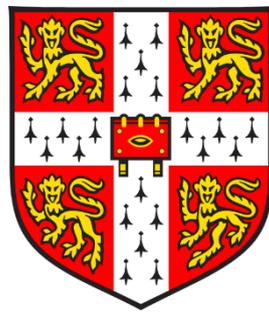


**Development of computer-based algorithms for
unsupervised assessment of radiotherapy contouring**



Huiqi Yang

Department of Oncology
University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Emmanuel College

Sep 2018

Development of computer-based algorithms for unsupervised assessment of radiotherapy contouring

Huiqi Yang

Abstract

INTRODUCTION: Despite the advances in radiotherapy treatment delivery, target volume delineation remains one of the greatest sources of error in the radiotherapy delivery process, which can lead to poor tumour control probability and impact clinical outcome. Contouring assessments are performed to ensure high quality of target volume definition in clinical trials but this can be subjective and labour-intensive.

This project addresses the hypothesis that computational segmentation techniques, with a given prior, can be used to develop an image-based tumour delineation process for contour assessments. This thesis focuses on the exploration of the segmentation techniques to develop an automated method for generating reference delineations in the setting of advanced lung cancer. The novelty of this project is in the use of the initial clinician outline as a prior for image segmentation.

METHODS: Automated segmentation processes were developed for stage II and III non-small cell lung cancer using the IDEAL-CRT clinical trial dataset. Marker-controlled watershed segmentation, two active contour approaches (edge- and region-based) and graph-cut applied on superpixels were explored. k-nearest neighbour (k-NN) classification of tumour from normal tissues based on texture features was also investigated.

RESULTS: 63 cases were used for development and training. Segmentation and classification performance were evaluated on an independent test set of 16 cases. Edge-based active contour segmentation achieved highest Dice similarity coefficient of 0.80 ± 0.06 , followed by graph-cut at 0.76 ± 0.06 , watershed at 0.72 ± 0.08 and region-based active contour at 0.71 ± 0.07 , with mean computational times of 192 ± 102 sec, 834 ± 438 sec, 21 ± 5 sec and 45 ± 18 sec per case respectively. Errors in accuracy of irregularly shaped lesions and segmentation leakages at the mediastinum were observed.

In the distinction of tumour and non-tumour regions, misclassification errors of 14.5% and 15.5% were achieved using 16- and 8-pixel regions of interest (ROIs) respectively. Higher misclassification errors of 24.7% and 26.9% for 16- and 8-pixel ROIs were obtained in the analysis of the tumour boundary.

CONCLUSIONS: Conventional image-based segmentation techniques with the application of priors are useful in automatic segmentation of tumours, although further developments are required to improve their performance. Texture classification can be useful in distinguishing tumour from non-tumour tissue, but the segmentation task at the tumour boundary is more difficult. Future work with deep-learning segmentation approaches need to be explored.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit of 60,000 for the Clinical Medicine and Clinical Veterinary Medicine for the PhD degree.

Huiqi Yang

Sep 2018

Acknowledgements

I am deeply indebted to my supervisors Dr Raj Jena and Prof Peter Hoskin whose guidance and insights have been instrumental in the development of my knowledge and skills. I have learnt extensively from the numerous stimulating and motivational discussions with Dr Jena, who has not only been sharing his expertise in the field but has also been highly supportive of new ideas. I am also extremely grateful to Prof Hoskin for his unwavering support, encouragement and advice throughout the course of my thesis, and it has been inspirational to work with him.

I am immensely grateful to Mrs Elizabeth Miles, Dr Karen Venables and the rest of the RTTQA team for giving me the opportunity to pursue this work as well as their support during my term as a clinical fellow. I would also like to thank the IDEAL-CRT and ISTART trial management groups for allowing me the use of the data for analysis.

I am extremely appreciative of Dr Angelica Aviles-Rivero and Dr Jingjing Zou for their advice on image processing techniques and study design, as well as Dr Gerry Lowe and Dr Andrew Shah on imaging physics. I would also like to sincerely thank Dr Antonios Georgantzoglou and Dr Mohammad Al Sa'd for assisting me with the initial scripting in MATLAB, and Mark Hayes for his generosity with access to the facilities at the Maxwell Centre.

Additionally, I would like to thank Michael Brown for his help with the treatment planning system as well as Richard Williams for his attempts at retrieving the data for when I encountered hard disk failure.

