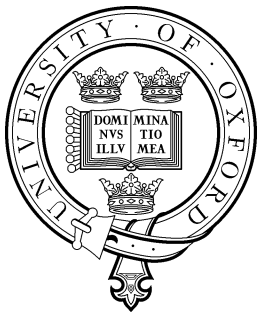
Quantifying Spatial Heterogeneity in Tumours from Dynamic Contrast-Enhanced MRI



Jolanta Mirecka

Mansfield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2019

I dedicate this thesis
to all the researchers, engineers and scientists,
whose selfless contributions, efforts and perseverance
constitute the body of work upon which we, students,
can lay the grounds for our work.

In the midst of green skies and spired horizons
I've hidden my angst towards
you, Science

For the quiet endurance
is taller within you
than the loudest fight

And in those lonely crowds we come
in search for your truth
and not our victory

So one of those nights
ONE way or the OTHER
we shall remain undefeated

Acknowledgements

First and foremost I would like to express my gratitude to my supervisors: Prof. Mark Jenkinson, Prof. Julia Schnabel and Prof. Michael Chappell. I am beyond grateful for all the possible forms of support you have provided, always ensuring I would not run out of resources, interest and inspiration to finish this work. I am grateful for your kind words of encouragement during the writing process, guidance, and above all, your support to follow my research interests. I would like to thank Julia for her efforts to continue regular meetings despite the supervisory change following her move to KCL. I would also like to thank Michael for his support especially taking over the administrative part of the suprevison following supervisory change. And finally, I am especially grateful to Mark, whose endless enthusiasm for science and life provided infinite inspiration.

I am grateful to all my colleagues and collaborators at IBME for providing such a wonderful work environment. Notably, I would like to thank Ben, Bartek and Adam for inspiring conversations constantly stimulating new ideas. I am grateful to the have been a part of the TUSCANI group, whose collective wisdom inspired a new outlook for this research. Furthermore, I am grateful to CRUK, EPRSC and the Oxford Cancer Imaging Centre for providing the necessary funding to complete this work.

I am furthermore thankful to my friends for endless conversations, debugging sessions and brain-storms inspiring the ideas in this thesis. Notably, I would like to thank my rubber-duck team: Szymon, Dan, Hania, Samantha and James - for your willingness to share your knowledge and support during the last few years. I would like to thank all my friends for the immense support and ability to put up with my disappearance during the thesis writing period.

I am eternally grateful to my parents for always inspiring me to pursue science and supporting my atypical lifestyle as a consequence. Without your love, support, and understanding I would not have been able to follow my passion. Your hard work has shaped me into who I am today, for which I am forever grateful.

I am thankful to the mountains for providing seemingly endless grounds to think through the concepts in this thesis, and the meaning of life in general. I would like to express my gratitude to the Oxford University Mountaineering Club, through which, in desperate times, I have discovered a different way of perceiving life that reshaped me as a person. I am thankful to all my adventure and climbing companions for sharing your passion for the outdoors and science. You made the research and writing seem like an adventure in itself in between countless benightments, belays, boulders, bikes, boots and bivvys that we shared together. If there is anything in the world to prepare you for the life of mountaineering, it certainly is the patience and uncertainty required to complete a PhD thesis.

And finally, I would like to thank Dan, whose endless support, inspiring conversations and shared adventures made my life fuller in every possible way leading to the completion of this thesis.

Statement of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at the University of Oxford or any other educational institution, except where due acknowledgement is made in the thesis.

Any contribution made to the research by others, with whom I have worked at the University of Oxford or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Jola Mirecka
April 30, 2020

Abstract

Cancerous tumours are characterized by an inherent perfusion heterogeneity that can change over time and following therapy. The process of angiogenesis often results in increasingly disorganised and leaky vasculature as tumours progress. This may lead to heterogeneous perfusion, which is known to produce locally varied subregions in tumours. Tracking such subregional differences might be the key to understanding and predicting tumour progression, which could be in turn applied to therapy response assessment. Dynamic contrast-enhanced MRI (DCE-MRI) can monitor such changes in perfusion and vascularity, which can then be quantified into a set of parameter maps. Subsequently, segmentation and classification methods are often applied to the maps to define tumour subregions that can be tracked over time. In this work we propose a texture-based Bag of Visual Words framework for the problem of assessing tumour perfusion heterogeneity from DCE-MRI images. We adapt and evaluate a number of shallow feature descriptors and encodings to estimate their suitability to the problem. Furthermore, we combine lightweight deep representations with shallow encodings and recent advances in encoder-decoder architectures to introduce a novel deep image representation for 3D segmentation and subsequent classification of images with no ground truth and limited training data. We show that even in extremely simple forms they are able to outperform the frequently used methods in tumour heterogeneity assessment. Finally, we demonstrate the suitability of the proposed methods in the more clinically-relevant context of therapy response assessment. We employ a number of strategies to improve confidence and transparency of the proposed methods.

Publications

- Jola Mirecka, Benjamin Irving, Danny Allen, Paul Kinchesh, Stuart Gilchrist, Ana Gomes, Veerle Kersemans, Sean Smart, Michael Chappell, Julia Schnabel, and Mark Jenkinson, **Variation-guided supervoxels for subregional tumour analysis in DCE-MRI**, ISMRM, 2018.

- Helen Winter, Jola Mirecka, Bartłomiej Papież, Benjamin Irving, Mark Jenkinson, Michael Chappell, Ricky A. Sharma, and Julia Schnabel, **Clinical feasibility of texture analysis in DCE MRI of patients receiving selective internal radiation therapy**, European Journal of Surgical Oncology 44: S30-S31, 2018. (Abstract)

- Jola Mirecka, Helen Winter, Bartłomiej Papież, Benjamin Irving, Mark Jenkinson, Michael Chappell, Ricky A. Sharma, and Julia Schnabel, **Clinical feasibility of texture analysis in DCE MRI of patients receiving selective internal radiation therapy**, NCRI, 2017.

- Benjamin Irving, Jola Mirecka, Ana Gomes, Danny Allen, Paul Kinchesh, Veerle Kersemans, Stuart Gilchrist, Sean Smart, Julia Schnabel, Michael Brady, Michael Chappell, **Perfusion-supervoxels for DCE-MRI based tumor subregion assessment**, ISMRM, 2017.

- Jola Mirecka, Benjamin Irving, Pavitra Kannan, Veerle Kersemans, Danny Allen, Sean Smart, Paul Kinchesh, Ruth Muschel, Mark Jenkinson, Julia Schnabel **Statistical texture modelling of tumour perfusion heterogeneity in DCE-MRI**, MICCAI workshop: Computational Methods for Molecular Imaging, 2015.

Contents

1 Introduction **1**
1.1 Thesis outline and contributions 3

2 Background **6**
2.1 Cancer and heterogeneity 6
2.1.1 Angiogenesis and perfusion heterogeneity 7
2.1.2 Terminology and scope 8
2.1.3 Heterogeneity in a clinical setting 9
2.2 Cancer imaging modalities 10
2.2.1 X-ray and Computed Tomography 11
2.2.2 Nuclear Imaging 12
2.2.3 Magnetic Resonance Imaging 12
2.2.4 Imaging heterogeneity 13
2.3 Heterogeneity biomarkers 15
2.3.1 Quantifying perfusion 15
2.3.2 Clinical standards 18
2.3.3 Non-spatial histogram-based methods 21
2.3.4 Model-based prior 23
2.3.5 Binary features and thresholding 24
2.3.6 Clustering methods 25
2.3.7 Texture analysis 29
2.3.8 Radiomics 32
2.4 Challenges and limitations 35
2.4.1 Limitations arising from imaging protocol 35
2.4.2 Methodological challenges 36
2.4.3 Clinical translation 38
2.5 Conclusions 38

3 Shallow image representations for segmentation and classification of tumour perfusion heterogeneity in DCE-MRI 41
3.1 Shallow image representations in texture . . . 43
3.1.1 Texture representations . . . 43
3.1.2 Bag of Visual Words pipeline . . . 46
3.2 Methods . . . 51
3.2.1 Texture feature descriptors . . . 51
3.2.1.1 Gabor filters . . . 52
3.2.1.2 Patch descriptors . . . 55
3.2.1.3 Local Binary Patterns . . . 55
3.2.2 Encodings . . . 58
3.3 Results . . . 60
3.3.1 Materials . . . 60
3.3.2 Qualitative evaluation . . . 65
3.3.3 Quantitative evaluation . . . 72
3.3.4 Segmentations . . . 73
3.3.5 Classifications . . . 77
3.3.6 Parameter evaluation . . . 81
3.3.7 Subregional validation . . . 84
3.3.8 Similarity and variation descriptors . . . 86
3.4 Conclusions . . . 88

4 Deep image representations for segmentation and classification of tumour perfusion heterogeneity in DCE-MRI 92
4.1 Deep image representations . . . 94
4.1.1 Deep texture representations for image classification . . . 97
4.1.1.1 Transfer learning . . . 98
4.1.1.2 Deep filter banks . . . 100
4.1.1.3 Fine-tuning . . . 101
4.1.1.4 Handcrafted networks . . . 103
4.1.2 Encoder-decoder architectures for voxel-wise segmentation . . . 105
4.1.2.1 Skip connections . . . 106
4.1.2.2 Upsampling . . . 107
4.2 Methods . . . 109
4.2.1 Deep texture feature descriptors . . . 110
4.2.1.1 Transfer UNet (T-UNet) . . . 110

4.2.1.2 Random UNet (Rand-UNet) . . . 111
4.2.2 Shallow encodings and classification . . . 115
4.3 Experiments . . . 115
4.3.1 Materials . . . 116
4.3.2 Segmentations . . . 117
4.3.3 Classifications . . . 120
4.3.4 Parameter evaluation . . . 123
4.4 Conclusions . . . 125

5 Clinical application: heterogeneity-based therapy response assessment with deep and shallow texture representations 129
5.1 Heterogeneity and therapy response . . . 130
5.1.1 Limitations of DCE-MRI biomarkers . . . 131
5.1.2 Confidence and transparence . . . 133
5.1.3 Ambiguity . . . 134
5.2 Methods . . . 135
5.2.1 Feature descriptors and encodings . . . 135
5.2.2 Classification . . . 136
5.3 Results . . . 136
5.3.1 Materials . . . 137
5.3.2 Segmentations . . . 142
5.3.3 Individual evaluation . . . 143
5.4 Conclusions . . . 146

6 Conclusions 150
6.1 Summary and discussion . . . 150
6.1.1 Technical . . . 150
6.1.2 Biological and clinical . . . 153
6.2 Limitations . . . 154
6.3 Outlook . . . 155
6.3.1 Shallow encoding evaluation . . . 156
6.3.2 Deep hybrid representations . . . 156
6.3.3 Clinical evaluation . . . 157

Bibliography 159

List of Figures

2.1 An axial slice of a DCE-MRI 4D volume (a) before enhancement, (b) during enhancement, (c) overlaid with K^{trans} perfusion parameter map. 18

2.2 Images of different textures (a) homogeneous, (b) locally homogeneous and (c) heterogeneous with identical grey-level distributions and bimodal histograms (image acquired from L. Alic [16]). 23

2.3 Summary of various simple methods to quantifying spatial intratumoural heterogeneity (image acquired from [1]). Such methods can be used either in isolation as described in Subsections 2.3.1-2.3.6, or as basic building blocks for further, more complex methods, e.g. clustering in Bag of Visual Words encodings (discussed in Chapter 3), features used for radiomics (Subsection 2.3.8) or binary features to estimate texture fractal dimension (Subsection 2.3.7) [14]. The last row corresponds to the parameter that is usually the basis for segmentation (BM - biomarkers, ADC - Apparent Diffusion Coefficient). 30

3.1 Modern texture representation - a schematic example of BoV pipeline. Feature detection step automatically identifies a number of keypoints or areas of interest. These are then input into a dense feature extractor to describe the image in terms of its local neighbouring area. A number of the most descriptive features are gathered into a set of visual words, which constitutes the basis for learning the global labelling. Finally, an encoding is generated describing a global distribution of the words, which is used as an input to a machine learning classifier. 47

3.2 Three examples of methods [96, 83, 88] in the context of the standard texture analysis pipeline, as shown in Fig. 3.1. X was put in place of a method if the step was skipped in the original methodology. 50

3.3 Top: Image superposition property, star pattern is decomposed into a number of bases with different base filters and orientations. Image taken from [105]. Bottom: A: Receptive field of a simple cell in V1, B: image corresponding to the highest response, C: Gabor function approximating the visual receptive field. Image taken from [104]. . . . 52
3.4 Left: 3D contour of a Gabor filter, middle three: overlapping planes overlayed with different views of the same Gabor filter, right: overlapping planes overlayed with a Gaussian filter. 54
3.5 Illustration of the differences between the chosen texture descriptors: a: filtering methods, b: patch-based methods, and c: local binary patterns. 58
3.6 Changes in tumour statistics in time (4 days*) for all three mice in D1. Top row showing changes in WTV, median and mean K^{trans} in that order. Bottom row showing changes in EF at 100s, 300s and 500s in that order. * Raw DCE data was missing for all mice on day 4, thus EF was only calculated until day 3. 66
3.7 Enhancement trace for each type of visual word (texton). Left column: sagittal slice across the tumour showing raw DCE image from dataset D1 sample M01 day 3 and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour. . . 67
3.8 Longitudinal change in subregional distributions for a single mouse (M03) in D1. Columns relate to days of progression, with the leftmost column being day 1 and rightmost column day 4. First row corresponds to raw DCE-MRI images, second row represents K^{trans} maps (overlaid on DCE image) and third row shows subregional segmentations (overlaid on DCE image). No DCE-MRI data for M03_day4 was present in the dataset (the segmentation for that sample was overlaid on gray-scale K^{trans} map instead). 68

3.9 A change in global frequency distribution of local perfusion patterns in longitudinal study (CaNT tumour model) for day 1 (top) and day 4 (bottom). This relates to our intuition about the change in tumour heterogeneity, homogeneous at first (only a limited number of similar textures frequently appearing), to highly heterogeneous (all textures frequently appearing). Bins in the histograms are corresponding in fixed order to the list of local perfusion patterns shown on the right. (M02 was missing a sample from day 4 due to noise in acquisition and during this analysis it was replaced with the sample from day 3). . . . 69

3.10 Global frequency distribution patterns in single scans categorised by tumour volume (below median volume value (top) or over median volume value (bottom) for MC38 tumour model). Bins in the histograms are corresponding in the fixed order to the list of local perfusion patterns shown on the right. Note that due to the larger tumour volumes, more bins are filled in the single timepoint histogram (increased area), but their shape (distribution) remains similar. 71

3.11 Enhancement trace for each type of visual word (texton). Left column: sagittal slice across the tumour showing raw DCE image from dataset D3 sample C96_R2_D6 and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour. . . 74

3.12 Subregional distributions of a growing tumour (columns relate to days of growth) shown for three different samples with Patch-KM. Pairs of rows (1, 2), (3, 4) and (5, 6) originated from three different tumours C96_R2, C97_NO and C98_R2 respectively (Table 3.4). Where a timepoint was missing from the series, labels for all individual mice were provided in top-bottom order. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). We are observing a change similar to the one observed in the trial study. 75

3.13 Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of feature descriptors in KMeans BoV encoding, from the top: Patch, LBP, Gabor, Vox. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). 78
3.14 Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of feature descriptors in GMM BoV encoding, from the top: Patch, LBP, Gabor, Vox. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). 79
3.15 Confidence heatmaps for segmentations. High values (black) correspond to high confidence, while low values (yellow) indicate low confidence in segmentations. Most voxels show an agreement of above 50% over different segmentation methods. 80
3.16 The effect of varying the k parameter (number of clusters) on the subregional distributions of the same sample. The rows correspond to $k = \{2 \times 2, 2 \times 3, 2 \times 4, 4, 6, 8\}$ respectively with the upper three generated with non-global and lower three with global variant of BoV encodings. Columns correspond to the days of growth. 83
3.17 Examples of similarities between the Patch-KM variant of the framework (top row) and manually labelled ground truth (GT) data from a single observer (bottom row) for three different samples. All three axes are shown for each of the samples and they correspond across the rows. Colors are arbitrary and relate to different visual words representing diffent tissue properties: enhancing tumour, non-enhancing tumour and necrosis. 85
3.18 Examples of failed segmentations (left), corresponding GT segmentations (middle) and corresponding T1W signal. In our method, a majorty of voxels were assigned a single label (maroon) corresponding to a non-enhancing tissue in the ground-truth (and thus likely a homogeneous one). T1W signal for the same area reveals mostly homogeneous tissue, which suggests that in this case the ground truth labels were not characterised with a degree of heterogeneity. 86

3.19 Comparison of different subregional segmentations on the same sample of a progressing tumour (columns relate to days of progression). First row represents visual words learned with supervoxels (local similarities). Second row relates to visual words learned from Patch (local variation). 88

4.1 An overview of the architectures used as feature extractors. Top: UNet, bottom RandNet. 113

4.2 Subregional distributions of a growing tumour (columns relate to days of growth) shown for two different samples with KM+TUNet and KM+RandUNet as compared to Patch+KM and baseline enhancement. Original image mid-way through enhancement (row 1 and 5), Patch+KM (row 2 and 6), KM+TUNet (row 3 and 7) and KM+RandUNet (row 4 and 8). Rows (1, 2, 3, 4) and (5, 6, 7, 8) originated from two different tumours C97_NO and C98_R2 respectively (Table 3.4). Where a timepoint was missing from the series, labels for all individual mice were provided in top-bottom order. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). Note the slices may bare differences due to difference in cropping, however slices are corresponding between the rows. 119

4.3 Subregional distributions of a growing tumour C96_R2 (columns relate to days of growth) shown for a number of shallow and deep representations with KMeans BoV encoding, from the top: LBP, Patch, UNet, RandNet. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of labels is constant and equal to 3). 121

4.4 Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of shallow and deep representations with GMM BoV encoding, from the top: LBP, Patch, UNet, RandNet. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of labels is constant and equal to 3). 122

4.5 Confidence heatmaps for segmentations. High values (black) correspond to high confidence, while low values (yellow) indicate low confidence in segmentations. Most voxels show an agreement of above 50% over different segmentation methods. 122
4.6 The effect of varying the d parameter (depth of the network) on the subregional distributions of the same sample. The rows correspond to $d = 2, 3, 4$ respectively, with the first row showing mid-point through enhancement. Columns correspond to the days of growth. 124

5.1 Changes in tumour statistics in time for all the patients in D4. Top left showing changes in WTV, top right showing changes in EF at 15s, bottom left 35s and bottom right 60s. No K^{trans} or T10 maps were available to calculate K^{trans} changes (no PK analysis was done for this chapter). 140
5.2 Enhancement trace for each type of visual word (texton). Left column: axial slice across the tumour showing raw DCE image from dataset D4 sample ART_120 Before and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour. 141
5.3 Sample 120 shown pre-therapy (top) and mid-therapy (bottom) for all the methods under evaluation. From left: mid-enhancement point, Vox, Gabor, Patch, LBP, UNet, RandNet, confidence heatmaps. . . . 144
5.4 Left: SAD values indicating percent of absolute change for each patient tumour (legend) and each method (x axis). Dashed lines indicate patients who have not achieved a complete response following 3 months of CRT. Right: Average SAD values across all methods with min/max error bars for each patient (x axis). 147

List of Tables

2.1 Summary of methods for tumour heterogeneity quantification outlined in section 2.3. 34

3.1 Each cohort used in this thesis. Details of the datasets D1, D2 and D3 can be found in Table 3.2 in this section. Details of the dataset D4 can be found in Table 5.1 in chapter 5 where analysis was performed. . . 61
3.2 Details of each dataset used in qualitative and quantitative study in Chapter 3 and 4. * 2 scans were unusable due to noise in acquisition, for details see Table 3.3. ** 4 scans were missing and 9 were unusable due to noise in acquisition, for details see Table 3.4. 62
3.3 D1 longitudinal data used in qualitative evaluation. 12 scans in total, 10 usable scans. Scans marked with × were unusable due to noise in acquisition. D2 is not listed as it contained 11 tumours all of which were used and it was not a longitudinal study. 62
3.4 D3 longitudinal data used in quantitative evaluation. 82 scans in total, 69 usable scans. Scans marked with × were unusable due to noise in acquisition; scans marked with ? were missing from the dataset. . . . 63
3.5 Classification accuracy scores and confidence DICE measures for various feature descriptors and codebook generators, all sharing the same BoV orderless normalised frequency encoding. 81
3.6 The effect of varying parameter k (number of clusters) on classification accuracy. Top row illustrates the results from non-global variant of BoV encoding, while the bottom row reflets on the global variant results. The columns refer to different value of k parameter. 84

4.1 Classification accuracy scores and confidence DICE measures for various methods and encodings, all sharing the same BoV pipeline. The best result are emphasised in bold. 123

4.2 The classification accuracy as a function of increasing the depth of a UNet architecture in the context of BoV encoding. 125

5.1 Details of the data used in this chapter. * Original study contained 23 patients, however 8 of them did not have enough timepoints to qualify for DCE-MRI pharmacokinetic analysis. ** 2 samples (4 scans) were unusable due to irrecoverable motion, 1 sample (2 scans) was unusable due to the lack of a correct ROI, for details see Table 5.2. 139
5.2 D4 clinical data used in this chapter. For details of the data see Table 5.1. 30 scans in total, 24 usable scans (12 pairs). Scans marked with × were unusable due to noise in acquisition or missing ROI. Patients marked in bold have not achieved a complete response at three months following CRT. 139
5.3 DICE scores for all methods under evaluation using the KM variant. . 143
5.4 SAD indicating percent of absolute change for each patients' tumour (rows) and each method (columns). Bold indicates patients who did not achieve a complete response at 3 months following CRT. 145

Chapter 1

Introduction

Tumours are characterised by spatial variation and heterogeneity in their physiological, bio-chemical and structural appearance, which is thought to be a reflection of underlying genetic heterogeneity [1] and hallmarks of cancer [2] such as angiogenesis. In tumours, the process of angiogenesis often results in increasingly more disorganised and leaky vasculature as tumours progress, consequently leading to increasingly heterogeneous perfusion [3, 4]. Recognizing such change in perfusion heterogeneity over time may prove prognostic[1] [5] in understanding tumour progression. Furthermore, as spatial heterogeneity can indicate poor clinical prognosis [6] and can change following therapy [7, 8], tracking the change in heterogeneity over time could constitute a basis to personalised, guided therapy or even prediction of therapy outcome. Finally, with the advent of anti-angiogenic drugs affecting tumour blood supply, some traditional biomarkers are no longer adequate. With the need for novel biomarkers, tumour heterogeneity has become a topic of interest as a potential predictive[2] [5] indicator [1].

[1]Prognostic - a biomarker used to identify likelihood of a clinical event, disease recurrence or progression; or a biomarker that provides information about the natural history of the cancer in an individual in the absence of treatment [5].

[2]Predictive - a biomarker used to identify individuals who are more likely to experience a favourable or unfavourable effect from a specific intervention or exposure [5].

Radiology-level imaging[3] depicts the apparent intra-tumoural variation, which can, in turn, be quantified. Although imaging is key in clinical diagnosis, staging and response assessment, most clinical studies are only limited to visual qualitative or weakly quantitative evaluation of parameters [9] such as tumour size (e.g. RECIST [10]) or average parameter values (e.g. median blood flow [11]). While visual assessment can be subjective and prone to inter- and intra-observer biases [12], more accurate assessment of the tumour size could be achieved by considering 3D measurements. Additionally, global averages are also suboptimal by discarding some of the useful spatial information contained in the images. This introduces a need for more spatial and quantitative methods that are capable of capturing heterogeneity.

Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) has shown great potential in evaluating tumour progression and response to therapy due to its ability to monitor local perfusion changes [13, 14]. Although its non-invasive nature poses no known side effects to the human body, DCE-MRI is not currently of frequent use in the clinic. Pharmacokinetic (PK) modelling is commonly used to quantify perfusion information from DCE-MRI images, resulting in a set of parameter maps. Although some parameters, such as median K^{trans} values, have already been identified as pharmacokinetic biomarkers [11, 15], they are suboptimal by not accounting for local perfusion patterns and discard valuable spatial information. As it has been shown [16] that different heterogeneous appearances can result in similar first order statistics (Fig. 2.2), it is necessary to include spatial methods in heterogeneity quantification. To quantify the heterogeneity of perfusion, spatial methods including texture analysis are sometimes performed on parameter maps.

Texture analysis is an active field of research, deeply rooted in computer vision. The research and development in texture representations was found to generalise well

[3]Radiology-level imaging includes plain radiography, ultrasound, computed tomography (CT), nuclear medicine (including positron emission tomography (PET)), magnetic resonance imaging (MRI) and its derivatives.

into other domains and so it gave rise to modern shallow image processing pipelines in computer vision in general, including numerous areas of application such as facial recognition, semantic segmentation, object recognition or scene classification, amongst many others [17]. The datasets used for evaluation of texture methods are typically less spatially complex than those used for object-level description, where spatial relationships between the objects are often of value [17, 18]. A similar trend occurs in biomedical images, which provide an abundance of repetitive patterns due to their three dimensional nature, however the relationships between those patterns are not as spatially complex as it is in the case for objects [19]. Texture research could therefore prove instrumental in providing solutions for field such as biomedical image analysis, particularly where there exists substantial background research demonstrating an inherent variation, or heterogeneity, reflected in the images. In this work, we therefore adapt and propose various texture-based solutions to the problem, considering texture as a proxy measure of heterogeneity.

1.1 Thesis outline and contributions

The goal of this project is to evaluate, adapt and propose new texture-based methods for growth and therapy response monitoring. In addition, we aim to test their ability to assess tumour heterogeneity and explore their suitability and potential as future imaging biomarkers[4]. This is designed to bridge a gap in tumour heterogeneity analysis by providing new ways to quantify heterogeneity. Furthermore, the contribution goes both ways, as it also serves as a testbed for currently available methods in computer vision to be tested "in the wild" on challenging data. The ultimate goal of this project is to obtain subregional proportion representation for tumours from heterogeneity-based subregions that could serve as a future biomarker for growth and

[4]Imaging biomarker - a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions [5].

therapy outcome enabling for comparisons to be made within and across multiple studies.

We begin by posing the problem in the clinical and pre-clinical settings in Chapter 2, and follow with a detailed review of the existing state-of-the-art methods used in the context of tumour heterogeneity assessment. The subsequent three chapters contribute to the field in three distinct ways: method comparison, technical advancement, and clinical relevance, respectively. Quantifying perfusion heterogeneity during natural history while unaffected by therapy, is important as a baseline for further therapy monitoring, planning and outcome prediction. We will therefore begin by posing the problem in terms of characterizing the change in heterogeneity during longitudinal growth of the tumour. We will evaluate the methods in the following chapters with respect to two tasks: 1) overall heterogeneity quantification; and 2) subregional tumour segmentation. In Chapter 3 we propose a Bag of Visual Words framework for the problem. We will adapt and evaluate a number of texture-based shallow image representations for quantifying longitudinal tumour variation. We will further examine the suitability of the introduced framework for the problem of assessing tumour perfusion heterogeneity. Chapter 4 will introduce novel deep texture representations that are particularly suitable for small datasets with no segmentation ground truth. We will show that even in such simple form the methods are capable of outperforming the shallow representations. Following this, Chapter 5 redefines the problem in terms of a more clinically-relevant question of heterogeneity-based therapy response assessment and explores the relevance of methods from the previous chapters in a clinical setting. Most importantly, we will outline a number of strategies taken in the design of the methods that could improve the transparency and confidence of the produced results. We believe that while the validation of DCE-MRI heterogeneity-based outcomes is an extremely challenging task, such level of transparency could provide an alternative way, contributing to their future potential for clinical translation. Finally,

Chapter 6 will provide conclusions and an outlook for future research.

We briefly summarise the key contributions of the thesis:

- We adapt a texture-based Bag of Visual Words framework to the problem of assessing tumour perfusion heterogeneity.
- We evaluate a number of shallow feature descriptors and encodings on a challenging problem in the real-world application domain in 3D.
- We evaluate the ability of the proposed framework to meaningfully describe the physiological processes in the tumour micro-environment.
- We introduce a novel framework for the analysis of small datasets with no segmentation ground truth in 3D as a fusion between deep and shallow representations joint with encoder-decoder architectures.
- We adapt light-weight deep image representations as feature extractors in the proposed framework, and extend them for the task of segmentation in 3D.
- We employ a number of strategies to improve the confidence and transparency in the presented framework to improve the likelihood of future clinical translation.
- We propose using the research presented in this work as an ensemble of methods capable of generating a set of likely hypotheses in order to aid the clinical decisions based on those results.

Chapter 2

Background

In this chapter, Sec. 2.1. briefly introduces cancer genetics and biology, with a particular focus on angiogenesis and its influence on perfusion heterogeneity. We further outline the role of intratumour heterogeneity in clinical therapy. This is followed by a discussion of the available tools to assess such heterogeneity in Sec. 2.2., principally focusing on routinely used imaging techniques. Subsequently in Sec. 2.3. we describe common methods for quantifying heterogeneity from such images (DCE-MRI in particular), including an outline of clinical standards. In Sec. 2.4. we discuss the challenges arising from both imaging heterogeneity as well as methods used to quantify heterogeneity and their translation to clinical practice. Finally Sec. 2.5 concludes the chapter.

2.1 Cancer and heterogeneity

Cancer is one of the leading causes of death worldwide, with approximately 18.1 million new cases and 9.5 million deaths in 2018 over 20 world regions [20]. There is a wide variety of cancer types; in the US in 2015 the most common ones included lung, breast (in females), prostate (in males) and colorectal cancer. Such diversity sets a foundation to eight biological hallmark capabilities characteristic to cancer, including:

sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, deregulating cellular energetics, avoiding immune destruction, as well as two characteristics such as genome instability and malnutrition, and tumour-promoting inflammation [2, 21]. There is an increasing amount of evidence that tumours are biologically complex [22, 23, 24], both on the genetic and microenvironmental levels. The manifestation of such complexity can be spatially correlated with some hallmarks of cancer, such as angiogenesis and related processes such as hypoxia [25, 26, 27].

2.1.1 Angiogenesis and perfusion heterogeneity

Cancer is an umbrella term for over 200 different types of diseases that share some common characteristics. All types of cancer are caused by multiple mutations in the DNA replication mechanism, resulting in a rapid, uncontrollable growth and spread of abnormal populations of cells [22]. Fast proliferation rate characteristic for tumours results in the depletion of oxygen to some regions of the tumour, the condition commonly referred to as hypoxia [28]. This is caused by an increasing distance of the cancerous cells from capillaries as the tumour grows. Hypoxia has also been previously identified to trigger the process of angiogenesis [29]. Angiogenesis means "vascular creation" (from Greek: *angeion* - vessel, *genesis* - creation) and refers to development of new vasculature in order to meet the increasing oxygen demands. However, the angiogenesis-induced neovasculature in tumours created in the process is abnormally structured and leaky, resulting in a high vascular heterogeneity of the tumour. This in turn leads to a high heterogeneity in perfusion [30], which is defined as the flow of fluid (e.g. blood flow) carried out via circulatory or lymphatic systems (e.g. vasculature) towards an organ, tissue or extravascular extracellular space (EES) in tumours. It is further known that such heterogeneity increases as the tumour grows [2, 21, 31]. Recognizing such change in perfusion heterogeneity over time could prove

informative in understanding tumour progression and, as a consequence, improve assessment of therapy response, therapy guidance or even prediction - leading to the development of personalised therapies.

2.1.2 Terminology and scope

Spatial heterogeneity in tumours is a large topic of interest in cancer biology, encompassing structural, metabolic, functional, genetic and micro-/macro-scale heterogeneity. Its definition can further vary between the studies and even imaging modalities used to assess it. In order to avoid ambiguity, here we outline the terminology used in this thesis and the scope that is of interest to this work.

The formulation of "spatial heterogeneity" can vary in terms of its scope. In [31], *intertumour* heterogeneity is defined as differences in structure and function between different tumours of a single patient. In contrast, *intratumour* heterogeneity captures the subregional variation within a single tumour. Finally, *interpatient* heterogeneity relates to tumour variation between different patients. In this work we are focusing principally on intratumoural heterogeneity and in part on interpatient heterogeneity.

In addition to differences in its scope, tumour spatial heterogeneity can also be defined differently between various types of studies. We differentiate between genetic, micro- and macro-scale heterogeneity. In this work we refer to spatial variation in genetic profile of the cell as *genetic* heterogeneity. Altered tumour microenvironment is a consequence of such genetic variation and such regional variation in tissue pathology is termed here as *micro-scale* heterogeneity. Imaging is capable of visualizing such apparent heterogeneity arising from genetic and tissue variation by representing an image as a discrete array of voxels at a lower spatial resolution, which is referred to as *macro-scale* heterogeneity. In contrast to genetic and tissue heterogeneity, imaging is performed at a different spatial scale, which has important implications, but also challenges (discussed later in Section 2.4 of this chapter) [1]. In this work we focus

solely on the macro-scale heterogeneity, but make some effort to also relate it back to micro-scale and underlying biology. For the sake of completion, the reader is referred to [32] for a review of heterogeneity assessment from genomic data, and to [33] for the assessment from digital pathology.

Finally, within the macro-scale, tumours can appear heterogeneous both in structure and function. For the reasons outlined in section 2.1.1 of this chapter, in this work we are predominantly interested in the spatial heterogeneity arising from underlying tumour perfusion.

There is certainly an advantage emerging through combining the information over different scales and scopes to enhance models describing tumour growth and response/resistance to therapy. However while genetic profiling is rarely performed, imaging has become a clinical standard and is performed repeatedly in diagnosis and response assessment [31, 34] due to its non-invasive nature as well as wide availability in hospitals. Furthermore, imaging is capable of capturing the information across the whole tumour, which is often a limitation in genetic and micro-scale sampling [31]. For those reasons, it is beneficial and of interest to develop methods capable of quantifying apparent heterogeneity directly from macro-scale imaging data, as well as further validating them against micro-scale and genetic studies.

2.1.3 Heterogeneity in a clinical setting

For many years morphological features such as tumour size have been (and still are) used and standardised as a primary guideline for tumour staging [35] and therapy response assessment [10, 36, 37]. However, the relationship between the tumour size and its stage or therapy response is in fact unreliable, and often irrelevant (as in case of anti-angiogenic therapies [38, 39]), rendering the metric insufficient on its own [40, 41, 42].

There has been a considerable amount of effort towards recognition of hetero-

geneity in various applications to cancer. As a consequence of the process of angiogenesis, the degree of heterogeneity is known to increase as the tumours progress [2, 3, 4, 21, 31]. Heterogeneity could further increase as well as decrease following therapy [8, 7], and could indicate a poor clinical prognosis in general [6]. Furthermore, it has been associated with resistance to therapy [43], for example hypoxic sub-populations of cells could prove radio-resistant [44], while irregular and leaky vasculature may lead to an increased difficulty during drug delivery [45]. The ability to recognise such change in both the overall degree of heterogeneity and its subregional distribution could therefore prove instrumental in the improved understanding of tumour growth and the subsequent assessment of response/resistance to therapy. Consequently, this could lead to better prediction of response to therapy and encourage development of personalised therapies. Finally, with the advent of anti-angiogenic drugs affecting tumour blood supply, some traditional (e.g. morphological) biomarkers are no longer entirely adequate. With the need for novel biomarkers, spatial intratumoural heterogeneity has become a topic of interest as a potential prognostic indicator [31].

2.2 Cancer imaging modalities

Imaging revolutionised modern medicine, in particular when it comes to cancer [46]. It allows for a non-invasive insight into a patient's body and has been widely used in cancer detection, diagnosis and staging [35], but also in therapy guidance and monitoring [47]. Some imaging modalities have the ability not only to visualise the apparent intratumoural variation, whether in structure or function of the tumour, but also to translate those observations into quantitative measurements. There are many types of imaging modalities, but in general they can be grouped into three distinct categories with respect to the nature of physical phenomenon and type of

electromagnetic spectrum involved in the formation of an image: (1) x-ray radiation modalities (Subsection 2.2.1), (2) gamma radiation modalities (Subsection 2.2.2), and (3) radio-magnetic modalities (Subsection 2.2.3).

2.2.1 X-ray and Computed Tomography

Medical radiographic images are produced by taking advantage of the relative inability of X-rays to pass through certain materials. An image can be therefore constructed by capturing differential X-ray attenuation through different substances. Different energy X-rays can be achieved through different tube anode materials, where tungsten is the most commonly used. Since the body is composed of tissues of varying density, X-rays are widely used in medical imaging, including plain radiography, fluoroscopy, mammography and Computer Tomography (CT). Notably, CT is a 3D extension of standard X-ray technology with a rotating tube. Attenuation measurements are then computationally combined over different angles. Such images can be regarded as attenuation maps, where bones with a higher absorption coefficient are represented as dense objects, while soft tissues with much lower absorption coefficients are represented as objects that are less visibly different from the surrounding air. X-rays are therefore used to measure the structural properties of the imaged tissue. The absorption property is often reflected in CT through Hounsfield Units (HU), which are calibrated accordingly to absorption coefficient of water having 0 HU as compared to air with -1000 HU. Plain radiography and CT are the most common and widely available imaging modality, partly due to the relatively low cost. For that reason CT is often performed in cancer diagnosis and staging. However, it delivers a low dose of ionizing radiation, which, over longer exposure, can be harmful to patients.

2.2.2 Nuclear Imaging

Similarly to CT, nuclear medicine involves radiation. However, unlike in CT where external radiation is applied, in nuclear imaging a radioactive tracer isotope is delivered to the tissues or organs inside the body. In Positron Emission Tomography (PET), photons originating from positron annihilation in the decaying tracer are captured with the means of external detectors such as gamma cameras. ^{18}F-FDG PET is the most commonly used example of such a process, where a positron-emitting radionuclide fluorine-18 (^{18}F) is delivered by the means of radioactive tracer based on fludeoxyglucose (FDG), an analog of glucose. Due to its ability to track glucose metabolic activity, it is very popular in cancer detection, tracking and evaluating metastatic spread. In addition to its extensive use in cancer, the true value of PET imaging can be seen in its chemical specificity. Different tracers can be designed in order to target different biologically active molecules, for example, amyloid or tau protein used to target dementia [48]. Furthermore, due to ^{18}F-FDG PET not being specific oncological tracer [49], some types of cancer cannot be monitored with it, which led to development of other specific tracers such as ^{11}C-Choline to target cell membrane metabolism used in prostate cancer [50] or ^{18}F-FMISO to target tumour hypoxia and oxygen metabolism in solid tumours [51]. The main limitation of PET radiotracers is their relatively short half-life (typically on the order of minutes), which is introduces a challenge of on-site production. ^{18}F half-life ($\sim$110 mins) is long enough to allow for manageable distribution and administration, which contributes to its popularity [52]. In addition, it is relatively expensive, which limits its widespread availability in the clinics.

2.2.3 Magnetic Resonance Imaging

Hydrogen atoms are abundant in all organic matter and are present in most human tissues in the form of water or fat molecules. Magnetic Resonance Imaging (MRI)

introduces a strong external magnetic field (typically between 1.5 and 7 Tesla) effectively polarizing the nuclei of hydrogen atoms (isolated protons) and aligning their magnetic moments, or spins. Subsequently, radio-frequency (RF) pulses are emitted, exciting the nuclear spin of the atoms, and thus changing their alignment by a certain flip angle (FA). Once the RF pulse is turned off, they relax back to alignment with respect to the magnetic field. The relaxation property is the basis for the contrast during the image formation. As the protons return to equilibrium, the spatial location of the resultant radio-frequency signal can be decoded by applying spatially-varying magnetic fields using gradient coils. Manipulating repetition time (TR) and echo time (TE) results in different image contrasts (e.g., T1- and T2- weighted images), based on the different structural tissue properties. However, other imaging protocols can be employed in order to also infer the functional properties. For instance, blood-oxygen-level dependent (BOLD) contrast in functional MRI (fMRI) exploits the differences in magnetic properties between oxygenated and deoxygenated blood. It can be used to measure aspects of the hemodynamic response in the brain, which is correlated with neuronal activity. Alternatively, Arterial Spin Labeling (ASL) captures the change in the magnetic equilibrium following the inflow of magnetically labeled water molecules from arterial blood. The labeling is achieved by inverting magnetic spin of water molecules in the neck. ASL is used to measure aspects of perfusion. MRI, although more expensive than CT, is still widely available in hospitals. Furthermore, it lacks the ionizing radiation present in both x-ray imaging and nuclear medicine.

