statistics	Range	1090	30	1088	0	match
configuration E	Statistics	Mean absolute deviation	125	6	125.3221	0	match
configuration E	Statistics	Robust mean absolute deviation	46.5	3.7	46.4508	0	match
configuration E	Statistics	Median absolute deviation	97.9	3.9	97.8686	0	match
configuration E	Statistics	Coefficient of variation	-8.28	0.95	-8.2802	0	match
configuration E	Statistics	Quartile coefficient of dispersion	0.795	0.337	0.79487	0	match
configuration E	Statistics	Energy	1580000000	140000000	1584085992	0	match
configuration E	Statistics	Root mean square	189	7	188.7068	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	Intensity histogram	Mean	21.7	0.3	21.7047	0	match
configuration E	Intensity histogram	Variance	30.4	0.8	30.4237	0	match
configuration E	Intensity histogram	Skewness	-2.29	0.07	-2.2894	0	match
configuration E	Intensity histogram	(Excess) kurtosis	4.4	0.33	4.405	0	match
configuration E	Intensity histogram	Median	24	0.2	24	0	match
configuration E	Intensity histogram	Minimum	1	0	1	0	match
configuration E	Intensity histogram	10th percentile	13	0.7	13	0	match
configuration E	Intensity histogram	90th percentile	25	0.2	25	0	match
configuration E	Intensity histogram	Maximum	32	0	32	0	match
configuration E	Intensity histogram	Mode	24	0.1	24	0	match
configuration E	Intensity histogram	Interquartile range	1	0.06	1	0	match
configuration E	Intensity histogram	Range	31	0	31	0	match
configuration E	Intensity histogram	Mean absolute deviation	3.69	0.1	3.6878	0	match
configuration E	Intensity histogram	Robust mean absolute deviation	1.46	0.09	1.4554	0	match
configuration E	Intensity histogram	Median absolute deviation	2.89	0.07	2.8934	0	match
configuration E	Intensity histogram	Coefficient of variation	0.254	0.006	0.25413	0	match
configuration E	Intensity histogram	Quartile coefficient of dispersion	0.0213	0.0015	0.021277	0	match
configuration E	Intensity histogram	Entropy	3.22	0.02	3.2214	0	match
configuration E	Intensity histogram	Uniformity	0.184	0.001	0.18362	0	match
configuration E	Intensity histogram	Maximum histogram gradient	6010	130	6010	0	match
configuration E	Intensity histogram	Maximum histogram gradient intensity	23	0.2	23	0	match
configuration E	Intensity histogram	Minimum histogram gradient	-6110	180	-6110	0	match
configuration E	Intensity histogram	Minimum histogram gradient intensity	25	0.2	25	0	match
configuration E	Intensity volume histogram	Volume fraction at 10% intensity	0.975	0.002	0.97455	0	match
configuration E	Intensity volume histogram	Volume fraction at 90% intensity	0.000157	0.000248	0.00015736	0	match
configuration E	Intensity volume histogram	Intensity at 10% volume	770	5	770	0	match
configuration E	Intensity volume histogram	Intensity at 90% volume	399	17	399	0	match
configuration E	Intensity volume histogram	Volume fraction difference between 10% and 90% intensity	0.974	0.001	0.9744	0	match
configuration E	Intensity volume histogram	Intensity difference between 10% and 90% volume	371	13	371	0	match
configuration E	Intensity volume histogram	Area under the IVH curve	0.663	0.006	0.66277	0	match
configuration E	GLCM (3D averaged)	Joint maximum	0.153	0.003	0.15302	0	match
configuration E	GLCM (3D averaged)	Joint average	22.1	0.3	22.1321	0	match
configuration E	GLCM (3D averaged)	Joint variance	24.4	0.9	24.4306	0	match
configuration E	GLCM (3D averaged)	Joint entropy	5.6	0.03	5.5967	0	match
configuration E	GLCM (3D averaged)	Difference average	1.7	0.01	1.6988	0	match
configuration E	GLCM (3D averaged)	Difference variance	8.22	0.06	8.22	0	match
configuration E	GLCM (3D averaged)	Difference entropy	2.39	0.01	2.3934	0	match