Last but not least, I am also thankful to have the support of my parents, sister, brother and my two nephews who have been tremendously understanding and encouraging throughout the course of this work and beyond.

Table of Contents

Table of Contents	i
List of figures.....	vii
List of tables.....	xvii
Abbreviations	xx
Chapter 1	1
Introduction.....	1
1.1 Advances in radiotherapy	1
1.2 Target definition in radiotherapy	1
1.3 Need for outlining assessment in radiotherapy	2
1.4 Radiotherapy trial quality assurance	3
1.5 The challenge of real time assessment of contouring	4
1.6 Lung cancer as tumour site focus for this study.....	5
1.7 Cross-sectional anatomy in the context of lung cancer	5
Aims.....	8
Overview of project.....	8
Literature review	9
1.8 General principles of image segmentation techniques.....	9
1.9 Segmentation of lung lesions – CT segmentation techniques	21
1.10 Segmentation of nodal lesions	35
1.11 PET-CT imaging	38
1.12 Commercial software products	39
Challenges of this project	40
Thesis structure	41
Chapter 2	42
Introduction.....	42
2.1 Use of MATLAB	42
2.2 Assessment of data heterogeneity	43
2.2.1 Imaging acquisition heterogeneity	43
2.3 Considerations for image enhancement	43

2.4 GTV and normal tissue density distribution	44
2.5 Summary of tasks.....	44
Methods.....	45
2.6 Clinical datasets	45
2.7 Study design.....	45
2.8 Clinical and imaging parameters.....	45
2.9 Gold-standard reference ROIs	46
2.10 Data handling	47
2.11 Image quality assessment	50
2.12 Statistics of GTV and surrounding tissues.....	51
Results of data assessment exercise	52
2.13 Task A.1 Data import and handling	52
2.14 Task A.2 Determination of heterogeneity and imaging quality of cases to assess need for image enhancement.....	56
2.14.1 Clinical characteristics of total data	56
2.14.2 Scanning parameters of total data	56
2.14.3 Image quality assessment.....	57
2.15 Task A.3 Determination of the descriptive statistics of the GTV and surrounding tissues	59
2.15 Task A.4 Determination of thresholding procedures.....	59
2.15.1 Solid versus non-solid lesions	62
Discussion.....	66
Conclusions.....	67
Chapter 3	68
Introduction.....	68
3.1 Edge-based segmentation techniques	68
3.2 Watershed segmentation	70
3.3 Active contour segmentation.....	72
3.4 Graph-cut segmentation.....	75
3.5 Summary of tasks.....	76
Methods.....	77
3.6 Pre-processing for watershed, active contour and graph-cut segmentation.....	77
3.7 Division of data.....	79

3.8 Watershed segmentation workflow	80
3.9 Active contour segmentation workflow	86
3.10 Graph-cut segmentation workflow	92
3.11 Assessment of segmentation	99
Results	102
3.12 Task B.1 Development and tuning of segmentation techniques on training dataset	102
3.12.1 Watershed segmentation	102
3.12.2 Chan-Vese active contour	117
3.12.3 Edge-based active contour	133
3.12.4 Graph-cut segmentation	149
Discussion.....	167
Chapter 4	168
Introduction.....	168
4.1 Summary of tasks.....	168
Results	169
4.2 Task C.1 Comparison of overall performance between different segmentation techniques on independent test dataset	169
4.2.1 Overall performance.....	169
4.2.2 Individual case performance.....	169
4.2.3 Computational time	175
4.3 Summary of performance for different segmentation approaches	175
Discussion.....	177
Conclusions.....	182
Chapter 5	183
Introduction.....	183
5.1 Summary of tasks.....	183
Methods.....	184
5.2 Gold-standard reference ROIs	184
5.3 Segmentation workflow and assessment	184
5.4 Datasets	184
5.5 Computational time	184