2.2.4 Imaging heterogeneity

While MRI- and CT-based images, in their simplest form, reflect the structural properties of the imaged tissue, nuclear imaging is capable of capturing its functional and metabolic profile. For that reason, it is frequently used in imaging heterogeneity. ^{18}F-FDG PET CT, using 2-deoxy-2fluorine-18fluoro-D-glucose as a tracer, enables

radiolabeling of glucose uptake and is therefore widely used in diagnosis and staging of some types of cancer via their heterogeneous metabolic activity [47]. While both, CT and MRI (with some exceptions) are typically used to image structure, function can also be inferred, often by means of a contrast agent. Dynamic imaging modalities enable monitoring of the tumour microenvironment. Perfusion CT (pCT) and Dynamic Contrast-Enhanced MRI share two steps in common: (1) injection of a bolus or an exogenous contrast agent (CA) traveling from intravenous line (IV) to EES of the tumour (Fig. 2.1 (b)), and (2) successive acquisition of volumetric images over time, which typically results in a 4D timeseries image. The CA fluid dynamics can be further analyzed to reflect perfusion in a qualitative or quantitative manner.

As described in the previous section, tumour vasculature is disorganised and abnormally structured as a consequence of angiogenesis. For that reason, contrast-based imaging modalities such as DCE-MRI are particularly well suited to the problem of heterogeneity assessment from perfusion. Due to the leaky vasculature, accumulation of CA (in case of DCE-MRI Gadolinium) in the EES results in an increased contrast (in case of DCE-MRI increased T1-weighted signal). DCE-MRI has already shown a great potential in evaluating tumour progression and response to therapy due to its ability to monitor local changes in perfusion [13, 14], even if simply by the means of image comparison before and after injection (Fig. 2.1 (a) and (b)). Furthermore, the time course at each voxel throughout the contrast take-up provides additional dynamic information that might be clinically important and can be quantified [53]. This is further discussed in the following section. Finally, in addition to being widely available in hospitals, its non-ionizing nature poses no known side effects to the human body unlike pCT. For the outlined reasons, DCE-MRI is of major interest to this thesis and all the analyses in the following chapters are performed with use of DCE-MRI data.

2.3 Heterogeneity biomarkers

In the previous section we outlined modalities and techniques available for imaging cancer, with a focus on DCE-MRI. In this section we cover strategies that are used to assess spatial intratumour heterogeneity from such images. First, we outline techniques to quantify perfusion from DCE-MRI. We then briefly discuss current clinical standards of assessing heterogeneity (in DCE-MRI and in general), which are largely limited to qualitative assessment. Finally, we follow with a review of recent attempts to establish more quantitative heterogeneity-based biomarkers, where we differentiate between spatial and non-spatial methods.

2.3.1 Quantifying perfusion

Before we can begin to consider any quantitative or qualitative measures in terms of perfusion heterogeneity, first we need the means to quantify the perfusion itself. As previously discussed in Section 2.1 of this chapter, angiogenesis is known to generate a large, chaotic network of leaky vasculature, consequently resulting in increasingly heterogeneous perfusion as the tumour grows. Such change can be subsequently captured with the means of dynamic imaging, resulting in a 4D image which is then reduced into a number of quantitative parameter maps to allow for meaningful analysis. Although a few dynamic imaging modalities share that step in common, a full review is considered outside of the scope of this thesis and we shall limit this review to parameter extraction methods solely used in the context of DCE-MRI.

As further discussed in Subsection 2.2.4, the nature of DCE-MRI enables it to be used to monitor changes in local perfusion. However, unlike in CT-based modalities where intensity is based on Hounsfield units, in MRI the relationship between tissue property and observed signal is non-linear and without a consistent, fixed scaling. A number of model-based [54, 55] and model free [56, 57] methods exist to quantify the

properties of perfusion in DCE-MRI.

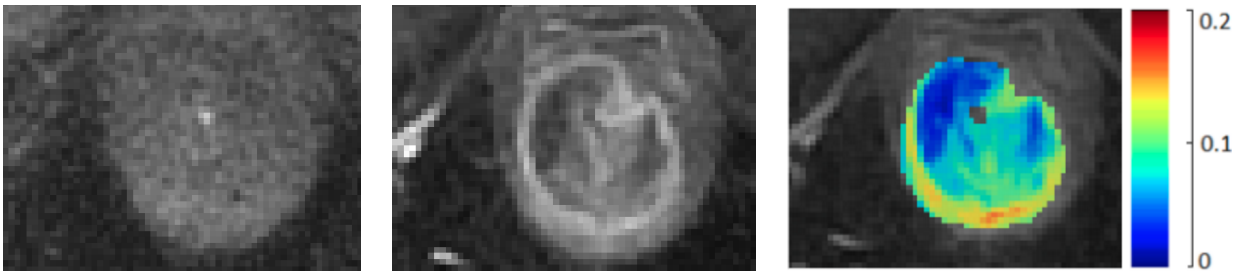
Model-based methods incorporate certain assumptions about the dynamics of the perfusion drawn from biology. The most common model-based approach is Pharmacokinetic (PK) modeling, which is centered around compartment representations of the underlying tissue, and the design of such is inspired by the biology and physics of perfusion. In PK modeling a compartment model is fit into the data, representing a function of contrast uptake over time for each voxel. This results in a number of parameter maps quantitatively describing the observed time intensity curves, hence these are also otherwise known as quantitative parameter maps (Fig. 2.1). Although many models are available [58], the two-compartment Toft's model (TM) [55] is the most popular model used in PK modelling and has become a standard in the analysis of DCE-MRI data. The model is defined as:

$$C_e(t) = K^{trans} C_p(t) \otimes e^{-k_{ep}t} \tag{2.1}$$

where C_p represents the Arterial Input Function (AIF) [59] and k_{ep} and K^{trans} are derived parameters referring to rate constant and transfer constant respectively [55]. C_e is the contrast uptake that is related to the signal enhancement (CA uptake in each voxel) by a function of the pre-contrast relaxation time (T_{10}), which is estimated from a separate variable flip angle set of images. A number of methods exist to evaluate the AIF, including population average models [60] or estimation from the dynamic images [61]. However, despite its popularity, mathematical analysis of TM (and the Extended TM) indicates that it is inaccurate in the case of highly vascular tissues with a very fast or slow tracer exchange [62] - which is a major concern when analysing heterogeneity. Furthermore, it has been shown [63] that quantitative parameters (such as K^{trans} or k_{ep}) in general have a greater variability than semi-quantitative parameters when it comes to reproducing results. Additionally, every assumption,

whether taken in form of AIF or PK model, has the potential of being incorrect and thus biassing the results. Finally, there is a general disagreement in terms of the PK model choice, which varies between the studies [64, 65]. Data-driven strategies could be employed to address these concerns.

Model-free methods, otherwise also known as semi-quantitative methods, derive a number of simple and heuristic parameters directly from the signal enhancement curve. Simple parameters include area under the curve (AUC) or gradient. Principal Component Analysis (PCA) [56] is one common example of a method to extract more complex heuristic parameters. PCA is an orthogonal linear transformation of the data in terms of its principal directions, as defined by the maximum covariance. In other words, it can find a lower dimensional subspace within the data that preserves such covariance [66]. In its simplest form, we formally define the covariance matrix of a mean-centred X data matrix as $X^T X$, which describes the covariance pairs and diagonal variance for each of the elements. We further perform eigendecomposition of such a covariance matrix into PDP^{-1}, where P is the matrix of eigenvectors corresponding to directions of the data, while D is a diagonal matrix describing the magnitude of each eigenvector. $X' = XP'$ is the new mean-centred data matrix as defined by the P' matrix of weighted eigenvectors with the largest eigenvalues. The method assumes that the principal directions in the data are well described by the covariance. Although sensitive to scaling, unlike PK modeling, PCA has an advantage of making no assumptions about the underlying biology. Nonetheless, correlations exist between model-based and model-free simple heuristic parameters [67].



(a) (b) (c)

Figure 2.1: An axial slice of a DCE-MRI 4D volume (a) before enhancement, (b) during enhancement, (c) overlaid with K^{trans} perfusion parameter map.

The process of parameter extraction is illustrated in Fig. 2.1. Such extracted parameter maps then commonly constitute a basis for any further qualitative or quantitative analysis. The following subsections will describe a number of qualitative and quantitative methods that have been used in assessing tumour heterogeneity.

2.3.2 Clinical standards

Since the 1950s imaging has made its way to clinical practice and is considered a standard. It has been widely used in cancer diagnosis, staging and repeatedly during therapy response assessment. As mentioned earlier, in Subsection 2.1.3, so far morphological assessment has been predominantly used to evaluate cancer stage or therapy response from medical images. Two types of clinical standards were developed on such basis, namely the bi-dimensional World Health Organization (WHO) criteria [36] and the one-dimensional Response Evaluation Criterion In Solid Tumours (RECIST) [37, 10]. WHO measures two maximal perpendicular directions of the tumour and assigns it to one of the four response categories based on percentage change. Following the increasing complexity of arising imaging modalities, RECIST revised the original WHO criteria, replacing it with one dimension while still using the same categories, although with different threshold percentage values [68]. However, while recent evidence suggests that subtle changes in size can be informative even in the

case of anti-angiogenic therapy [69], many researchers regarded such traditional measurements as insufficient on their own [38, 39, 40, 41, 42]. Instead (or as well as), measuring heterogeneity as a prognostic factor has become a topic of interest, especially with a recent development of anti-angiogenic drugs. Nonetheless, it is important to highlight some notable exceptions in the matter, mainly, PERCIST [70] response criteria assessment developed as a consequence of PET popularity in cancer detection and tracking, which allows for such assessment in the absence of anatomic change.

While a quantified measure of heterogeneity has not been of major consideration in clinical standards so far, it is important to acknowledge that there are exceptions where it has been recognised as a qualitative aid. Tracer kinetic analysis over extracted parameter maps has been adopted in clinical practice in terms of visual assessment [14]. Tumours consist of highly vascular as well as necrotic or hypoxic areas often accompanied with edema, which are reflected in their structural and, mainly, functional and metabolic appearance. Many imaging protocols have been developed to visualise such changes, such as ^{18}F FDG PET CT or DCE-MRI, which have been further used to drive and guide therapy (for example, non-uniform radiation dose delivery over tumour area termed as dose painting). A common example of such qualitative interpretation of heterogeneity is identifying a tumour 'hotspot', related to the area with the most increased abnormality, for example SUV_{max} (maximum standardised uptake value) indicatory of glucose metabolism in ^{18}F-FDG PET-CT [31]. Another use-case of 'hotspot' analysis is recognition of a tumour in terms of a so-called 'onion' model with enhancing rim and necrotic core [1, 71]. However, it is important to note that at the basis of such qualitative descriptions is the tumour subregional distribution, based on its functional or metabolic appearance. Furthermore, spatial structural heterogeneity has been long accepted as an indicative component in the BI-RADS classification scheme established to evaluate potential malignancy in breast lesions [72]. It is even further understood as indicative in risk prediction of

cancer development based on the spatial variation present in the structure of fibroglandular and fatty breast tissue. Consideration of heterogeneity in these contexts has been mainly qualitative (with a notable exception of breast risk assessment which received considerable amount of attention in recent years [73]).

It is important to appreciate that such a visual assessment has a number of crucial limitations. The human eye is, by far, not a perfect mechanism. When assessing an image we are sensitive to a number of factors, including lighting conditions, contrast or colour scale, to list only a few [12]. Furthermore, visual assessment is prone to inter- and intra-observer biases. Finally, it is dependent on many external variables not only between different imaging modalities, but also acquisition-related parameters, even within the same modality or scanner. Quantitative analysis could provide clinicians and doctors with additional robust information to guide their decisions and provide them with the kind of image information that might prove invisible or elusive to a human eye.

Recently, there has been much interest in developing heterogeneity-based clinical imaging biomarkers from such qualitative cues. Biomarkers are defined in general as a "characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions" [5] that carries a diagnostic or prognostic value. Specifically an imaging biomarker is such a measure that is derived from medical images [31]. Although their validation and translation into clinical practice still remains a challenge due to the "translational gap" [5], non-clinical biomarkers are still of great interest to therapy development [74, 75].

Even though heterogeneity has been such a major qualitative aid in the assessment of tumours from DCE-MRI, studies have shown that for a number of reasons none of the emerging quantitative DCE-MRI biomarkers have been clinically translated [76], although even simple histogram-based statistics show promise. However, all of

the studies in consideration used model-based parameter maps. Furthermore, the majority of the studies, whether applied to the whole image or subregional analysis, were defined on a voxel-wise basis and no sophisticated feature extraction techniques were applied. There might yet be a potential in establishing clinically relevant DCE-MRI heterogeneity-based biomarkers from spatial features derived over model-free parameter maps, which is further discussed later in this section along with a summary of available metrics and biomarkers.

In summary, spatial intratumour heterogeneity is an undeniably important element in understanding tumour progression and evaluation of therapy response. However, although qualitatively acknowledged in a few specific applications, it has been largely overlooked in clinical studies and standards. It is, therefore, of interest to develop measures that can a) explicitly account for heterogeneity and variation in lesions, b) conduct quantitative spatial evaluation of heterogeneity and c) further apply such measures in clinical decisions. As previously mentioned (in Subsection 2.1.3 of this chapter), there is an interest to establish imaging biomarkers in terms of overall global degree of heterogeneity as well as local subregions of tumour, both of which are of interest in this work. In the following sections we outline recent efforts in the design of such methods and indicate their application in both of these contexts.

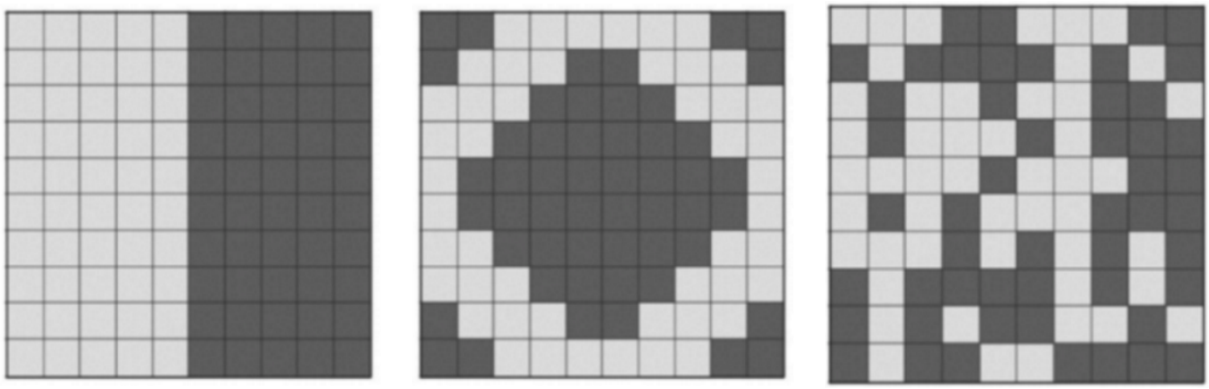
2.3.3 Non-spatial histogram-based methods

An image can be represented by a frequency histogram of individual voxels, and thus in terms of its intensity (or signal) distribution. A number of descriptors can be extracted from such a histogram, including simple metrics such as mean or median. However, simple heterogeneity-related metrics could also be extracted, namely standard deviation, variance, skewness, kurtosis or interquartile range. Such measures could then be correlated with clinically valuable information, such as response to therapy. In DCE-MRI, such distributions are normally defined over parameter maps,

such as K^{trans}, and have been identified as prognostic factors [77]. It is important to note, however, that such methods discard the valuable spatial information contained in the images. For that reason, they can only be used in the context of overall degree of heterogeneity.

In [11], non-spatial methods were found to be most commonly used in tumour diagnosis, response monitoring as well as outcome prediction (although closely followed by spatial co-occurrence matrices covered in Subsection 2.3.7). Similarly, in [1] the authors found that non-spatial, specifically, histogram analysis methods accounted for approximately half of the reviewed studies. There are many software solutions available for histogram analysis, which contributed to the popularity of the method.

In [76] the authors suggest an increased benefit of intratumour heterogeneity assessment from such non-spatial methods, especially when compared to average values. However, it is important to highlight that when applied in the context of quantifying heterogeneity such methods share a fundamental flaw, as entirely different heterogeneous appearances could result in similar or identical intensity distribution [16]. This can be easily illustrated when comparing a highly heterogeneous tumour (Fig. 2.2 (c)) to a locally homogeneous tumour composed of necrotic core and enhancing rim (Fig. 2.2 (b)). Furthermore, some of the statistics calculated from histograms (e.g. kurtosis) could be difficult to relate to biology [31]. For those reasons, the problem of quantifying tumour heterogeneity should be considered in terms of spatial analysis.



(a) (b) (c)

Figure 2.2: Images of different textures (a) homogeneous, (b) locally homogeneous and (c) heterogeneous with identical grey-level distributions and bimodal histograms (image acquired from L. Alic [16]).

2.3.4 Model-based prior

Tumours could further be viewed in the light of prior knowledge. Drawing from observations in tumour biology, some studies quantify heterogeneity by representing tumours with a prior model reflective of their biological structure or function. Notably, two types of models have been considered: 1) composed of concentric circular subregions [71], and 2) based on the "onion" tumour model with enhancing rim and necrotic core [78]. Such methods were mainly used in terms of subregional analysis, however the frequency distribution of such regions could also be used in the context of assessing the overall degree of heterogeneity.

Tumours, however, can take many complex shapes, rendering it difficult to fit such models to real clinical data (e.g. colorectal). Additionally, not all tumour models will display such concentric characteristics, deeming the method suitable only in a subset of studies. Furthermore, the geographical arrangement of such subregions is a subject to change as the tumour grows. Such methods are therefore limited to modelling rather simple heterogeneity. They are also likely to be reflective of a single state in the tumour growth or therapy monitoring. Finally, introducing a prior might not be

biologically reflective of the tumour microenvironment and could be misleading. The majority of these issues could be partly mitigated by data-driven methods.

2.3.5 Binary features and thresholding

One of the simplest ways to consider the problem in a data-driven manner is in terms of thresholding. This can be either done automatically, or manually. The problem in general is defined as:

$$f(x) = \begin{cases} 1 & \text{if } x \geq T, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

where T is an arbitrarily or automatically selected threshold value. The simplest manual approach is to view a tumour in terms of presence or lack of signal with an arbitrary threshold signal value. There is also a wide range of automatic thresholding methods that could be used for the task. In addition, thresholding is often used as a basis for further analysis, as in [14], where the authors have used thresholding to define enhancing voxels, that were then, in turn, applied as a features for fractal analysis (discussed later in this section) to characterise tumour enhancing fraction. Thresholding methods are usually considered in the context of subregional assessment.

Automatic thresholding methods are typically very sensitive to the changes in absolute of intensity values, which is an important consideration in terms of repeatability of studies, particularly with varying image acquisition parameters. Furthermore, automatic thresholding methods often assume a bimodal distribution of the data, which might not reflect reality, particularly in noisy imaging observations. The choice of a threshold point is also a challenge. Whether arbitrarily or automatically chosen, the points might not have a meaningful relation to the underlying biology.

2.3.6 Clustering methods

In machine learning, clustering is an umbrella term for many data-driven techniques that generally indicate grouping of similar data points together. The similarity can mean something different in different contexts, however is normally expressed over some descriptive features associated with data points called feature vectors. In this case, data points are commonly understood as individual voxels in a 3D volume, and described with multi-parametric feature vectors consisting of values from a number of extracted parameter maps. Clustering methods are mainly applied in the context of subregional analysis in tumours, however overall degree of heterogeneity could also be quantified with frequency distributions of voxel subregional labeling.

Many kinds of clustering methods exist, but can be differentiated between based on the notion of a "cluster". Clusters could be defined, for example, in terms of distances between the feature vectors or distribution densities of the clusters - both of which are the at the heart of most popular clustering algorithms. Although many other types of clustering methods exist, in this review we limit ourselves to the two most popular types, covered here in greater detail, as they are extensively used in the latter chapters of this thesis. In this work, however, they shall be used in a different context - constituting the basis for image feature encodings widely applied in Chapters 3 and 4. Both of those algorithms can be described by the means of the Expectation Maximization (EM) algorithm [79], that although generally used to model complex distributions with the means of simpler components, can also be applied to the problem of clustering.

K-means

K-means (via Lloyds algorithm) [80] is an example of a clustering that belongs to the first type of clustering group. It is defined on the basis of distances between the

data points and centroids, where a centroid is a centre point or a mean of a cluster (that does not have to be a member of the dataset). It is an optimization problem defined as minimizing the cost function - a sum of within-cluster distances (typically Euclidean sum of squared distances (SSD), thus also understood in terms of a cluster variance):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2.3}$$

where μ_i is the mean of points in cluster S_i, S is a set of data points (e.g. voxels in a 3D image), $S = \{S_1, S_2, ..., S_k\}$ and k is a pre-specified number of clusters. The algorithm iteratively refines the definition of the cluster sets, calculating a new mean for each updated cluster set S_i:

$$\mu_i^{new} := \frac{1}{|S_i|} \sum_{x \in S_i} x \tag{2.4}$$

The definition of set memberships are refined by members' proximity (defined by distance) to newly chosen means. Subsequently, the means are updated until the algorithm converges to a minimum (as often defined by SSD), although a global minimum is not guaranteed.

As a limitation, K-means represents the data through hard classifications and does not allow for cluster overlap. Furthermore, as a consequence of its dependence on the means and Euclidean distances, the algorithm has a tendency to create clusters of similar size (in terms of cluster volume in the feature space) as the nearest centroid is always preferred. Another consequence of Euclidean distance dependence is that K-means is sensitive to outliers. In K-means the cluster centre is represented by a mean which is easily influenced by extreme values and can lead to incorrect representation of the cluster centre as a consequence. This is particularly pronounced in medical imaging where (if defined on a voxel-wise basis) per-voxel signal measurements can be noisy or affected by motion. There are a number of ways to deal with outliers, for

example the problem can be reconsidered with (1) alternative distance measures, or (2) different clustering approach, for example - distribution-based.

Gaussian Mixture Models

The Gaussian Mixture Model (GMM) approach is a popular example of distribution-based clustering. The method attempts to fit a fixed number (mixture) of Gaussian distributions:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2.5}$$

where $p(x)$ is the likelihood of x represented by a linear combination of k Gaussians, $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ is a multivariate Gaussian with a mean μ_k and covariance matrix Σ_k, and π_k is a mixing coefficient of that Gaussian constrained to sum-up to 1 across clusters. The subsequent maximization is done with the EM algorithm. In the first expectation step we compute a likelihood matrix:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n \mid \mu_i, \Sigma_i)} \tag{2.6}$$

where γ is a matrix of likelihood for each point n represented by a Gaussian k. Similarly to the K-means clustering, in the next maximization step we iteratively update the parameters μ_k, Σ_k and π_k as follows:

$$\pi_k^{new} := \frac{N_k}{N} \tag{2.7}$$

$$\mu_k^{new} := \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n \tag{2.8}$$

$$\Sigma_k^{new} := \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{2.9}$$

The advantage of GMMs is that they allow for soft probabilistic cluster assign-

ments and so clusters can overlap. However, a number of studies have shown that real world data is rarely well represented by a normal distribution [81], and therefore using GMM imposes a strong assumption on the data that might not reflect reality.

There are some issues associated with clustering methods in general. One of the main challenges is the choice of k (whether centroids or Gaussians). Altogether this is not a trivial problem, as it is dependent on the definition of a cluster in the first place. Iteratively exploring the k parameter space is computationally expensive. There are a number of automatic methods that can be used to estimate k, for instance the elbow method, however the 'elbow' point might be difficult to determine unambiguously. The number of centroids k could also be defined with some prior knowledge. In biomedical applications this is often the case, where k is set to a number of expected biological components in the clustered domain, for example hypoxic, necrotic or healthy tissue. Another limitation of clustering methods is overfitting. In general, the term refers to a problem in machine learning when the learned hypothesis represents the training data exactly, but will not generalise past training to new set of observations. This is often the case when the model is too complex, thus adapting to exactly represent some noise or error measurements, and could be a consequence of, for instance, excessively long training time or lack of dropout connections in deep learning. This is of particular issue in biomedical imaging, as real world data is often noisy. The problem of overfitting is an active research problem of its own, and it is considered out of the scope of this chapter. Clustering methods are also highly sensitive to initialization. For instance, a random initialization cannot guarantee unique solutions, which is of particular issue when it comes to reproducibility of the results. Finally, considering the clustering problem in depth, algorithms such as K-means or GMM can also suffer the curse of dimensionality. This phenomenon refers to the increased sparsity of data as its dimensionality increases, therefore the low-dimensional intuition about Gaussian distributions or Euclidean distances is not

directly applicable in high dimensions. For example, the Gaussian distribution in higher dimensions is thought to concentrate in the tail of the Gaussian bell, while hyperspheres concentrate a large number of points around the shell of the hypersphere, thus likely biasing the distance measure. This is not generally an issue in the context of a voxel-based subregional segmentation task, in which effective dimensionality is relatively low (equal to the number of parameter maps). However, it could become an obstacle if more sophisticated multi-scale features are used to describe the voxels, or raw dynamic data with high temporal resolution is considered.

The majority of these methods used in the context of subregional heterogeneity segmentations are usually defined on a voxel-wise basis [31]. As a consequence, the areas grouped together are understood as similar - which is an issue when attempting to analyze variation or heterogeneity that is of particular importance in tumours. This is further discussed in Section 2.5 of this chapter. One way to address this problem is to explicitly account for the present variation with the means of texture analysis.

2.3.7 Texture analysis

Texture is a property that is very intuitive to humans, however difficult to quantify or formalise mathematically. By definition, although it partly relates to geometric patterns in artificial images, it also refers to characterizing the local variation in natural surfaces. For that reason, it has been a large topic of interest in assessing intratumoural heterogeneity. Texture, after all, could be referred to as a visual reflection of the variation contained in the structural, functional or metabolic profile of the tumour as acquired via imaging. In contrast to any previous methods covered in this section, which were defined on a voxel-wise basis and where no sophisticated feature extraction techniques were applied, texture, by definition, relates to locally spatial appearance (although clustering could also be defined spatially as covered in later

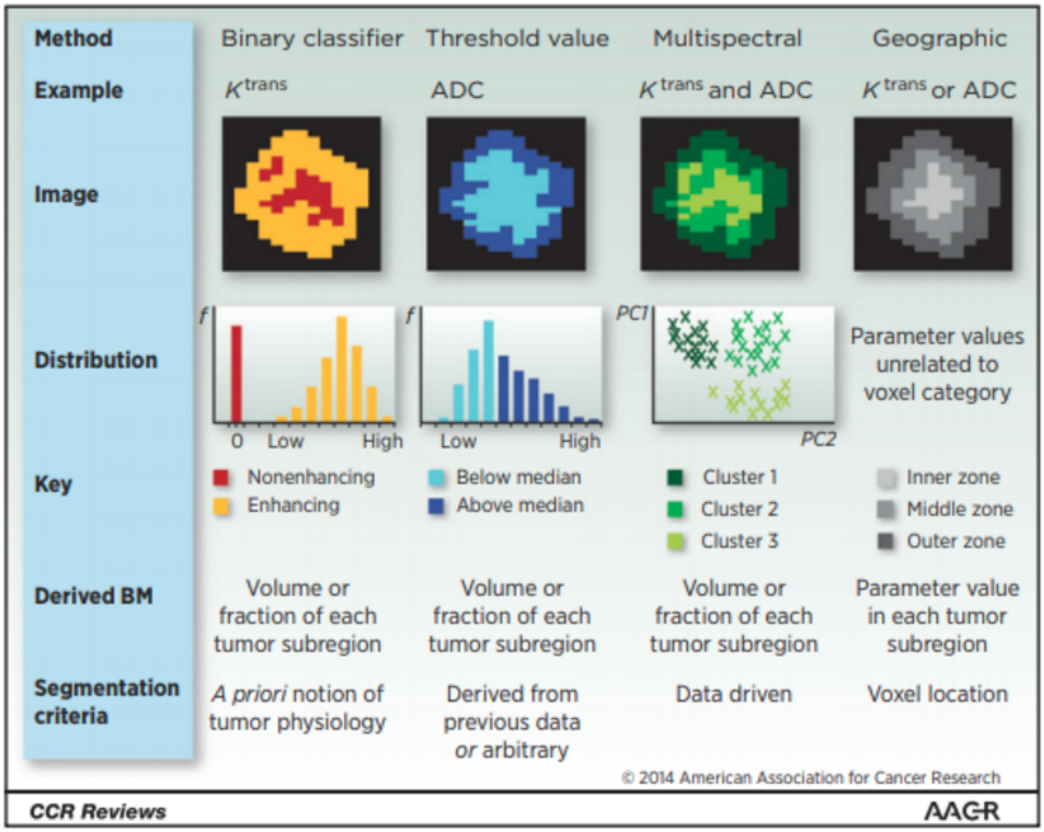


Figure 2.3: Summary of various simple methods to quantifying spatial intratumoural heterogeneity (image acquired from [1]). Such methods can be used either in isolation as described in Subsections 2.3.1-2.3.6, or as basic building blocks for further, more complex methods, e.g. clustering in Bag of Visual Words encodings (discussed in Chapter 3), features used for radiomics (Subsection 2.3.8) or binary features to estimate texture fractal dimension (Subsection 2.3.7) [14]. The last row corresponds to the parameter that is usually the basis for segmentation (BM - biomarkers, ADC - Apparent Diffusion Coefficient).

chapters). Texture is a major topic of interest of this thesis, therefore it is covered in greater detail in Chapter 3, and this review is limited only to its popular applications in assessing tumour heterogeneity. In general, texture can be used both in the context of subregional analysis as well as overall degree of heterogeneity (mainly as a consequence of Bag of Visual Words pipelines discussed in the following chapter), however in the majority of research it has been used primarily to quantify the overall heterogeneity.

Both in [11] and [1] the authors observed that the majority of spatial methods used to quantify intratumour heterogeneity relate to texture analysis, and in particular, to co-occurrence matrices, otherwise known as the Haralick method [82]. In simple terms, the method is a second-order statistic describing frequency distribution of all possible co-occuring value pairs, where value is dependent on the domain of application and is most frequently defined as a gray-level intensity (otherwise known as Grey-Level Co-Occurrence Matrix GLCM). Formally, it is defined as:

$$C_{\Delta x,\Delta y}(i,j) = \sum_{x=1}^{n}\sum_{y=1}^{n} \begin{cases} 1 & \text{if } I(x,y) = i \text{ and } I(x+\Delta x, y+\Delta y) = j \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

where i, j are pixel values and and x, y are image positions in an image or a matrix I of size $n \times m$. Each value in the C matrix will therefore be the number of times the value of i and j co-occurred in an image, as defined by an offset $(\Delta x, \Delta y)$. A number of radial coordinates as well as multiple value ranges could be considered for a more complex representation. An image with p unique intensity values will therefore produce a matrix C of size $p \times p$. This poses some computational challenges as the size of an image will grow quadratically with an increase in a number of intensity levels. Additional angles of direction and multi-scale steps will further increase the computational complexity.

Fractal analysis is another texture-related method that is popular [11, 1] in the context of assessing tumour heterogeneity. The aim is to analyze the complexity of a pattern by recursively breaking it down to smaller components. In its simplest form, box counting dimension ($d0$) aims at counting the number of equally-sized grid elements at different scales that are occupied by a pattern or a feature. Feature choice depends on the domain of application, and in medical imaging is often defined as a binary feature (e.g. threshold, presence or lack of signal) over the box area. Other fractal measures could also be considered, such as the information dimension ($d1$) or correlation dimension ($d2$).

The decades of research in texture analysis, however, yielded a vast number of methods that could also be considered in this context. Notably, Local Binary Patterns (LBPs) [83], filtering [84] and patch-based [85] texture methods were considered major milestones in the field of texture analysis, and are covered in more detail in Chapter 3 of this thesis.

2.3.8 Radiomics

The majority of the methods outlined above are typically used in similar ways. There are a number of extracted features, some constituting a basis for another (e.g. tracer kinetic threshold used to measure $d0$ fractal dimension [14]). Depending on the number of produced features, either all are used to evaluate their prognostic value, or a feature selection strategy is employed, which can be automatic or manual. This process is on a small scale, similar to the more recent and increasingly popular approach of radiomics [34].

With an increasing amount of available imaging data there comes the potential to divert from a traditional view of medical images as a visual aid, towards considering them as minable databases [34]. The power of radiomics lies simply in the nature of data-mining from large quantities of extracted features. The types of extracted

features include all of the techniques we covered earlier in this section, starting from simple and easy to understand by human visual cues (e.g. shape or size), to increasingly complex second- and higher-order statistics. Of all the extracted features, manual feature selection is very non-intuitive and might prove difficult to a human. A large number of features are therefore extracted and subsequently automatic data-mining techniques are employed instead, to determine the features which are most descriptive of the data or the ones with the most predictive power. In addition to extensive feature extraction, selection and classification, radiomics also emphasises data normalization and attempts to unify the standards across the field in an effort to improve repeatability of studies.

Despite efforts to unify the process (including image normalization, feature normalization, region of interests etc.), radiomic studies still remain largely independent [31]. Furthermore, one of the major disadvantages associated with such techniques is the high dimensionality of features, many of which are redundant, as often demonstrated by the subsequent feature selection step. This can lead to curse of dimensionality and overfitting. In addition, this may render the feature extraction step computationally ineffective. Multi-scale feature extraction followed by feature selection (although at a smaller scale) has been used in the past in computer vision [83, 84]. However, it was shown that, as opposed to large quantities of extracted features, even simple raw patch information at multiple scales could be descriptive enough if applied in both the local and global context of the image [85]. This only became possible with the introduction of vector-quantization and encodings (discussed later in this work), rooted in the formal definition of a texton in 2001 [84].

Method family	Spatial?	Data-driven?	Advantages and disadvantages
Histogram	N	Y	**Advantages:** - data-driven nature - widely available software **Disadvantages:** - non-spatial nature can lead to misclassification of similar global distributions with vastly different appearances
Prior-based	Y	N	**Advantages:** - spatial nature - prior-knowledge could be an advantage for some models **Disadvantages:** - in general, prior assumptions might be misleading and non-reflective of underlying biology
Thresholding	Y	Y	**Advantages:** - spatial and data-driven **Disadvantages:** - sensitive to variation - often dependent on bimodal distribution of frequency histograms which might be unlikely to occur
Clustering	Y	Y	**Advantages:** - spatial and data-drivien - can be defined over more sophisticated features **Disadvantages:** - typically used to define similarity, thus could disregard spatial variation Sensitive to: - outliers - overfitting - curse of dimensionality - initialization
Texture	Y	Y	**Advantages:** - spatial and data-driveni - accounts for variation - can be defined on a local and global basis **Disadvantages:** - some methods are produce a large number of features posing a computational burden
Radiomics	Y	Y	**Advantages:** - spatial and data-driven - large number of features increases the likelihood of finding good predictors - automatic feature selection step **Disadvantages:** - large number of quantitative features poses a computational burden

Table 2.1: Summary of methods for tumour heterogeneity quantification outlined in section 2.3.

2.4 Challenges and limitations

In the previous sections we outlined both available imaging modalities, as well as methods frequently used to assess spatial heterogeneity from such images. In this section we will highlight some important limitations. There are a number of challenges that are associated with assessing heterogeneity. Generally, they could be divided into three groups: (1) imaging limitations - caused by the physical constraints of imaging modalities and protocols, (2) methodological challenges - related to the mathematical properties and limitations of the approaches described, and finally (3) clinical translation issues - arising from the traditional medical background and practicalities of translating biomarkers into clinical practice.

2.4.1 Limitations arising from imaging protocol

As previously mentioned, capturing heterogeneity at a macro-scale via medical imaging - while widely available and beneficial - implies a number of challenges. Firstly, the information displayed in a single voxel is merely an average of underlying structure or physiology at a smaller scale. At the same time, signals from neighbouring voxels may not be independent. This is particularly pronounced when imaging and interpreting variation, but could also be evident at boundaries of structural or functional elements. Such effects are referred to as partial volume averaging. Some of these issues could be partly mitigated by considering each voxel in association with its neighbourhood information. Adding larger scale context to local information could later improve its descriptive power and is of particular interest in this thesis.

Images acquired by means of the same scanner, protocol or even modality can differ widely in terms of their absolute intensity values. It is important to highlight that intensity normalization strategies should be employed prior to analysis. The intensity in a single voxel can be further compromised and prone to misinterpretation

due to patient or organ movement. This applies to single scan images, but even more so in dynamic imaging, where multiple images are acquired over time. This could be addressed by motion correction, a minor disadvantage of which is that it could introduce artificial deformations in the image that might not be reflective of the true biology.

Finally, resolution of the image is a particularly important factor when analyzing heterogeneity. In order to accurately reflect the state of underlying micro-environment of the tumour, the spatial resolution must be high. Furthermore, a sufficiently high temporal resolution is required in the case of dynamic imaging, in order to accurately image the behaviour of the tumour, for example, the dynamics of the contrast agent. However, while higher resolution is beneficial to analysis, it calls for a compromise between image quality and practicalities of scanning or patient comfort. Furthermore, acquiring images at higher spatial resolution implies longer scanning time, which in dynamic imaging may come at a cost of loss in temporal information. An inevitable compromise must always be achieved between spatial and temporal resolution.

2.4.2 Methodological challenges

When it comes to methodological challenges, the majority of outlined available strategies have one thing in common: they are defined on voxel-wise basis, which has some important limitations. Primarily, it disregards some useful locally spatial information contained in the images. Furthermore, voxel-wise segmentation or classification tasks are highly dependent on the accuracy of information contained in a single voxel, which can be particularly problematic if patient movement is involved. The majority of those concerns can be answered in part by considering voxel neighbourhood information to provide local context. Furthermore, in terms of subregional analysis, tracking subregions over time may prove challenging as lesions vary in shape and size over time. This can be partly addressed by applying image registration methods,

however this introduces a possibility of introducing deformations that are unreflective of the true biology. Another way to address this is to consider the change in subregional proportions over time instead of tracking individual subregions. This, however, discards the information about the spatial arrangement of subregions.

Another issue inherent to biomedical imaging is a lack of data. Biomedical datasets are usually acquired in-house during clinical or pre-clinical trials and cannot be shared without permission. This makes them more difficult to obtain compared to machine learning/computer vision benchmark datasets, which are widely available online. This is of particular importance when it comes to image analysis methods, as the final predictive or discriminative property is usually determined with the means of machine learning, that by nature requires large quantities of data. This issue is even more pronounced in deep learning, which requires thousands of images to allow for training. Both deep learning and traditional machine learning also demand a large amount of data that has been labeled. This is, again, especially problematic in biomedical imaging, particularly in the task of segmentation as it requires hours of laborious human annotations to obtain these gold-standard labeled datasets.

Finally, relating the findings back to underlying tumour biology, especially its microenvironment and associated genetic information, is one of the biggest challenges remaining. In [1] the authors find that only a few studies integrate the spatial heterogeneity information over different scales. The authors further highlight that such integration could prove either in agreement (different scales measuring the same biology), or disagreement (measuring different biology) - and therefore complementary to each other. While both options open interesting prospects for the future, the reasons so few studies combine such information is due to the challenges associated with it. Mainly, relating imaging data to histopathology is a very difficult task [86], largely due to the fact that pathological samples only reflect a small subset of the lesion in the corresponding image [31], while analysis of genetic data is both financially and

computationally expensive.

2.4.3 Clinical translation

In addition to imaging challenges, it is important to mention issues arising from social and historical contexts. Despite decades of research into the applications of machine learning in medicine, such methods are only starting to emerge in clinical practice. Many important challenges arise from which, including the fear of replacing the clinicians or a public distrust towards machine-based "judgment". Although currently society is undergoing a slow shift in opinion in favour of machine learning in medicine particularly pronounced in medical imaging, it is still of crucial importance to provide systematic reviews and extensive statistical testing of such methods in clinical applications.

Clinical translation of research biomarkers is a slow process, and nearly all promising biomarkers never make it to clinical practice [5]. This is primarily due to the fact that in order to be routinely used in clinical settings they need extensive biological and clinical validation [31], that, as mentioned in the previous subsection, is a particularly challenging task. One of many reasons for that is the lack of repeatability and reproducibility of studies. At the heart of the problem is also a small number of public benchmarks that could allow testing for the reproducibility of studies. Recently, however, there has been a considerable amount of effort into development of such datasets thanks to initiatives such as Kaggle or MICCAI challenges.