configuration E	GLCM (3D averaged)	Sum average	44.3	0.4	44.2641	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	GLCM (3D averaged)	Sum variance	86.6	3.3	86.5761	0	match
configuration E	GLCM (3D averaged)	Sum entropy	3.96	0.02	3.9639	0	match
configuration E	GLCM (3D averaged)	Angular second moment	0.0638	0.0009	0.063763	0	match
configuration E	GLCM (3D averaged)	Contrast	11.1	0.1	11.1464	0	match
configuration E	GLCM (3D averaged)	Dissimilarity	1.7	0.01	1.6988	0	match
configuration E	GLCM (3D averaged)	Inverse difference	0.608	0.001	0.60798	0	match
configuration E	GLCM (3D averaged)	Normalised inverse difference	0.955	0.001	0.95514	0	match
configuration E	GLCM (3D averaged)	Inverse difference moment	0.576	0.001	0.57643	0	match
configuration E	GLCM (3D averaged)	Normalised inverse difference moment	0.99	0.001	0.99044	0	match
configuration E	GLCM (3D averaged)	Inverse variance	0.41	0.004	0.41006	0	match
configuration E	GLCM (3D averaged)	Correlation	0.771	0.006	0.77067	0	match
configuration E	GLCM (3D averaged)	Autocorrelation	509	8	508.6877	0	match
configuration E	GLCM (3D averaged)	Cluster tendency	86.6	3.3	86.5761	0	match
configuration E	GLCM (3D averaged)	Cluster shade	-2070	70	-2072.4674	0	match
configuration E	GLCM (3D averaged)	Cluster prominence	68900	2100	68901.0958	0	match
configuration E	GLCM (3D averaged)	Information correlation 1	-0.181	0.003	-0.18057	0	match
configuration E	GLCM (3D averaged)	Information correlation 2	0.813	0.004	0.8126	0	match
configuration E	GLCM (3D merged)	Joint maximum	0.153	0.003	0.15312	0	match
configuration E	GLCM (3D merged)	Joint average	22.1	0.3	22.1312	0	match
configuration E	GLCM (3D merged)	Joint variance	24.4	0.9	24.4431	0	match
configuration E	GLCM (3D merged)	Joint entropy	5.61	0.03	5.6143	0	match
configuration E	GLCM (3D merged)	Difference average	1.7	0.01	1.6957	0	match
configuration E	GLCM (3D merged)	Difference variance	8.23	0.06	8.234	0	match
configuration E	GLCM (3D merged)	Difference entropy	2.4	0.01	2.398	0	match
configuration E	GLCM (3D merged)	Sum average	44.3	0.4	44.2625	0	match
configuration E	GLCM (3D merged)	Sum variance	86.7	3.3	86.6628	0	match
configuration E	GLCM (3D merged)	Sum entropy	3.97	0.02	3.9669	0	match
configuration E	GLCM (3D merged)	Angular second moment	0.0635	0.0009	0.063526	0	match
configuration E	GLCM (3D merged)	Contrast	11.1	0.1	11.1095	0	match
configuration E	GLCM (3D merged)	Dissimilarity	1.7	0.01	1.6957	0	match
configuration E	GLCM (3D merged)	Inverse difference	0.608	0.001	0.60828	0	match
configuration E	GLCM (3D merged)	Normalised inverse difference	0.955	0.001	0.95521	0	match
configuration E	GLCM (3D merged)	Inverse difference moment	0.577	0.001	0.57676	0	match
configuration E	GLCM (3D merged)	Normalised inverse difference moment	0.99	0.001	0.99046	0	match
configuration E	GLCM (3D merged)	Inverse variance	0.41	0.004	0.41004	0	match
configuration E	GLCM (3D merged)	Correlation	0.773	0.006	0.77275	0	match
configuration E	GLCM (3D merged)	Autocorrelation	509	8	508.6804	0	match
configuration E	GLCM (3D merged)	Cluster tendency	86.7	3.3	86.6628	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	GLCM (3D merged)	Cluster shade	-2080	70	-2076.0801	0	match
configuration E	GLCM (3D merged)	Cluster prominence	69000	2100	69042.6344	0	match
configuration E	GLCM (3D merged)	Information correlation 1	-0.175	0.003	-0.1754	0	match
configuration E	GLCM (3D merged)	Information correlation 2	0.813	0.004	0.81253	0	match
configuration E	GLRLM (3D averaged)	Short runs emphasis	0.776	0.001	0.77615	0	match
configuration E	GLRLM (3D averaged)	Long runs emphasis	3.55	0.07	3.5494	0	match