Results	185
5.6 Task D.1 Comparison of segmentation techniques on isolated peripheral primary disease	185
5.6.1 Performance of training dataset	185
5.6.2 Performance of independent test dataset	188
5.7 Task D.2 Comparison of segmentation techniques on isolated peripheral primary disease with dataset from different trial source	189
5.7.1 Clinical and scanning parameters of ISTART dataset	189
5.7.2 Performance of ISTART dataset	190
5.8 Task D.3 Comparison of segmentation of advanced lung tumour versus isolated peripheral lung tumours.....	193
Discussion.....	195
Conclusions.....	195
Chapter 6	196
Introduction.....	196
6.1 Summary of tasks.....	197
Methods.....	199
6.2 Study design - Feature selection and classification using LDA in MaZda.....	199
6.3 Processing of CT images.....	199
6.4 Generation of ROIs – BMP files	200
6.5 Generation of ROIs – ROI files	200
6.6 Generation of texture features	201
6.7 Feature reduction.....	202
6.8 Classification – COST B11 programme.....	203
6.9 Classification - MATLAB.....	204
6.10 Classifier assessment.....	206
6.11 Experiments.....	207
Results	208
6.12 Feature selection.....	208
6.13 Task E.1 Classification of GTV versus adjacent non-tumour tissue with multiple texture features.....	211
6.14 Task E.2 Classification of GTV versus adjacent non-tumour tissue with single texture feature (most discriminatory feature and mean)	214
6.14.1 Sum Variance (0,4) as sole feature	214

6.14.2 Mean as sole feature.....	215
6.14.3 Comparison of k-NN classification for multiple texture feature set versus sum variance and mean value as sole feature	216
6.15 Task E.3 Classification of GTV versus tissue at a distance away	218
Discussion.....	221
Conclusions.....	222
Chapter 7	223
Introduction.....	223
7.1 Summary of tasks.....	224
Methods.....	225
7.2 Generation of ROIs – BMP files	225
7.2 Generation of texture features	226
7.3 Feature selection.....	226
7.4 k-NN classification and classifier assessment	227
7.5 Study design – classification using k-NN classifier	227
7.6 Experiments	228
Results	229
7.7 Parameter tuning: Nested cross-validation	229
7.7.1 Task F.1 Assessment of classification of GTV versus non-tumour tissue at non-boundary region	229
7.7.2 Task F.2 Assessment of classification of GTV versus non-tumour tissue at boundary region.....	233
7.8 Estimated performance of classification models with optimised parameters.....	238
7.9 Re-training of final classification models.....	241
7.10 Independent test set	243
Discussion.....	245
Conclusions.....	247
Chapter 8	248
Project overview	248
Overall conclusions	249
Proposed future developments.....	249

References	252
Appendix	280
Appendix A	280
Appendix A.1 Plots for k-nearest neighbours classification optimisation using multiple texture feature set	280
Appendix A.2 Plots for k-nearest neighbours classification optimisation using most discriminatory texture feature Sum Variance (0,4)	282
Appendix A.3 Plots for k-nearest neighbours classification optimisation using mean as single feature	284
Appendix B	286
Appendix B.1 Plots for re-optimisation of final k-nearest neighbour classification models ..	286