2.5 Conclusions

Investigation of intratumoural perfusion heterogeneity has a potential to provide more insight into tumour growth, and subsequently, therapy assessment. Imaging is capable of depicting such variations in tumours, which are thought to reflect the underlying

heterogeneity in the tumour microenvironment. Visual assessment of such variation is subjective and prone to error, therefore automatic methods should be employed. A summary of existing methods applicable in the context of spatial perfusion heterogeneity assessment is presented in Table 2.1.

The majority of such methods work on a voxel-wise basis. In these, tumours are seen as globally heterogeneous, composed of discrete subpopulations of cells grouped together into locally homogeneous subregions that display similar features or characteristics. Such a grouping is typically achieved by means of, for example, thresholding, prior-knowledge or clustering. Tumours, however, in addition to being globally heterogeneous, can also exhibit a local variation - a reflection of, for instance, vascular changes in the tumour microenvironment. Such local information could be homogeneous (Fig. 2.2 (b)) as well as heterogeneous (Fig. 2.2 (c)), however in either of these circumstances it should be accounted for during the analysis step. The majority of methods used for assessing heterogeneity are, however, only focused on the global aspect of heterogeneity via voxel-wise image analysis methods, with an exception of some texture-based approaches. Additionally, those existing methods that are defined in terms of more complex features (in contrast to voxel-wise basis), extract large quantities of features, many of which are redundant, thus imposing a significant computational burden.

Modern image analysis pipelines, inspired by decades of progress in computer vision, provide us with powerful tools to view images in both local and global contexts. Local information is captured via the design of robust neighbourhood-dependent features, often acquired over multiple scales. Such local information is then grouped together (vector-quantised) to reduce the image representation into only its most dominant patterns, which partly addresses the problem of high dimensionality emerging from large quantities of extracted features. The distribution of such local dominant patterns, referred to in computer vision as encoding, is then used as a global repre-

sentation of an image. In this context, we could well describe complex images that are locally heterogeneous but globally homogeneous, or vice versa. Adding such an extra dimension and depth to the image representation has the potential to better represent the true complexity underlying heterogeneous tumours and alleviate some of the issues associated with the existing methods, thus it constitutes the focal point of investigation in this thesis.

In the following chapter, feature-based methods capable of capturing local and global variation are explored, along with their application to tumour perfusion heterogeneity. Developing and validating methods for therapy response assessment is a challenging task, mainly due to the lack of available public benchmark data. However, understanding the tumour progression on its own, unaffected by therapy, is also important and can be used as a baseline for further therapy assessment. We therefore begin this work by posing a problem in a simpler context of characterizing only the change in heterogeneity during tumours' progression.

Chapter 3

Shallow image representations for segmentation and classification of tumour perfusion heterogeneity in DCE-MRI

In this chapter we introduce a comprehensive framework to the problem of assessing perfusion heterogeneity from DCE-MRI of a murine tumour model, based on the Bag of Visual Words (BoV) framework and spatial features. After extracting pharmacokinetic parameters, their heterogeneity is measured via frequency distribution models of texture primitives (textons, or visual words) over a set of extracted spatial variation-based features. We first show that the framework is capable of visualizing a growth-induced change through both local and global perfusion heterogeneity on a small longitudinal dataset. We further investigate its performance over a larger longitudinal dataset in order to provide support for statistical testing. We assess a number of popular texture-based feature extraction methods on the task of subregional tumour distribution (semantic segmentation) and overall tumour quantification

(longitudinal classification). We evaluate the methods against one another as well as against a popular voxel-based method used for assessing degree of heterogeneity in DCE-MRI. Furthermore we show indirectly that the heterogeneity information described by the proposed framework can be related to underlying tissue properties via an alternative publicly available benchmark dataset (due to the lack of ground-truth information). Finally we explore the the influence of different feature descriptors along with their assumptions (variation- and similarity-based) over the assessment outcome and assess its significance in the context of heterogeneous tumours.

While a large part of this chapter is devoted to the introduction of the framework in the specific context of perfusion heterogeneity, it should be understood that the presented methods are well-established in the field of computer vision and the novelty of this work arises from their specific application domain. The aim of the chapter is therefore to explore the potential and suitability of such methods as future biomarkers in order to bridge a gap in tumour heterogeneity analysis. However, parts of this chapter also serve as a quantitative review and a testbed for such methods "in the wild" on uncontrolled and unpredictable real-world data.

The key novel contributions of this chapter are as follows:

- Introducing a Bag of Visual Words framework to the problem of assessing tumour perfusion heterogeneity in DCE-MRI;
- Combining and adapting a number of existing literature solutions into a single pipeline suitable for analysing three dimensional data (Filter Banks, Patch Descriptors, Local Binary Patterns as well as K-Means and Gaussian Mixture Models as the basis for building visual vocabulary).

3.1 Shallow image representations in texture

In this section we introduce the background of shallow texture representations and place the previously described methods from Chapter 2 in a structured context. Sec. 3.2 describes in detail the feature extraction, codebook generation and encoding methods used in this work. We further outline experiments and their results in Sec. 3.3, where we address both the accuracy performance of methods and the question of their suitability to the problem. Finally Sec. 3.4 concludes the chapter with a short discussion.

3.1.1 Texture representations

Texture, by definition, is a "visual and especially tactile property of a surface" [87], or "the characteristic physical structure given to a material (...)" [87]. It represents a spatial organisation of visual primitives (or textons) and can be thus understood both, in terms of regularly repeated geometric artificial patterns, but also as a projection (whether in a human eye or a camera) of visual properties and stochastic patterns in natural surfaces, whether regular or irregular. Varma and Zisserman [88] defined texture as a property of which we only have an intuitive understanding, and which, unlike colour, lacks a formal mathematical definition, also incidentally highlighting the challenges and difficulties associated with quantifying it. Texture, being a fundamental characteristic of many types of images, can be particularly pronounced in medical images amongst many other domains. For that reason textural representations have been under extensive investigation in the field of computer vision since the late 1970s. With further influences from natural language processing [17], the research and development in texture representations was found to generalise well into other domains and so it gave rise to modern shallow image processing pipelines in computer vision in general, including numerous areas of application such as facial recognition, semantic

segmentation, object recognition and scene classification amongst many others.

In the early days of texture research, methods were limited to simple non-spatial first-order statistics of grey-level distributions. Image histograms were popular due to computational simplicity as well as rotational invariance [89]. A number of simple statistical features could be extracted from histograms, such as standard deviation, skewness or range, and used as an input to a classifier. This type of method was also found to be the most commonly adopted in assessing tumour perfusion heterogeneity [11, 1]. In contrast to first-order statistics that reflect on the image solely in the context of its global representation, second-order statistical features assess the relationship between pixels pairs [89]. One of the most popular type of such representation is the Grey Level Co-Occurrence Matrix (GLCM) described in the previous chapter. A number of statistical features are typically extracted from such a matrix, including entropy or energy [82] as two examples out of the 14 metrics originally described in the 1973 GLCM paper.

One of the main challenges associated with texture analysis is the high intra-class variance. There could be many reasons for it, including changes in viewpoint, scale, rotation, translation, illumination conditions, occlusions or shadows to mention only a few. This has motivated researchers in the 1990s to a search for universal features robust to such intra-class variance. Furthermore, higher-order statistical features assessing pixel relationships in the global context and beyond pixel pairs became a point of interest due to their low sensitivity to image noise [89]. It is generally agreed [89, 17] that the choice of texture features is an important design choice, as even the most advanced classifiers would fail given non-descriptive features. Robust feature design has therefore been a large topic of interest in the field of texture analysis and gave rise to methods such as Local Binary Patterns [83], Patch Descriptors [85, 88] and many filtering methods [84] (e.g. Gabor), all of which are used in this chapter and described later in Sec. 2.2.

The breakthrough came with the introduction of textons, leading to Bag of Textons [84, 85] and later Bag of Words [90, 91] encodings. The concept arises from B. Julesz's early work on theory of human visual perception [92], culminating in a 1980s formal definition of the texton theory [93]. It described textons as smallest repetitive patterns composed of local features, such as blobs or corners, constituting human basis for texture recognition and discrimination. While there were many previous attempts at formalising the idea with blob or corner extractors [94], the concept remained largely theoretical until an important innovation was made in 2001, when Leung and Malik [84] established a mathematical definition of a texton as a cluster centre in a feature space resulting in the first Bag of Textons approach. This was later revised by Varma and Zisserman [85, 88] to account for varying scale and illumination conditions within classes, represented directly by the visual words (the most dominant patterns) as opposed to features. Furthermore, through vector-quantizing the feature space, such representations reduce the spatial and computational search complexities, which introduced the possibility of using more complex information-rich features, such as raw image patches [88].

The final piece of innovation leading to modern image processing pipelines was influenced by the field of natural language processing with the introduction of Bag of Words encoding. The idea behind encodings is to construct a global representation over the local image features and thus map an image to a vector representation in such a way that brings similar data points (images) close together in the feature space, therefore enabling analysis with a linear classifier such as SVM [95]. Bag of Words encoding has been initially applied to text in sentences or documents by encoding the number of word occurrences in a set (or bag), disregarding the order of occurrence. Initially the encoding was introduced in texture analysis in a similar manner over a set of textons, or visual words [85]. The first generalised application of such an encoding in computer vision was introduced in VideoGoogle [90] and later formalised

in [91]. More advanced image encodings followed in subsequent years [95].

3.1.2 Bag of Visual Words pipeline

The purpose of texture representation is to transform an image into a set of vectors descriptive of texture properties such that they could be effectively utilised by a classifier [17]. Decades of research have led to the the modern formulation of BoV pipelines, as illustrated in Fig. 3.1. The main building blocks include 1) local texture descriptors (including feature detection and extraction step), 2) global texture representations (including visual words codebook generation and encodings) and 3) a classifier. Any texture representation could be described using the pipeline as presented in Fig. 3.2, and what is more, this systematic framework can also be extended to understand deep representations as described in Chapter 4. This work introduces the concept of Bag of Visual Words pipeline as a novel approach to analysing spatial heterogeneity in DCE-MRI. A number of literature solutions, as described in the following subsection, have been adapted and combined into a single pipeline suitable for this problem. The pipeline can be considered in greater detail as follows.

Feature detection

A set of N patches of size $n \times n$ is extracted across an image I of size $m \times m$, either from pre-defined or automatically extracted keypoints (areas of interest), or densely across an image, sometimes with a stride s. In a special case when $s = n$, it becomes a fixed grid representation.

Although we refer to this part as 'detection', in practice it is a choice between dense or sparse local representation. In texture research, dense features generally dominate [17]. However, a number of methods considered the use of sparse representations, including [96]. While dense representations are capable of encoding more

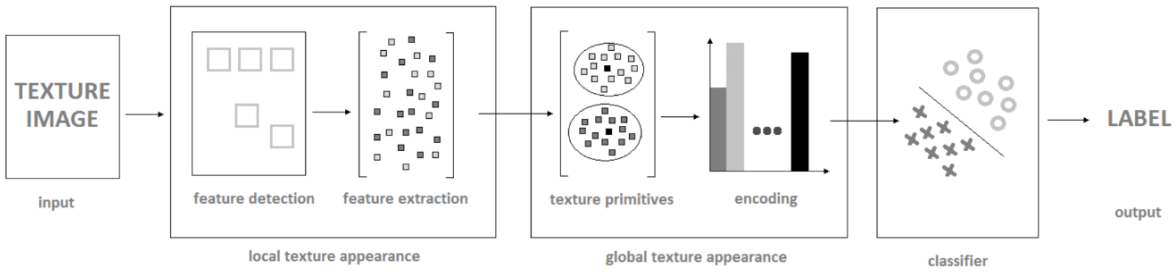


Figure 3.1: Modern texture representation - a schematic example of BoV pipeline. Feature detection step automatically identifies a number of keypoints or areas of interest. These are then input into a dense feature extractor to describe the image in terms of its local neighbouring area. A number of the most descriptive features are gathered into a set of visual words, which constitutes the basis for learning the global labelling. Finally, an encoding is generated describing a global distribution of the words, which is used as an input to a machine learning classifier.

local information, they are also computationally expensive, thus sparse representations are advantageous where memory is a constraint. Initially such representations reflected on visual cues such as blobs, corners or edges as areas of visual interest [94]. More recently, sparse features were also introduced in advanced stereo vision and image registration techniques as a basis for keypoint matching between frames, giving rise to methods such as DAISY [97] or SIFT [98].

Feature extraction

A number of texture descriptors are applied to the pool of extracted patches N (or a whole image if $N = m^2$). The resulting is a set of feature vectors such that $\{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^D$, where D is the dimensionality of the feature vector (e.g. if a descriptor is an 8-point pixel neighbourhood, $D = 8$).

As the development of robust and invariant texture descriptors is generally accepted to be crucial to the success of the representation, research in this area has received much attention over the years. They could be derived either densely across the image, or over a set of pre-extracted sparse regions of interest. The most popular

local feature descriptors can be divided into two groups [89]: 1) wavelet- and filter-based and 2) distribution-based. The first group includes filters and filter banks such as Gabor [99], Maximum Response 8 [85], Leung and Malik [84] and more. The resultant feature vectors from such methods typically concatenate convolution products from various pre-designed filters at various scales and orientations. The second group includes Patch [88], Local Binary Patterns [83] or Random Projections [100] descriptors, and represents the features in terms of spatial distributions of pixel/voxel values.

Visual words

Textures are by definition somewhat repetitive. Feature space, composed of local feature descriptors $\{x_i\} \in \mathbb{R}^D$, is therefore iteratively partitioned, or vector quantised, into a set of the most dominant visual patterns, or words, such that $\{w_i\}_{i=1}^{K}$, $w_i \in \mathbb{R}^D$. The visual words are therefore of the same dimensionality as the feature space, and K refers to the number of components in the generated codebook. The original vectors are then assigned one (hard assignment) or more (soft probabilistic assignment) visual word references.

Common methods used to partition the feature space typically draw from clustering and include K-means, fuzzy c-means or Gaussian Mixture Models. Some representations do not require codebook generation, for example in the case of Local Binary Patterns, however it is important to note that this is only feasible in case of a small image dimensionality, i.e. 2D. Finally, while visual words codebook generation is typically used as a stepping stone towards the global image representation, it can conveniently produce image segmentations, that although methodologically trivial, can give further insight and aid understanding of the final results.

Encodings

Spatial features represent an image as an invariant and versatile, however often overly redundant feature space [17], especially in the case of dense feature extraction methods. Based on the texture definition, it is expected that some local patterns would be repeated, therefore a set of dominant patterns (texture primitives) could be found resulting in a more spatially compact global encoding representation. Encodings aim to describe an image globally from a set of local dominant patterns [95] in a way that reflects a visual similarity between the image/patch pairs through their distance proximity in the feature space. They aggregate statistics or distributions of local patterns as represented by the prior vector-quantised feature space resulting in a set of vectors $\{v_i\}_{i=0}^{N}$, $v_i \in \mathbb{R}^K$ (dimensionality depends on the encoding, for example $\mathbb{R}^{2KD}$ in case of Fisher Vector [101]). Popular approaches include orderless frequency-based Bag of Visual Words (BoV) [91], Vector of Locally Aggregated Descriptors (VLAD) [102], or Fisher Vector (FV) [101] and Improved Fisher Vector (IVF) [103].

Classifier

Such defined global representations of images or image patches are then fed into a machine learning classifier or regressor, depending on the nature of the task. While it is a field of research of its own and a review of such is considered out of scope of this work, the most popular algorithms used in conjunction with shallow image representations include Support Vector Machines (SVMs), and also linear or logistic regression, decision trees, random forests, or simple dense neural network architectures.

In this work we propose an application of the BoV pipelines to tackle the problem

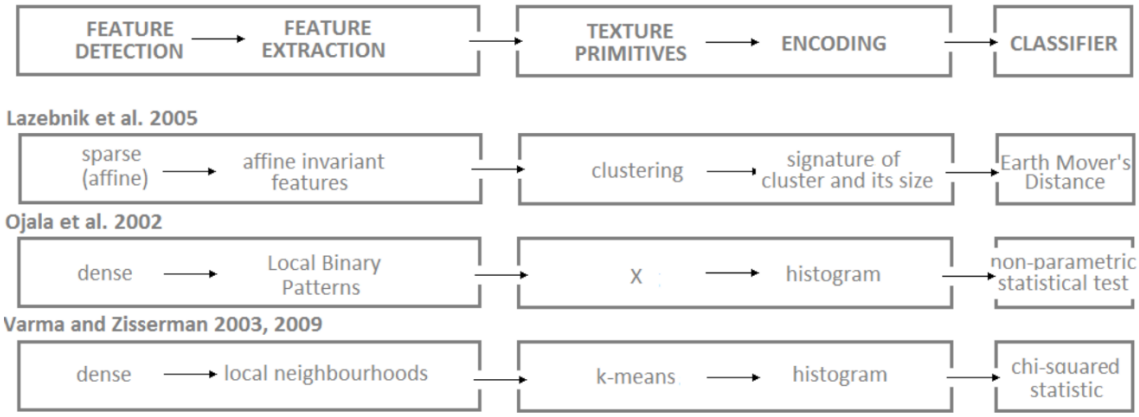


Figure 3.2: Three examples of methods [96, 83, 88] in the context of the standard texture analysis pipeline, as shown in Fig. 3.1. X was put in place of a method if the step was skipped in the original methodology.

of assessing spatial perfusion heterogeneity from DCE-MRI for a number of reasons. Firstly such frameworks are, by design, capable of accounting for local image variation (contained in local neighbourhoods) as well as its global distribution across the whole image. For that reason, they are capable of processing and visualising the more complex variations and patterns that likely to occur in real-world data. We expect the frameworks to be particularly suitable to the problem of assessing tumour perfusion heterogeneity, which may be appearing both in the local and global context of the image. Furthermore, the vector-quantisation step allows for a more compact representation of the original feature space (consequently allowing for more features) and improves computational complexity of classification by working in a quantised feature space, thus reducing the computational burden posed by some existing methods. The vector-quantisation is also performed over the whole dataset (as opposed to per single image), which provides stable and balanced visual words that could improve repeatability of the studies, possibly even without the need of re-training. Moreover, BoV pipelines are modular and can be easily decomposed into a number of discrete processing steps, thus providing an in-depth look into the workings of the algorithm. This is unlike many recent black-box deep representations and could help understand

the theoretical basis of the resulting assessments. This advantage makes them particularly appealing in the context of heterogeneity on the macro-scale as it may yield additional valuable insights into the rationale behind the computationally determined result and could provide clinicians with a degree of certainty and trust. Furthermore, they allow not only for a quantitative analysis, but also provide a much richer qualitative description that could further aid the clinicians. Finally, as MRI introduces variability associated with acquisition parameters resulting in sometimes unintentionally varied intra- and inter-sample tissue appearances, we evaluate a number of robust features designed over the years to address the problem of high inter-class variance, including invariance to scale, rotation and illumination conditions.

3.2 Methods

Modern shallow image processing pipelines (including those used in texture analysis) are characterised by both local and global components. In this work, we consider a number of local feature descriptors in connection with global encodings for the task of assessing spatial perfusion heterogeneity in tumours. While there are many technically advanced ways to analyse, segment and classify medical images, we restrict the investigation to the those particularly suitable to the task, and thus explore a number of texture or variation-based features as a proxy for a measure of heterogeneity.

3.2.1 Texture feature descriptors

In this work we compare three popular feature descriptors: Gabor filters, Patch descriptors and Local Binary Patterns (LBP). Furthermore, we extend those descriptors into 3D. The differences between the descriptors are visually explained in Fig. 3.5

3.2.1.1 Gabor filters

Filter banks are collections of band-pass filters, usually at different scales and orientations. Convolving an image with a filter bank separates it into a number of components equal to the number of filters, where each component is represented by only a subset of frequencies from the original image. In this sense, images could be understood as a superposition of filter responses at all frequencies and orientations (Fig. 3.3 top). The ability of filtering methods to emphasise important information while at the same time filtering out low level frequencies and noise has made them very popular in texture research. Although many types of filter designs exist, in this work we use Gabor filters as a representative example, which are thought to well approximate the function of a mammalian visual receptive field (Fig. 3.3 bottom) [104]. We further extend the notion to 3-dimensional Gabor functions.

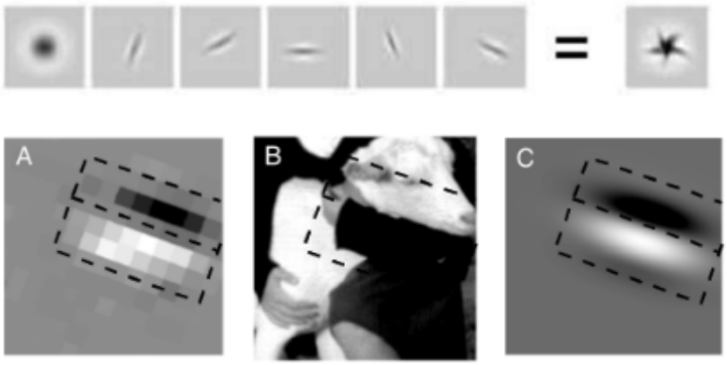


Figure 3.3: Top: Image superposition property, star pattern is decomposed into a number of bases with different base filters and orientations. Image taken from [105]. Bottom: A: Receptive field of a simple cell in V1, B: image corresponding to the highest response, C: Gabor function approximating the visual receptive field. Image taken from [104].

Given a 2D Gabor filter defined as [106]:

$$\phi_{\theta,\sigma,\gamma,\lambda,\varphi}(x,y) = \exp\left(-\frac{x'^2 + (\gamma y')^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \varphi\right) \tag{3.1}$$

where the exponent term is a Gaussian envelope spatially restricting the cosine func-

tion. Additionally, λ represents the wavelength of the cosine and φ is the phase offset. Parameters x' and y' are defined as:

$$x' = x\cos\theta + y\sin\theta \tag{3.2}$$

$$y' = -x\sin\theta + y\cos\theta \tag{3.3}$$

and x and y are the coordinates of the grid of size $n \times n$. We therefore define a 3D Gabor filter as follows:

$$\phi_{\theta,\sigma,\gamma,\lambda,\varphi}(x,y,z) = \exp\left[-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2} + \frac{(\gamma y')^2}{\sigma_y^2} + \frac{(\gamma z')^2}{\sigma_z^2}\right)\right]\cos\left(2\pi\frac{x'}{\lambda} + \varphi\right) \tag{3.4}$$

where x', y' and z' are determined by a product with a rotation matrix in Euclidean space in order to achieve different orientations:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R(\theta)\begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.5}$$

The rotation matrix is defined as a product of three separate rotation matrices over three axes as follows:

$$R(\theta) = R_x R_y R_z \tag{3.6}$$

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} R_y = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} R_z = \begin{bmatrix} 0 & \cos\theta_z & -\sin\theta_z \\ 0 & \sin\theta_z & \cos\theta_z \\ 0 & 0 & 1 \end{bmatrix} \tag{3.7}$$

The filter bank is then constructed by aggregating Gabor kernels at a 4 orientations over the range of $\theta = (0, \pi)$ for each of three dimensions, plus a single Gaussian blob filter, 65 filters in total (Fig. 3.4).

The features are generated via convolution with the filter bank (Fig. 3.5 a), formally:

$$GBR(i,j,k) = \phi * I(i,j,k) = \sum_{p=-n/2}^{n/2} \sum_{r=-n/2}^{n/2} \sum_{s=-n/2}^{n/2} \phi(p,r,s) I(i-p, j-r, k-s) \quad (3.8)$$

where $x(i,j,k)$ is the filter response at a position (i,j,k), ϕ is the previously defined kernel and $I(i,j,k)$ is the intensity of the original 3D volume at the same position. This results in a set of features $\{x_i\}_{i=0}^{N}$, where $x_i \in \mathbb{R}^D$ and $D = 65$.

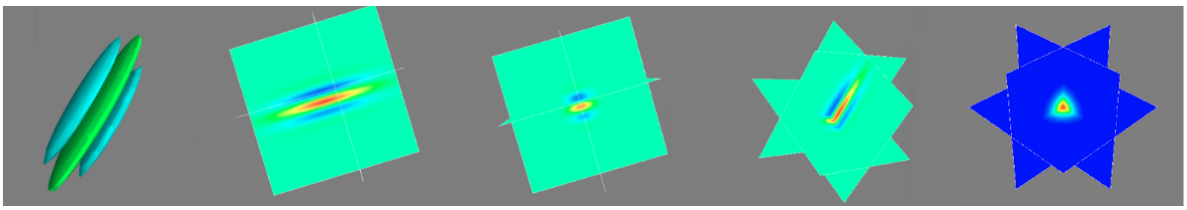


Figure 3.4: Left: 3D contour of a Gabor filter, middle three: overlapping planes overlayed with different views of the same Gabor filter, right: overlapping planes overlayed with a Gaussian filter.

Although shown to be powerful and effective in texture analysis, there are some downsides for filtering methods. Firstly, while the ability to filter out low-frequency noise details is usually considered an advantage, it is worth noting that it could also disregard some significant low-frequency image information. Furthermore, filters need to be pre-designed by humans, which introduces the possibility of bias and emphasises the previous downside as they might not be a good representation of the observed data.

3.2.1.2 Patch descriptors

The design of patch descriptors challenged the supremacy of filtering methods. With the downside of filter responses only retaining high-frequency information and possibly disregarding low-frequency details [107], there was a need for richer texture descriptors. Varma and Zisserman showed that a superior classification accuracy could be achieved by utilising the whole neighbourhood information from local patches. In terms of filter banks, this operation could be understood as a convolution with a filter bank of zeros, except for ones in position of neighbours of interest at each filter. In other words, we are collecting raw intensity vectors densely across an image (as in a convolution) (Fig. 3.5 b), with N neighbourhoods/patches of size $n \times n \times n$, and thus modelling a joint distribution of pixel/voxel intensities with respect to the central pixel. In this work, due to the low resolution of input images we are constrained by the maximum scale of $n = 3$ and N is equal to the number of pixels (dense representation). This results in feature vectors $\{x_i\}_{i=0}^{N}$, $x_i \in \mathbb{R}^D$ and $D = 27$.

3.2.1.3 Local Binary Patterns

Local Binary Patterns (LBPs) are another example of distribution-based feature descriptors. Olaja et al. [83] proposed that much of the spatial information in a local environment is encoded in the differences between pixel intensities, and taking this further they argued that local structure could be encoded only by the sign of the difference. In other words, the local neighbourhood is thresholded by the central pixel value and subsequently binarised (Fig. 3.5 c). Such binary information could be then converted into a decimal value as a local descriptor for the central pixel. Formally, LBP is defined as:

$$LBP(i,j,k) = \sum_{p=-n/2}^{n/2} \sum_{r=-n/2}^{n/2} \sum_{s=-n/2}^{n/2} s(I(i-p, j-r, k-s) - I(i,j,k))2^{prs} \qquad (3.9)$$

where $LBP(i, j, k)$ is the LBP feature vector extracted at a position (i, j, k), $I(i, j, k)$ is the intensity of the original 3D volume at the same position representing the central voxel, 2^{prs} is a decimal conversion and the function $s(x)$ is the binary sign of the difference defined as:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.10}$$

The drawback of the original fundamental definition is that it only keeps the information regarding the structure of the pattern disregarding important information contained in the magnitude of differences. In this work, we therefore opted for one of many variation of LBP, the Extended LBP (ELBP) [108]. ELBP is advantageous in the sense that it exploits extra information given by the local as well as global intensities, and thus puts the pattern structure in the context of neighbourhoods as well as the image. Furthermore, instead of only keeping the sign, we retain the information about the magnitudes of differences, thus no longer using binary information. We define ELBP as a concatenation of two separate descriptors: ELBP_CI and ELBP_NI. Originally, ELBP_CI is defined as the pattern of local differences with the mean of the image, thus giving it a global context:

$$\text{ELBP_CI}_p(x_c) = s(x_c - \beta) \tag{3.11}$$

where x_c is a central voxel of a patch p, previously corresponding to the notation of $I(i, j, k)$, β is the mean of a whole image and $s(x)$ function returns the sign of x. The ELBP_NI is further originally defined as the local differences with the mean of local neighbourhood, thus giving it a local context:

$$\text{ELBP_NI}_p(x_c) = \sum_{i=1}^{n^3} s(x_i - \beta_p)2^i \tag{3.12}$$

where β_p is the mean of the local intensities across the neighbourhood p. Originally, the intensity information is collected on a circular neighbourhood with the radius size of r around the central pixel $(-r\sin(2\pi i/p), r\cos(2\pi i/p))$. Due to the resolution and scale restrictions we simplify this to the trivial $n \times n \times n$ neighbourhood extraction.

There are a number of design decisions and changes we made to adapt LBPs to our application domain. Due to the restrictions on scale imposed by the low image resolution, we found that the structural information carried by sign alone was too sensitive to the noise, which is ubiquitous in such an application domain. Furthermore, the spatial complexity of the standard LBP encoding increases together with an increase in image dimensionality, with 256 dimensional vectors from 2D images, but 1.3×10^8 dimensional vectors in case of 3D images for a single scale only. We therefore opted to use a non-binary difference-based vectorised version of LBP, which both carries more local information than signs alone and has a lower dimensionality. Formally this is defined as:

$$\text{ELBP_CI}_p(x_c) = \{x_i - \beta\}, i \in N(p) \tag{3.13}$$

$$\text{ELBP_NI}_p(x_c) = \{x_i - \beta_p\}, i \in N(p) \tag{3.14}$$

where p is an index of the local neighbourhood expressed as a set of differences between each voxel and the local or global mean, and $N(p)$ represents the set of neighbours of point p. The descriptors are then concatenated together as a single feature descriptor, resulting in a set of features $\{x_i\}_{i=0}^{N}$ with $x_i \in \mathbb{R}^D$ and $D = 54$ (concatenation of two 27-point neigbourhoods).

Furthermore, the original ELBP, apart from global and local intensity information, also included ELBP_RD which encoded radial difference information related to different scales, however due to low data resolution we were limited to the use of single scale and thus excluded this particular descriptor. For the same reasons we did not

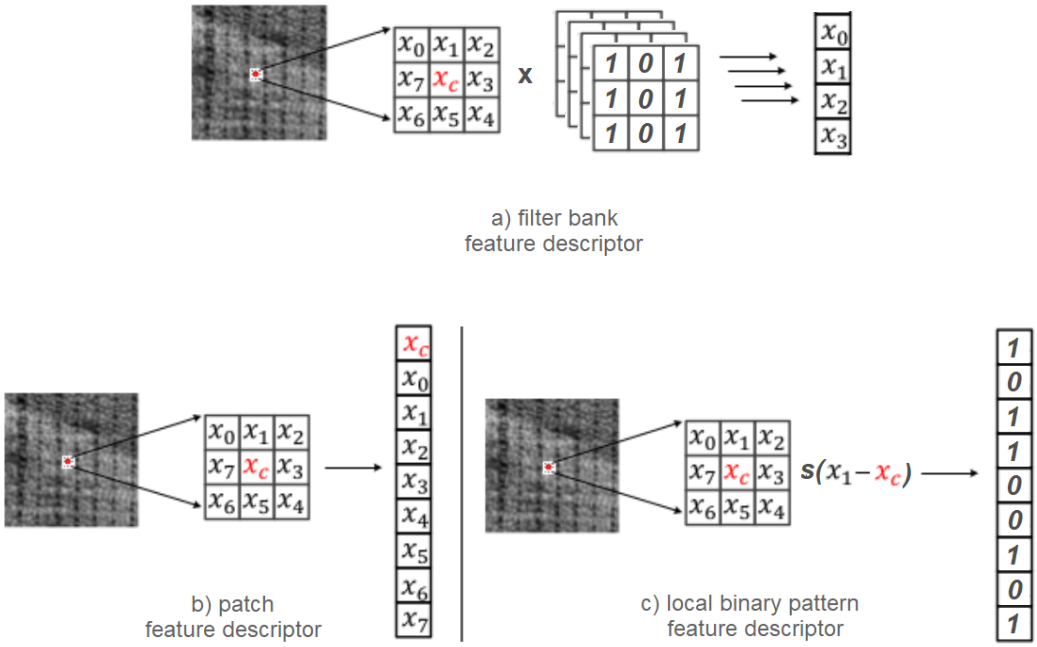


Figure 3.5: Illustration of the differences between the chosen texture descriptors: a: filtering methods, b: patch-based methods, and c: local binary patterns.

consider the use of RELBP or MRE_LBP in this work, as they both operate over multiple scales. Finally, although the standard LBP is not used in conjunction with BoV encodings, in this work retain the vectorised version of the differences rather than a single decimal value encoding. For that reason, BoV encodings are used, resulting in information rich and computationally light representations in 3D.

3.2.2 Encodings

In this work we evaluate two variants of the standard BoV encoding focusing on codebook generation: 1) with KMeans clustering and 2) with Gaussian Mixture Models.

Initially, we reduce the spatial complexity of the feature space by vector-quantizing it. This generates the codebook of most dominant words, or texture primitives. The partitioning of the feature space is achieved by these two popular codebook generation methods: KMeans and GMMs. As described in Chapter 2, KMeans aims to minimise

the sum of squared differences between data points and cluster means for each of the clusters:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{3.15}$$

This is achieved by initialising the data with some prior and iteratively updating the means until convergence is reached as described in Section 2.3.6 of Chapter 2. In contrast, GMM aims to describe the data by a likelihood of it being represented by a linear combination of Gaussian components:

$$\sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \;\; \sum_{k=1}^{K} \pi_k = 1 \tag{3.16}$$

The final assignments, via highest probabilities for all data points, are achieved iteratively through the EM algorithm as described in Section 2.3.6 of Chapter 2. As in the original paper by Varma and Zisserman [85], in this work the codebook generation is performed on a per-class basis to account for all possible variations of the perfusion enhancement patterns within each time point.

We adopt the original BoV orderless frequency encoding to model the image global representation. Given a set of feature vectors $\{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^D$ and a visual codebook $\{w_i\}_{i=1}^{K}$, $w_i \in \mathbb{R}^D$, for each of the classes k a global image representation v_k of that class is defined as:

$$v_k(w_k, x) = \sum_{i=1}^{N} f(x_i, w_k) \tag{3.17}$$

where $f(x, w)$ is a frequency function defined as:

$$f(x_i, w_k) = \begin{cases} 1, & \text{if } l(x_i) = w_k \\ 0, & \text{otherwise.} \end{cases} \tag{3.18}$$

and the function $l(x)$ returns the label assigned for feature x.

While the focus of this chapter lies in evaluating features, we attempt to keep the

rest of the pipeline relatively constant. Although there are many advanced shallow encoding methods, for simplicity and clarity while evaluating features we use the original BoV frequency encoding with hard assignments. For the sake of completion, in-depth evaluation of encoding methods can be found in [95].

3.3 Results

We conduct a number of experiments to evaluate the proposed framework. Initially we qualitatively assess the suitability of the method to the domain and the problem on two small pre-clinical longitudinal and single timepoint datasets. We further use a larger longitudinal dataset to support statistical testing of the results, and evaluate the data with respect to two tasks: segmentation and classification. Quality of segmentations is assessed with three types of qualitative and quantitative measures: 1) visually assessed contiguity of segmentations, 2) confidence of segmentations, along with visual maps as a confidence aid, and 3) classification accuracy itself. We further show, via a proxy non-perfusion dataset, that such segmentations (and thus representations) can be related to physiological features on a macro-scale as assessed via manually acquired ground-truth data. Finally, we investigate the assumption of spatial variation as a measure for heterogeneity. To that end, we compare it with similarity-based features to assess the influence the choice of features, along with their assumption, has on the final outcome.

3.3.1 Materials

A number of datasets were used for the evaluation of the method. All the cohorts and corresponding data used in this thesis are presented in the Table 3.1. Initially, in order to determine the potential of the method as a future biomarker, we used small longitudinal and single timepoint datasets. In order to enable statistical testing, we

experiment type	qualitative (trial) study	quantitative (evaluation) study	clinical (application) study
data used	D1 and D2	D3	D4
methods employed	Patch + KM	Vox + KM/GMM Filters+KM/GMM Patch+KM/GMM LBP+KM/GMM UNet+KM/GMM RandNet+KM/GMM	Vox + KM/GMM Filters+KM/GMM Patch+KM/GMM LBP+KM/GMM UNet+KM/GMM RandNet+KM/GMM
chapter in the thesis	Chapter 3 Section 3.3.2	Chapter 3 Chapter 4	Chapter 5

Table 3.1: Each cohort used in this thesis. Details of the datasets D1, D2 and D3 can be found in Table 3.2 in this section. Details of the dataset D4 can be found in Table 5.1 in chapter 5 where analysis was performed.

used a larger longitudinal dataset for method comparison and evaluation. Finally, we used a clinical data in Chapter 5 to attempt answering more clinically relevant questions.

Qualitative evaluation: trial study

We tested the method on two pre-clinical datasets of female C57/BI6 mice (D1 and D2, Table 3.2). Images were acquired at 7T DCE-MRI with 500 repeats of a 3D gradient echo scan (TR=1.15ms, TE=0.6ms, FA=5°) and isotropic resolution of 0.411 x 0.411 x 0.411 mm using a custom-made single turn surface coil loop that was integrated into a custom-made tumour support apparatus [109]. T1(0) mapping was performed using a variable flip angle implementation of the DCE type scan described above but operating with 14 flip angles in the range 1-6 degrees (nominal), with B1 inhomogeneities accounted for using images acquired with the actual flip angle method [110]. The longitudinal (D1) study was used with adenocarcinoma NT (CaNT) tumour models with DCE-MRI scans acquired over four days for three mice. Two time

DATA	NUMBER OF SUBJECTS	NUMBER OF SCANS	RES	TUMOUR MODEL	METHODS
D1	3 mice (female C57/B16)	4 scans each (one per day) 12 scans total 10 usable scans*	0.41mm isotropic × 500s	CaNT	Patch+KM (trial study)
D2	11 mice (female C57/B16)	1 scan each 11 scans total 11 usable scans	0.42mm isotropic × 500s	MC38	Patch+KM (trial study: control)
D3	10 mice (female CBA)	7-10 scans each (one per day) 82 scans total 69 usable scans**	0.42mm isotropic × 60s	CaNT	all methods (quantitative)

Table 3.2: Details of each dataset used in qualitative and quantitative study in Chapter 3 and 4.
* 2 scans were unusable due to noise in acquisition, for details see Table 3.3.
** 4 scans were missing and 9 were unusable due to noise in acquisition, for details see Table 3.4.

D1	M01	M02	M03
day 1	✓	✓	✓
day 2	×	✓	✓
day 3	✓	✓	✓
day 4	✓	×	✓

Table 3.3: D1 longitudinal data used in qualitative evaluation. 12 scans in total, 10 usable scans. Scans marked with × were unusable due to noise in acquisition. D2 is not listed as it contained 11 tumours all of which were used and it was not a longitudinal study.

D3	C96_R1	C96_R2	C96_R3	CR96_R1L1	C98_R1
day 1	✓	✓	✓	✓	✓
day 2	✓	✓	✓	✓	✓
day 3	✓	×	✓	✓	✓
day 4	✓	✓	✓	✓	✓
day 5	✓	✓	×	×	✓
day 6	✓	✓	✓	✓	✓
day 7	×	✓	✓	✓	✓
day 8		✓		✓	✓
day 9					?
day 10					?
D3 c.d.	C98_R3	C97_NO	C97_R3	C98_NO	C98_R2
day 1	✓	✓	✓	✓	✓
day 2	✓	✓	×	×	✓
day 3	✓	✓	✓	✓	✓
day 4	✓	✓	✓	×	✓
day 5	✓	✓	✓	✓	✓
day 6	✓	×	✓	✓	✓
day 7	✓	✓	✓	✓	×
day 8	✓	✓	✓	✓	✓
day 9	?				
day 10	?				

Table 3.4: D3 longitudinal data used in quantitative evaluation. 82 scans in total, 69 usable scans. Scans marked with × were unusable due to noise in acquisition; scans marked with ? were missing from the dataset.

points from two different mice were unusable due to noise in acquisition (Table 3.3). The non-longitudinal study (D2) contained a range of eleven tumour volumes (80-290mm^3) available with MC38 tumour models captured at the same growth stage. As the study was not longitudinal, tumour volume was used as a mock indicator of the growth stage to test the effect of scaling in the tissue characteristics. To segment the tumour tissue we used semi-automatic thresholding based on the values of flip angle.