configuration E	GLRLM (3D averaged)	Low grey level run emphasis	0.0204	0.0008	0.020378	0	match
configuration E	GLRLM (3D averaged)	High grey level run emphasis	471	9	471.1768	0	match
configuration E	GLRLM (3D averaged)	Short run low grey level emphasis	0.0187	0.0007	0.01866	0	match
configuration E	GLRLM (3D averaged)	Short run high grey level emphasis	346	7	346.4779	0	match
configuration E	GLRLM (3D averaged)	Long run low grey level emphasis	0.0313	0.0016	0.031305	0	match
configuration E	GLRLM (3D averaged)	Long run high grey level emphasis	1900	20	1903.9705	0	match
configuration E	GLRLM (3D averaged)	Grey level non-uniformity	4000	10	3997.5752	0	match
configuration E	GLRLM (3D averaged)	Normalised grey level non-uniformity	0.135	0.003	0.13519	0	match
configuration E	GLRLM (3D averaged)	Run length non-uniformity	16600	300	16559.9259	0	match
configuration E	GLRLM (3D averaged)	Normalised run length non-uniformity	0.559	0.001	0.55942	0	match
configuration E	GLRLM (3D averaged)	Run percentage	0.664	0.003	0.66383	0	match
configuration E	GLRLM (3D averaged)	Grey level variance	39.8	0.9	39.7538	0	match
configuration E	GLRLM (3D averaged)	Run length variance	1.26	0.05	1.2633	0	match
configuration E	GLRLM (3D averaged)	Run entropy	4.87	0.03	4.8683	0	match
configuration E	GLRLM (3D merged)	Short runs emphasis	0.777	0.001	0.77709	0	match
configuration E	GLRLM (3D merged)	Long runs emphasis	3.52	0.07	3.5221	0	match
configuration E	GLRLM (3D merged)	Low grey level run emphasis	0.0204	0.0008	0.020352	0	match
configuration E	GLRLM (3D merged)	High grey level run emphasis	471	9	471.3789	0	match
configuration E	GLRLM (3D merged)	Short run low grey level emphasis	0.0186	0.0007	0.018648	0	match
configuration E	GLRLM (3D merged)	Short run high grey level emphasis	347	7	347.1823	0	match
configuration E	GLRLM (3D merged)	Long run low grey level emphasis	0.0311	0.0016	0.031149	0	match
configuration E	GLRLM (3D merged)	Long run high grey level emphasis	1890	20	1888.785	0	match
configuration E	GLRLM (3D merged)	Grey level non-uniformity	51900	200	51949.1185	0	match
configuration E	GLRLM (3D merged)	Normalised grey level non-uniformity	0.135	0.003	0.13532	0	match
configuration E	GLRLM (3D merged)	Run length non-uniformity	215000	4000	215058.7889	0	match
configuration E	GLRLM (3D merged)	Normalised run length non-uniformity	0.56	0.001	0.56021	0	match
configuration E	GLRLM (3D merged)	Run percentage	0.664	0.003	0.66383	0	match
configuration E	GLRLM (3D merged)	Grey level variance	39.7	0.9	39.7226	0	match
configuration E	GLRLM (3D merged)	Run length variance	1.25	0.05	1.2528	0	match
configuration E	GLRLM (3D merged)	Run entropy	4.87	0.03	4.8705	0	match
configuration E	GLSZM (3D)	Small zone emphasis	0.676	0.003	0.6764	0	match
configuration E	GLSZM (3D)	Large zone emphasis	58600	800	58563.985	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	GLSZM (3D)	Low grey level emphasis	0.034	0.0004	0.034	0	match
configuration E	GLSZM (3D)	High grey level emphasis	286	6	285.8554	0	match
configuration E	GLSZM (3D)	Small zone low grey level emphasis	0.0224	0.0004	0.022435	0	match
configuration E	GLSZM (3D)	Small zone high grey level emphasis	186	4	186.016	0	match
configuration E	GLSZM (3D)	Large zone low grey level emphasis	105	4	105.0228	0	match
configuration E	GLSZM (3D)	Large zone high grey level emphasis	33600000	300000	33559401.88	0	match
configuration E	GLSZM (3D)	Grey level non-uniformity	231	6	231.2609	0	match
configuration E	GLSZM (3D)	Normalised grey level non-uniformity	0.0414	0.0003	0.041378	0	match
configuration E	GLSZM (3D)	Zone size non-uniformity	2370	40	2367.8517	0	match