List of figures

1.1	Illustration of relationships between gross tumour volume (GTV), clinical target volume (CTV), internal target volume (ITV) and planning target volume (PTV), in relation to the treated and irradiated volumes.....	2
1.2	Anatomy of cross-sectional imaging of the thorax in at the level of the a) aortic arch, b) carina and c) left pulmonary vein.....	6
1.3	Histogram showing three apparent classes where dotted lines represent identification of threshold between the three classes.....	10
1.4	Illustration of edge detection by first and second derivative operations for two images.....	11
1.5	Scale profile of image data shown in black lines, with watershed segmentation whereby the local maximum pixel scale values define the watershed lines (denoted by red lines), and local minimum pixel scale values define the catchment basins (denoted by blue stars).....	13
1.6	Example of a directed weighted graph of a 3 x 3 image, with each pixel represented by grey nodes, and edges represented by yellow (n-links) and red (t-links) arrows. In the image, s and t denote the location of the source and sink respectively, with the green line representing the segmentation results.....	16
1.7	Deformable registration of diagnostic PET-CT (orange uptake) with planning CT in three orthogonal planes. Green arrow denotes pathological lymph node.....	38
2.1	Design of pre-segmentation projects.....	46
2.2	Axial image illustrating placement of ROIs for evaluation of image quality. Red ROIs denote selection in soft tissue musculature, blue ROIs denote selection in background subcutaneous fat.....	51
2.3	Example of images generated for an axial slice for one of the datasets with indirect import through generation of STRUCT file. a) CT axial slice b) Contour superimposed on CT axial slice c) Close up of contour points on CT d) Binary mask generated for ROI.....	54
2.4	Example of display of GTV contour for the same axial slice in a) MATLAB and b) VODCA..	55
2.5	Example of display of GTV contour for the same axial slice in a) MATLAB and b) VODCA, showing multiple contours on a single axial slice.....	56
2.6	Display of ROI selection for two subjects with differences in anatomy, as well as presence of artifacts from localisation markers (green arrow). Red ROIs denote selection in soft tissue musculature, blue ROIs denote selection in background subcutaneous fat.....	59
2.7	Histogram plots displaying distribution of Hounsfield Units for the GTV, bone, ipsilateral lung and vessels across 18 subsample training cases collectively.....	60
2.8	Boxplots showing HU distribution for the GTV and other normal tissue across 18 subsample training cases collectively. Pink dash-dot line denotes the use of 158 HU as the absolute upper threshold; green dash-dot line denotes the use of -500 HU as the absolute lower threshold.....	62
2.9	Mean and 95% confidence interval of lower percentiles for GTV between non-solid and solid lesions for 18 subsample training cases.....	63
2.10	Mean and 95% confidence interval of lower percentiles for GTV between non-solid and solid lesions for total training dataset (63 cases).....	65
3.1	Edge-based detection techniques applied on CT image is shown in figure 3.1f, with the reference contours of the tumour indicated in yellow outline. a) Sobel approach (threshold 0.0567) b) Prewitt approach (threshold 0.0560) c) Roberts approach (threshold 0.0546) d) Laplacian of Gaussian approach (sigma 2, threshold 0.0019) e) Canny approach (sigma $\sqrt{2}$, threshold – upper: 0.0313, lower: 0.0125).....	69

3.2	Results of segmentation with lowering of threshold using Canny edge detection technique (sigma $\sqrt{2}$, threshold – upper: 0.01, lower: 0.009), displaying increase detection of weaker edge boundaries and persistent limitation of incomplete edge linkage.....	70
3.3	a) Axial slice of CT image b) Application of watershed segmentation, with result of watershed lines denoted by black boundary, and their corresponding white catchment basins.....	70
3.4	a) Gradient magnitude (Sobel) of CT image b) Application of watershed segmentation, with result of watershed lines denoted by the black boundary, and their corresponding white catchment basins.....	71
3.5	Representation of grey triangle region to be segmented from the white background, and the initiation of the Chan-Vese active contour curve as the red circular outline.....	73
3.6	Examples of possible evolutions for the segmentation curve based on the Chan-Vese active contour model.....	74
3.7	Summary of initial pre-processing workflow for different segmentation techniques.....	77
3.8	Diagram on data division for segmentation processes.....	79
3.9	Workflow for watershed segmentation.....	80
3.10	Synthetic images for evaluation of watershed segmentation process. a) Original image (also used as assessment reference); b) Degraded image (Gaussian edge smoothing with SD = 2.5, and gaussian noise with SD equivalent to 30 HU, mean = 0).....	83
3.11	User interface for semi-automated approach. A magnified axial slice is shown with mediastinal windowing. The red marker denotes the centroid of the GTV for the slice. The dashed line depicts the point placement and the linear connection between the points, to generate a polygon. Note that the point placement was performed at a distance away from the GTV boundary...	85
3.12	Workflow for active contour segmentation.....	86
3.13	Comparison of Chan-Vese active contour segmentation in relation to different initial masks a) Smallest convex hull bounding the submitted contour; b) Smallest rectangle bounding the submitted contour c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask. (Parameter settings – Iteration number: 300, Contraction bias: 0.7, Smoothing factor: 0).....	88
3.14	Workflow for graph-cut segmentation.....	92
3.15	Application of SLIC0 algorithm for $k = 500$ and $k = 1000$ respectively (10 iterations). Aqua lines denote superpixel boundaries. Both examples show poor adherence of the superpixels to the medial boundary of the tumour, as well as the mediastinal structures.....	96
3.16	Application of SLIC algorithm for $m = 10$, $m = 15$ and $m = 20$ respectively ($k = 2000$, 10 iterations). Aqua lines denote superpixel boundaries. The regularity of the superpixels increased with increasing m , which can be seen in the lung parenchyma as well as in the chest wall.....	97
3.17	GUI used for the qualitative assessment of the segmentation results, showing a segmentation result on the left, and the reference contours on the right. In addition to inbuilt magnification and pan functionalities, windowing levels could also be adjusted. The slider function enabled efficient visualisation of serial CT slices.....	101
3.18	Watershed segmentation using intermediate gradient operator for 2000-pixel circle against background equivalent to lung parenchyma; a) Final results of watershed segmentation on uncorrected gradient magnitude image; b) Uncorrected gradient magnitude image; c) Final results of watershed segmentation on gradient magnitude image corrected for 4-pixel connectivity; d) Corrected gradient magnitude image.....	103
3.19	Effect of a) increasing ROI boundary blurring and b) increasing image noise, on mean Dice similarity coefficient using watershed segmentation.....	103