Quantitative evaluation: main study

A larger longitudinal data set was used to allow for quantitative evaluation of the framework (D3, Table 3.2). The scans were acquired with a 4.7T DCE-MRI using 60 repeats of cardio-respiratory gated RF and gradient spoiled 3D echo scans every 10-15 seconds (TR=1.4ms, TE=0.64ms FA=5^o) with isotropic resolution of 0.42 x 0.42 x 0.42 mm. T1(0) mapping was determined from a B1-corrected variable flip angle (VFA) scan with 16 FAs ranging from 1^o to 7^o. The data set was acquired for 10 female CBA mice implanted with CaNT tumour models scanned longitudinally over 8 to 10 days, although some samples were missing up to two data points, yielding a total of 69 usable volumes spanning 8 time points (Table 3.4).

Analysis parameters

In all cases we were restricted by a low resolution to use only the smallest possible scale of $n = 3$ and therefore kept the scale parameter constant across experiments. The choice of words/mixing components parameter k is always an important decision when it comes to clustering, and while there are a few ways to estimate it (e.g. elbow), we employed an informed trial-and-error strategy. We were aiming at keeping

the number of ks meaningful in line with our knowledge of tumour microenvironment, as well as relatively low to allow for visually meaningful segmentations. Initial prototyping (Sec. 3.3.6, Fig. 3.16) revealed the optimal value of $k = 4$ that also neatly corresponds to a likely number of tissues associated with tumour microenvironments.

As it is not possible to accurately estimate developmental boundaries for a growing tumour, in all classification experiments we kept the number of classes constant and equal to $c = 2$ (early and late stage[1]) with a suitable margin of two samples in between to avoid biasing the classifier with intermediate patterns. We performed classification with Leave-One-Out cross validation (LOOCV) on a per-subject basis, which is equivalent to k-fold cross validation on a per-sample basis (k equal to the number of samples) to avoid biasing classifier predictions with previously seen tissues.

3.3.2 Qualitative evaluation

We evaluate the suitability of the proposed framework to the problem as well as the application domain. For this, we tested the method on two pre-clinical datasets. One longitudinal study was used with adenocarcinoma NT (CaNT) tumour models with DCE-MRI scans acquired for three mice over four days. A non-longitudinal study with MC38 tumour models was used as a comparison control, consisting of eleven scans acquired at the same or a similar developmental stage. The analysis was performed over quantitative parameter maps, acquired as described in Section 2.3.1 of Chapter 2. EES space enhancement was modelled with Toft's model and population AIF [111]. Volume transfer coefficient K^{trans} was derived and used for this analysis as suggested in [54]. No prior-motion correction was required due to the custom-made tumour support apparatus. We performed the analysis on both datasets separately.

[1]Early and late stages refer to the first and the last scans in the longitudinal series for each tumour. The first scans were acquired as soon as the tumour became visible with a naked eye (as qualitatively assessed by a radiobiologist in charge of the acquisition), the last scans were obtained on last day of the longitudinal period (day 4 or earlier if day 4 was missing from the longitudinal series as in the case of M02).

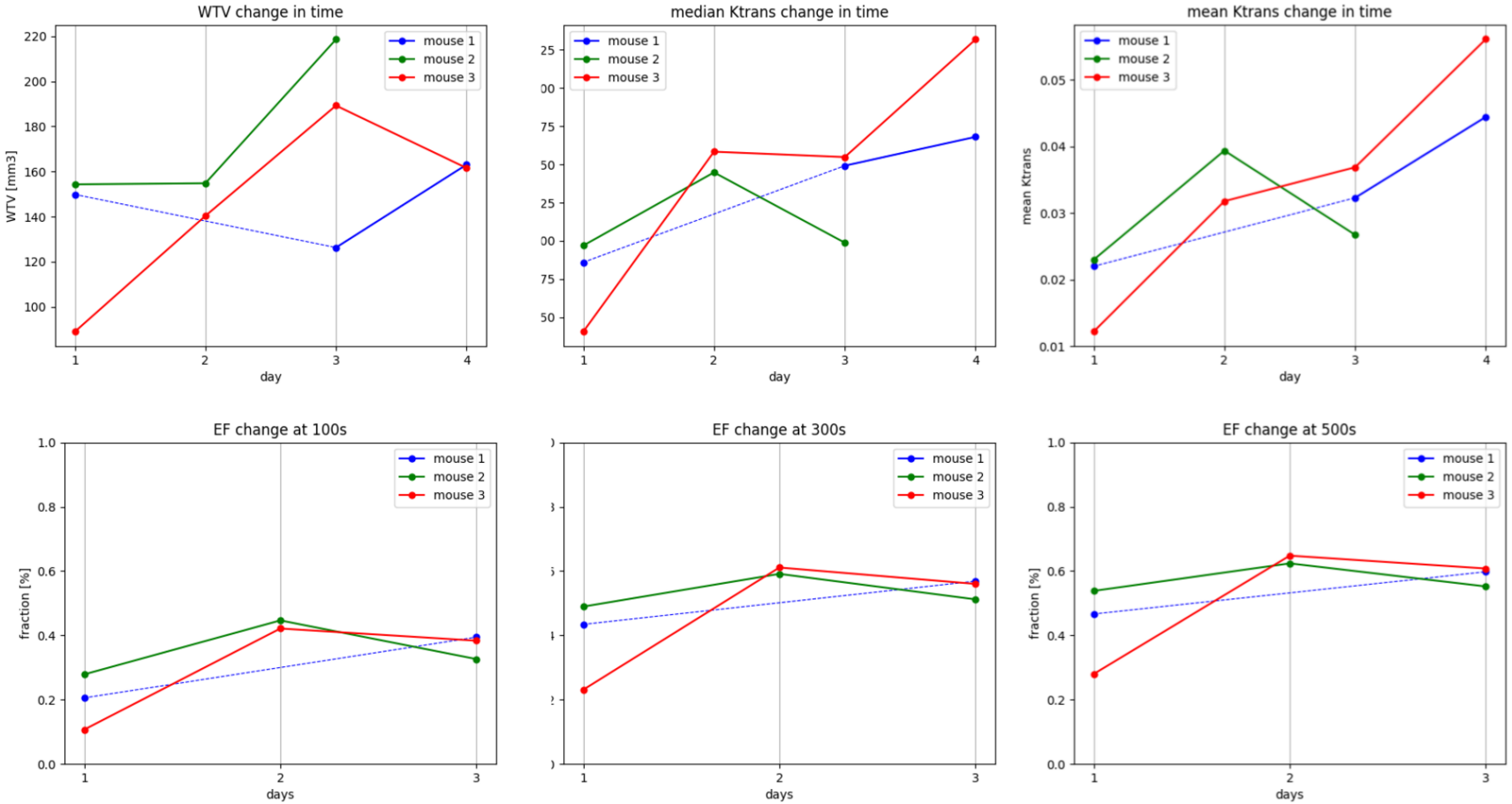


Figure 3.6: Changes in tumour statistics in time (4 days*) for all three mice in D1. Top row showing changes in WTV, median and mean K^{trans} in that order. Bottom row showing changes in EF at 100s, 300s and 500s in that order.
* Raw DCE data was missing for all mice on day 4, thus EF was only calculated until day 3.

We grouped the longitudinal data into two classes: early and late growth stage. As the definition of the growth stage is somewhat arbitrary due to the lack of formal definition of stage boundaries, we aimed to allow for a suitable margin of at least two days between the classes and therefore due to the small number of time points we only used the first and the last time point. The non-longitudinal dataset was sorted by tumour volume and grouped into two classes by using a threshold equal to the median value of the volumes. Due to the low number of data points (both longitudinally and in terms of sample size) we evaluated the framework qualitatively and visually to assess its fitness for the domain and the problem. All of the initial analysis was performed with a Patch+KM variant of the BoV framework (patch features and KMeans-based visual codebook in the frequency based encoding) as described in the original paper [85].

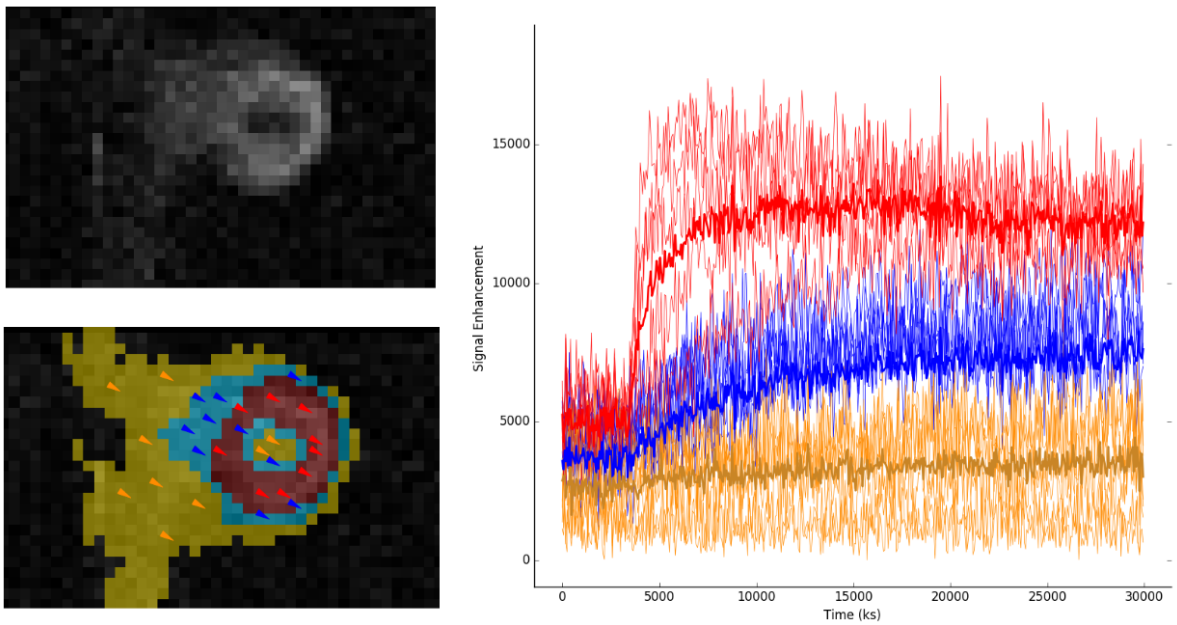


Figure 3.7: Enhancement trace for each type of visual word (texton). Left column: sagittal slice across the tumour showing raw DCE image from dataset D1 sample M01 day 3 and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour.

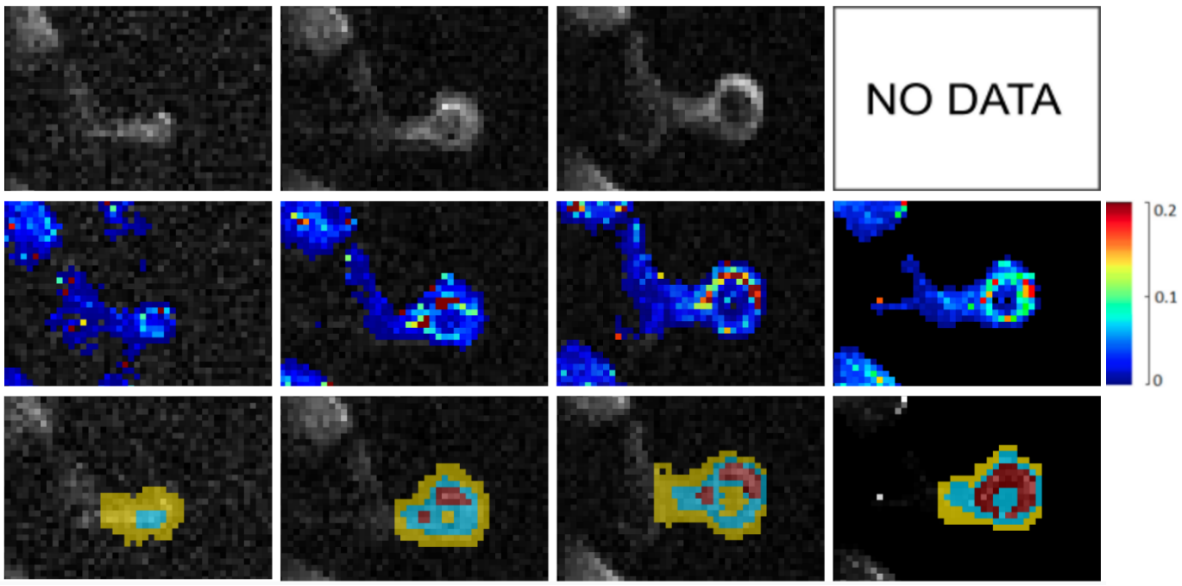


Figure 3.8: Longitudinal change in subregional distributions for a single mouse (M03) in D1. Columns relate to days of progression, with the leftmost column being day 1 and rightmost column day 4. First row corresponds to raw DCE-MRI images, second row represents K^{trans} maps (overlaid on DCE image) and third row shows subregional segmentations (overlaid on DCE image). No DCE-MRI data for M03_day4 was present in the dataset (the segmentation for that sample was overlaid on gray-scale K^{trans} map instead).

Statistical summary of the data is shown in Fig. 3.6, including change in whole tumour volume (WTV), median and mean K^{trans} values, and enhancing fraction (EF) at 100s, 300s and 500s during contrast uptake. A threshold for enhancing voxels was defined as a signal change greater than 4 standard deviations of signal values over the WTV in the pre-contrast images as in [112]. Due to the large amount of noise present, the s.d. was taken from all pre-contrast images (first 50 timepoints). A fraction of voxels exceeding that threshold was calculated at 100, 300 and 500s. As the data contained a large amount of noise (Fig. 3.7), the value for each voxel at each timepoint was taken as a mean of 10 neighbouring timepoints for that voxel.

The method was capable of clearly visualising subregional longitudinal change in the tumours (Fig. 3.8). Mean enhancement traces as well as enhancements for individual voxels were visually distinct between different subregions of the tumour

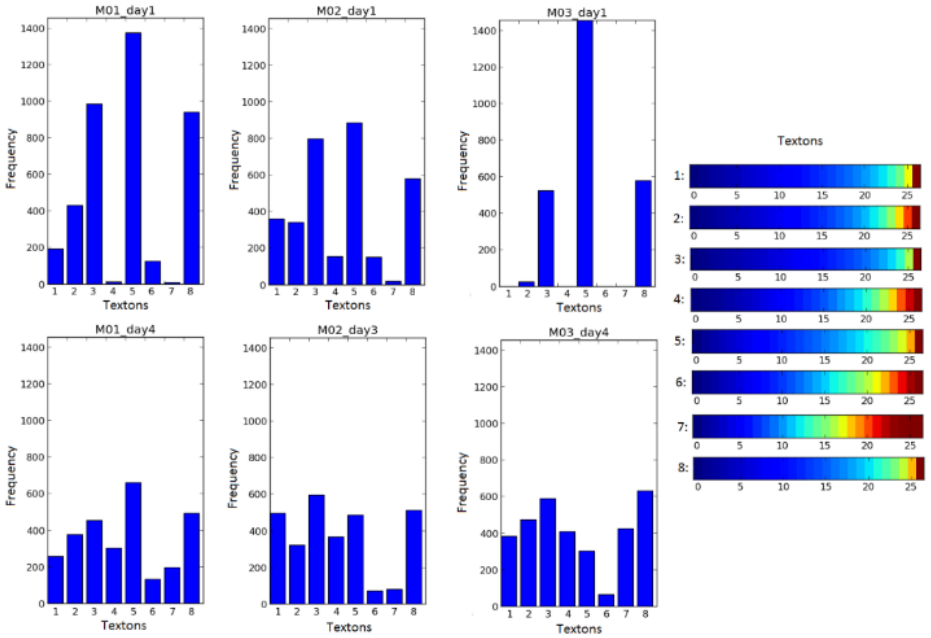


Figure 3.9: A change in global frequency distribution of local perfusion patterns in longitudinal study (CaNT tumour model) for day 1 (top) and day 4 (bottom). This relates to our intuition about the change in tumour heterogeneity, homogeneous at first (only a limited number of similar textures frequently appearing), to highly heterogeneous (all textures frequently appearing). Bins in the histograms are corresponding in fixed order to the list of local perfusion patterns shown on the right. (M02 was missing a sample from day 4 due to noise in acquisition and during this analysis it was replaced with the sample from day 3).

(Fig. 3.7). Mean K^{trans} values for the subregions were 0.01, 0.02 and 0.07 for yellow, blue and red subregions respectively. Variance in K^{trans} values for the subregions was 0.01, 0.03 and 0.05 for yellow, blue and red subregions respectively. The red subregion had both the highest mean as well as the highest variance, and visually appeared to describe areas of high contrast uptake. Additionally, longitudinal change in the fraction of visual word 3 (red) correlated with change in EF, but the correlation was not significant (Pearson, p=0.07). Yellow subregion had the lowest mean K^{trans} value, and related to areas of low contrast uptake which could describe areas of low perfusion or necrosis. The maximum and minimum K^{trans} value was 0.17 and 0.0, and this held true for every visual word. This suggests (and visual analysis confirms this) that the method is robust to local inhomogeneities in signal and noise, which is often a problem in voxel-based type analyses.

The proposed method was capable of visualizing a change in global frequency distributions of local perfusion patterns in the longitudinal study (Fig. 3.9). While in the early stages tumours were globally homogeneous with certain patterns (typically more homogeneous) appearing more frequently, in the late developmental stages we observed more uniform distributions of local perfusion patterns, thus suggesting globally quite heterogeneous tumours. It is important to note though, that such local patterns could be either heterogeneous or homogeneous. It should be highlighted that boundary effects could be reflected in the distributions, particularly when using larger numbers of visual words. Experiments using k=3 words provided reasonably stable features (with no obvious boundary effects) across different tumours as well as the datasets (Fig. 3.7, 3.11, 5.2), but it was not assessed for statistical significance. Furthermore, the change was discriminative enough to differentiate between the growth stages with SVM, with only a single misclassified sample. Although it should be noted that due to the small sample size, such classification results should be taken with caution and the experiment was executed only to gain some initial

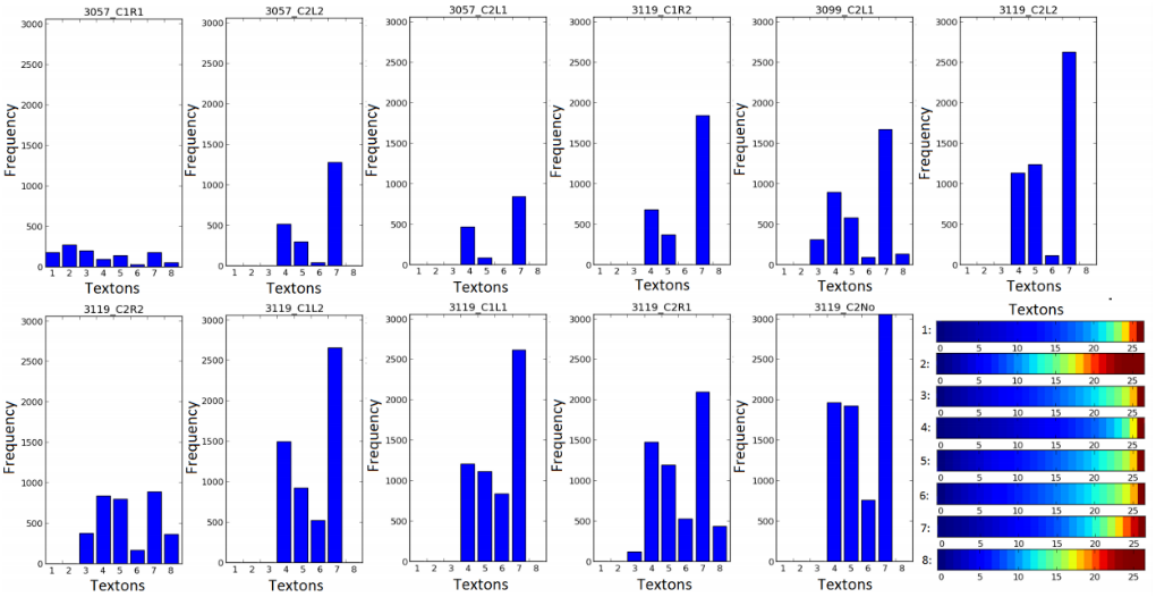


Figure 3.10: Global frequency distribution patterns in single scans categorised by tumour volume (below median volume value (top) or over median volume value (bottom) for MC38 tumour model). Bins in the histograms are corresponding in the fixed order to the list of local perfusion patterns shown on the right. Note that due to the larger tumour volumes, more bins are filled in the single timepoint histogram (increased area), but their shape (distribution) remains similar.

intuition for further directions. It is important to highlight that mice also received a small dose of radiation during a CT scan which could have had a therapeutic effect.

We further evaluated the method on the non-longitudinal data (Fig. 3.10). While the non-normalised histogram areas increased due to the increase in tumour volume, there was no apparent change of distribution in between the volumes (as confirmed through classification), suggesting that the observed changes in the previous dataset were not due to the scaling of local patterns. The initial experimental results constitute a preliminary confirmation of our intuition about the influence of change in microvasculature and perfusion onto the macro-appearance of the tumour, as measured by DCE-MRI.

3.3.3 Quantitative evaluation

As mentioned earlier, one of the main aims of this chapter is to develop and understanding of the changes in perfusion heterogeneity of the tumour during standard growth as unaffected by therapy, which could in turn give further insight into therapy data. For that reason, we evaluate the framework on a larger longitudinal data set, consisting of a wide range of time points per sample. Furthermore, a larger number of samples also allows for a quantitative evaluation of the framework and feature extraction method comparison.

We evaluate the framework on a longitudinal study of 10 mice scanned over 10 days. Due to some samples missing time points, there was a total of 69 data points available for evaluation. We use the Tofts model and population derived AIF to model EES enhancement and derive K^{trans} parameters as a measure of perfusion as recommended in [54]. Motion artefacts were minimised with cardiac synchronisation and subsequent immediate re-acquisition.

We evaluate the method on two tasks: 1) subregional assessment via segmentations, and 2) quantifying the change in the overall degree of heterogeneity via growth-based classification. In the subregional assessment, we used all available data for segmentations. In the task of quantifying the change in the overall degree of heterogeneity, we classify the dataset into two classes of early and late growth stage. As opposed to the therapy response data, the growth stage definition remains somewhat arbitrary. For that reason we aimed to reserve a suitable margin between the classes to ensure minimal bias. In addition, due to the small volumes restricted by relatively low spatial resolution, we reject the initial two time points from classification. Due to the large number of longitudinal timepoints we aimed to select at least two volumes per sample per class. Therefore, in most cases the early stages (class 1) contained day 3 and 4 of tumour growth and late stages (class 2) contained day 7 and 8. While mostly following the outlined rule, due to incompleteness of data, some samples were

reassigned the next closest time point. To avoid computational overhead, the images were cropped to the smallest possible window capable of containing the largest tumour across all images.

In both the segmentation and classification tasks, the feature extraction methods under evaluation were Filters (Gabor), Patch and LBPs as compared to a standard voxel-based method. In addition we used two variants of codebook generation in BoV orderless frequency encoding: KMeans and GMM. Outside of method comparison, as in the previous subsection, Patch+KM remains our baseline method for other experiments. We evaluate the methods on the task of segmentation and classification with a number of metrics. As spatially unstructured and non-contiguous labels are not useful in a clinical context, we primarily seek segmentations with a high visual coherence and contiguity. Furthermore, to assess the confidence in segmentations, motivated by the lack of validation endpoint we introduce a confidence-based ground-truth and evaluate the segmentations with the DICE coefficient. In the classification task we evaluate the methods with the measure of classification accuracy. Finally, to evaluate the overall fitness of the method to the problem domain, we consider all three metrics in conjunction, as we are aiming for segmentations that are both descriptive of heterogeneity and visually coherent, to optimise the clinical relevance.

3.3.4 Segmentations

The introduced framework was capable of visualising the changes in subregional tumour distributions over the course of tumour growth with visually distinct enhancement traces (Fig. 3.11). The results from the same method (Patch+KM) for three exemplar samples of growing tumours can be seen in Fig. 3.12. As an estimate of the corresponding slices in a growing tumour, we show the largest slice through the sagittal plane. The labels are arbitrarily assigned to the learned visual words, representing different perfusion patterns (between 0 and 4) and are corresponding across

all images.

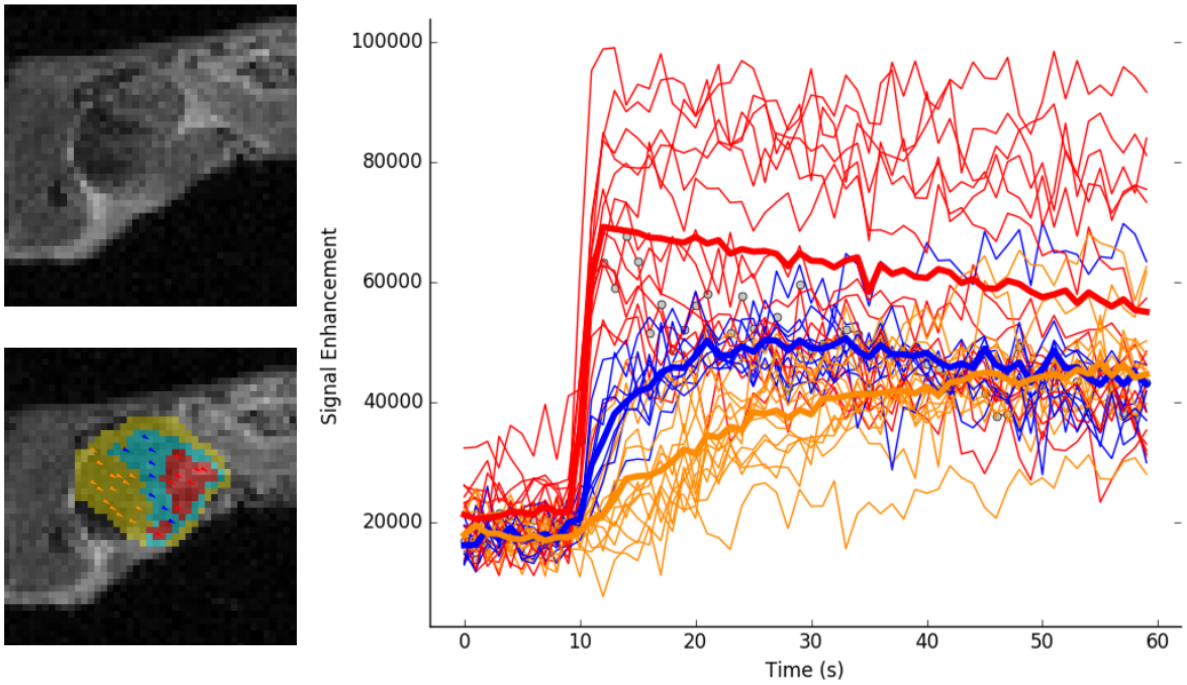


Figure 3.11: Enhancement trace for each type of visual word (texton). Left column: sagittal slice across the tumour showing raw DCE image from dataset D3 sample C96_R2_D6 and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour.

The observed visual distributions of perfusion patterns reveal a similar, but opposite change compared to that seen in the trial study. The tumours appear initially heterogeneous with a number of uniformly frequent local perfusion patterns, becoming globally homogeneous with a single dominant perfusion pattern (green). However, it is important to highlight that such difference observed between the two studies could be caused by the fact that the tumours in this study were scanned for over twice as long as in the trial data, and the early time points in this study roughly align with the last time points of the previous study. Segmentations from those days in both studies do indeed reveal similarly heterogeneous tumours.

While there is no segmentation ground truth available for this dataset with respect

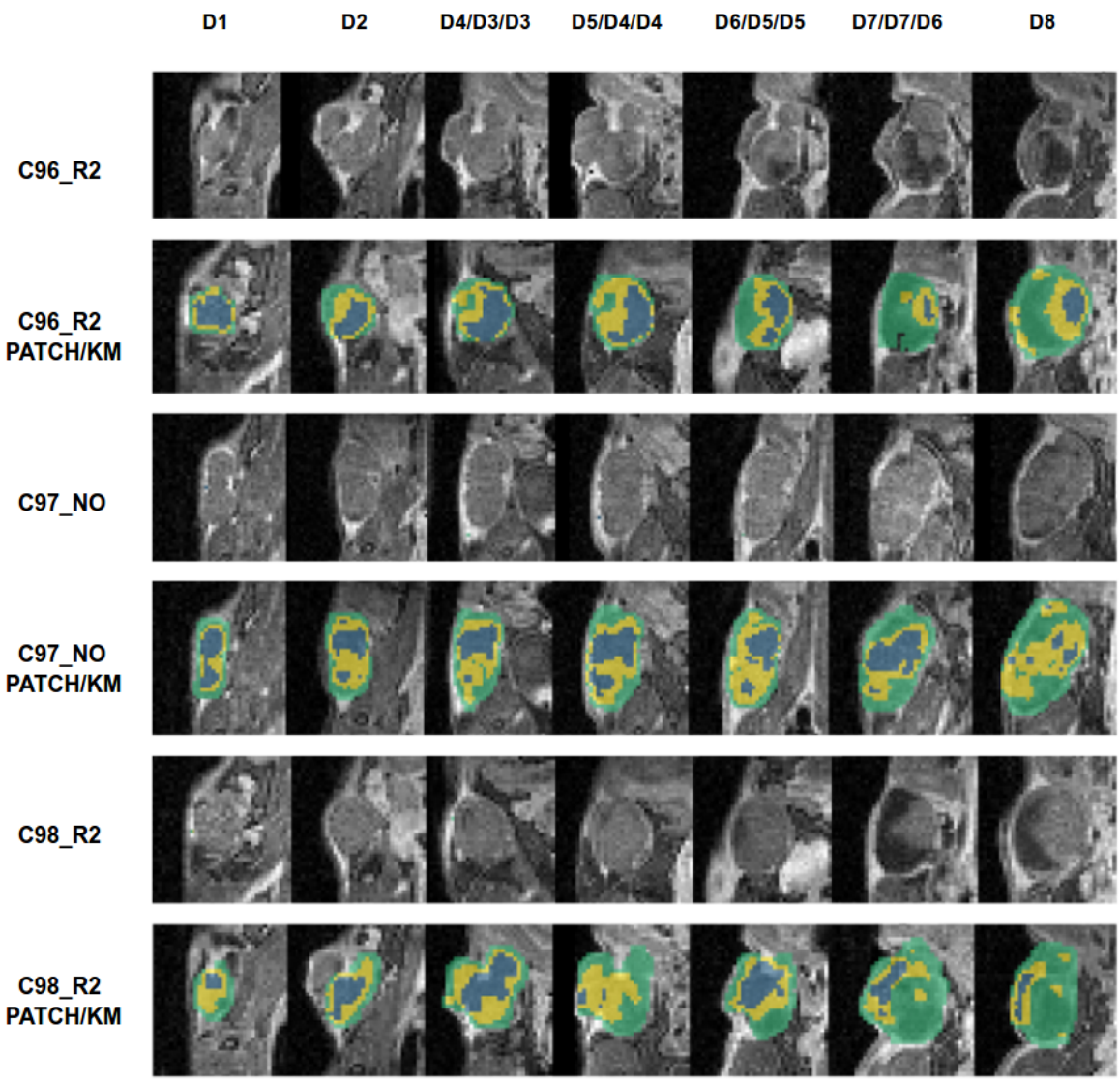


Figure 3.12: Subregional distributions of a growing tumour (columns relate to days of growth) shown for three different samples with Patch-KM. Pairs of rows (1, 2), (3, 4) and (5, 6) originated from three different tumours C96_R2, C97_NO and C98_R2 respectively (Table 3.4). Where a timepoint was missing from the series, labels for all individual mice were provided in top-bottom order. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). We are observing a change similar to the one observed in the trial study.

to subregional distributions, we would like to have a way to, at least, estimate the confidence in the segmentation results. We therefore we define a confidence-based DICE coefficient metric. We formulate a consistency measure based on an assumption that human confidence in segmentation accuracy increases together with the number of similarities between the segmentations from different methods. We therefore define a ground truth based on per-voxel mode across segmentations from different methods for each of the tumours, indicating per-voxel agreement. We then use the DICE metric to measure the agreement of each segmentation method with proposed consistency ground-truth formally defined as:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{3.19}$$

where TP stands for a true positive, FP stands for a false positive and FN stands for false negative.

Furthermore, we can represent the generated consistency ground truth as segmentation confidence maps, where a higher number of label disagreements per-voxel across segmentations indicates low confidence, while a high number of agreements indicates high confidence.

Naturally there are some drawbacks associated with this approach. For example DICE scores could be high if all segmentations appeared similarly incorrect, or the score of the correct segmentations would be low in case of a large number of incorrect ones influencing the consistency ground truth. This metric is therefore to be used with caution, together with a visual inspection of the data and confidence maps.

Segmentations from different descriptors for the same exemplar sample of a growing tumour can be seen in Fig. 3.13 for the KM variant of visual codebook and Fig. 3.14 for the GMM variant. Average DICE scores across the methods are presented in Table 3.5, indicating that most methods (with exception of Vox) produced

segmentations that were similar for between 60% and 96% of the tumour volume. Patch and LBP were capable of creating the most visually coherent segmentations with both encodings. Filters and Vox in general produced visually non-contiguous segmentations with both encodings, but with the GMM variant of codebook generation they were slightly better. DICE scores confirm our visual estimates. The scores were generally higher for both Patch and LBP, although lower for KM, due to more disagreements across the feature descriptors than in GMM. In general, segmentations in GMM agreed more, both, visually and with the DICE score.

The segmentation confidence heatmaps can be seen in Fig. 3.15, with the high values (black) representing high confidence (all methods agreeing on a label) and low values (yellow) representing low confidence (most methods disagreeing on a label). In general, they reveal confident segmentations, with above 50% agreement for most voxels. As expected, the lowest confidence was visible in the early stages, which could be explained by their small size as restricted by the low resolution, resulting in high disagreement.

Due to the lack of ground truth data, segmentation confidence and DICE scores alone are not sufficient to evaluate the methods, as segmentations on their own are a convenient by-product of the pipeline. We therefore assess the segmentations in conjunction with classification accuracy, which is the primary intention and final goal for segmentations.

3.3.5 Classifications

The encodings were evaluated by a SVM classifier with parameters kept constant and using the defaults in the scikit-learn [113] implementation. The accuracy was estimated with a Leave-One-Out Cross Validation (LOOCV). To avoid the introduction of biases associated with intra-subject tissue similarity into the classifier, the LOOCV was performed on a per-subject basis, thus 4 time points per fold with the number of

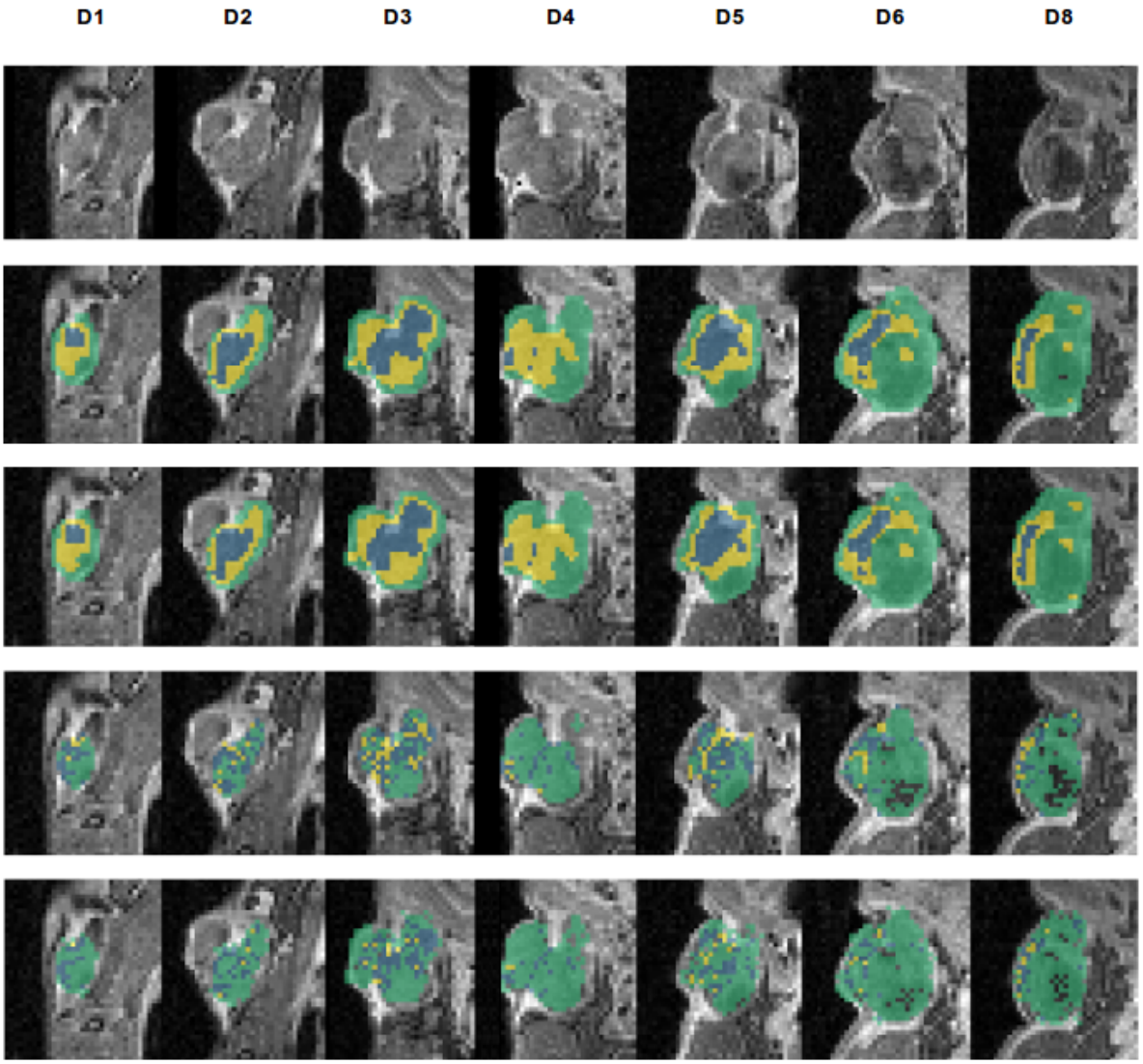


Figure 3.13: Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of feature descriptors in KMeans BoV encoding, from the top: Patch, LBP, Gabor, Vox. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3).

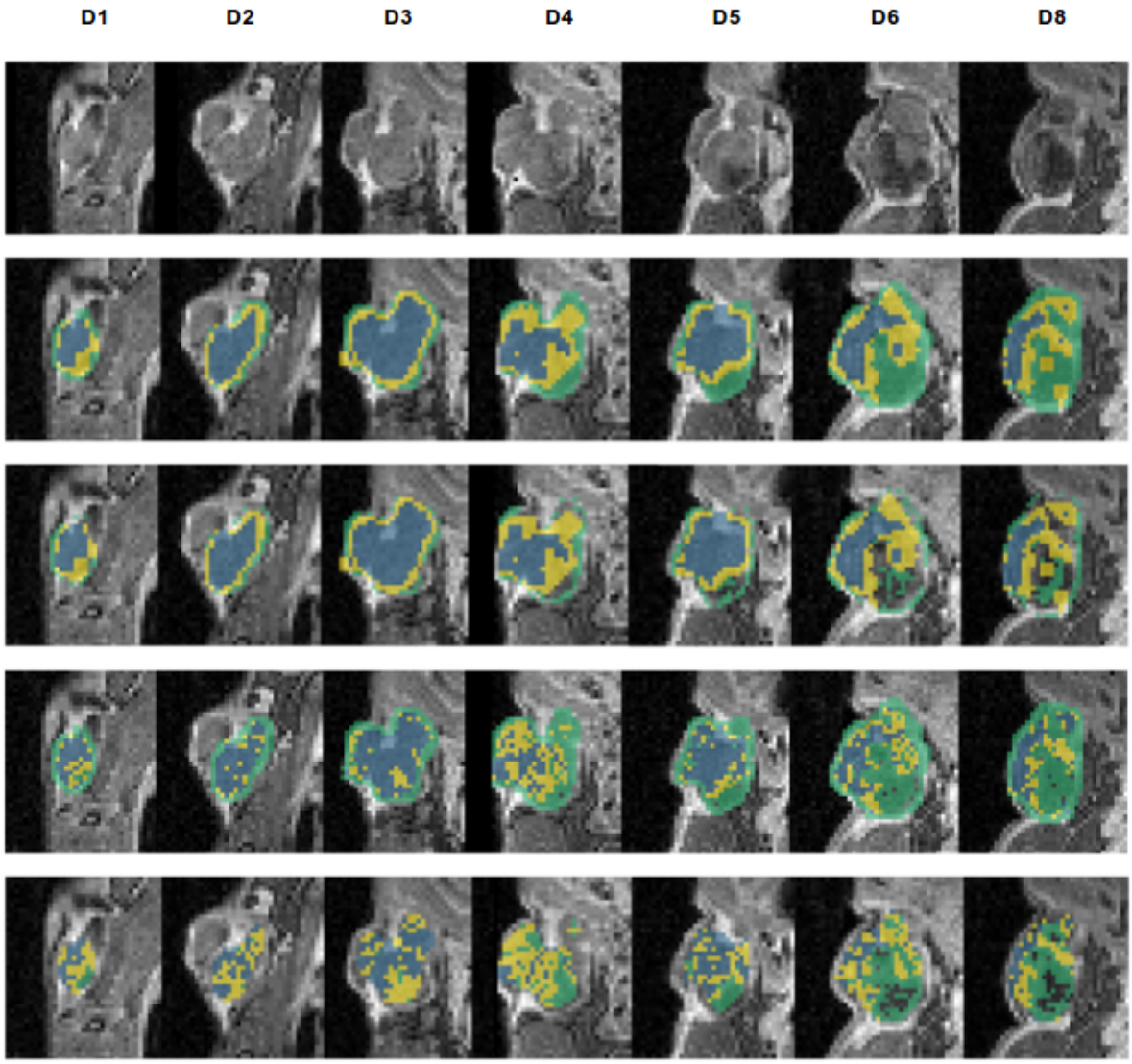


Figure 3.14: Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of feature descriptors in GMM BoV encoding, from the top: Patch, LBP, Gabor, Vox. Colors are arbitrary, but relate across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3).