configuration E	GLSZM (3D)	Normalised zone size non-uniformity	0.424	0.004	0.42366	0	match
configuration E	GLSZM (3D)	Zone percentage	0.126	0.001	0.12564	0	match
configuration E	GLSZM (3D)	Grey level variance	50.8	0.9	50.7992	0	match
configuration E	GLSZM (3D)	Zone size variance	58500	800	58500.636	0	match
configuration E	GLSZM (3D)	Zone size entropy	6.57	0.01	6.5652	0	match
configuration E	GLDZM (3D)	Small distance emphasis	0.527	0.004	0.52687	0	match
configuration E	GLDZM (3D)	Large distance emphasis	12.6	0.1	12.5666	0	match
configuration E	GLDZM (3D)	Low grey level emphasis	0.034	0.0004	0.034	0	match
configuration E	GLDZM (3D)	High grey level emphasis	286	6	285.8554	0	match
configuration E	GLDZM (3D)	Small distance low grey level emphasis	0.0228	0.0003	0.022762	0	match
configuration E	GLDZM (3D)	Small distance high grey level emphasis	136	4	136.2357	0	match
configuration E	GLDZM (3D)	Large distance low grey level emphasis	0.179	0.004	0.17853	0	match
configuration E	GLDZM (3D)	Large distance high grey level emphasis	4850	60	4853.9504	0	match
configuration E	GLDZM (3D)	Grey level non-uniformity	231	6	231.2609	0	match
configuration E	GLDZM (3D)	Normalised grey level non-uniformity	0.0414	0.0003	0.041378	0	match
configuration E	GLDZM (3D)	Zone distance non-uniformity	1500	30	1502.7134	0	match
configuration E	GLDZM (3D)	Normalised zone distance non-uniformity	0.269	0.003	0.26887	0	match
configuration E	GLDZM (3D)	Zone percentage	0.126	0.001	0.12564	0	match
configuration E	GLDZM (3D)	Grey level variance	50.8	0.9	50.7992	0	match
configuration E	GLDZM (3D)	Zone distance variance	5.56	0.05	5.5573	0	match
configuration E	GLDZM (3D)	Zone distance entropy	7.06	0.01	7.0637	0	match
configuration E	NGTDM (3D)	Coarseness	0.000188	0.000004	0.00018849	0	match
configuration E	NGTDM (3D)	Contrast	0.0752	0.0019	0.075247	0	match
configuration E	NGTDM (3D)	Busyness	4.65	0.1	4.6497	0	match
configuration E	NGTDM (3D)	Complexity	574	1	574.2282	0	match
configuration E	NGTDM (3D)	Strength	0.167	0.006	0.16743	0	match
configuration E	NGLDM (3D)	Low dependence emphasis	0.118	0.001	0.11823	0	match
configuration E	NGLDM (3D)	High dependence emphasis	134	3	134.3185	0	match
configuration E	NGLDM (3D)	Low grey level count emphasis	0.0154	0.0007	0.015389	0	match

Table C1: Cardiff results for every configuration tested in the IBSI consensus study continued.

Dataset	Family	Radiomic Feature	Benchmark	Tolerance	Cardiff	Difference	Check
configuration E	NGLDM (3D)	High grey level count emphasis	502	8	501.5158	0	match
configuration E	NGLDM (3D)	Low dependence low grey level emphasis	0.00388	0.00004	0.0038829	0	match
configuration E	NGLDM (3D)	Low dependence high grey level emphasis	36.7	0.5	36.659	0	match
configuration E	NGLDM (3D)	High dependence low grey level emphasis	0.457	0.031	0.45668	0	match
configuration E	NGLDM (3D)	High dependence high grey level emphasis	76000	600	76003.5803	0	match
configuration E	NGLDM (3D)	Grey level non-uniformity	8170	130	8168.0847	0	match
configuration E	NGLDM (3D)	Normalised grey level non-uniformity	0.184	0.001	0.18362	0	match
configuration E	NGLDM (3D)	Dependence count non-uniformity	2250	30	2246.5056	0	match
configuration E	NGLDM (3D)	Normalised dependence count non-uniformity	0.0505	0.0003	0.050501	0	match
configuration E	NGLDM (3D)	Dependence count percentage	1	0			
configuration E	NGLDM (3D)	Grey level variance	30.4	0.8	30.4237	0	match
configuration E	NGLDM (3D)	Dependence count variance	39.4	1	39.4423	0	match
configuration E	NGLDM (3D)	Dependence count entropy	7.06	0.02	7.0643	0	match
configuration E	NGLDM (3D)	Dependence count energy	0.0106	0.0001	0.010622	0	match