3.20	Dice similarity coefficients for different gradient operators showing effect of ROI edge blurring. a) Gaussian noise equivalent to 10 HU; b) Gaussian noise equivalent to 20 HU; c) Gaussian noise equivalent to 30 HU.....	104
3.21	Segmentation results using degraded image with noise equivalent to 10 HU and ROI edge blurring with a SD of 1.5 (mean = 0). a) Degraded image; b) Roberts gradient operator; c) Sobel gradient operator; c) Prewitt gradient operator; d) Central difference gradient operator; e) Intermediate difference gradient operator.....	106
3.22	Hounsfield units at the vicinity of the GTV boundary. Each of the 18 subsample training cases is denoted by each individual plot. Positive values for the distance from GTV boundary represent the direction towards the centre of the GTV, negative values represent the direction away from the centre of the tumour.....	107
3.23	Mean gradient magnitude in relation to the distance from the GTV boundary as computed through the Prewitt and Sobel gradient operators, showing the mean of the 18 subsample training cases (error bars denote the SD).....	107
3.24	Watershed segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case).....	109-110
3.25	Example of overestimation of the tumour region by the watershed approach. a) Watershed segmentation with inclusion of the left pulmonary artery (blue arrow), and extension towards the vertebral body posteriorly (green arrow). b) Gold standard contour. c) Gradient magnitude (Sobel operator). d) Superimposition of exclusion mask (blue regions) on gradient magnitude. Presence of edge in gradient image between the left pulmonary artery and the tumour depicted by orange arrow, and red arrows indicate locations of competing gradient magnitude.....	111
3.26	Same case and axial slice as figure 3.25, with corresponding exclusion masks (blue regions) overlaying gradient magnitude. Red outline – watershed segmentation; yellow outline – reference contours.....	112
3.27	Atlas-based normal tissue segmentation. Cyan – bone; Red – lung; Light green – trachea; Pink – spinal cord; Orange – chestwall; Dark green – vessels; Dark blue – mediastinal soft tissue; yellow – heart. (a – f from case 1; g – i from case 2).....	113
3.28	Dice similarity coefficients for three runs of semi-automated watershed segmentation.....	114
3.29	Same case and axial slice as figures 3.25 and 3.26, with corresponding exclusion masks (blue regions) and manually placed points (linearly connected) overlaying gradient magnitude. Red outline – watershed segmentation; yellow outline – reference contours.....	115
3.30	Mean performance of watershed segmentation on each fold of the validation datasets (error bars represent standard deviation).....	116
3.31	Mean precision vs Mean recall plots for 18 training cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Convex polygon with 4-pixel erosion, b) Circle with 4-pixel erosion, c) Circle with no erosion (Chan-Vese active contour segmentation).....	118
3.32	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (Convex polygon initial mask with 4-pixel erosion; Chan-Vese active contour algorithm).....	119
3.33	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (Circle initial mask with 4-pixel erosion; Chan-Vese active contour algorithm).....	120
3.34	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (circle initial mask with no erosion; Chan-Vese active contour algorithm)...	121