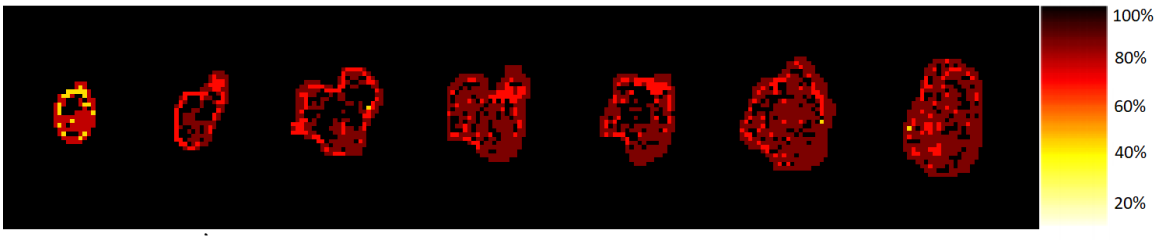


Figure 3.15: Confidence heatmaps for segmentations. High values (black) correspond to high confidence, while low values (yellow) indicate low confidence in segmentations. Most voxels show an agreement of above 50% over different segmentation methods.

folds equal to the number of samples (10), giving 40 samples in total.

Accuracy

The best classification results with KM were achieved by LBP with 75% accuracy. The best classification results for GMM were achieved by the traditional Vox with 82.5% accuracy. The Vox-GMM also scored the best overall in accuracy, followed by LBP-KM. Furthermore it is important to note that some of the misclassifications could also be due to the somewhat arbitrary developmental class assignments caused by a human bias or missing samples, and could thus perform better given a definitive labelling (for example, therapy response data). This could be particularly the case for samples which were consistently misclassified by all the methods.

Accuracy scores alone, taken without the context of underlying segmentations, should not be taken as a definitive evaluation, as underlying characteristics driving the black-box classifier could be clinically meaningless. We therefore perform the final evaluation in conjunction with segmentation scores. Overall the best performing methods (detailed view in Table 3.5) were Patch+GMM (70% Acc., 0.958 DICE), LBP+GMM(70% Acc., 0.963 DICE) and LBP+GMM (75% Acc., 0.723 DICE). Despite its high classification score, Vox+GMM did not obtain good segmentation agreement with other methods of comparable accuracy.

	Vox	**Filters**	**Patch**	**LBP**
	Acc. | DICE	Acc. | DICE	Acc. | DICE	Acc. | DICE
KM	65% | 0.599	70% | 0.604	70% | 0.715	**75%** | **0.723**
GMM	**82.5%** | 0.516	70% | 0.754	**70%** | **0.958**	**70%** | **0.963**

Table 3.5: Classification accuracy scores and confidence DICE measures for various feature descriptors and codebook generators, all sharing the same BoV orderless normalised frequency encoding.

The framework obtained sufficiently good scores to be considered suitable for the task; mainly it was capable of visualising a change in subregional distributions that was descriptive enough for successful classification. It should be noted that while Vox-GMM method performed unexpectedly well, a closer inspection reveals highly non-contiguous segmentations that are difficult to interpret (Fig. 3.14). Furthermore, some more necrotic samples in the later growth stages were segmented entirely as background, which has proved sufficient for discrimination with a classifier, but is not in general meaningful nor useful and could result in overfitting. Segmentations are considered an important aid to understanding the change in distributions and thus such low contiguity and visual coherence renders the method less useful. Furthermore, the lack of visual consistency in segmentations could also potentially reflect on the fallacy of a black-box machine learning classifier. This highlights the importance of BoV pipelines, which through their transparency could benefit the field by aiding clinicians and doctors with additional validation for machine learning techniques potentially bridging the translational gap with respect to trust-related issues [1, 5].

3.3.6 Parameter evaluation

We evaluate the influence of parameters on the overall classification accuracy. While we are restricted to a scale of size $n = 3$, we manipulate the number of clusters k.

In addition we explore two variants of codebook generation: 1) as described in the original [85] paper, learning the visual words on per-class basis to allow for all possible variations of the perfusion patterns, and 2) learning the visual words globally. We keep the method constant and use the Patch+KM variant of the framework.

Segmentations produced by varying the k parameter in the same sample of a growing tumour can be seen in Fig. 3.16. An increase in k produces increasingly non-contiguous segmentations, and with higher values of k the segmentations became unintelligible. The global variant of codebook generation also produces less visually coherent segmentations than a non-global alternative. This is mainly due to the fact that the generated words did not correspond to all possible tissue perfusion pattern types within each class, and thus were unable to relate the same tissue variants across the time points (Fig. 3.16 row 4 column 4).

Although higher ks produced increasingly unintelligible segmentations and are therefore less useful for a human, the most visually coherent results are not necessarily the best distribution to be used by a machine learning classifier. We therefore further evaluate the influence of the k parameter on classification accuracy. The results of varying k parameter across all the samples with LOOCV ran on a per-subject basis (4 samples per fold) and with two classes are presented in Table 3.6. Quantitative results show that the global variant of codebook generation with $k = 8$ performed the best, followed by $k = 6$. However as shown in Fig. 3.16, the resultant segmentations were not relatable across all classes and time points. The best result with visually coherent segmentations was generated with a non-global variant of $k = 4$.

While a non-global variant of BoV encoding was capable of generating the most visually coherent segmentations, it is the global variant that produced the best classification accuracy. This highlights an important design decision to be made when using the framework. While global codebooks increase the classification accuracy, spatially uncorrelated labels are not as clinically meaningful or useful, while at the

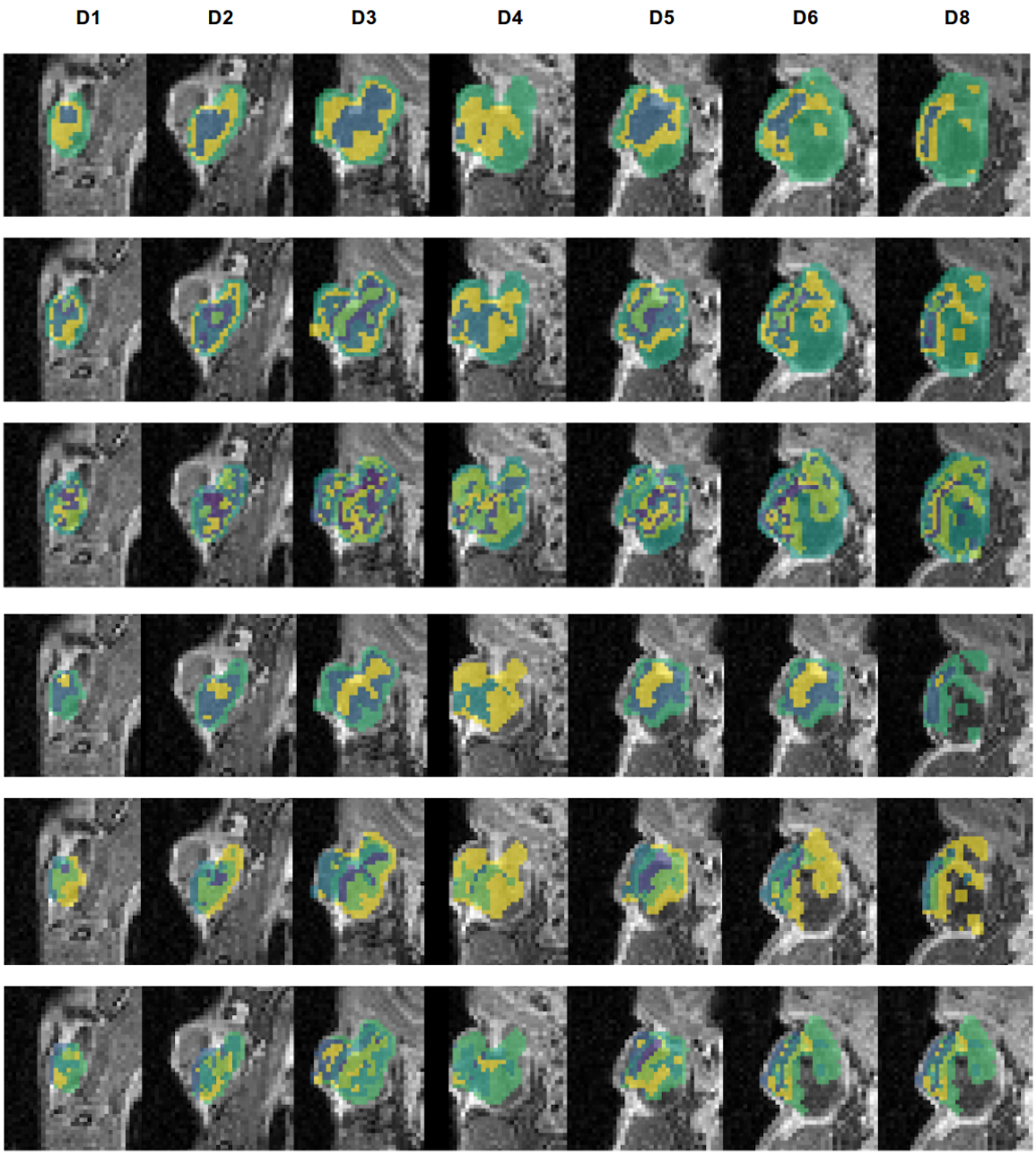


Figure 3.16: The effect of varying the k parameter (number of clusters) on the subregional distributions of the same sample. The rows correspond to $k = \{2 \times 2, 2 \times 3, 2 \times 4, 4, 6, 8\}$ respectively with the upper three generated with non-global and lower three with global variant of BoV encodings. Columns correspond to the days of growth.

	k=4	**k=6**	**k=8**
non-global $(k/2)$	70%	55%	60%
global	67.5%	77.5%	**82.5%**

Table 3.6: The effect of varying parameter k (number of clusters) on classification accuracy. Top row illustrates the results from non-global variant of BoV encoding, while the bottom row reflets on the global variant results. The columns refer to different value of k parameter.

same time meaningful segmentations are not necessarily the most descriptive for a machine classifier and higher accuracies can be achieved by increasing k globally.

3.3.7 Subregional validation

In the previous subsections we have shown that the proposed framework is capable of both qualitatively and quantitatively assessing the growth-induced change in tumour heterogeneity. However, as the datasets we evaluated on lack ground truth data, we cannot definitively attribute those changes to physiological features of the macro scale arising from tumour microenvironment.

Due to the lack of availability of similar DCE-MRI data containing subregional annotations, we attempt to assess some level of segmentation accuracy via an alternative publicly available data set from a BRATS 2013 challenge set for MICCAI (braintumorsegmentation.org). It is crucial to note that the dataset consists of T1W and T2W images and is therefore highly suboptimal for validation of physiological DCE-MRI subregions. It is, however, useful in assessing whether the method is capable of describing visually distinct regions, whether relating to physiological or anatomical data. Nevertheless, this is a limitation and a validation with a more appropriate data should be considered for future work.

The dataset consists of 20 pre-therapy scans of high-grade glioma and low-grade glioma in addition to further synthetic data. To avoid biasing the classifier we evaluate

the framework on high grade gliomas only. The manual segmentations were derived by 4 different observers from T1W and T2W images and details on scanning protocol parameters can be found in [114]. The labels were related to four different types of tissue: necrosis, edema, non-enhancing and enhancing tumour. In this evaluation we have excluded the surrounding areas of edema and only focused on the tumour area as in our original evaluation, therefore limiting the number of tissues to three: non-enhancing tumour, enhancing tumour, necrosis, as well as background.

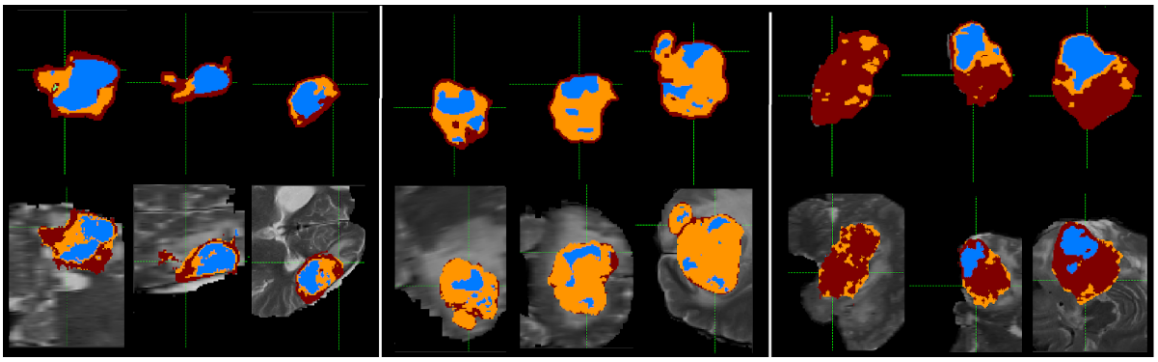


Figure 3.17: Examples of similarities between the Patch-KM variant of the framework (top row) and manually labelled ground truth (GT) data from a single observer (bottom row) for three different samples. All three axes are shown for each of the samples and they correspond across the rows. Colors are arbitrary and relate to different visual words representing diffent tissue properties: enhancing tumour, non-enhancing tumour and necrosis.

It is important to keep in mind that we cannot expect the method to perform flawlessly as the dataset is not characterised by an inherent heterogeneity, and it is not dynamic data. In other words, the manually segmented tissues could be, but are not necessarily, characterised by a varying degree of heterogeneity or variation present, and any such heterogeneity is not related to dynamic perfusion information. For some of the samples, the framework appears to be capable of picking out tissues with varying degrees of heterogeneity that align well with the manual tissue segmentations, however, in some cases it classifies the whole sample with a single, homogeneous or heterogeneous, label. We therefore decided not to use DICE to score the segmenta-

tions as over the whole dataset the metric is not useful for explaining the relation of segmentations to the degree of heterogeneity. It is rather indicatory of the relation to the types of the tissue present in high-grade gliomas, which do not necessarily differ based on the degree of heterogeneity present. Instead, through visual comparisons with manually acquired ground-truth data we show for a subset of the samples that the segmentations (and thus global representations) derived over macro-scale tumour appearance could be related to physiological properties of the underlying tissue (Fig. 3.17). We also show some examples of cases where the method has failed to align with the ground truth (Fig. 3.18) and argue that this could be related to the fact that in such cases the manually segmented tissues could not be represented by a varying degree of variation or heterogeneity.

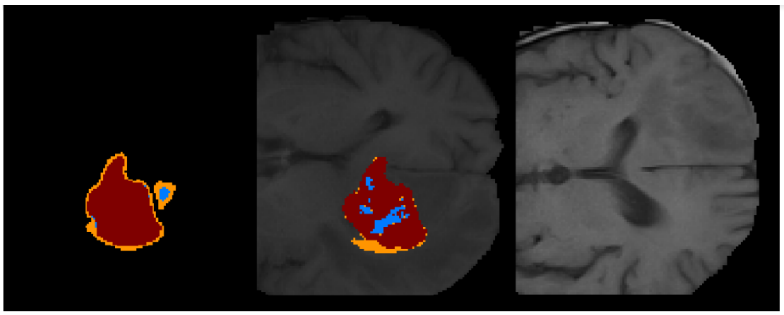


Figure 3.18: Examples of failed segmentations (left), corresponding GT segmentations (middle) and corresponding T1W signal. In our method, a majorty of voxels were assigned a single label (maroon) corresponding to a non-enhancing tissue in the ground-truth (and thus likely a homogeneous one). T1W signal for the same area reveals mostly homogeneous tissue, which suggests that in this case the ground truth labels were not characterised with a degree of heterogeneity.

3.3.8 Similarity and variation descriptors

One of the main drawbacks associated with shallow image representations is that the feature descriptors are human designed to fit the domain of application, and are not learned from the data. For that reason, such features are often called 'handcrafted'. This introduces many assumptions in the design of the descriptors that might not

necessarily be true or reflective of the data they are applied to. For that reason, when evaluating the shallow representations, it is of crucial importance to raise the question of the validity of chosen methods for the application domain. One way to evaluate this is to assess the performance of the methods on the particular problem, as we have done in the previous sections. However, in some cases this could be insufficient, particularly when using black-box machine learning classifiers. We therefore make an attempt to assess the influence the choice of segmentation, along with its assumptions, has on the final results of the analysis.

Different segmentation methods carry certain assumptions about the composition and distribution of data. The choice of feature descriptor, whether in segmentation or classification, can therefore highly influence the analysis outcome. In this work, we have focused on the variation-based feature descriptors assuming such variation, or heterogeneity, to be a descriptive characteristic of tumour progression. To challenge this assumption we compare our results to those from supervoxels (SV) - a feature descriptor based on the assumption of local similarity.

As in the previous sections, we evaluate the experiment with the Patch+KM variant of the framework. For similarity-based features, instead of local neighbourhoods, mSLIC [115] is used on the parameter maps to group similarly perfused areas (supervoxels) and mean values from supervoxels are used as features. In order to retain the valuable variation information we could not average over the local neighbourhoods in the Patch+KM method, however, it is possible in the case of similarity-based supervoxels, which highlights an important computational advantage of the SV method. The rest of the pipeline remains constant with the KM codebook generation step. In order to evaluate the method we introduce the contiguity metric defined in [116], as a proportion of 26 neighbouring voxels with a different label to a central voxel. Formally given as:

$$c_i = \frac{\left|\left\{j \middle| l_j \neq l_i \wedge j \in neigh(i)\right\}\right|}{26} \tag{3.20}$$

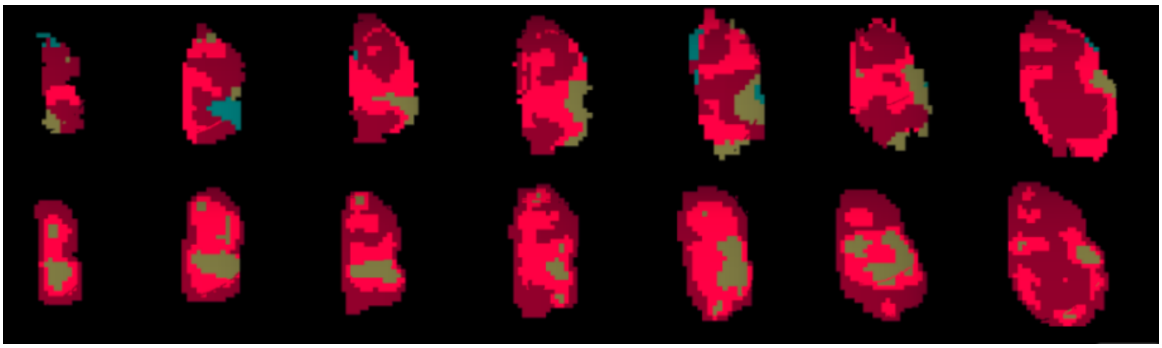


Figure 3.19: Comparison of different subregional segmentations on the same sample of a progressing tumour (columns relate to days of progression). First row represents visual words learned with supervoxels (local similarities). Second row relates to visual words learned from Patch (local variation).

where i is a voxel index, $neigh$ is a set of voxel indices from the 26-point neighbourhood of i and l_i, l_j denote labels of voxels with indices i and j. The average measure for the whole image is taken ranging between 0 (contiguous) and 1 (non-contiguous).

The average contiguity measure for the whole cohort was 0.39 for the Patch+KM and 0.40 for SV (as reported in [116]). The segmentation results are illustrated in Fig. 3.19 for both methods over the same sample of a growing tumour. Both the contiguity scores as well as visual evaluation reveal that both types of descriptors were capable of visualising similarly contiguous and visually coherent segmentations. Interestingly though, while there are visible differences across the segmentations, there are also some similarities, indicating some fundamental tumour characteristics independent of texture descriptors. However, as the distributions of labels differ between the methods, it is important to note that different type of tumours/tumour models may benefit from different descriptors.

3.4 Conclusions

In the previous section we have shown that the introduced framework was capable of visualising a longitudinal tumour progression with a change in subregional dis-

tributions. We have used a small trial dataset for visual assessment, however it is important to note that the observed change in distribution could also be due to the small dose of radiation delivered to mice during a brief CT scan that could have had a therapeutic effect. Furthermore, we have shown that no change was present in the non-longitudinal data, thus rejecting the possibility of growth-induced scaling affecting the local perfusion patterns, however it is important to highlight that the different tumour type could also be a factor. The initial experimental results were a preliminary confirmation of our intuition about the influence of change in microvasculature and perfusion on the macro-appearance of the tumour. At such low resolutions and small scales it is crucial to mention the possible influence of the noise, however as it was kept constant across all the images it is unlikely to be a major factor. We have further tested the method on a larger longitudinal study to understand the changes in a standard growth process as unaffected by therapy, but also to support statistical testing and machine learning. The subregional distributions as visually observed in the segmentations follow a similar, but an opposite trend to those from the trial study. However, this could be due to the fact that the second longitudinal study was acquired for much longer (twice as long) than the first one, and thus the second study goes beyond the growth representations from the first study. Furthermore, it is important to note that the use of the same tumour model (CaNT) could also be a factor behind the observed similarities. In addition, we have shown that the changes, as represented by the framework, were descriptive enough to characterise the longitudinal growth of a tumour automatically with a trained classifier. We have tested a number of feature descriptors and visual codebook generators in the Bag of Visual Words encodings as measured by visual coherence, confidence-based DICE scores and classification accuracy together, as the metrics on their own should not be considered sufficient for reasons outlined in the previous section. The Patch and LBP variants of the encodings, particularly in conjunction with the GMM variant

of the codebook generation provided the best representations to characterise tumour growth, and were also the most visually coherent. The popular variant of Vox with GMM-based encoding produced surprisingly good classification results, however upon a closer inspection, some later stage segmentations proved to be classified entirely as background, which although descriptive enough for a machine learning classifier, is not useful or meaningful clinically. Furthermore, some variants of the method, in particular LBP and Patch in both encodings together with Filters in GMM, achieved a high agreement between segmentations. Despite the lack of ground truth validation data, we have also made an attempt to relate the subregional patterns generated by the framework (and thus the global representations) to the characteristics of underlying tissue indirectly via a proxy publicly available dataset with a ground-truth. As the data was non-dynamic and not related to perfusion heterogeneity we cannot conclusively evaluate the relation, however we use it as an indication. Through this proxy we have shown that the generated results can be physiologically meaningful and for a subset of the dataset the generated distributions reasonably aligned with manually segmented tissues. Finally, we have made an attempt to evaluate the influence of the assumptions that different feature representations have on the final outcome of the analysis. We have shown that although different assumptions (e.g. similarity vs. variation) result in some segmentation differences, interestingly there are also some fundamental similarities existing in the tumour irrelevant of the feature design. However, for accurate analysis, this highlights the importance of choosing the right representation for the right tumour type, e.g. homogeneous or heterogeneous, when using shallow image representations. Some tumour types could therefore benefit more from certain descriptors.

In conclusion, the framework obtained sufficiently good scores to be considered suitable for the task. Mainly it was capable of visualising changes in subregional distributions that were descriptive enough for successful classification with visually

coherent segmentations, particularly when using Patch and LBP descriptors. However, as demonstrated in Subsection 3.3.8 of this chapter, shallow representations have a tendency to introduce priors associated with the design of the descriptors. Furthermore, deep image representations have recently largely outperformed the shallow representations increasing the classification accuracy to that of near human level [117]. The next chapter will therefore evaluate deep image representations for the task of tumour perfusion heterogeneity assessment in DCE-MRI.

Chapter 4

Deep image representations for segmentation and classification of tumour perfusion heterogeneity in DCE-MRI

Growing tumours are characterised by an increase in perfusion heterogeneity, that also changes following therapy [3, 6, 7, 8]. In this work we would like to leverage that information to understand the behaviour of tumours during the growth process, which could be in turn applied to understanding their therapy response. To this end, we search for automatic methods that are capable of characterising a change in heterogeneity, or variation, present in the DCE-MRI images of tumours that could be a reflection of underlying physiological processes in the tumour microenvironment.

In the previous chapter we have introduced a BoV pipeline for the problem of assessing tumour perfusion heterogeneity from DCE-MRI images. We have shown both qualitatively and quantitatively that the framework is capable of both discriminating and meaningfully visualising tumour subregions based on perfusion patterns

as captured with DCE-MRI. We have further shown that such subregions have the potential to relate to physiological processes in the tumour tissue, where those tissues are characterised by varying degrees of heterogeneity. We have observed a change in the distributions of the subregions (as visualised through segmentations and histograms), and the change was found to be significant enough to discriminate between different growth stages. The findings confirm our initial intuition about the influence of change in microvasculature of the tumour that can be detected at the resolution used in the DCE-MRI scans, that agrees with the available knowledge of the processes in the tumour microenvironment.

While the framework has shown sufficient potential in its applications to the problem, it is important to note that shallow representations have several drawbacks. They introduce human designed, or 'handcrafted', features that incorporate prior assumptions, which might not be truly reflective of the data, and thus could introduce a human bias [118]. They are also relatively simple as a two level structure of local and global information, although sometimes aided with multi-scale features and pooling. Deep image representations have recently largely outperformed the shallow representations, increasing the performance accuracy to that of near human ability [117].

In this chapter we leverage the recent advancements in deep texture representations and extend them towards the task of segmentation with no ground truth and from limited training data. The novelty of this chapter is therefore twofold: 1) inspired by advances in encoder-decoder networks [119], we extend recent advances in deep texture representations for the task of voxel-wise segmentation; 2) inspired by [120] we further combine deep and shallow learning architectures in our representations for optimally lightweight image representations that require little or no training. We examine whether such architectures could be applied to the problem of assessing tumour perfusion heterogeneity with no segmentation ground truth at all and very limited training data. We show that even in a very simple form, these methods are capable

of matching the performance of those from the previous chapter that used shallow representations. Furthermore, we back this claim by demonstrating their ability to produce visually coherent and contiguous segmentations that quantitatively match those from the previous chapter.

This chapter is also an analog to the previous one on shallow representations. While we are using deep image representations for segmentation and classification of the tumours instead of the shallow ones, the structure and experiments remains similar for a fair comparison.

The key novel contributions of this chapter are as follows:

- Adapting deep representations to a introduce novel lightweight deep pipeline capable of: a) performing classification with little training data; and b) generating segmentations with no ground truth.

4.1 Deep image representations

Until 2012, Bag of Visual Words approaches described in the previous chapter, with a number of competing feature descriptors and encodings [95], were firmly considered the state-of-the-art in many areas of computer vision. Since then, the research in the field has been largely dominated by deep learning methods, which sparked much enthusiasm in the public imagination. Papers flourished and many types of architectures emerged, notably including Convolutional Neural Networks as originally described by AlexNet and variations of such, but also Recurrent Neural Networks, including Long Short Term Memory (LSTM), encoder-decoder architectures, autoencoders, Generative Adversarial Networks (GAN) and more. All such networks are rooted in a chain of inventions, starting with a definition of a perceptron by Rosenblatt [121], later enforced by the development of back-propagation by Hinton [122], leading to the introduction of convolutional networks by LeCun [123]. Finally, more

recent algorithmic [124], architectural and computational improvements along with the appearance of large public benchmark datasets [117] led to the modern deep architectures, such as demonstrated by Krizhevsky with AlexNet [118]. With many layers of non-linear feature extractors followed by a number of fully-connected layers, such architectures are therefore considered 'deep', as opposed to previously explored 'shallow' pipelines [125].

Such architectures are substantially more intricate and deeper than any previously explored methods, and thus they are able to represent increasingly abstract information about the image and model complex relationships beyond those that are possible in shallow processing pipelines and traditional machine learning methods [126]. Such levels of abstraction can be further extracted and exploited even in other domains of application. At the same time, such architectures provide the advantage of a large number of parameters learned directly from the data and thus decrease the likelihood of introducing prior bias, as in the case of handcrafted features in shallow representations.

Deep architectures are comprised of a number of layers, or building blocks, organised in a graph-like structure that differs between the architectures. To list some of the examples, RNNs introduce a recursive structure, while GANs, typically, run two sequential nets in parallel. CNNs are one example of simpler sequential architectures that we use frequently in this work. While the arrangement of the blocks, or layers, differs between architectures, the structure of the blocks often bears similarities across the methods. CNNs demonstrate two frequently occurring types of layers: convolutional and fully-connected. Convolutional layers are composed of a number of convolutions, followed by activation function (e.g. ReLU) and pooling (or un-pooling) operations. With the exception of learned filters, this is in many ways similar to the the process in standard BoV pipelines where filter descriptors were used and many researchers took note of the analog [17, 120, 18]. In fact, in [120] the

authors noted that CNN architectures can be seen in the light of the BoV pipelines, where convolutional layers correspond to the feature extraction stages in shallow representations of increased complexity (achieved by depth), while fully-connected layers serve as encodings. With this hindsight, CNNs have been widely used as a feature extractor, and in [120] the authors even explored mix-and-match between the deep and shallow representations.

In the recent years deep architectures significantly outperformed shallow image representations on the task of image classification and segmentation [118, 127] on challenging benchmarks such as ImageNet [117], often also outperforming human experts. A similar trend has been observed in the field of medical image analysis [128], where the amount of information available in 3D medical images is abundant to the point that it would be difficult for a human use the entirety of it for evaluation, however it is easily exploited by the automatic nature of deep networks [129].

In our domain of application we are facing three major obstacles limiting the potential exploit of advantages arising from deep learning. Firstly, as outlined in the previous chapters, we would like to be able to explicitly account for the variation reflected in the images in order to truthfully describe the apparent macro-scale heterogeneity arising from physiological processes in the tumours, which is difficult to achieve with the black-box nature of deep networks. Secondly, while training a deep network typically requires large quantities of imaging data, the lack of training data is a known problem in the biomedical image analysis field. Finally, in the case of voxel-wise segmentations, deep learning typically requires fully or at least sparsely annotated images (training labels). However, in biomedical imaging acquiring manual annotations is not only laborious, but also often ambiguous. The annotations are usually done on a per-slice basis, which is inefficient as neighbouring information is repetitive and could be leveraged for automatic annotation [19]. Furthermore, opinions between the experts could be contradictory and non-conclusive [129].

In the previous chapter we explored texture representations to explicitly account for the spatial variation as a measure representative of tumour heterogeneity in perfusion. In this chapter, by analog, in order to address the first of the outlined issues we investigate the recent efforts in the design of deep representations particularly suitable for the task of texture analysis [17]. Secondly, to address the problem of lack of training data, we focus on adapting only such deep representations that require minimal or no training as in the case of domain transfer [120] or handcrafted networks [130, 131]. Furthermore, inspired by the recent advancements in encoder-decoder architectures [119] we extend those representations for the task of segmentation. Finally, to address the lack of segmentation ground truth, inspired by [120], we combine those architectures with shallow image encodings. Due to their unsupervised nature we achieve complete pipelines with no requirement for segmentation ground truth or much training data.

In the remainder of this section we briefly describe the motivation for this work and background in the field of deep image representations. We outline the notable advancements that are most relevant and influential to this work. Section 4.2 describes the methods used in this work, including the novel contributions. We further outline the results of our experiments in Sec. 4.3, evaluating the suitability of the framework to the task of assessing tumour perfusion heterogeneity. Finally, we conclude the chapter in Sec. 4.4 with a discussion.

4.1.1 Deep texture representations for image classification

While deep networks have demonstrated astonishing performance improvements on challenging datasets, training deep networks amounts to tuning millions of parameters on many GPUs and requires large numbers of labelled images [17], which is an issue in the field of biomedical image analysis. Furthermore, most of the deep architectures largely focus on representing object-level of abstraction in the images (e.g.

faces), while such complex representation might not be necessary in the context of tumours. In this work, therefore, while we would like to leverage the advantages of deep representations, we would like to do so with two constraints in mind. Firstly, we will look for deep representations that are particularly suitable for describing variation, or texture. We will show that for our purpose such architectures do not necessarily require as much depth as is required for object recognition. Secondly, we will only adapt representations that are minimally demanding of training data.

Automatic recognition, segmentation and classification of textures has traditionally been performed with shallow BoV pipelines, which were described in the previous chapter. With the deep learning revolution, naturally, attempts have been made to adapt such methods to perform optimally in texture-like domains. A recent detailed review on texture representations for image classification outlines three distinct approaches to the problem present in the current literature: 1) fine-tuned networks, 2) transfer learning approaches, and 3) hand-crafted networks. In addition, naturally, there are end-to-end trained models, however, they require substantial amounts of training data. Furthermore, it must be noted that there already is a considerable amount of effort in biomedical imaging devoted to the design of such optimally trainable architectures suitable to limited amounts of data. In fact, some architectures were so generally optimal to be applicable beyond the context of biomedical imaging, as it was in the case of a UNet [19]. For those reasons, end-to-end trainable models are not of major interest to this work.

4.1.1.1 Transfer learning

The impressive performance of deep learning has been largely attributed over the years to the development of large labelled datasets as well as an increasing computing power. However, both of these factors, and especially the prior one concerning large labelled data, are relatively difficult to obtain in certain domains of application. Typical

problems in biomedical imaging inherently suffer from a lack of training data, which limits their ability to utilise the advantages arising from deep learning. Indeed, the true potential of deep learning was only realised after the introduction of large labelled datasets such as ImageNet [117, 118]. This introduces questions of how much labelled information is required to truthfully and generally represent the problem, and whether increasing the manual supervision will yield ever increasing performance, as raised by the authors in "Have I seen enough?" paper [132].

It has been shown [120] that the knowledge acquired during training a model on one problem is reusable and could be successfully transferred to another problem, astonishingly even if their domains are different (e.g. cars vs. tracks). This has been attributed to the impressive ability of deep networks to represent increasingly abstract information about the world that is general and not problem-specific. The process is called transfer learning, also frequently referred to as domain adaptation. The success of transfer learning is naturally dependent on providing source data that is varied enough to truthfully (or closely) describe and represent the world, and thus the filters learned from the source data should be general enough to represent the target problem. This raises the question of the suitable similarity of the source to the target data. Large complex datasets with many categories and large inter-class variance (e.g. occlusions, varied lighting conditions) proved to generalise well to many different domains, as is in the case of ImageNet with 1000 classes and 1.2 million images. Utilizing similarities of a smaller dataset [17, 133], whose inter-class variance and local appearance resembles that of the target domain could also generalise well. It was also found that deeper models tend to generalise better due to their ability to represent more abstract information [17, 120].

One of the problems associated with transferring the general knowledge learned with a deep model is that the final dense classification layers are tailored to the source domain (e.g. the final classification layer usually contains the number of channels

equal to the number of labels for the dataset). Two common ways of exploiting the abstract information encoded in the deep layers of the network are: 1) using convolutional layers as a feature extractor; and 2) strategies involving fine-tuning. Both are covered in the following sections.

4.1.1.2 Deep filter banks

Many researchers observed the similarity between deep and shallow representations [18, 17, 120], with the difference of convolutional filters learned directly from the data, rather than handcrafted, in a hierarchical architecture of feature extractors that enabled learning more complex and abstract representations. In CNNs, convolutional layers are often compared to the feature extraction step in shallow pipelines, while fully connected layers are seen as encodings [17, 120].

A number of evaluations exist examining the effect of transfer from CNNs towards texture domains [120, 134], demonstrating that transferability from fully connected layers (regarded as encoding and classification step) is poorer than from convolutional layers. Consequently, convolutional layers should carry less domain-specific information [17], which motivates their use as a feature extractor for further shallow processing in the same sense as filter banks are used in shallow representations [120, 135]. In contrast, the output of fully connected layers could also be utilised in this sense [125, 120], however convolutional features were shown to be more robust to transformations [17]. In [120] the authors explored the mix-and-match between encoding and feature extractor components of deep and shallow pipelines and found that Fisher Vector encoding jointly with CNN feature extractors proved optimal for classification of textures (with ImageNet domain transfer).

Inspired by the outlined research, in this work we adopt the approach of [120], however in order to improve the performance we perform the knowledge transfer from a similar radiological domain. In addition, for a fair comparison with other methods

from the previous chapter, we replace the FV encoding with the BoV encoding we used previously. One of the limitations of the approaches outlined in [17, 120] is that they were designed specifically for the problem of classification. In the following Section 4.1.2 we outline the extension of these representations to the task of segmentation.

4.1.1.3 Fine-tuning

Another way to exploit the advantages of deep learning without the need for heavy training is to use fine tuning during domain adaptation. Fine tuning a network tailors it to the task-specific domain, and thus provides an advantage that the whole network would be utilised. During the process, the final fully connected layer is modified and trained to the number of channels required by the new domain. Furthermore, following fine-tuning, the low-level information encoded in the first layers also adapts to better represent that of the target domain [136, 17]. The process has been shown to considerably improve the performance over the pre-trained methods [125]. Furthermore, the final global representation is generated in hierarchical end-to-end manner, which is advantageous over adding an extermal encoding stage as is the case for transfer learning used as a feature extractor. There are many approaches to fine-tuning, including re-training entire models or freezing convolutional layers to list a few, and the choice of approach depends on the size and similarity of source and target datasets. In addition to improved performance, fine-tuning has a number of advantages as outlined in the following paragraphs. It is worth noting that, in a similar way to end-to-end trained models, it is difficult to evaluate all of the advantages arising from fine tuning for texture domains due to the lack of datasets of the same scale as ImageNet [17].

With fine-tuning, the resulting representations are generated with both local and global information encoded directly by the network and so do not require further encodings or classifiers, which is a considerable advantage over shallow representa-

tions and domain transfer with no fine-tuning. Furthermore, such information arising from the final convolutional layers is aggregated as a spatially sensitive feature vector followed by order sensitive pooling [17], which provides the advantage of adding some spatial context as opposed to orderless shallow encodings. However, it has been shown that to represent textures it might not be necessary to encode as much spatial context as is required for object-level representations (e.g. faces) [17, 18, 120, 137] as we search repeated patterns of low complexity [18]. Consequently, orderless encodings, such as FV-CNN [120, 17], have been shown to perform well in texture domains. This has also motivated a number of researchers to design architectures that incorporate elements of orderless encodings either by replacing the fully-connected layers with orderless pooling layers (e.g. bilinear pooling) [138], or integrate the traditional orderless encodings (e.g. VLAD) into the layers of the network to enable end-to-end architecture [139]. They were, however, shown either to be computationally expensive, or unable to outperform the simple FV-CNN-type approaches [17].

While the initial layers of the network are understood to encode the most general geometric features (e.g. blobs and corners) that could be transferable across the domains, these are also likely to not transfer well without fine-tuning if the source and the target carry major differences in the low-level fine information. In [136] the authors note that when visualising the gradients during fine-tuning of FCN [119] between domains it becomes apparent that the first layer gradients receive the most adjustment. It is a consequence of the fact that the first layer encodes the low-level geometric information that is, in that case, dataset-dependent. However, arguably, given a domain transfer from a dataset bearing maximum similarity in the low-level information to its target domain (e.g. two radiological datasets), the advantages arising from fine-tuning for those layers could be partially alleviated. Indeed, the authors in [18] observed that fine-tuning during a transfer learning between two texture domains achieved better accuracy than fine-tuning transfer between object-specific

(e.g. ImageNet) and texture-specific domains of the same size. By analogy, we argue that transfer learning without fine-tuning from a similar domain will provide superior results to the generic pre-trained ImageNet transfer to a different domain.

Despite its proven advantages, it is also worth noting that fine-tuning still requires a considerable effort in re-training the network, which in the case of segmentations would require ground-truth data. In this work, motivated by the reasons outlined in the previous paragraphs, we therefore propose a number of compromises to alleviate the necessity for fine-tuning and labelled data, while still retaining some of the advantages of the process. Firstly, while fine-tuning provides the advantage of adding global and possibly spatial context to the resulting classification/segmentation that is generated in an end-to-end manner, it has been shown that it is not necessarily an advantage when it comes to representing textures. We therefore argue that, although more spatially complex than images of textures, biomedical images might not require as much spatial context as objects, for example faces, and thus orderless encodings could be sufficient for such representations. Secondly, while the fine-tuning during transfer learning provides the advantage of adapting the initial low-level dataset-specific information contained in the first layers, we believe that it could be imitated by performing a transfer from a larger dataset in the same domain as the target data. Finally, while fine-tuning enables re-training the final classification layers and thus adapting it to the required number of channels (labels), using shallow pooling and encodings eliminates the need for it, while also eliminating the need for segmentation ground-truth labels. In [125] the authors reported that even without fine-tuning they already observed an accuracy increase of 10%.

4.1.1.4 Handcrafted networks

The convolutional layers in the standard CNN are typically expressed with a combination of convolutions, activation functions and pooling operators. As mentioned

before, this is similar to the processing steps of shallow representations, with the exception of trainable filters. Decades of research in the texture domain for optimal feature descriptors yielded a number of highly accurate descriptors particularly useful for describing textures. Naturally a question arises, whether utilizing such features in the deep representations would bring improvement in accuracy over certain domains for which the features were designed. It is important to note, that trainable filters are not the only advantage of deep representations. Notably, the networks are able to generalise so well due to their ability to represent increasingly abstract and complex concepts beyond anything that classical machine learning methods were capable of, as a consequence of their hierarchical deep structure [126]. In this sense, replacing learnable filters with carefully designed technical features could potentially improve the performance for specialist datasets that are not large enough to enable training or fine-tuning, even if only as a consequence of the depth alone, but also potentially due to the specialised filters.

Indeed, researchers have explored the idea, although not much research has been devoted to the topic. Notably, Mallat's influential ScatNets [131], composed of a hierarchical cascade of invariant wavelets followed by modulus non-linearities and average pooling, which has a structure that is very similar to the standard CNN convolution block, only with a fixed filter and a different activation. Notably, such networks do not require any training to yield good performance. The authors suggested that it is not optimal to expand the networks to such depths as are typically seen in deep learning, and suggested a depth of two convolutional layers. In [130] the authors proposed a similar concept, however based on much simpler trainable PCA-based filters, and used LBP-based histogram encoding. The authors further proposed two variants of the network with the same topology, using different filter weights: RandNet and LDANet. In RandNet the filters were initialised to Random Projection matrices as in [100], while in LDANet the filters were learned from Linear Discriminant Analysis

(LDA). In [17] the authors found that the performance of PCANet and its variants is poorer than that of ScatNet, however at the same time they provide considerable improvement in computational complexity and allow for training-free representations.

4.1.2 Encoder-decoder architectures for voxel-wise segmentation

The downside of the approaches described in the previous subsection is that they were designed specifically in the context of whole image classification. However, due to the encoding part of shallow representations in the previous chapter we were able to achieve segmentations as a by-product of the pipeline. We argue that introducing such transparency could prove instrumental in adaptation of such methods into clinical practice, as they could provide some explanation as of the results of the method and inspire confidence in those results. For that reason, we would also like to extend the deep texture representations described in the previous section for the task of segmentation.

One of the main limitations of the deep texture representations outlined earlier is their compact representation. Due to the nature of convolutions with padding and strides along with the pooling operators that we find in CNNs, the output of the convolutional layers is typically a compact representation of the original image, restricted to a feature vector with a number of channels. Such features, while representative of the whole image, have discarded valuable spatial information contained in the original data. In order to restore it, one could simply upsample the image, however this could prove very inaccurate, as the information about the spatial arrangement of pixels was lost during pooling. In order to address this problem we seek inspiration in encoder-decoder architectures, which are widely used in segmentation research.

Voxel-wise multi-class semantic segmentation has been an active field of research for decades, deeply rooted in texture segmentation, most recently accelerated by

the task of autonomous driving. Recently, as in any other image processing field, it has been largely driven by the advances in deep learning neural networks, with encoder-decoder architectures at the forefront of the field. The idea behind encoder-decoder architectures is to supplement a convolutional section of the network, referred to as encoder, with a corresponding decoder architecture, where pooling operations are replaced with upsampling operations. This restores the original resolution of the image. However, the issue with upsampling, as mentioned earlier, is that as a consequence of pooling in the encoder stage the spatial information is lost. There are many types of architectures that deal with the problem, where a majority of them use skip-connections in order to restore the missing spatial information. They differ in the ways they upsample the images and combine the skip connection information.

4.1.2.1 Skip connections

The idea behind skip connections is to make a link between the encoder and the decoder, that carries over some details about the spatial information lost as a consequence of pooling. Typically, this is done in a symmetric manner. Upsampling operations can be seen in this sense as the opposite of pooling, where one restores, while the other loses spatial information. In order to fully restore the original resolution, the decoder network typically contains the same number of layers as used in the encoder, and thus the connections between pooling and upsampling can be made.

Many different architectures exist, employing different strategies to join the skip connection information. The three most popular architectures include FCN [119], SegNet [140] and UNet [19]. The FCN network was the first of its kind to use both skip connections and the upsampling for pixel-wise classification. The original FCN, in the contrast to other encoder-decoder architectures, retained the fully connected layers, however "convolutionised" them to produce a heatmap of activations, thus the name of Fully Convolutional Network. For the skip connection they employed

a strategy of fusing information from pooling layers in the convolutional part with "convolutionised" predictions using a summing operation. In SegNet the authors exploited the skip connections with "unpooling" guided by information about the spatial location of individual pixels from corresponding symmetric pooling layers. Finally, in UNet, as a skip strategy the authors simply employed a concatenation between the upsampled predictions in the decoder and corresponding high resolution features from the encoder, followed by convolution with a halved number of channels to restore the symmetry.

4.1.2.2 Upsampling

As mentioned in the previous section, in addition to competing strategies employed by skip connections, the architectures also differ in the upsampling approaches. The most straightforward way is to simply upsample images by repeating the local information prior to the convolutions. This is however suboptimal, providing very coarse results. Another way is to employ interpolation, where bilinear interpolation is the most popular choice in decoder architectures. Finally, some architectures use "transpose convolutions" or "strided convolutions", where the information is computed in the same manner as convolutions, however backwards. This is achieved with the use of a convolution matrix with learnable weights.

In FCN the authors employed a single layer of deconvolution as an upsampling strategy (or strided convolution) [119], initializing the trainable weights to bilinear interpolation. In SegNet [140], as mentioned in the previous section, the strategy involved an "unpooling" operation, which upsampled the high frequency features in the decoder stage with their corresponding location in the pooling layer from the encoder stage (otherwise suplemented with zeros). Finally, depending on the implementation of UNet, they typically use trainable up-convolutions with concatenated high resolution information from the encoder stage, followed by a convolution that

halves the number of channels [19] to restore the original channel number.

In this work, we would like to use encoder-decoder architecture to expand deep texture representations for classification, as described in the previous subsection. However, as we have chosen for our problem only the light-weight representations that require no training, we would also like to use light-weight decoders. There are two obstacles limiting the use of decoders in our problem: 1) the upsampling layers require training; and 2) the final pixel-wise classification layer requires domain adaptation. To deal with the first issue, transpose convolutions could be replaced with interpolation with fixed weights. However, in order to obtain better accuracy, we deal with it in the same manner as convolution weights in the encoder stage, which are transferred from a corresponding domain. In order to overcome the second outlined limitation, we introduce shallow encodings as in the case of FV+CNN described in the previous subsection. The unsupervised nature of encoding generation (KMeans/GMM) allows for producing segmentations, where frequency distributions are later used as encoding for whole-image classification.

Although in this work we primarily rely on encoder-decoder type of deep architecture, and specifically on UNet, autoencoders are another approach to the problem that is the closest related to the proposed method. Autoencoders are another family of architectures that unlike encoder-decoder networks, attempt to extract a set of significant features in an unsupervised manner, which replaces the need for transfer learning and handcrafted networks strategies employed in this work. Although traditionally used for compressing and restoring encoded information, autoencoders have also been shown successful in the task of segmentation when aided by a subsequent classifier [141].

To conclude this section, we have described the major innovations in the fields of deep texture representations and semantic segmentations that are the most influential to this work. The next section will outline the architectures and encodings used in

this chapter. Finally, before progressing to the next section, a potential ambiguation should be cleared up between two terms frequently used in this work: encoding and encoder. We refer to encodings in shallow (typically, but also in deep) representations as a way of a compactly representing an image as a feature vector reflecting the similarity of instances with proximity in the feature space, and thus enabling classification. On the other hand, in this work an encoder is in this work understood as a convolutional section of the network that downsamples the image into a compact representation, that is typically followed with a decoder stage in the task of segmentations, or fully connected layers in the task of classification.

4.2 Methods

In this work we attempt to evaluate the suitability of deep image representations for the task of tumour perfusion heterogeneity assessment inherently characterised by increasing heterogeneity and suffering from lack of ground-truth and limited training data. In order to achieve this, we would like to examine deep texture representations, where we consider texture as a proxy measure of such variation. To this end, we evaluate two approaches outlined in the previous sections: 1) transfer learning, and 2) handcrafted architectures, that are both lightweight representations with minimal requirements for training. Furthermore, we adapt both approaches to the task of segmentation with encoder-decoder architectures. Finally, we combine the architectures with shallow encodings as is frequently done in the case of transfer learning. Due to their unsupervised nature, we achieve complete pipelines that allow for segmentations from unlabelled data and classification with minimal constraints on the amount of training data. The novel contribution of this work is therefore threefold: 1) we adapt two deep texture representations suitable for classification from limited training data and expand them to 3D, 2) we extend these methods to the task of segmentation

with no ground-truth data, and 3) we assess their suitability to the problem of assessing tumour perfusion heterogeneity from DCE-MRI. In the following sections we describe the details of the introduced methods. As we use deep representations as feature extractors, the first subsection will outline the two types of architectures used. Following that, we will describe the remainder of the pipeline.

4.2.1 Deep texture feature descriptors

In this work we utilise transfer learning in combination with encoder-decoder architectures. We introduce two types of voxel-wise deep feature extractors, based on transfer learning (T-UNet) and handcrafted representations (Rand-UNet). We also give further theoretical context to motivate the use of Random Projection filters in the latter architecture as in [130].

4.2.1.1 Transfer UNet (T-UNet)

In order to well describe the heterogeneous properties in DCE-MRI images of tumours we combine the advantages of deep texture representations with the advancements in semantic segmentations. The details of the architecture can be seen in Fig. 4.1. In the encoder stage, we utilise the training-less nature of transfer learning as a feature extractor (as in FV+CNN [120]). The architecture of the encoder is the same as that of a UNet, and similar to VGG16 [142] with the exception that we use 2 sets of convolution and ReLU activation followed by max pooling at each depth level (as opposed to 2x2 and 2x3 convolution blocks as in VGG16). The kernel sizes and number of channels are kept constant across the experiments and the same as in [19]. As we are not applying fine-tuning in order to retain the network's ability to generate segmentations with no ground-truth and to improve the performance, the knowledge transfer is applied from a similar domain of the BRATS dataset that consists of brain MRI images. In the decoder stage, in order to restore the lost spatial information, we

utilise UNet-style [19] skip connections between corresponding pooling and upsampling layers in the symmetric architecture, with simple concatenation between them, as in the case of a UNet. In a mirror-like manner to the encoder stage, upsampling and concatenation (instead of pooling) is followed with two sets of convolutions. The upsampling is performed with deconvolutions and weights are obtained during the transfer learning stage as in the encoder. In contrast to the original work we do not train the final post-upsampling convolutional layers or the final pixel-wise classification layer for segmentation. Instead, again as in FV+CNN for texture classification, we use shallow orderless BoV encoding for pixel-wise classification of the abstract and upsampled convolutional features. As typical for BoV, a compact vector frequency representation is used that is suitable for whole-image classification. Also, as in the previous chapter, we explore two variants of encoding generation, resulting in KM+TUNet and GMM+TUNet. Finally, in the same way we replaced the voxel-wise activation layer with traditional shallow encoding over the abstract up-sampled features, we also use a traditional SVM classifier over the whole-image encodings to produce the final supervised classifications. In this sense, we use the network as a feature extractor as in [120], however due to the decoder side, we perform classification on a per-voxel basis as opposed to for the whole image. While the resultant network contains trainable parameters for convolutions, they were all obtained during the transfer learning from the BRATS dataset, resulting in no parameter training and reduced computational complexity.

4.2.1.2 Random UNet (Rand-UNet)

It has been previously shown in the shallow context that using raw patch information instead of convolution outputs from pre-designed filters can significantly improve the performance. Although this is partly attributed to the 'handcrafted' nature of the filters, it is important to note that this could also be a consequence of the fact that raw

patch information is capable of capturing low-frequency information that is filtered out with convolutions. It has been later shown that the use of Random Projections (RP) further improved the performance of Patch descriptors in the texture-specific context [100] due to their ability to filter out noise in the high-dimensional features. RP were later shown to improve the results over Patch descriptors also in the biomedical domain for assessing the degree of variation in mammographic images [143]. This is of particular interest to this work as, similarly, we aim to describe the change in tumour heterogeneity based on the apparent variation reflected in DCE-MRI images. Random Projections, motivated in [100] by the theory of compressed sensing, refer simply to projecting the highly dimensional space into a random low-dimensional subspace. This is done by creating an RP matrix, which is sampled from a zero-mean unit-variance normal distribution. In the shallow representations the resulting feature descriptors are computed simply by a matrix multiplication of the original patch $I \in \mathbb{R}^{pxn}$ with the RP matrix $M \in \mathbb{R}^{nxm}$, where p is the number of pixels, n is the initial dimensionality and m is the target dimensionality. The result is a matrix $Y \in \mathbb{R}^{pxm}$ of the desired dimensionality m.

In the previous section we described and motivated the use of handcrafted architectures that use fixed weights and thus require no training, such as ScatNet, PCANet or RandNet, and were shown to perform well in texture-like domains. As outlined in the previous paragraph, RP previously demonstrated optimal performance in radiological domains on variation-motivated problems, which is why we would like to explore the suitability of RandNet to the problem of assessing perfusion heterogeneity in DCE-MRI images. To that end we introduce Rand-UNet (Fig. 4.1), where we combine handcrafted architecture in the encoder stage with UNet-like decoder and traditional shallow encoding/classifier, very much like in the previous subsection. The main differences between the two approaches are: 1) weights, and 2) depth. Unlike in T-UNet, where we transferred the weights from a different domain, and in contrast

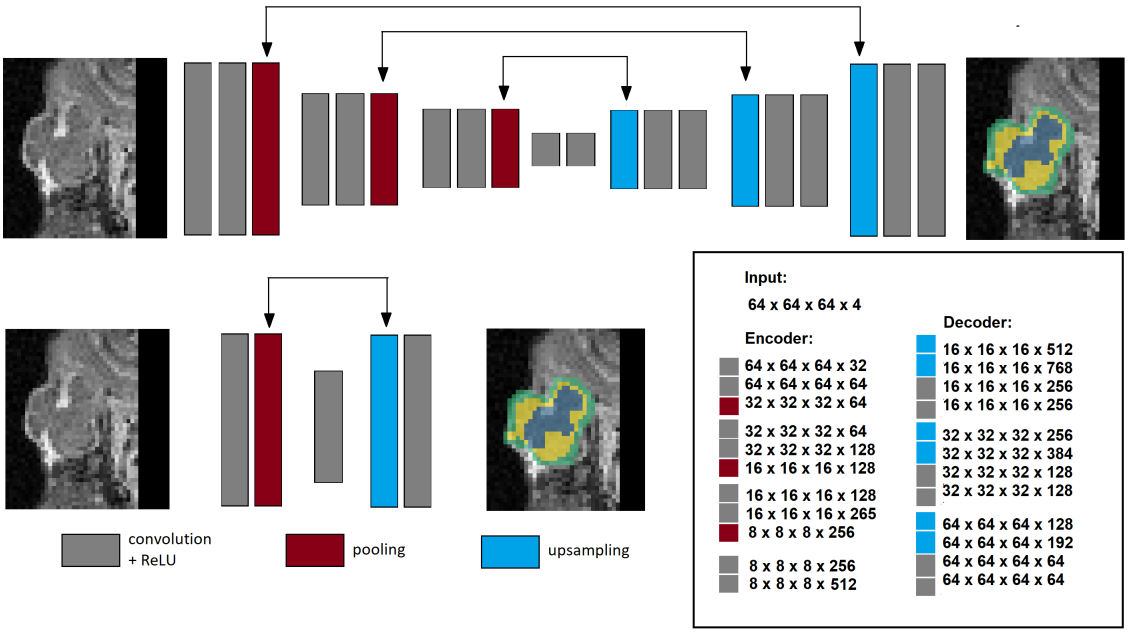


Figure 4.1: An overview of the architectures used as feature extractors. Top: UNet, bottom RandNet.

to ScatNet, where the weights were fixed to Gabor filters, we use a much simpler version presented in [130]: RandNet with pre-set RP matrices extended to 4D tensors (3D images plus channels). We set the desired dimensionality to m=5 channels as in the case of ScatNet, resulting in computationally light representations. While scattering nets originally used modulus non-linearities to compute the joint invariants of successive filter transforms, for simplicity of implementation in this work we use ReLU activation instead, which is more commonly used nowadays. We further extend the network with a decoder step and skip connections as in T-UNet, thus the topology of the network remains similar. The only difference in the decoding stage is that we employ a simple scaling strategy for upsampling, instead of deconvolutions. We further use the network again as a feature extractor in connection with BoV encoding, resulting in KM-RandUNet and GMM-RandUNet variants. While the resultant network contains trainable parameters for convolutions, they were all set to the values of random projection matrices, resulting in no parameter training and

a reduced computational complexity. While the use of random weights may initially seem counterintuitive, we will show, that even in such an extremely simple form the representations are capable of generating meaningful results.

For the UNet architecture, we experiment with three levels of depth $d = \{2, 3, 4\}$. As mentioned in the first section of this chapter, in order to well describe texture-like data, which is considerably less complex than high level object information such as faces, we set the depth to $d = 2$ as our baseline method and hypothesise that increasing depth will have a detrimental effect on both the quality of segmentations, and consequently, the classification accuracy. The depth of 2 has also been suggested by the authors of the ScatNet handcrafted network, which is kept constant across the experiments. The resulting dimensionality of the feature vectors generated by T-UNet and Rand-UNet deep feature extractors equals 64 and 5, respectively.

As mentioned in the previous section, while we do not use fine-tuning, in order to retain some advantages of it and improve performance we opted for the following design choices. We performed the knowledge transfer from a biomedical image domain that, although different in the context of the global information (T1- and T2- MRI images of brain tumours), arguably carries similar fine properties in the first layer. Similarly, Random Projections used in the handcrafted network have been previously shown to perform well in the context of radiological data, specifically mammographic images. Furthermore, they were used to correctly estimate the risk of developing breast cancer based on the variation present in the image, which is also the aim of this work. Finally, we use a shallow encoding and a classifier instead of the domain-adapted final fully-connected sigmoid layer. This enables a light-weight analysis with no segmentation ground-truth or large quantities of training data.

4.2.2 Shallow encodings and classification

While in this chapter we are using deep image representations for segmentation and classification of the tumours instead of shallow ones, the structure of the rest of the pipeline remains the same for a fair comparison. Initially, the set of feature vectors, as represented by deep feature extractors described in the previous section, is vector-quantised in order to find the most dominant patterns, or visual words. As was the case in the previous chapter, we evaluate two variants of the shallow BoV encoding: 1) with KMeans clustering and 2) with Gaussian Mixture Models. The resulting labelling is then represented by an orderless frequency encoding, which constitutes the basis for further classification. This part of the pipeline remains exactly the same as in the previous chapter and the details of its implementation can be found in Subsection 3.2.2 of the previous chapter.

4.3 Experiments

We execute a set of experiments to evaluate the proposed methods, that are compared to the shallow representations introduced in the previous chapter as well as to the baseline method frequently used in the context of tumour perfusion heterogeneity assessment. For fair comparison, we not only used the same pipeline setup (with the exception of feature extractors), but we also evaluate the methods on the same dataset as in the previous chapter. We test the proposed representations on the task of segmentation and classification as in the previous chapter, and employ the same metrics for evaluation: 1) visually assessed contiguity, 2) confidence of segmentations, and 3) classification accuracy. For the final evaluation, we combine all three metrics together for a meaningful conclusion.

The methods under evaluation are split into three groups: 1) baseline, 2) shallow, and 3) deep. We therefore compare KM and GMM variants of: Vox (baseline), LBP

(shallow), Patch (shallow), UNet (deep) and RandNet (deep). Vox is considered as it is commonly used for assessing tumour perfusion heterogeneity, while Patch and LBP were the optimally performing shallow representations from the previous chapter.

4.3.1 Materials

A longitudinal study was used, as in the previous chapter (D3, Table 3.2), with scans acquired with a 4.7T DCE-MRI and 60 repeats of cardio-respiratory gated RF and gradient spoiled 3D echo scans every 10-15 seconds (TR=1.4ms, TE=0.64ms FA=5°) with isotropic resolution of 0.42 x 0.42 x 0.42 mm. The data set was acquired for 10 female CBA mice implanted with CaNT tumour models scanned longitudinally over 8 to 10 days, although some samples were missing up to two data points, yielding a total of 69 usable volumes spanning 8 time points (Table 3.4). We use the Tofts model and population derived AIF to model EES enhancement and derive K^{trans} parameters as a measure of perfusion, as in the previous chapter. Motion artefacts were minimised with cardiac synchronisation and re-acquisition.

We evaluate the methods on two tasks: 1) subregional assessment via segmentations, and 2) quantifying the change in the overall degree of heterogeneity via growth-based classification. All available data was used to calculate segmentation agreements. For the classification task, due to the arbitrary definition of the growth stage we aimed at grouping the data into two classes (early and late stage) with a minimum of two data points of margin between them to avoid class overlap. More details on the grouping can be found in subsection 3.3.3 of the previous chapter. Finally, to alleviate the computational burden, the images were further cropped to the smallest possible window capable of containing the largest tumour across all images (plus margin).

The details of the network architectures are described in subsection 4.2.1 of this chapter and are illustrated in Fig. 4.1. We experimented with a set of network depths

$d = \{2, 3, 4\}$, with $d = 2$ as a baseline comparison method for reasons outlined in the previous section. The target dimensionality of m=5 was kept constant across the experiments with RP matrices in Rand-UNet. For T-UNet the transfer was performed from the BRATS dataset, details of which can be found in subsection 3.3.7 of the previous chapter. The choice of word/mixing components was the same as in the previous chapter for fair comparison and equal to $k = 4$, that aligns with the number of tissues likely to appear in the tumour microenvironment. Initial prototypes showed that both deep representations performed better with a global variant of the encodings, which is a setting we kept constant through the rest of the experiments. We believe this is a consequence of deep networks learning the global context. The comparison methods from the previous chapter were also chosen according to their best performing specification. We performed the classification evaluation with LOOCV on a per-subject basis; more details can be found in section 3.3.1 of the previous chapter. All the encodings were classified with Support Vector Machines (SVM) with parameters kept constant and using the default settings from the scikit-learn [144] implementation. Finally, all deep architectures were implemented with Keras [145].

4.3.2 Segmentations

The introduced methods were capable of meaningfully visualising the change in subregional distributions of the tumour over the time. In Fig. 4.2 we can observe similarities between the segmentations for two different tumours over time, as compared to Patch-KM, one of the best methods from the previous chapter. As an estimate of the corresponding slices in a growing tumour, we show the slice through the sagittal plane containing the largest area across. The labels are arbitrarily assigned to the learned visual words, representing different perfusion patterns (between 1 and 4) and are consistent across all images.

The distributions generated by deep representations appear to follow the same

trend, heterogeneous at first, the increasingly homogeneous as the tumours grow. The similarities of segmentations coming from different methods and approaches increase our confidence in the results. Astonishingly, even the Rand-UNet representation with random weights was capable of generating segmentations that were visually (Fig. 4.2) and quantitatively (Table 4.1) relatable to those from the previous chapter. While it is noticeable that KM+RandUNet appears to lose fine details, the general trend in individual tumours, as well as over time, remains similar.

Given that there is no segmentation ground truth, in the previous chapter we introduced a confidence metric/maps. We calculate the DICE coefficient over a ground-truth generated as a per-voxel agreement of all segmentations. In this sense, the segmentations are understood as increasingly confident with increasing number of different methods yielding similar results. More detail of the implementation can be found in section 3.3.4 of the previous chapter. Segmentations from different methods, as compared to the best performing methods from the previous chapter, can be seen in Fig. 4.3 for the KM and in Fig. 4.4 for the GMM variant of encoding. Average DICE scores are outlined in Table 4.1 alongside classification accuracy. Note that with a higher number of available segmentations, DICE scores were also updated for the previous shallow methods, which is reflected in Table 4.1. The segmentations achieved a high agreement with the previously generated methods, thus increasing the confidence in the generated results. Interestingly, deep representations with GMM considerably improved the contiguity of segmentations over the shallow ones, which resulted in a drop of overall agreement due to differences. At the same time, DICE scores increased from the previous chapter in the case of the KM variant in the light of further agreements. As expected, the RandNet variant in general performed worse than UNet, however, even with such a trivial implementation it already produced more contiguous segmentations than some methods in shallow representations. The segmentation confidence heatmaps can be seen in Fig. 4.5, with the high values

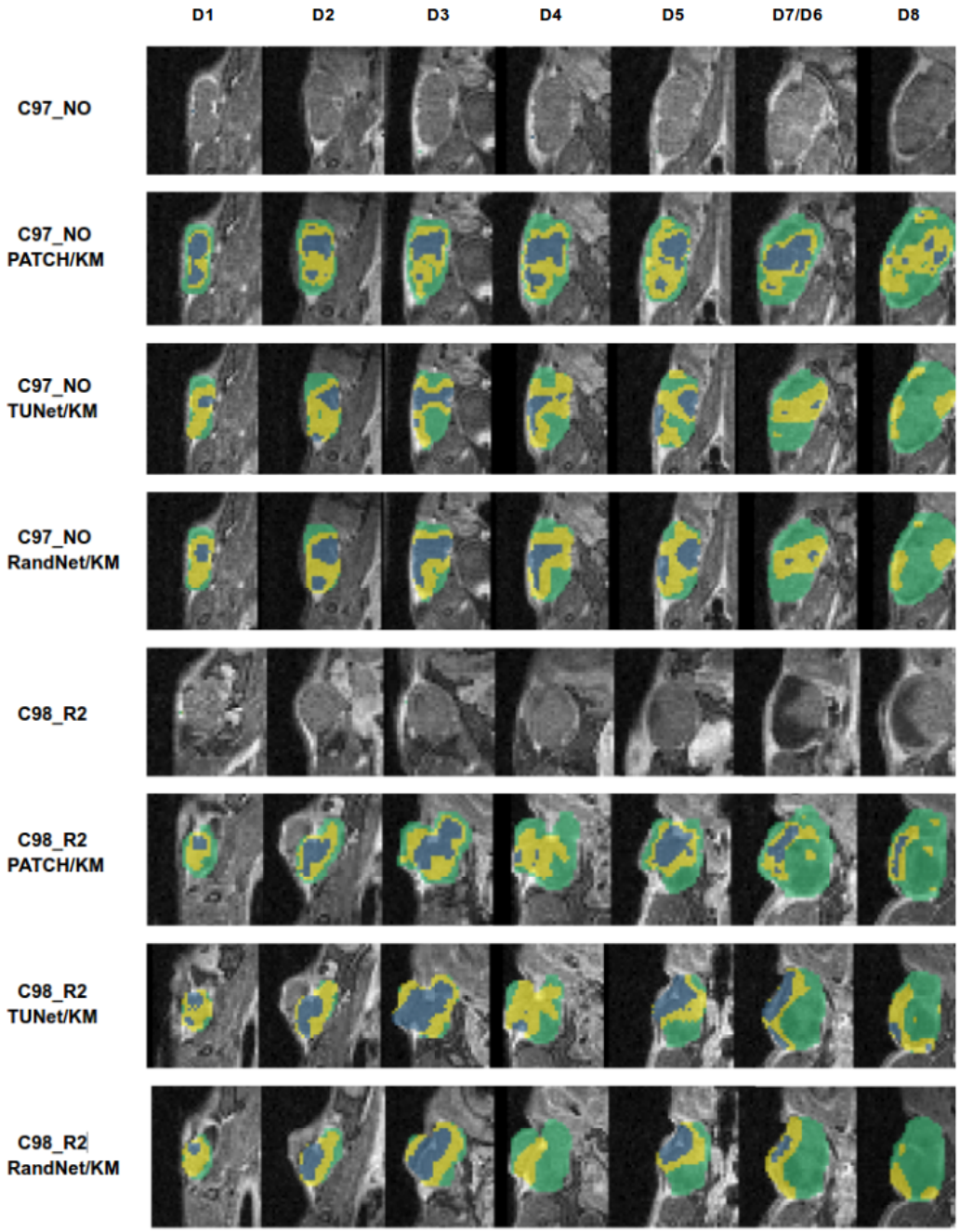


Figure 4.2: Subregional distributions of a growing tumour (columns relate to days of growth) shown for two different samples with KM+TUNet and KM+RandUNet as compared to Patch+KM and baseline enhancement. Original image mid-way through enhancement (row 1 and 5), Patch+KM (row 2 and 6), KM+TUNet (row 3 and 7) and KM+RandUNet (row 4 and 8). Rows (1, 2, 3, 4) and (5, 6, 7, 8) originated from two different tumours C97_NO and C98_R2 respectively (Table 3.4). Where a timepoint was missing from the series, labels for all individual mice were provided in top-bottom order. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of tissues is constant and equal to 3). Note the slices may bare differences due to difference in cropping, however slices are corresponding between the rows.

(black) representing high confidence (all methods agreeing on a label) and low values (yellow) representing low confidence (most methods disagreeing on a label). In general, the segmentations had an agreement of 50% or more, as in the case of shallow methods. Note that our joint confidence maps also increased in confidence now that a greater number of similar segmentations have been added. As observed in the previous chapter, the early stages of growth resulted in the lowest agreement across segmentations. We believe that this is due to the low resolution as a limiting factor.

Due to the lack of GT data, as in the previous chapter, we consider segmentation scores alone as insufficient for overall evaluation. For that reason we choose to compare the results with classification accuracy, which we consider as an endpoint and ultimate goal of the process, as it would be in the case of therapy outcome assessment.

4.3.3 Classifications

The results of classification accuracy are presented in Table 4.1. We compare the outcome of deep representations introduced in this chapter (KM/GMM + TUNet and RandUNet) with the best performing methods from the previous chapter on shallow representations (Patch and LBP + KM/GMM) and the baseline Vox + KM/GMM method.

The results show that even without any training or fine-tuning at all, deep representations were capable of outperforming and matching the performance of the best performing methods from the previous chapter. KM+TUNet was able to outperform any previously explored methods with a margin of 5% (with the exception of Vox+GMM, however we determined in the previous chapter that this was due to overfitting). GMM+TUNet was also generally found to perform better than the best performing methods using shallow representations (Patch+KM/GMM and LBP+GMM: 7.5% better, LBP+KM: 2.5% better). In addition, we, somewhat surprisingly, found that even an extremely simple implementation of RandUNet was

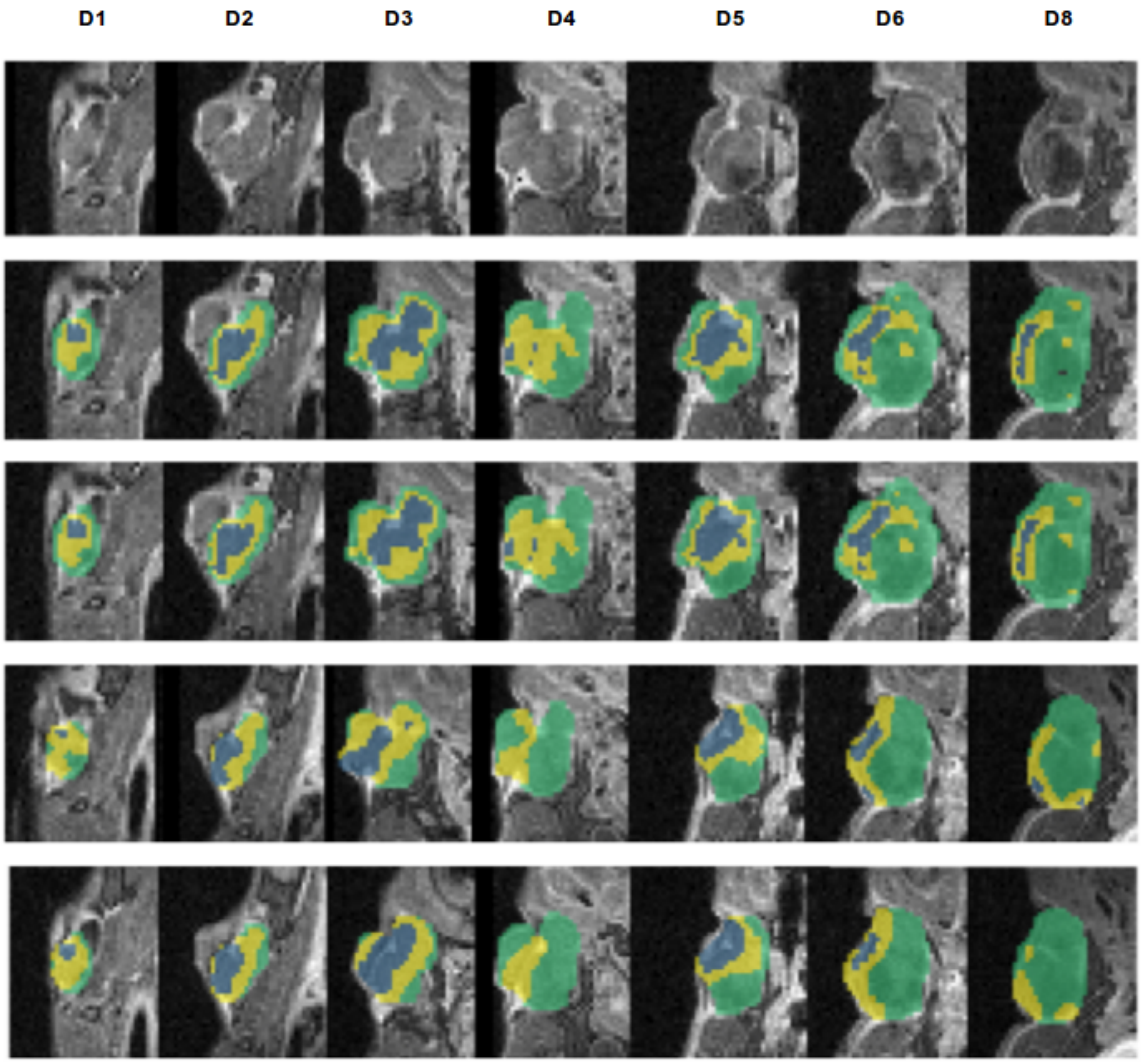


Figure 4.3: Subregional distributions of a growing tumour C96_R2 (columns relate to days of growth) shown for a number of shallow and deep representations with KMeans BoV encoding, from the top: LBP, Patch, UNet, RandNet. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of labels is constant and equal to 3).

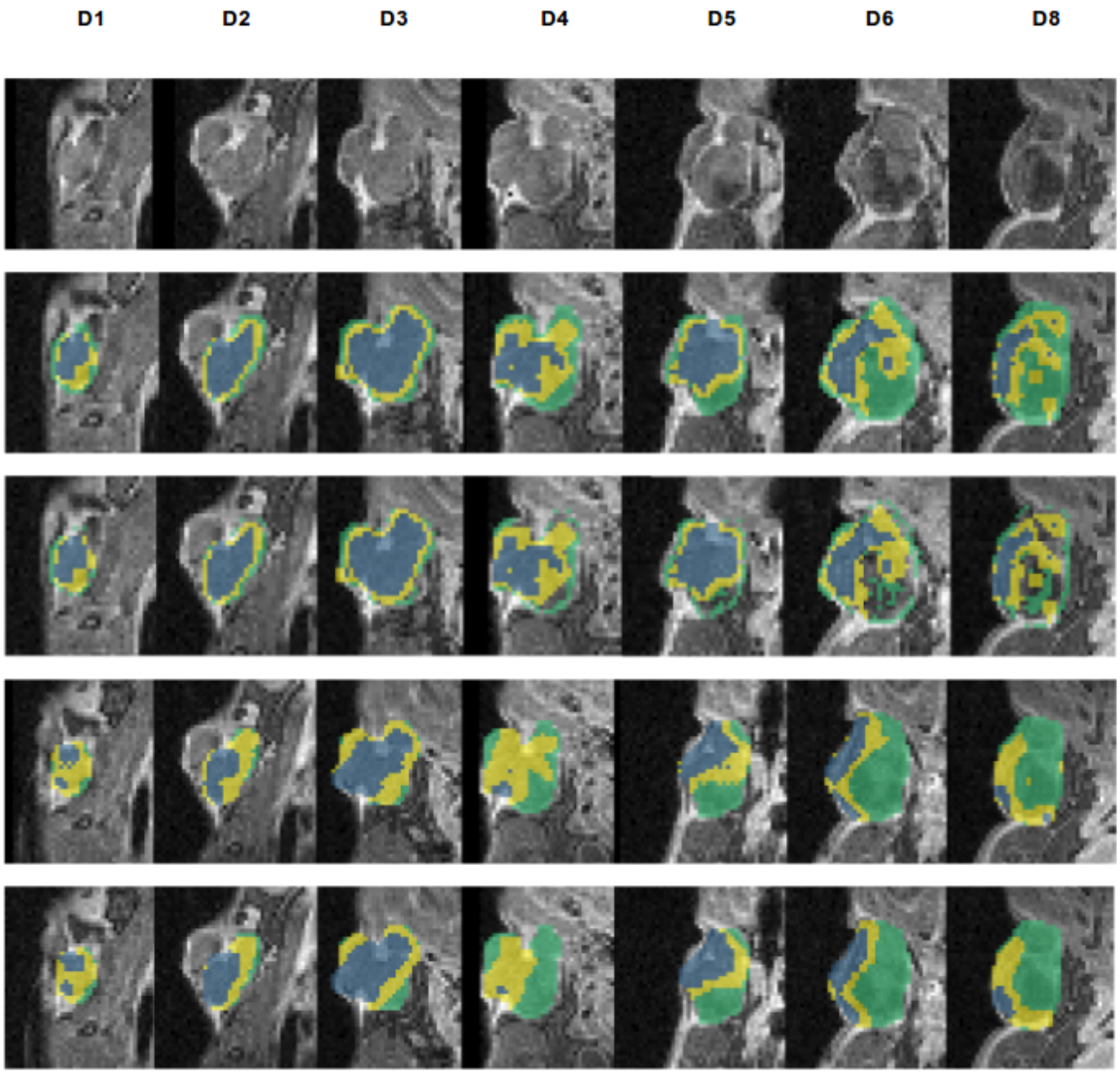


Figure 4.4: Subregional distributions of a growing tumour C98_R2 (columns relate to days of growth) shown for a number of shallow and deep representations with GMM BoV encoding, from the top: LBP, Patch, UNet, RandNet. Colors are arbitrary, but are consistent across the images and describe different visual words representing different tissue properties (number of labels is constant and equal to 3).

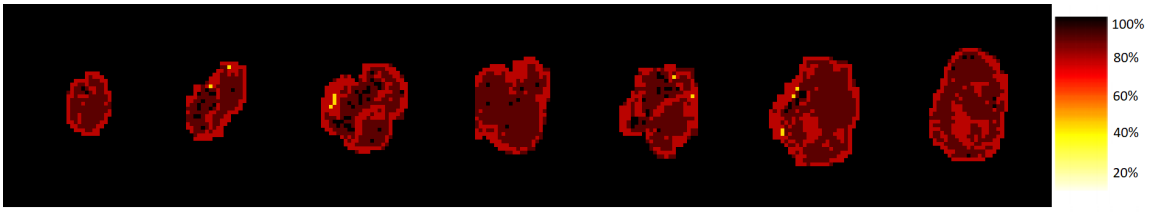


Figure 4.5: Confidence heatmaps for segmentations. High values (black) correspond to high confidence, while low values (yellow) indicate low confidence in segmentations. Most voxels show an agreement of above 50% over different segmentation methods.

	Vox	Patch	LBP	UNet	RandUNet
	Acc. | DICE	Acc. | DICE	Acc. | DICE	Acc. | DICE	Acc. | DICE
KM	65% | 0.480	**70%** | **0.798**	**75%** | **0.806**	**80%** | **0.825**	**75%** | **0.729**
GMM	**82.5%** | 0.638	**70%** | 0.562	**70%** | 0.586	**77.5%** | **0.783**	**75%** | **0.715**

Table 4.1: Classification accuracy scores and confidence DICE measures for various methods and encodings, all sharing the same BoV pipeline. The best result are emphasised in bold.

already able to match the performance of the best methods using shallow representations (LBP+KM). Furthermore, it is worth noting that the upsampling layer in Keras [145] used in RandUNet is not an interpolation, but simply repeats the local information, therefore the performance could be potentially improved if deconvolutions were used. For an average across KM and GMM, the best classification accuracy was obtained by TUNet with 78.75%.

For the reasons outlined in the previous chapter, accuracy scores should be considered in the light of the underlying segmentations. For that reason, we perform the final evaluation in conjunction with DICE and visual assessment. Interestingly, all deep representations were capable of producing considerably more contiguous segmentations than their shallow counterparts, certainly in the case of the GMM variant, which could be of more interest to clinicians for reasons outlined in the previous chapter. While all the deep methods were reasonably contiguous with high DICE scores, KM+UNet produced the best performance when considered with segmentation agreement as well as classification accuracy (80% Acc., 0.825 DICE).

4.3.4 Parameter evaluation

We would like to test the influence of depth of the network on the results as represented both by segmentations and classification accuracy. As it is typical for a UNet to contain 4 layers, while 2 is a typical size of a handcrafted network, we evaluate the

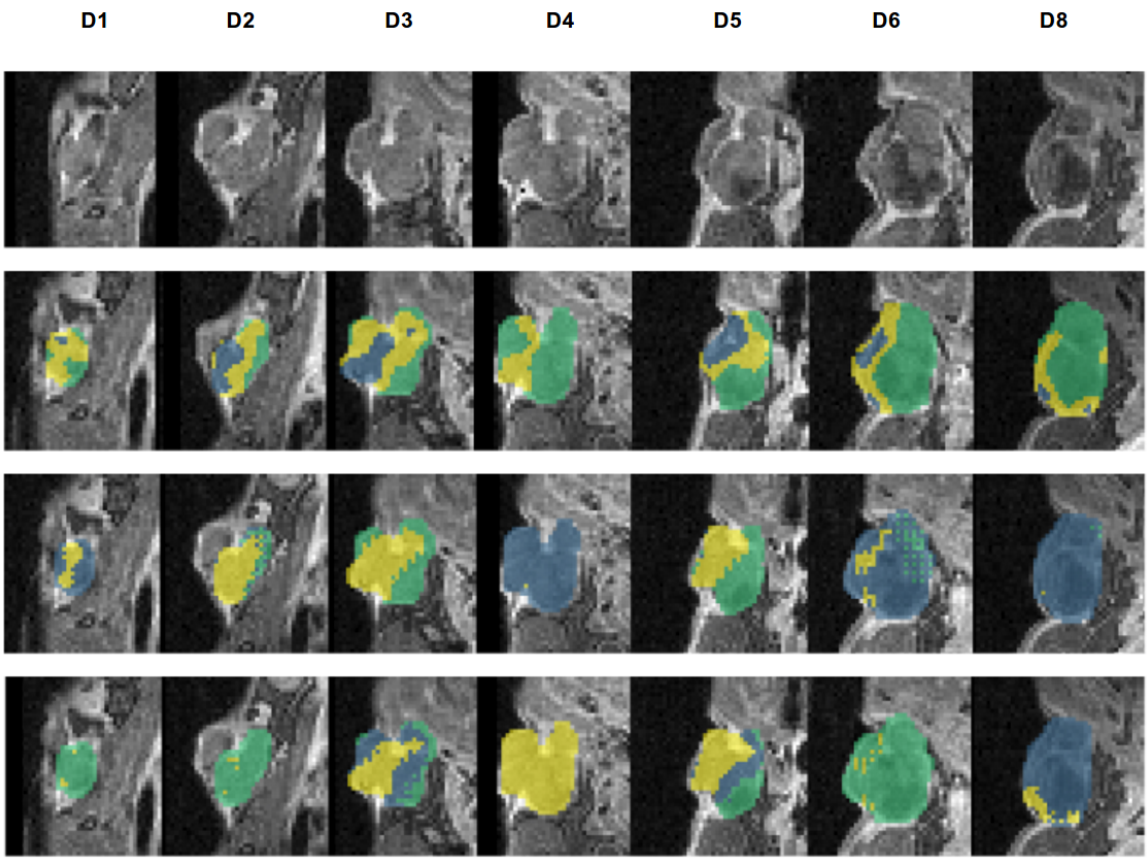


Figure 4.6: The effect of varying the d parameter (depth of the network) on the subregional distributions of the same sample. The rows correspond to $d = 2, 3, 4$ respectively, with the first row showing mid-point through enhancement. Columns correspond to the days of growth.

effects of altering the depth of a UNet. We explore three variants of the network, with 2 (small), 3 (medium) and 4 (large) downsampling blocks. The segmentated images can be seen in Fig. 4.6. The classification results are presented in Table 4.2.

It becomes apparent from the segmentations that increasing depth leads to deterioration of the quality and contiguity. We can observe the introduction of upsampling artefacts (characteristic checkered pattern). Furthermore, the labels become inconsistent across the growth, an effect that was similar to what we observed in shallow representations when introducing global encodings. Interestingly though, the general distribution still follows the pattern we found in all previous experiments, which further consolidates our confidence in the outcome.

	small	medium	large
UNet (acc.)	72.5%	55%	57.5%

Table 4.2: The classification accuracy as a function of increasing the depth of a UNet architecture in the context of BoV encoding.

As hypothesised in the previous sections, increasing the depth of the network also has a negative effect on the classification accuracy. This is expected to be a consequence of the fact that biomedical images, although more complex than textures, do not require as much complexity in their representations as object-level images. Interestingly, the large variant of a UNet performed slightly better than the medium variant, which is also noticeable in the segmentations with slightly better label consistency.

4.4 Conclusions

In this work, we aimed to assess the suitability of deep representations to the problem of assessing perfusion heterogeneity in tumours, from limited training data and no ground truth. We have shown that even in such an extremely simple form, these methods are capable of matching and outperforming those from the previous chapter that used shallow representations. Furthermore, we back this claim by demonstrating their ability to produce visually coherent and contiguous segmentations that quantitatively match the shallow segmentations from the previous chapter. In subsection 3.3.7 we have tested the ability of the shallow representations to meaningfully reflect the physiological information through subregional distributions. As the deep segmentations achieved high overlap agreement with the shallow segmentations, we argue that they are also suitable to meaningfully reflect on the subregional processes in the tumour microenvironment. The similarity that we were able to observe across so many segmentations improves our trust in the results. It also suggests, that there is an

inherent change in the tumour growth process which, as we have demonstrated, can be visualised and quantitatively described. Finally, we have shown that the change is again sufficient to discriminate between the early and late stages of growth. Our findings suggest that, with more sophistication, deep representations could be successfully utilised in the context of assessing tumour perfusion heterogeneity even for datasets with limited resolution and a small number of images.

From the technical side, we have suggested a lightweight image representation solution particularly suited to datasets that are too small to train or fine-tune and with no segmentation ground-truth. Even with the simple nature of the representations used in this chapter, we were able to match and outperform the methods from the shallow representations. It is expected, therefore, that with more sophistication, the architectures could achieve even better performance. This, however, remains out of the scope of this work and is designated for future experimentation. Astonishingly, we have also shown that with no training at all, no ground truth and random weights, our representations were capable of approximating the segmentations obtained from more sophisticated shallow methods, and matching the performance of methods commonly used for assessing perfusion heterogeneity. This could be due to the introduction of specialised filters designed and evaluated specifically for the domain. However, in this case we believe the major advantage is in exploiting the hierarchical nature of deep architectures which are capable of modelling more complex relations than any previously explored machine learning methods [126]. In this sense, specialised handcrafted cascading architectures could constitute a valuable alternative for problems with small datasets that are unsuitable for training or fine-tuning, in particular if there was an existing body of research using handcrafted descriptors. This could possibly be combined with the shallow-deep nature of hybrid methods and adapted towards the task of segmentation.

While we have shown that even such simple deep models were capable of match-

ing the performance of shallow representations, we would still like to make a case for shallow pipelines. Although deep representations generally tend to outperform shallow ones, unlike in shallow representations, deep architectures are often accompanied with improvements such as GPU training, large scale datasets, or data augmentation. Some researchers argue that it is therefore possible for the magnitude of the performance gap to be partially explained by the lack of such sophisticated computational and data augmentation techniques not previously explored in shallow representations [125]. Furthermore, shallow representations, as shown in the previous chapter, are transparent and easy to interpret, which is an important advantage for clinical applications. Finally, without the shallow pipelines, we would lack the tools to enable the transparency of deep representations proposed in this chapter. For those reasons, it is worth noting that despite of the dominance of deep representations, shallow representations are still firmly relevant in the field and should not be abandoned entirely in favour of deep representations.

In summary, we have shown that deep representations can be applied to the problem of assessing tumour perfusion heterogeneity, yielding similar results to those from shallow representations. We were able to visualise changes in the subregions of the tumour over time, and those changes were significant enough to enable a machine learning characterization of early and late stages of growth. Growth stages are, however, a somewhat arbitrary measure with unclear definition of the class boundaries, which is suboptimal for the sound definition of the problem. Furthermore, while it is important to understand the subregional distributions of a tumour during the standard growth process, as unaffected by therapy, we would also like to apply this to the context of therapy response. In the next section, we will therefore evaluate the suitability of the proposed shallow and deep approaches on a clinical therapy dataset. Furthermore, we will utilise the tools and metrics outlined in these chapters to enhance the confidence and transparency of the process, which could contribute to the

future clinical potential of these tools.

Chapter 5

Clinical application: heterogeneity-based therapy response assessment with deep and shallow texture representations

In the previous two chapters we looked at the problem of assessing the change in perfusion heterogeneity in the context of longitudinal tumour growth. To this end, in Chapter 3 we introduced a BoV pipeline for problem, additionally evaluating a number of shallow image representations on the task. In the Chapter 4, we further introduced novel deep representations that, uncharacteristically for deep learning, were able to perform segmentations with no ground-truth and classify images with limited training data. We found that both deep and shallow representations were capable of characterizing the change in heterogeneity, and that such change was significant enough to differentiate between the early and late stages of tumour growth. In this chapter, we evaluate whether the introduced framework along with shallow and deep representations is capable of characterizing the problem of therapy response

assessment.

The key novel contributions of this chapter are as follows:

- Applying the Bag of Visual Words framework developed in previous chapters (both shallow and deep) to evaluate and flag early signs of lack of response in a clinical therapy setting.

5.1 Heterogeneity and therapy response

Tumours are characterised with heterogeneity in their macro-appearance that is believed to be a consequence of the underlying processes in their microenvironment. Notably, the process of angiogenesis is known to produce increasingly disorganised vasculature, that, as a consequence, increases the heterogeneity of perfusion [3, 4]. Furthermore, such heterogeneity has been found to change following therapy [7, 8]. Tracking such changes could therefore prove predictive in understanding tumour progression, which could, in turn, be applied to understanding response to therapy in the clinical setting.

As real clinical data is often complex and trials are comparatively small[1], in the previous chapter we opted to use a larger pre-clinical dataset to evaluate the proposed methods between each other, as well as their suitability to the problem defined in terms of growth. We found that the methods were capable of visualizing the change in heterogeneity and that such change was significant enough to differentiate between the stages of growth. The problem of assessing tumour growth while unaffected by therapy is important, as the findings could be further applied in the context of therapy response. However, the question of therapy response assessment itself is much more valuable clinically. For that reason in this chapter we redefine a problem

[1]Although larger DCE-MRI datasets do exist. Due to the lack of availability, this work explores a smaller dataset, however this is a limitation and any future work should consider evaluating a larger clinical dataset.

to assess whether the change in heterogeneity is characteristic enough to identify and flag early response to therapy, or lack of thereof. Having previously introduced and evaluated a number of deep and shallow texture representations, and their suitability to the problem of growth-based perfusion heterogeneity assessment, we now would like to use: a) the methods, b) transparency of the framework, and 3) the knowledge we acquired in the previous chapter and apply it to the important and meaningful context of therapy response assessment.

Perhaps the single most important open problem in the clinical setting is that even the most sophisticated method will remain meaningless without extensive validation. In [31] the authors note that most biomarkers fail to get translated into clinical practice, which is termed as a "translational gap". The term describes failure in implementing new products and approaches into clinical practice. There are a number of reasons, however the most relevant for imaging is the lack of technical and biological/chemical validation, which becomes increasingly difficult without ground-truth data and with the black-box nature of machine learning techniques. We must note that while the value of rigorious validation cannot be replaced, in this work we have designed our methods to deliberately address some of the limitations for clinical adoption. We outline those limitations in the following paragraphs.

5.1.1 Limitations of DCE-MRI biomarkers

In addition to the existence of a clinical gap, in [76] the authors state that none of the DCE-MRI variables has yet been accpted as a validated biomarker in clinical practice. They suggest this is due to a number of reasons. Firstly, pharmacokinetic variables are a composite of physiological processes, which makes it difficult to establish an accurate relation with the tumour microenvironment. Secondly, pharmacokinetic models consistently overestimate the values of K^{trans}, which is typically recommended as an end-point for DCE-MRI analysis [54]. Thirdly, there is a significant variation

in reproducibility of pharmacokinetic parameters. Fourthly, histological validation of DCE-MRI measurements is not complete. The authors suggest that this is partly due to the difficulties arising from comparison of information at radically different scales between histology and MRI. Furthermore, it has been shown that alignment of such information that has been deformed and is likely to lie on different axes is a challenging task [86]. Finally, only a limited number of studies do in fact present evidence for correct estimation of clinical outcome, and those that do, lack comprehensive validation. The last two issues relate to the nature of biologcial and technical validation respectively.

It is worth noting that the first three of the outlined drawbacks are a consequence of limitations of quantitative pharmacokinetic modelling. To address this concern, in this work we make a deliberate choice to perform the analysis with semi-quantitative pharmacokinetic modelling based on Principal Component Analysis (PCA). The advantage of this method is that it is entirely data-driven, thus it makes no assumptions about the underlying data. The details of the method can be found in Section 2.3.1 of Chapter 2.

The issues of technical and biological validatation still, however, remain a major problem that proves difficult to address. Comparison of histology and macro-scale images such as MRI is a challenging task due to the large difference in the scales they individually correspond to. In addition, histological images typically reflect on a single slice through the tumour, which might be difficult to correlate with the standard axial view of the tumour. Furthermore, disection of tumours could introduce challenging deformations that would be difficult to represent with available image registration methods. Technical validation and reproducibility also remain challenging tasks in the light of such high inter-patient heterogeneity. While the issue of validation may be difficult to overcome, in the following paragraphs we describe a number of strategies we employed to improve the transparency and confidence of our results, which we

believe could increase their potential for clinical translation.

5.1.2 Confidence and transparence

Segmentations are an important aid for medical image assessment. Not only can they be used to guide clinicians in their diagnosis, but they also provide a visual aid to make sense of the, often black-box, machine learning methods. Furthermore, segmentations in this work constitute a basis for the subsequent classification. In this sense, we can estimate segmentation quality based on the accuracy metric for classification, which is their final purpose. Conversely, we can also estimate the quality of classification (e.g. identify the cases of overfitting as in the previous chapters) with visual assessment of the contiguity of the segmentations. However, as we have no ground-truth segmentations available for this data, the definitive assessment of their quality remains out of reach. For that reason, we have introduced a number of steps in order to increase confidence and trust in the results. We aid the segmentations with confidence maps, which are based on the voxel-wise agreement across all methods. This is based on the assumption that human confidence in the decision effectively increases with the number of corresponding observations. In this sense, voxels with agreements across all the methods would score high confidence (black) and disagreeing pixels would score low (white). We further use this information to compute our simulated "ground truth", where the label of the pixel is determined by the median of all values for that voxel accross segmentations. Finally, we compute DICE overlap scores with the ground-truth to determine the degree of misalignment of each segmentation from the most frequent (and thus we assume most likely) outcome.

Furthermore, the black box nature of machine learning methods makes it difficult to support the validity of the findings. In the design of our methods we made deliberate decisions to address some limitations of clinical translation. Mainly, we advocate a multi-step pipeline where the results can be visualised and inspected. We believe

that such transparency could be a stepping stone towards an increased acceptance of such methods in clinical practice.

5.1.3 Ambiguity

Segmentations of medical images are ambiguous [129]. This is a fact well reflected not only in the varied results of automatic methods, but, perhaps even more so, in the variety of segmentations resulting from a number of human experts. This is particularly a problem in the task of tumour subregional assessment. While the tumour itself often remains clearly visible (especially in DCE-MRI due to the contrast agent), a definitive differentiation between neigbouring tissues often remains ambiguous. As a consequence, a clinical decision based on a deterministic outcome of a machine-generated results could lead to a sub-optimal treatment, which is a point raised by the creators of UNet [19, 129]. Furthermore, they suggest that probabilistic segmentations (segmentations with voxel-wise probabilities) discard the information about possible co-variances between results from different voxels. For those reasons, they argue that there is an advantage in having an ensemble of hypotheses that could be further propagated into the pipeline and even used as suggestions for additional tests to narrow down the set of likely hypotheses.

This is a concept which we already briefly touched upon in Section 3.3.8 of the chapter on shallow representations. We compared the output segmentations from two feature descriptors based on opposing underlying assumptions: similarity and variation. While we observed some differences, interestingly the main underlying pattern of subregional distributions remained similar during the growth process, which inspired more confidence in our results. In this chapter, we take it a step further and provide a full ensemble of methods to characterise the change in the distributions of perfusion patterns over time. In this sense, complimentary results coming from many different segmentations could provide more confidence in the outcome and help explain

the black-box classification results. At the same time, contradicting segmentations could provide the radiologists with an ensemble of likely outcomes which could further aid their clinical decision. In addition, as a result of machine generated outcomes, such an ensemble would also provide results that could be unlikely to occur to a human expert, which might be beneficial for clinical decisions.

5.2 Methods

In this chapter we would like to use our knowledge of tumour perfusion heterogeneity during normal tumour growth process unaffected by therapy, and apply it to therapy response assessment. We would also like to evaluate the suitability of two types of methods we introduced for the problem in the previous chapters: deep and shallow, as compared to a baseline method frequently used in tumour heterogenetiy assessment. The methods under evaluation are, therefore, as follows: Vox (baseline), Filters (shallow), Patch (shallow), LBP (shallow), T-Unet (deep), Rand-UNet (deep). All of the methods are also evaluated with the KM encoding variant of visual codebook generation, as previous experiments indicated it performed slightly better than GMM.

5.2.1 Feature descriptors and encodings

In this work we employ the same BoV pipeline as described in the previous two chapters. We perform shallow and deep feature extraction on a per-voxel basis. Following that, the feature space is vector quantised in order to find the most dominant perfusion patterns, referred to as visual words. The result is a segmentation representing subregional distributions of the perfusion patterns. Finally, a proportion encoding of such visual words represents the overall degree of heterogeneity in the tumour. Such encoding is then fed into a classifier.

5.2.2 Classification

While in the previous chapters we used the encodings to classify between the early and late stages of tumour growth, in this chapter we are aiming at evaluating whether we can identify and flag early non-responders, therefore classification is no longer adequate. In addition, the dataset is highly imbalanced with only 3 cases of patients achieving a full response at 3 months following CRT, making the data unsuitable for machine learning classifiers. Instead, we will use the transparency of the introduced pipeline to explore the similarities and differences in the subregional distributions, as well as similarities in the overall change in the subregional distributions (encodings) for each tumour. Additionally, we will employ an ensemble of methods to evaluate this to increase the set of likely hypotheses as described in Sec. 5.1.3.

In order to identify potential early non-responders, we define each patient as a percent of absolute change in tumour subregions. In other words, each patient is represented with a vector of absolute differences between the mid- and pre-therapy scans with each of the visual words representing the proportion of a perfusion pattern. To indicate the overall change, we sum the absolute differences:

$$SAD = \sum_{i=0}^{k} abs(h_1(i) - h_2(i)) \tag{5.1}$$

where $h(i)$ represents the proportion of the tumour where visual word indexed by i was present, the 1 and 2 subscripts indicate for pre- and mid-therapy scan respectively and k denotes the number of visual words.

5.3 Results

In this section we evaluate the methods from the previous chapters on a clinical dataset. Following the final comparison of both shallow and deep representations in the previous chapter, it was shown that the KM variant of the pipeline performed

slightly better than GMM, and thus we use it in this work. The remainder of the pipeline remains the same for a fair comparison. The methods under evaluation are split into three groups: 1) baseline, 2) shallow, and 3) deep. We therefore compare KM variants of: Vox (baseline), Filters (shallow), Patch (shallow), LBP (shallow), Unet (deep), RandNet (deep). These constitute all of the methods we have introduced or compared to in this work.

As in the previous chapters, we evaluate the methods on the task of segmentation, however no classification is performed, instead we examine the differences and similarities between the encodings. We employ the same metrics for evaluating segmentation as in the previous chapters: 1) visual contiguity assessment, and 2) confidence of segmentation. The differences between individual patients are evaluated with 3) SAD metric, indicating overall absolute change in the fractions of tumour subregions.

In addition we provide a number of strategies (described in Section 5.1 of this chapter) to increase the confidence and transparency of the results. To increase the confidence we: 1) provide confidence heatmaps for the segmentations based on the per-voxel agreement across all the methods; 2) we perform the final examination based on an ensemble of hypotheses resulting from all the methods introduced in this work.

5.3.1 Materials

Data from the ART trial was used to evaluate the clinical potential of the introduced framework and methods (Table 5.1). The trial was designed as an observational study for patients undergoing chemo-radio therapy (CRT) treatment for histologically confirmed invasive primary anal squamous cell carcinoma. Acceptance criteria included tumour stage $\geq$ T2N0. One of the study endpoints was to identify %ACD change in early therapy. In this work, we use the data to identify and flag early response to therapy (or lack of thereof). The patients received 3 scans: 1 day prior to receiving

CRT (here termed as Before), following fraction 8-10 CRT day 10-12 (here termed as Mid), and three months following the CRT. The number of patients and scans they each had is presented in Table 5.2. The details of the scans are presented in Table 5.1. The motion assessment was made subjectively by the author of this work and clinical collaborators working on the trial.

Amongst other modalities, DCE-MRI images were acquired. The dataset contained two timepoints per sample: 1) pre- and 2) mid-therapy. Although there were a total of 15 usable DCE-MRI scans[2] (Table 5.2), some contained considerable motion. There were two types of motion in the data, rectal motion and patient motion. Two samples were rejected from the analysis due to the irrecoverable motion damage. One sample was rejected due to the lack of a correct ROI. The remaining 12 were motion corrected with Demons [146] implemented using fast guided filtering from [147].

All available data was used for the evaluation. Prior to the analysis, the images were cropped to the smallest possible window capable of containing the largest tumour across all images (plus margin). All the methods kept the parameters we found optimal in the previous sections.

Statistical summary of the data is shown in Fig. 5.1, including change in whole tumour volume (WTV) and enhancing fraction (EF) at 15s, 35s and 60s during contrast uptake. A threshold for enhancing voxels was defined as a signal change greater than 5 standard deviations of signal values over the WTV in the pre-contrast images as in [112]. Due to the large amount of noise present, the s.d. was taken from all pre-contrast images (first 5 timepoints). A fraction of voxels exceeding that threshold was calculated at 15, 35 and 60s. As the data contained a large amount of noise (Fig. 5.2), the value for each voxel at each timepoint was taken as a mean of 10 neighbouring timepoints for that voxel.

[2]The original study contained 23 patients, however 8 of them did not have enough timepoints to be considered for DCE-MRI pharmacokinetic analysis.

DATA	NUMBER OF SUBJECTS	NUMBER OF SCANS	RES	TUMOUR MODEL	METHODS
D4	15 patients*	2 scans each (1 day before CRT and following 8-10 fractions of CRT - 10-12 days) 30 scans total 24 usable scans*	1.47 × 1.47 × 2.5mm × 65s	squamous carcinoma of the anus (stage: ≥T2N0)	all methods

Table 5.1: Details of the data used in this chapter.
* Original study contained 23 patients, however 8 of them did not have enough timepoints to qualify for DCE-MRI pharmacokinetic analysis.
** 2 samples (4 scans) were unusable due to irrecoverable motion, 1 sample (2 scans) was unusable due to the lack of a correct ROI, for details see Table 5.2.

D4	**ART_108**	**ART_109**	ART_111	**ART_112**	ART_113
before	✓	✓	✓	✓	✓
mid	✓	✓	✓	✓	✓
D4 c.d.	ART_114	ART_116	ART_117	ART_120	ART_122
before	✓	×	×	✓	×
mid	✓	×	×	✓	×
D4 c.d.	ART_123	ART_124	ART_125	ART_127	ART_128
before	✓	✓	✓	✓	✓
mid	✓	✓	✓	✓	✓

Table 5.2: D4 clinical data used in this chapter. For details of the data see Table 5.1. 30 scans in total, 24 usable scans (12 pairs). Scans marked with × were unusable due to noise in acquisition or missing ROI. Patients marked in bold have not achieved a complete response at three months following CRT.

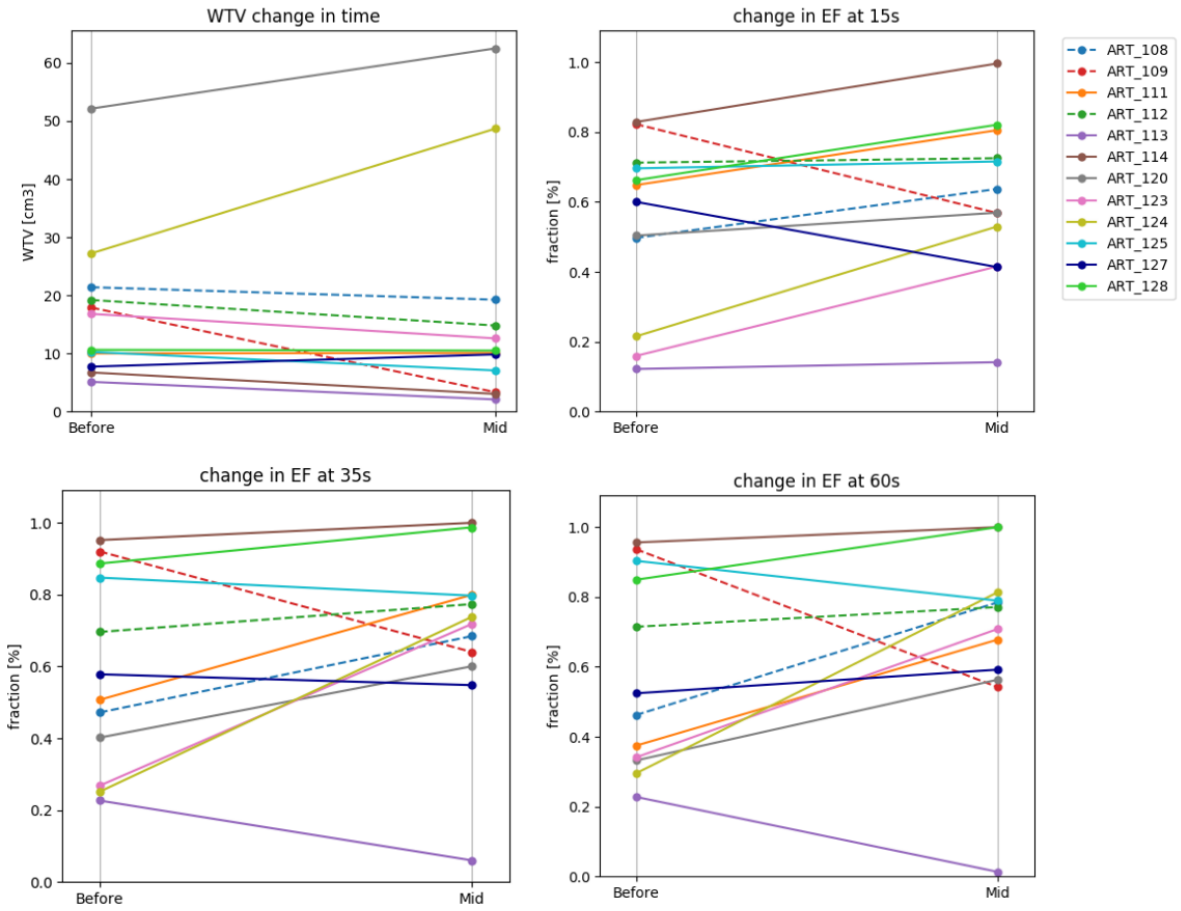


Figure 5.1: Changes in tumour statistics in time for all the patients in D4. Top left showing changes in WTV, top right showing changes in EF at 15s, bottom left 35s and bottom right 60s. No K^{trans} or T10 maps were available to calculate K^{trans} changes (no PK analysis was done for this chapter).

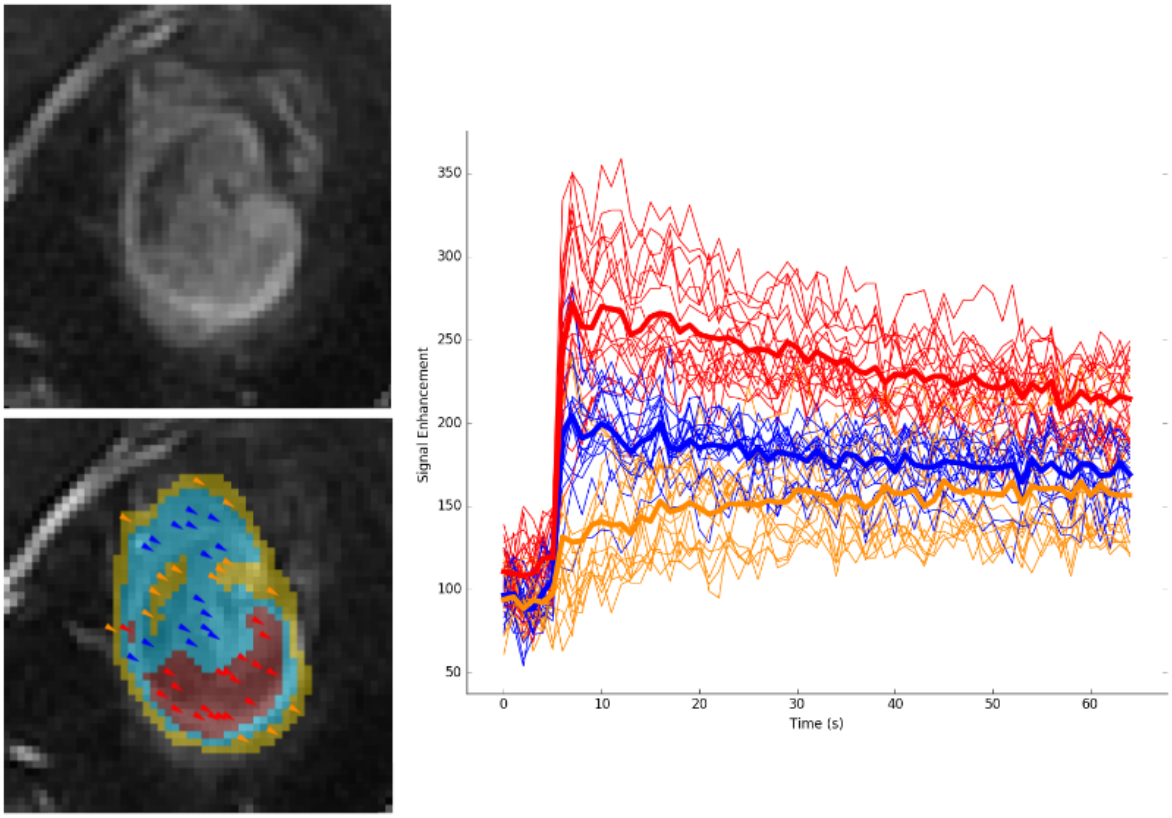


Figure 5.2: Enhancement trace for each type of visual word (texton). Left column: axial slice across the tumour showing raw DCE image from dataset D4 sample ART_120 Before and corresponding segmentation. Right image: a set of enhancement traces for a number of voxels indicated by arrows on the segmentation (colours of the arrows, plots and subregions are corresponding). The thick plot for each subregion is the mean enhancement trace for each subregion in the whole tumour.

5.3.2 Segmentations

Details of a segmentation for ART_120_Before are presented in Fig. 5.2. Individual subregions appear to describe physiologically distinct subregions and change in the fraction of visual word 3 (red) had a significant correlation with change in EF (Pearson, p=0.008).

The results of the longitudinal segmentations are outlined in Fig. 5.3. We can observe that all of the methods were capable of visualising the subregions of the tumours. Additionally, the subregions were visually corresponding across the methods. In contrast to previous chapters, none of the methods, even the baseline Vox, produced visually non-contiguous results. We believe that this could be attributed to the model-free pharmacokinetic representation. However, it becomes apparent upon closer inspection that the deep representations seem to produce more smoother results, which could be due to the multiple sets of convolutions. Furthermore, while Gabor filters failed in the previous chapters in the pre-clinical context, the results of the Gabor method seem to align well with the other methods in this work. Interestingly, some regions are consistently classified as the same by all the methods, as indicated on the confidence heatmaps representing the degree of misalignment across methods as black regions. Additionally, although in the later experiments the sample presented in Fig. 5.3 was ranked as one of the samples with the least overall change in the subregions (Fig. 5.4), we can still observe spatial redistribution of perfusion patterns. This fact highlights the importance of the introduced transparency aspect, which could be useful in clinical decisions.

The DICE scores presented in Table 5.3 confirm our initial visual assessment. Deep representations show less agreement with the shallow ones. It is worth noting that this is only indicatory of a joint agreement and could also simply mean that the low scores in deep methods were a consequence of the smaller number of deep representations in the joint ground-truth. Interestingly, we observed a good agreement

Vox	Filters	Patch	LBP	UNet	RandNet
0.76	0.68	0.73	0.74	0.58	0.56

Table 5.3: DICE scores for all methods under evaluation using the KM variant.

of all shallow representations, with Filters scoring slightly lower.

5.3.3 Individual evaluation

The SAD scores for all individual samples and methods are presented in Table 5.4. Visual representation of the data is presented in Fig. 5.4.

For the majority of the patients, the proposed method indicated no clear consensus on the degree of overall change in the tumour and different methods ranked the amount of change differently. Figure 5.1 shows that most smaller tumours ($\leq$ 20cm^3) received a decrease in WTV following 8-10 fractions of CRT, while large tumours increased in WTV, however the proposed method (Fig. 5.4) indicates no considerable differences between the groups of large and small tumours. In general, the EF metric remained inconclusive (Fig. 5.1) and no clear pattern emerged to conclusively indicate early lack of response.

A single patient's (ART_109, non-responder) overall absolute change in the fraction of tumour subregions was significantly different (t-test, p=0.002) to all other patients. The patient received a large amount of change and all proposed methods were in agreement (Fig. 5.4). In addition, a large decrease in subregion 2 was observed, which was shown to describe areas with high contrast uptake (Fig. 5.2). Patients WTV decreased more than the rest of the tumours, however the change was not significant (t-test, p>0.05) as compared to the rest of the patients in the smaller tumour group as well as all patients. The EF metric remained inconclusive, as while the patient received a decrease in EF (Fig. 5.1), other patients (and responders) received a similar decrease (e.g. ART_127, ART_113).

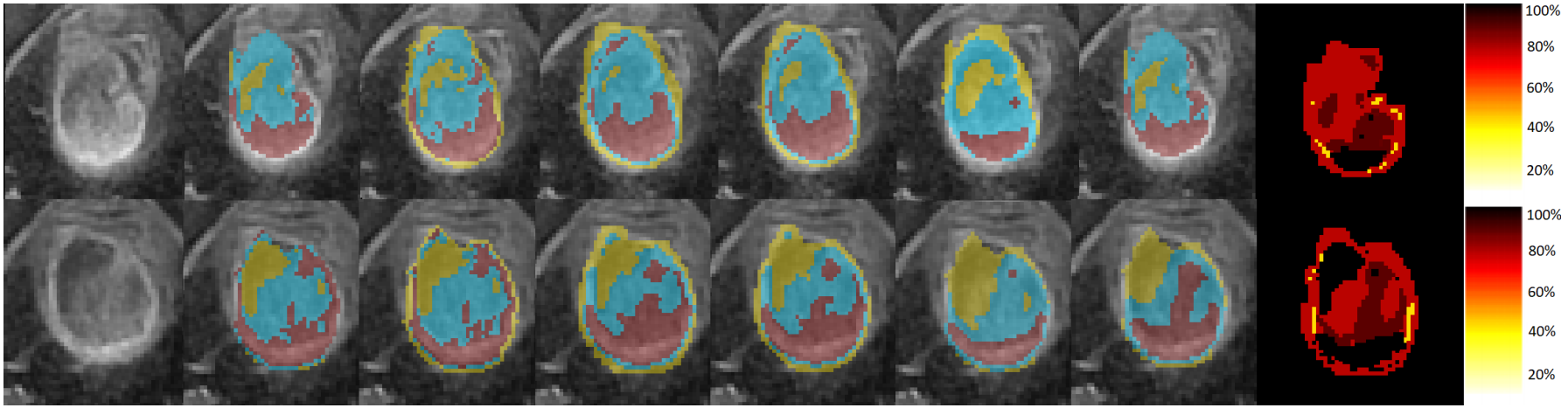


Figure 5.3: Sample 120 shown pre-therapy (top) and mid-therapy (bottom) for all the methods under evaluation. From left: mid-enhancement point, Vox, Gabor, Patch, LBP, UNet, RandNet, confidence heatmaps.

ID	Vox	Filters	Patch	LBP	UNet	RandNet
112	3.95	5.08	8.67	8.21	25.42	21.7
120	35.84	28.83	15.44	21.60	8.35	30.29
111	29.81	20.00	20.32	22.33	39.14	32.04
127	28.21	17.87	10.12	10.14	50.40	63.42
108	58.03	44.20	35.40	34.03	15.15	13.20
124	34.77	25.78	37.57	35.10	54.67	82.51
114	37.95	16.57	45.74	42.75	66.79	64.16
123	42.95	26.57	39.78	39.30	83.48	88.43
113	104.87	40.33	57.46	49.26	22.62	48.36
125	103.41	61.32	31.28	33.45	48.05	50.40
128	84.64	58.98	34.25	35.20	85.13	59.70
109	119.03	81.21	69.30	73.32	161.75	154.22

Table 5.4: SAD indicating percent of absolute change for each patients' tumour (rows) and each method (columns). Bold indicates patients who did not achieve a complete response at 3 months following CRT.

In the case of the other 2 non-responders (ART_108 and ART_112) no significant difference to other patients in the cohort was observed with the proposed method (although one patient (ART_112) was ranked by most methods as the least changed). Change in WTV also did not appear to be different for these two patients as compared to the rest of the cohort.

While the proposed method was able to identify a single non-responder as opposed to other traditional metrics, the other two patients who did not receive a complete response 3 months followith CRT were found to be no different than the rest of the cohort by both the proposed method as well as traditional metrics. This, however, does not exclude the possibility that a change occurred later throughout the therapy, as the evaluated scans were taken at a very early point during CRT (10-12 days). This suggests that while early prediction can correctly identify some non-responders, measurements should be taken frequently and throughout the therapy as the signs of early response to therapy may not necessarily lead to a complete response.

Additionally, while assessing the degree of heterogeneity during natural history and as unaffected by therapy in the previous chapters, we found that the tumours undergo a non-linear change in heterogeneity which varies dependening on the stage of the tumour. It is possible that the change in heterogenetiy following a therapy would also be different depending on the initial stage of the tumours. Although the acceptance criteria of this study included tumour stage $\geq$ T2N0, this still spans a number of possible stages affecting the interpretation of the data.

5.4 Conclusions

In the previous chapters we demonstrated that the introduced methods have a good ability to characterise the change in heterogeneity on controlled pre-clinical data. However, clinical studies are characterised with a large inter-sample variety arising

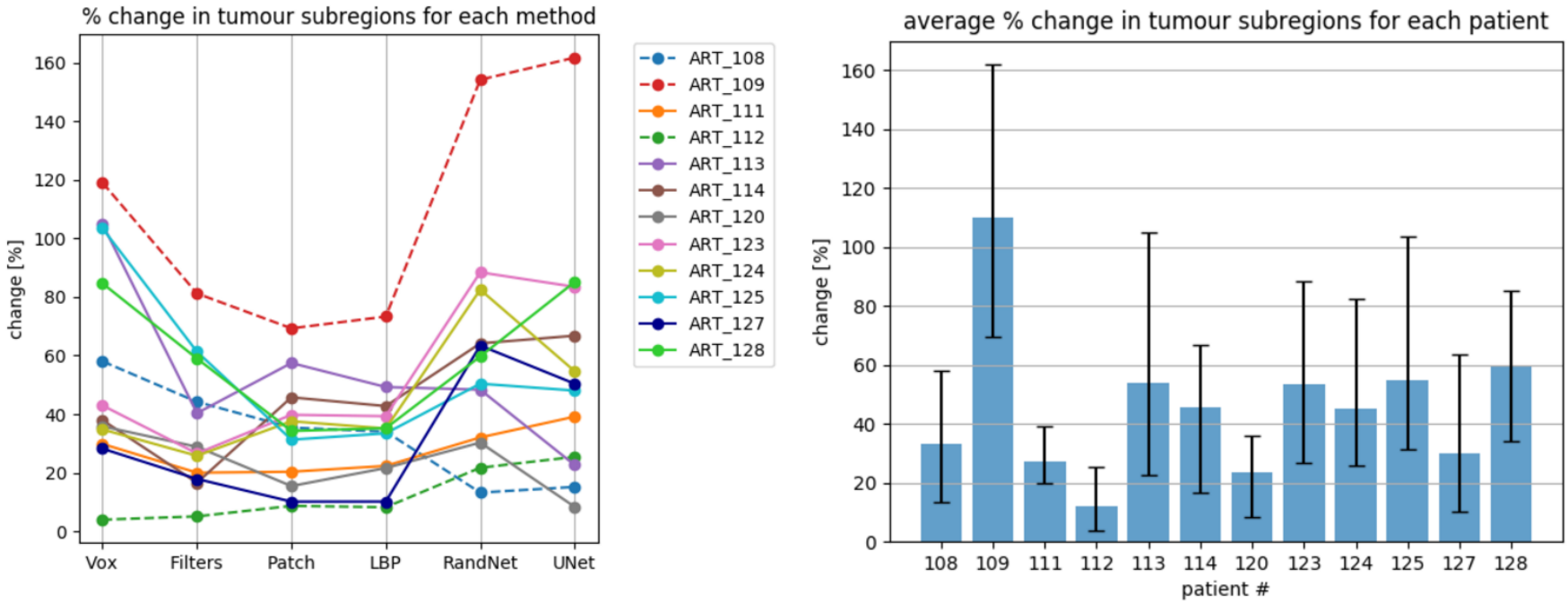


Figure 5.4: Left: SAD values indicating percent of absolute change for each patient tumour (legend) and each method (x axis). Dashed lines indicate patients who have not achieved a complete response following 3 months of CRT. Right: Average SAD values across all methods with min/max error bars for each patient (x axis).

from a number of sources. For that reason, consistent characterization of heterogeneity across patients becomes significantly more challenging. In this chapter we attempted to identify and flag early signs of lack of response following fraction 8-10 of CRT based on the total change in distributions of heterogeneous patterns. We have demonstrated the potential of the method to: a) characterise the degree of heterogeneity as well as subregional distributions of it; and b) flag signs of lack of early response in some patients. More importantly, however, we provided strategies to make the analysis as transparent as possible and to increase confidence in the results. Furthermore, we highlight in this work a number of important factors that should be considered prior to the analysis, such as the stage of the tumour. This becomes a basis for future directions and suggestions for controlled trials in the next chapter - conclusions.

While the introduced measurements could be related to the physiological processes in the tumour environment, a definitive validation remains to be conducted. While such validation is difficult for a number of reasons outlined earlier in this chapter (Sec. 5.1.1), we therefore provided a set of strategies to make the introduced tools as transparent as possible without the tools required for comprehensive validation.

We have shown that the transparency and confidence characteristics of our pipelines are beneficial in the analysis of such an extremely varied datasets. Firstly, the individual methods were not always consistent in their results. However, due to the ensemble strategy we took in this work we were able to characterise this change with greater confidence for the patients where all methods agreed. In addition, due to the transparency of our pipeline, we were able to observe the subtle changes in the segmentations even when overall change was minimal. Finally, with the knowledge gathered from characterizing heterogeneity during the natural history as unaffected by therapy, we suggest that initial state of the tumour should be a pre-requisite for further analysis and trials.

In conclusion, while the therapy response results are promising, the challenging problem of automatic therapy response assessment remains open. The task of therapy response assessment is extremely challenging and thus, in order to bridge the tranlational gap, the unparalleled value of extensive and rigorous validation of such results on large studies remains crucial and irreplaceable. However, we hope that this work provides a solid start towards mediating this process by demonstrating transparency and confidence strategies that can improve clinical decisions, which we have indeed found useful in this analysis. We believe that such strategies provide more information, confidence and transparency about the results, and aided by further validation and larger studies, could potentially pave the road for such pipelines to be translated to clinical practice. To that end, we hope our demonstration of sophisticated and robust pipelines that work well in the pre-clinical context with small datasets and no ground-truth could further contribute to future research in that context.

Chapter 6

Conclusions

In this work we applied texture-based deep and shallow image representations using the Bag of Visual Words pipeline to the problem of tumour heterogeneity assessment. In the following sections we first summarise the contributions of this thesis, and then provide an outlook for future work.

6.1 Summary and discussion

In the following section we summarise the contributions of this thesis. Sec. 6.1.1 outlines the technical contributions, while sec. 6.1.2 focuses on the biological findings that are more relevant clinically.

6.1.1 Technical

The first technical contribution of this thesis is the adaptation of a Bag of Visual Words (BoV) pipeline to the problem of assessing tumour perfusion heterogeneity. Additionally, we evaluated a number of feature descriptors and encoding variants together with the baseline voxel-based clustering methods often used in the challenging context of tumour heterogeneity assessment. Furthermore, we chose descriptors suit-

able for texture analysis as a proxy measure of heterogeneity. We demonstrated the potential of the method to visualise subregional distributions of different perfusion patterns that can be later quantified for assessing the overall degree of heterogeneity. We have shown that the introduced pipeline was capable of differentiating between different stages of growth. We further showed indirectly that such perfusion patterns could be related to the underlying processes in the micro-enviroment, where those processes can be characterised with varying degrees of heterogeneity. Finally, we demonstrated that methods with different underlying assumptions can have different results over the same dataset. However, we observed that such changes were subtle and interestingly the underlying perfusion patterns remain similar.

The second technical contribution of this work is the introduction of novel deep representations that, through a connection with shallow pipelines, allow for segmentations with no ground truth and, through adapting lightweight deep architectures, can perform classification with little training data. We demonstrated their ability to assess the change in tumour heterogeneity. Even in such a trivial form and with no training at all, the introduced methods were capable of outperforming the shallow representations. Furthermore, we have shown that by using specialised handcrafted feature descriptors as weights in a cascading architecture of a neural network, with no training, we were able to match the results of the methods from shallow representations. It is expected, that with more sophistication, such deep representations could achieve even better performance. Furthermore, such architectures could constitute a valuable alternative for problems with small datasets that are unsuitable for training or fine-tuning and with no ground truth. This could be particularly true for domains with a history of research on specialised feature descriptors. Finally, we have shown that while deep representations can easily outperfom the shallow ones, they should not be replaced entirely, as the transparent nature of shallow representations makes them more valuable in the clinical context.

The third and final contribution of this thesis is the demonstration of the potential of these methods on the more meaningful and clinically-relevant question of therapy response assessment. We evaluated all the previously reviewed methods on a clinical dataset. We observed that while deep and shallow BoV pipelines considerably improved the results over the standard voxel-based clustering in the pre-clinical dataset, they performed similarly to shallow representations when evaluated on the clinical dataset. This could be due to the different nature of the data (e.g. tumour type), but could also be a consequence of a different pharmacokinetic modelling approaches between the chapters. For that reason, we suggest that further work should be done to evaluate the performance of voxel-based clustering on semi-quantitative parameter maps over different tumour types. Furthermore, we demonstrated a consistently good performance of shallow and deep BoV pipelines in both clinical and pre-clinical settings, which demonstrates their superiority to the commonly used voxel-based clustering that in our experiments performed worse in the pre-clinical data. However, their ultimate evaluation in the clinical setting cannot be complete due to the small-sized dataset that does not support statistical testing. To this end, we consider the final clinical chapter more in an exploratory character and the value of its contribution arisies from its clinical findings. We used the knowledge about tumour growth we gathered in the previous chapters to understand the nature of response the therapy in the clinical data. The clinically-relevant findings are presented in the following section. In addition, we provided the strategy of using the methods introduced in the previous sections as an ensemble, thus allowing for generating multiple hypotheses that could guide clinical decisions better than a single, potentially erroneous, outcome.

6.1.2 Biological and clinical

In this work we have shown that there is a growth-related change in tumour perfusion patterns, that we motivated and used as a proxy-measure of heterogeneity at a macro-scale. Furthermore, we demonstrated that the change is consistent across the samples and significant enough to differentiate between the early and late growth stages. We further demonstrated the potential of this method to identify and flag early signs of lack of response in some patients, which is a more clinically-relevant question. While the quantitative results look promising, we also demonstrated that the true value of the introduced methods lies in their transparency, which can be exploited to better understand the results. This brought to our attention a number of important factors that should be considered in the analysis of heterogeneity that we oultine in the following paragraphs.

Cancer is an extremely complex disease as demonstrated by numerous studies. As a consequence, the arising heterogeneity, which is a composite measure of multiple processes in the underlying microenvironment of the tumour, is also very complex. It is therefore unlikely that one type of change in heterogeneity across many tumour types, stages and patients could be foundand used for ultimate evaluation. However, the majority of clinical studies of tumour heterogeneity do not account for those factors, which could be contributing to the current lack of repeatability of studies. For that reason, in order to better understand the relationship of heterogeneity and the outlined factors, an effort should be made to create a number of controlled studies. We make a few suggestions for these in the following paragraphs.

In Chapter 5 of this work we attempted to take the lessons learned in the evaluation of heterogeneity in the pre-clinical dataset, and apply it to clinical therapy response assessment. We made an observation that the change in heterogeneity during the growth process was varied in a non-linear way, which allowed us to hypothesise that such change following therapy response asssessment also varies in a way that de-

pends on the stage of the tumour. However, it is important to note that it remains unclear whether there is a constant relation between the growth-based change in a clinical and pre-clinical data. For that reason, it is important to provide future evaluation assessing the correlation between clinical and pre-clinical datasets. While such assessment would be impossible in terms of growth in clinical study, one could indirectly reason a relation by comparing pre-clinical growth and therapy data, with clinical therapy data. In addition, comparisons across different tumour types should be performed. Finally, we suggest that controlled studies evaluating growth stage correlation with the type of change in heterogeneity following therapy would be beneficial for better understanding the nature of heterogeneity in tumours.

In summary, following the findings of this work we make suggestions for future trials that evaluate a number of factors and their correlation. Mainly, evaluating the relationship between the change in heterogeneity following therapy in clinical and pre-clinical settings could help establish a better understanding of the effect of therapy. In addition, a pre-clinical trial evaluating the change in heterogeneity between different tumour types could be beneficial to better understanding the complexity of heterogeneity during growth between different types of cancer. Finally, a controlled study looking at differences in therapy response based on the growth stage of the tumour could prove useful for understanding the complexity of heterogeneity following therapy.

6.2 Limitations

While we demonstrated the potential of the method in both pre-clinical and clinical settings, there are limitations associated with it.

From a technical point of view, due to small size of some pre-clinical tumours, we were limited to the use of very small feature spaces (3x3), which is an issue particularly

in the case of filters. Additionally, due to the use of algorithms with random elements (e.g. K-means), this work lacks good repeatability. While the methods prove to be reasonably accurate, they have lower precision, especially when it comes to boundaries between tumour subregions.

From the clinical point of view, the methods were evaluated on datasets with a small number of mice/patients, which is a major limitation as it restricts the possibilities of statistical testing. In future work, a larger DCE-MRI dataset should be considered to enable statistical testing and validate the work. Finally, proper validation of the subregional segmentations is missing from this work. In future work, histology could be considered to validate it, however many problems exist with this type of validation. Primarily, they result from the large differences in the scale they reflect, slice misalignments as well as potential deformations introduced by tumour dissections.

6.3 Outlook

In this section we present recommendations for future work that could improve both the technical aspects of the methods, as well as our knowledge about tumour perfusion heterogeneity.

In this work, in order to address the large scale question of assessing perfusion heterogeneity, we assessed a wide range of approaches spanning many deep and shallow representations. However, at the cost of covering much breadth, future extensions of this work would benefit from in-depth evaluations outlined in the following paragraphs. Furthermore, we suggest further clinical trials that could be beneficial for better understanding of the problem.

6.3.1 Shallow encoding evaluation

Although the introduced methods on their own have shown to be sufficient to characterise the tumour growth with variation-based features, a number of further improvements could be done to the shallow representations. In this work, we have shown that even with simple orderless frequency encodings we can achieve a good growth-based tumour classification accuracy. While outside the scope of this work, the framework introduced here could further benefit from evaluation of more advanced feature encodings. A number of encoding methods, including VLAD or Fisher Vector, have been shown to outperform traditional BoV and significantly improve classification accuracy [95]. The authors of [95] argue that it is the introduction of advanced encodings that truly contributed to the improvement in accuracy performance in computer vision prior to the introduction of deep representations.

Furthermore, we have not explored the advantage of GMM soft-assignments in the context of BoV encodings which could be advantageous, especially in the context of noisy and uncertain visual data such as those arising from MRI images. In this sense, statistical assumptions of the MAP in the GMM could also be used to estimate the accuracy with which the assignments were made.

6.3.2 Deep hybrid representations

We demonstrated the ability of relatively simple deep respresentations to outperform the shallow ones. Furthermore, we have shown the potential of using specialised feature descriptors in a cascading structure of a neural networks for problems with small datasets and those lacking ground-truth. Naturally, the question arises whether combining the two approaches would bring an improvement in performance accuracy.

Such hybrid networks have previously been explored in [148] and [149], where the shallow information was introduced as one of the channels in the network. In this work, the performance of the methods could be further improved by aiding the

advantages arising from the hierarchical nature of deep representations with further information. Such representations, as mentioned in Chapter 4, have the advantage of learning directly from the data, therefore preventing human bias that assumptions in the design of handcrafted feature descriptors could introduce. At the same time, a large body of research on handcrafted descriptors could improve the performance for the domains they were designed for. Combining the two approaches has the benefit of retaining the advantages of both. For this work, we suggest that enhancing the deep representations with texture information could be beneficial specifically for assessing tumour heterogeneity, by explicitly accounting for that variation in addition to learning from the data directly. Furthermore, the deep information could also be enhanced with further pre-processing, such as supervoxels, that arguably could improve the results in noisy observations.

This work could further benefit from an in-depth evaluation and introduction of more sophisticated techniques beyond the proposed deep representations. The representations investigated in this work were only introduced in a very simple form that was already capable of matching the accuracy of shallow representations. It is therefore expected that with further sophistication, further improvement could be observed. Varying the depth of RandUNet could provide further insight. In addition, different types of handcrafted feature descriptors could be explored, and evaluating the method on other datasets could widen their potential, especially in domains that suffer from small datasets and/or lack of ground-truth.

6.3.3 Clinical evaluation

Finally, while in this work we demonstrated the potential of the introduced methods on the challenging problem of therapy response assessment and improved their clinical relevance by a number of transparency and confidence strategies, ultimately the translational gap cannot be bridged without further validation. To this end, we

suggested a number of controlled trials in Sec. 6.1.2 of this chapter. In summary, establishing the relation of heterogeneity in the context of growth between different tumour types, clinical and pre-clinical datasets, as well as in the context of therapy response is of crucial importance to better explain the complexity of the heterogeneity as a composite of underlying processes. Additionally, evaluation of the change of heterogeneity in different tumour stages following therapy could be of great value. Furthermore, a thorough validation of these methods with histological data would greatly improve the value of the methods in a clinical context. To that end, we hope that our demonstration of the potential of the introduced methods in the pre-clincial and clinical setting, as well as strategies we employed to improve transparency of the research could contribute to future interest and further work in introduced methods based on deep and shallow BoV representations to assess the spatial heterogeneity.

References

[1] James PB O'Connor, Chris J Rose, John C Waterton, Richard AD Carano, Geoff JM Parker, and Alan Jackson. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clinical Cancer Research*, 21(2):249–257, 2015.

[2] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[3] Clifford J Eskey, Alan P Koretsky, Michael M Domach, and Rakesh K Jain. 2H-nuclear magnetic resonance imaging of tumor blood flow: spatial and temporal heterogeneity in a tissue-isolated mammary adenocarcinoma. *Cancer research*, 52(21):6010–6019, 1992.

[4] Kjetil G Brurberg, Jon-Vidar Gaustad, Camilla S Mollatt, and Einar K Rofstad. Temporal heterogeneity in blood supply in human tumor xenografts. *Neoplasia*, 10(7):727–735, 2008.

[5] James PB O'Connor, Eric O Aboagye, Judith E Adams, Hugo JWL Aerts, Sally F Barrington, Ambros J Beer, Ronald Boellaard, Sarah E Bohndiek, Michael Brady, Gina Brown, et al. Imaging biomarker roadmap for cancer studies. *Nature reviews Clinical oncology*, 14(3):169–186, 2017.

[6] Michail Shipitsin, Lauren L Campbell, Pedram Argani, Stanislawa Weremowicz, Noga Bloushtain-Qimron, Jun Yao, Tatiana Nikolskaya, Tatiana Serebryiskaya,

Rameen Beroukhim, Min Hu, et al. Molecular definition of breast tumor heterogeneity. *Cancer cell*, 11(3):259–273, 2007.

[7] Robert A Gatenby, Ariosto S Silva, Robert J Gillies, and B Roy Frieden. Adaptive therapy. *Cancer research*, 69(11):4894–4903, 2009.

[8] L Simpson-Herren, PE Noker, and SD Wagoner. Variability of tumor response to chemotherapy ii. contribution of tumor heterogeneity. *Cancer chemotherapy and pharmacology*, 22(2):131–136, 1988.

[9] James PB O'Connor, Alan Jackson, Marie-Claude Asselin, David L Buckley, Geoff JM Parker, and Gordon C Jayson. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *The lancet oncology*, 9(8):766–776, 2008.

[10] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[11] Lejla Alic, Wiro J Niessen, and Jifke F Veenland. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PloS one*, 9(10):e110300, 2014.

[12] David Borland and Russell M Taylor Ii. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[13] Mark A Zahra, Kieren G Hollingsworth, Evis Sala, David J Lomas, and Li T Tan. Dynamic contrast-enhanced mri as a predictor of tumour response to radiotherapy. *The lancet oncology*, 8(1):63–74, 2007.

[14] JPB O'Connor, CJ Rose, A Jackson, Y Watson, S Cheung, F Maders, BJ Whitcher, C Roberts, GA Buonaccorsi, G Thompson, et al. Dce-mri biomarkers of tumour heterogeneity predict crc liver metastasis shrinkage following bevacizumab and folfox-6. *British journal of cancer*, 105(1):139–145, 2011.

[15] James PB O'connor, Alan Jackson, Geoff JM Parker, Caleb Roberts, and Gordon C Jayson. Dynamic contrast-enhanced mri in clinical trials of antivascular therapies. *Nature reviews Clinical oncology*, 9(3):167–177, 2012.

[16] Lejla Alić. *Quantification of Tumour Heterogeneity in MRI*. Rotterdam: Erasmus Universiteit Rotterdam, 2013.

[17] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019.

[18] Vincent Andrearczyk and Paul F Whelan. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69, 2016.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[20] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[21] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[22] Robert Weinberg. *The biology of cancer*. Garland science, 2013.

[23] Gloria H Heppner and Bonnie E Miller. Tumor heterogeneity. *Cancer and Metastasis Reviews*, 2(1):5–23, 1984.

[24] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338, 2013.

[25] Rakesh Kumar, Hiroki Kuniyasu, Corazon D Bucana, Michael R Wilson, and Isaiah J Fidler. Spatial and temporal expression of angiogenic molecules during tumor growth and progression. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 10(6):301–311, 1998.

[26] Anne Eberhard, Sebastian Kahlert, Valentin Goede, Bernhard Hemmerlein, Karl H Plate, and Hellmut G Augustin. Heterogeneity of angiogenesis and blood vessel maturation in human tumors: implications for antiangiogenic tumor therapies. *Cancer research*, 60(5):1388–1393, 2000.

[27] Maria Picchio, Roswitha Beck, Roland Haubner, Stefan Seidl, Hans-Jurgen Machulla, Timothy D Johnson, Hans-Jurgen Wester, Gerald Reischl, Markus Schwaiger, and Morand Piert. Intratumoral spatial distribution of hypoxia and angiogenesis assessed by 18f-faza and 125i-gluco-rgd autoradiography. *Journal of Nuclear Medicine*, 49(4):597–605, 2008.

[28] M Christiane Brahimi-Horn, Johanna Chiche, and Jacques Pouysségur. Hypoxia and cancer. *Journal of molecular medicine*, 85(12):1301–1307, 2007.

[29] Naoyo Nishida, Hirohisa Yano, Takashi Nishida, Toshiharu Kamura, and Masamichi Kojiro. Angiogenesis in cancer. *Vascular health and risk management*, 2(3):213–219, 2006.

[30] Xiangyu Yang and Michael V Knopp. Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review. *BioMed Research International*, 2011:1–12, 2011.

[31] James PB O'Connor. Cancer heterogeneity and imaging. In *Seminars in cell & developmental biology*, volume 64, pages 48–57. Elsevier, 2017.

[32] Philippe L Bedard, Aaron R Hansen, Mark J Ratain, and Lillian L Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, 2013.

[33] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory investigation*, 95(4):377–384, 2015.

[34] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.

[35] Mahul B Amin and Stephen B Edge. AJCC cancer staging manual. 2017.

[36] AB Miller, BFAU Hoogstraten, MFAU Staquet, and A Winkler. Reporting results of cancer treatment. *Cancer*, 47(1):207–214, 1981.

[37] Patrick Therasse, Susan G Arbuck, Elizabeth A Eisenhauer, Jantien Wanders, Richard S Kaplan, Larry Rubinstein, Jaap Verweij, Martine Van Glabbeke, Allan T van Oosterom, Michaele C Christian, et al. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216, 2000.

[38] Rakesh Kumar, Hiroki Kuniyasu, Corazon D Bucana, Michael R Wilson, and Isaiah J Fidler. Spatial and temporal expression of angiogenic molecules during tumor growth and progression. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 10(6):301–311, 1998.

[39] Rakesh K Jain et al. Antiangiogenic therapy for cancer: current and emerging concepts. *Oncology-Melville*, 19(3):7–16, 2005.

[40] Perry W Grigsby, Carlos A Perez, KS Clifford Chao, Alaa Elbendary, Thomas J Herzog, Janet S Rader, and David G Mutch. Lack of effect of tumor size on the prognosis of carcinoma of the uterine cervix stage ib and iia treated with preoperative irradiation and surgery. *International Journal of Radiation Oncology* Biology* Physics*, 45(3):645–651, 1999.

[41] Tobias Klatte, Jean-Jacques Patard, Michela De Martino, Karim Bensalah, Gregory Verhoest, Alexandre De La Taille, Clément-Claude Abbou, Ernst Peter Allhoff, Giuseppe Carrieri, Stephen B Riggs, et al. Tumor size does not predict risk of metastatic disease or prognosis of small renal cell carcinomas. *The Journal of urology*, 179(5):1719–1726, 2008.

[42] William D Foulkes, Matthew J Grainge, Emad A Rakha, Andrew R Green, and Ian O Ellis. Tumor size is an unreliable predictor of prognosis in basal-like breast cancers and does not correlate closely with lymph node status. *Breast cancer research and treatment*, 117(1):199–204, 2009.

[43] Ash A Alizadeh, Victoria Aranda, Alberto Bardelli, Cedric Blanpain, Christoph Bock, Christine Borowski, Carlos Caldas, Andrea Califano, Michael Doherty, Markus Elsner, et al. Toward understanding and exploiting tumor heterogeneity. *Nature medicine*, 21(8):846–853, 2015.

[44] Melissa R Junttila and Frederic J de Sauvage. Influence of tumour microenvironment heterogeneity on therapeutic response. *Nature*, 501(7467):346–354, 2013.

[45] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine*, 366(10):883–892, 2012.

[46] M Angell, JP Kassirer, and AS Relman. Looking back on the millennium in medicine [editorial]. *N Engl J Med*, 342(1):42–49, 2000.

[47] Sally F Barrington, N George Mikhaeel, Lale Kostakoglu, Michel Meignan, Martin Hutchings, Stefan P Müeller, Lawrence H Schwartz, Emanuele Zucca, Richard I Fisher, Judith Trotman, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. *Journal of clinical oncology*, 32(27):3048–3058, 2014.

[48] Leonardo Iaccarino, Arianna Sala, Silvia Paola Caminiti, and Daniela Perani. The emerging role of pet imaging in dementia. *F1000Research*, 6, 2017.

[49] Egesta Lopci, Cristina Nanni, Paolo Castellucci, Gian Carlo Montini, Vincenzo Allegri, Domenico Rubello, Franca Chierichetti, Valentina Ambrosini, and Stefano Fanti. Imaging with non-fdg pet tracers: outlook for current clinical applications. *Insights into imaging*, 1(5-6):373–385, 2010.

[50] Kathryn L Wallitt, Sairah R Khan, Suraiya Dubash, Henry H Tam, Sameer Khan, and Tara D Barwick. Clinical pet imaging in prostate cancer. *RadioGraphics*, 37(5):1512–1536, 2017.

[51] JS Lewis and MJ Welch. Pet imaging of hypoxia. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging*, 45(2):183–188, 2001.

[52] Mian M Alauddin. Positron emission tomography (pet) imaging with 18f-based radiotracers. *American journal of nuclear medicine and molecular imaging*, 2(1):55–76, 2012.

[53] Noboru Hara, Mina Okuizumi, Hiroshi Koike, Makoto Kawaguchi, and Vladimir Bilim. Dynamic contrast-enhanced magnetic resonance imaging (dce-mri) is a useful modality for the precise detection and staging of early prostate cancer. *The Prostate*, 62(2):140–147, 2005.

[54] Paul S Tofts. Modeling tracer kinetics in dynamic gd-dtpa mr imaging. *Journal of magnetic resonance imaging*, 7(1):91–101, 1997.

[55] Paul S Tofts, Gunnar Brix, David L Buckley, Jeffrey L Evelhoch, Elizabeth Henderson, Michael V Knopp, Henrik BW Larsson, Ting-Yim Lee, Nina A Mayr, Geoffrey JM Parker, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted mri of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(3):223–232, 1999.

[56] Erez Eyal, Daria Badikhi, Edna Furman-Haran, Fredrick Kelcz, Kevin J Kirshenbaum, and Hadassa Degani. Principal component analysis of breast dce-mri adjusted with a model-based method. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 30(5):989–998, 2009.

[57] Tong San Koh, Choon Hua Thng, Juliana TS Ho, Puay Hoon Tan, Helmut Rumpel, and James BK Khoo. Independent component analysis of dynamic

contrast-enhanced magnetic resonance images of breast carcinoma: A feasibility study. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 28(1):271–277, 2008.

[58] Steven P Sourbron and David L Buckley. Classic models for dynamic contrast-enhanced mri. *NMR in Biomedicine*, 26(8):1004–1027, 2013.

[59] Melanie Heilmann, Christine Walczak, Julien Vautier, Jean-Luc Dimicoli, Carole D Thomas, Mihaela Lupu, Joël Mispelter, and Andreas Volk. Simultaneous dynamic t 1 and t 2* measurement for aif assessment combined with dce mri in a mouse tumor model. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 20(4):193–203, 2007.

[60] Matthew R Orton, James A d'Arcy, Simon Walker-Samuel, David J Hawkes, David Atkinson, David J Collins, and Martin O Leach. Computationally efficient vascular input function models for quantitative kinetic modelling using dce-mri. *Physics in Medicine & Biology*, 53(5):1225–1239, 2008.

[61] Benjamin Irving, Lydia Tanner, Monica Enescu, Manav Bhushan, Esme J Hill, Jamie Franklin, Ewan M Anderson, Ricky A Sharma, Julia A Schnabel, and Michael Brady. Personalised estimation of the arterial input function for improved pharmacokinetic modelling of colorectal cancer using dcemri. In *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pages 126–135. Springer, 2013.

[62] Steven P Sourbron and David L Buckley. On the scope and interpretation of the tofts models for dce-mri. *Magnetic resonance in medicine*, 66(3):735–745, 2011.

[63] Susan M Galbraith, Martin A Lodge, N Jane Taylor, Gordon JS Rustin, Søren Bentzen, J James Stirling, and Anwar R Padhani. Reproducibility of dynamic

contrast-enhanced mri in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR in Biomedicine*, 15(2):132–142, 2002.

[64] Chong Duan, Jesper F Kallehauge, G Larry Bretthorst, Kari Tanderup, Joseph JH Ackerman, and Joel R Garbow. Are complex dce-mri models supported by clinical data? *Magnetic resonance in medicine*, 77(3):1329–1339, 2017.

[65] Jesper Folsted Kallehauge, Kari Tanderup, Chong Duan, Søren Haack, Erik Morre Pedersen, Jacob Christian Lindegaard, Lars Ulrik Fokdal, Sandy Mohamed Ismail Mohamed, and Thomas Nielsen. Tracer kinetic model selection for dynamic contrast-enhanced magnetic resonance imaging of locally advanced cervical cancer. *Acta Oncologica*, 53(8):1064–1072, 2014.

[66] Harry Strange and Reyer Zwiggelaar. *Open Problems in Spectral Dimensionality Reduction*. Springer, 2014.

[67] Lejla Alic, Martin van Vliet, CF Van Dijke, AMM Eggermont, JF Veenland, and WJ Niessen. Heterogeneity in dce-mri parametric maps: a biomarker for treatment response? *Physics in Medicine & Biology*, 56(6):1601–1616, 2011.

[68] Samina Khokher, Muhammad Usman Qureshi, and Naseer Ahmad Chaudhry. Comparison of who and recist criteria for evaluation of clinical response to chemotherapy in patients with advanced breast cancer. *Asian Pacific Journal of Cancer Prevention*, 13:3213–3218, 2012.

[69] Gordon C Jayson, Cong Zhou, Alison Backen, Laura Horsley, Kalena Marti-Marti, Danielle Shaw, Nerissa Mescallado, Andrew Clamp, Mark P Saunders, Juan W Valle, et al. Plasma tie2 is a tumor vascular response biomarker for vegf

inhibitors in metastatic colorectal cancer. *Nature communications*, 9(1):1–14, 2018.

[70] Richard L Wahl, Heather Jacene, Yvette Kasamon, and Martin A Lodge. From recist to percist: evolving considerations for pet response criteria in solid tumors. *Journal of nuclear medicine: official publication, Society of Nuclear Medicine*, 50(Suppl 1):122S, 2009.

[71] Jon-Vidar Gaustad, Ilana C Benjaminsen, Bjørn A Graff, Kjetil G Brurberg, Else-Beate M Ruud, and Einar K Rofstad. Intratumor heterogeneity in blood perfusion in orthotopic human melanoma xenografts assessed by dynamic contrast-enhanced magnetic resonance imaging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 21(6):792–800, 2005.

[72] Susan G Orel, Nicole Kay, Carol Reynolds, and Daniel C Sullivan. Bi-rads categorization as a predictor of malignancy. *Radiology*, 211(3):845–850, 1999.

[73] Matthias Elter and Alexander Horsch. Cadx of mammographic masses and clustered microcalcifications: a review. *Medical physics*, 36(6Part1):2052–2068, 2009.

[74] Paul Workman, Eric O Aboagye, Yuen-Li Chung, John R Griffiths, Rachel Hart, Martin O Leach, Ross J Maxwell, Paul MJ McSheehy, Pat M Price, and Jamal Zweit. Minimally invasive pharmacokinetic and pharmacodynamic technologies in hypothesis-testing clinical trials of innovative therapies. *Journal of the National Cancer Institute*, 98(9):580–598, 2006.

[75] John C Waterton and Liisa Pylkkanen. Qualification of imaging biomarkers for oncology drug development. *European journal of cancer*, 48(4):409–415, 2012.

[76] Alan Jackson, James PB O'Connor, Geoff JM Parker, and Gordon C Jayson. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. *Clinical Cancer Research*, 13(12):3449–3459, 2007.

[77] Rakesh K Jain, Dan G Duda, Christopher G Willett, Dushyant V Sahani, Andrew X Zhu, Jay S Loeffler, Tracy T Batchelor, and A Gregory Sorensen. Biomarkers of response and resistance to antiangiogenic therapy. *Nature reviews Clinical oncology*, 6(6):327–338, 2009.

[78] David Checkley et al. Use of dynamic contrast-enhanced mri to evaluate acute treatment with zd6474, a vegf signalling inhibitor, in pc-3 prostate tumours. *British journal of cancer*, 89(10):1889, 2003.

[79] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[80] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[81] Theodore Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, 105(1):156–166, 1989.

[82] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[83] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[84] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44, 2001.

[85] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *International journal of computer vision*, 62(1-2):61–81, 2005.

[86] Andre Hallack, Bartłomiej W Papież, James Wilson, Lai Mun Wang, Tim Maughan, Mark J Gooding, and Julia A Schnabel. Correlating tumour histology and ex vivo mri using dense modality-independent patch-based descriptors. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 137–145. Springer, 2015.

[87] Dictionary.com. https://www.dictionary.com/browse/texture. Accessed: 2015-06-02.

[88] Manik Varma and Andrew Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *European Conference on Computer Vision*, pages 255–271. Springer, 2002.

[89] Xianghua Xie and Majid Mirmehdi. A galaxy of texture features. In *Handbook of texture analysis*, pages 375–406. World Scientific, 2008.

[90] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, page 1470. IEEE, 2003.

[91] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[92] Bela Julesz. Visual pattern discrimination. *IRE transactions on Information Theory*, 8(2):84–92, 1962.

[93] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[94] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[95] Ken Chatfield, Victor S Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, volume 2, page 8, 2011.

[96] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[97] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2010.

[98] David G Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, page 1150. Ieee, 1999.

[99] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 18(10):959–971, 1996.

[100] Li Liu and Paul Fieguth. Texture classification from random features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):574–586, 2012.

[101] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[102] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716, 2012.

[103] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[104] Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

[105] Song-Chun Zhu, Cheng-En Guo, Yizhou Wang, and Zijian Xu. What are textons? *International Journal of Computer Vision*, 62(1-2):121–143, 2005.

[106] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 18(10):959–971, 1996.

[107] Manik Varma and Andrew Zisserman. Texture classification: Are filter banks necessary? In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–691. IEEE, 2003.

[108] Li Liu, Lingjun Zhao, Yunli Long, Gangyao Kuang, and Paul Fieguth. Extended local binary patterns for texture classification. *Image and Vision Computing*, 30(2):86–99, 2012.

[109] Veerle Kersemans, Pavitra Kannan, John S Beech, Russell Bates, Benjamin Irving, Stuart Gilchrist, Philip D Allen, James Thompson, Paul Kinchesh, Christophe Casteleyn, et al. Improving in vivo high-resolution ct imaging of the tumour vasculature in xenograft mouse models through reduction of motion and bone-streak artefacts. *PloS one*, 10(6):e0128537, 2015.

[110] Vasily L Yarnykh. Actual flip-angle imaging in the pulsed steady state: a method for rapid three-dimensional mapping of the transmitted radiofrequency field. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 57(1):192–200, 2007.

[111] Melanie Heilmann, Christine Walczak, Julien Vautier, Jean-Luc Dimicoli, Carole D Thomas, Mihaela Lupu, Joël Mispelter, and Andreas Volk. Simultaneous dynamic t 1 and t 2* measurement for aif assessment combined with dce mri in a mouse tumor model. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 20(4):193–203, 2007.

[112] SB Donaldson, David L Buckley, JP O'connor, Susan E Davidson, Bernadette M Carrington, Andrew P Jones, and Catharine ML West. Enhancing fraction measured using dynamic contrast-enhanced mri predicts disease-free survival in patients with carcinoma of the cervix. *British journal of cancer*, 102(1):23–26, 2010.

[113] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[114] Bjoern Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elisabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian Avants, Nicholas Ayache, Patricia Buendia, Louis Collins, Nicolas

Cordier, Jason Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Cagatay Demiralp, Christopher Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan Iftekharuddin, Raj Jena, Nigel John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, S. J. Price, Tammy Riklin-Raviv, Syed Reza, Michael Ryan, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Silva, Nuno Sousa, Nagesh Subbanna, Gabor Szekely, Thomas Taylor, Owen Thomas, Nicholas Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, pages 1993–2024, 2014.

[115] Benjamin Irving. maskslic: regional superpixel generation with application to local pathology characterisation in medical images. *arXiv preprint arXiv:1606.09518*, 2016.

[116] Benjamin Irving, Jola Mirecka, Ana Gomes, Danny Allen, Paul Kinchesh, Veerle Kesermans, Stuart Gilchrist, Sean Smart, Julia Schnabel, Mike Brady, and Michael Chappell. Perfusion-supervoxels for dce-mri based tumor subregion assessment. In *In proceedings of the 25th Annual Meeting of ISMRM*, 2017.

[117] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[118] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[119] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[120] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836, 2015.

[121] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[122] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[123] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[124] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[125] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[126] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[127] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[128] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[129] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

[130] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.

[131] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1233–1240, 2013.

[132] David Novotny, Diane Larlus, and Andrea Vedaldi. I have seen enough: Transferring parts across categories. 2016.

[133] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[134] Paolo Napoletano. Hand-crafted vs learned descriptors for color texture classification. In *International Workshop on Computational Color Imaging*, pages 259–271. Springer, 2017.

[135] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014.

[136] Sarath Shekkizhar. Fcn tensorflow. https://github.com/shekkizh/FCN.tensorflow. Accessed: 2018-08-20.

[137] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015.

[138] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[139] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[140] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[141] Cheryl Sital, Tom Brosch, Dominique Tio, Alexander Raaijmakers, and Jürgen Weese. 3d medical image segmentation with labeled and unlabeled data using autoencoders at the example of liver segmentation in ct images. *arXiv preprint arXiv:2003.07923*, 2020.

[142] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[143] Jolanta Mirecka, Erika Denton, Reyer Zwiggelaar, and Harry Strange. Random projections for mammographic texture segmentation and high/low risk classification. In *In Proceedings of MICCAI Breast Image Analysis (BIA)*, 2015.

[144] Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[145] François Chollet et al. Keras. https://keras.io, 2015.

[146] Bartłomiej W Papież, James M Franklin, Mattias P Heinrich, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Gifted demons: deformable image registration with local structure-preserving regularization using supervoxels for liver applications. *Journal of Medical Imaging*, 5(2):024001, 2018.

[147] Bartłomiej W Papież, Mattias P Heinrich, Jérome Fehrenbach, Laurent Risser, and Julia A Schnabel. An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration. *Medical image analysis*, 18(8):1299–1311, 2014.

[148] Amarjot Singh, Devamanyu Hazarika, and Aniruddha Bhattacharya. Texture and structure incorporated scatternet hybrid deep learning network (ts-shdl) for brain matter segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1181–1188, 2017.

[149] Amarjot Singh and Nick Kingsbury. Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2991–2995. IEEE, 2018.