3.35	Example showing similar Chan-Vese active contour segmentation results using different initialisations a) Convex hull with 4-pixel erosion; b) Circle with 4-pixel erosion; c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.....	122
3.36	Examples (cases a to c) showing variation in Chan-Vese active contour segmentation results using different initialisations; suffix _1) Convex hull with 4-pixel erosion; suffix _2) Circle with 4-pixel erosion; suffix _3) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.....	123
3.37	Comparison of Chan-Vese active contour segmentation in relation to different initial masks for the same axial slice, where the optimal parameter settings were used for figures a to c, whilst contraction bias of 0.6 and smoothing factor of 0.7 were used for figures d to f. a,b) Convex hull with 4-pixel erosion; b,e) Circle with 4-pixel erosion; a,d) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.....	124
3.38	Example showing splitting of generated contour with Chan-Vese segmentation a) Convex hull with 4-pixel erosion; b) Circle with 4-pixel erosion; c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.....	125
3.39	Mean precision-recall curves for Chan-Vese segmentation as function of contraction bias, for varying smoothing factors. a) Training Run 1; b) Training Run 2; c) Training Run 3.....	126
3.40	Mean Dice similarity coefficient for Chan-Vese segmentation of training run 1 displaying impact of different a) contraction bias b) smoothing factor.....	127
3.41	Mean Dice similarity coefficient for Chan-Vese segmentation of training run 2 displaying impact of different a) contraction bias b) smoothing factor.....	128
3.42	Mean Dice similarity coefficient for Chan-Vese segmentation of training run 3 displaying impact of different a) contraction bias b) smoothing factor.....	129
3.43	Chan-Vese active contour segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case).....	131-132
3.44	Mean performance of Chan-Vese active contour segmentation on each fold of the validation datasets (error bars represent standard deviation).....	132
3.45	Mean precision versus recall plots for 18 subsample cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Convex polygon with 4-pixel erosion, b) Circle with 4-pixel erosion, c) Circle with no erosion (Edge-based active contour segmentation).....	134
3.46	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Convex polygon initial mask with 4-pixel erosion; Edge-based active contour algorithm).....	135
3.47	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Circle initial mask with 4-pixel erosion; Edge-based active contour algorithm).....	136
3.48	Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Circle initial mask with no erosion; Edge-based active contour algorithm).....	137
3.49	Examples (cases a to e) showing variation in edge-based active contour segmentation results using different initialisations; suffix _1) Convex hull with 4-pixel erosion; suffix _2) Circle with 4-pixel erosion; suffix _3) Circle with area equivalent to the submitted contour. Red	

	outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.....	139-140
3.50	Mean precision vs recall plots for training cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Training run 1, b) Training run 2, c) Training run 3 (Edge-based active contour segmentation).....	142
3.51	Mean Dice similarity coefficient for edge-based active contour segmentation of training run 1 displaying impact of different a) contraction bias b) smoothing factor.....	143
3.52	Mean Dice similarity coefficient for edge-based active contour segmentation of training run 2 displaying impact of different a) contraction bias b) smoothing factor.....	144
3.53	Mean Dice similarity coefficient for edge-based active contour segmentation of training run 3 displaying impact of different a) contraction bias b) smoothing factor.....	145
3.54	Edge-based active contour segmentation results (red outlines) for six representative training cases (a – f) versus reference contours (yellow outlines), with corresponding DSC for each case. Green and orange arrows indicate regions of poor conformity. (suffix _1 to 3 represent different axial slices for each case).....	147-148
3.55	Mean performance of edge-based active contour segmentation on each fold of the validation datasets (error bars represent standard deviation).....	148
3.56	Impact of desired number of superpixel regions on a) mean boundary recall and b) mean undersegmentation error for different SLIC compactness.....	149
3.57	Superpixels generated for one image slice showing impact of variation of number of regions a) 3000; b) 7000; c)11000; d) 15000). Suffix _1) SLIC compactness 5; suffix _2) SLIC compactness 10; suffix_3) SLIC compactness 12; suffix_4) SLIC 0; Fig. 3.57_5) Reference outlines in yellow.....	150
3.58	Superpixels generated for one image slice with two tumour regions showing impact of variation of number of regions a) 3000; b) 7000; c)11000; d) 15000). Suffix _1) SLIC compactness 5; suffix _2) SLIC compactness 10; suffix _3) SLIC compactness 12; suffix _4) SLIC 0; Fig. 3.58_5) Reference outlines in yellow.....	151
3.59	Size (number of pixels) of each tumour ROI in the axial plane across the 18 subsample cases.....	152
3.60	Performance of lazysnapping segmentation using superpixels in relation to the number of superpixel regions and different SLIC algorithms, showing results of a) Mean Dice similarity coefficient; b) Mean recall; c) Mean precision.....	153
3.61	Representative images slices of individual cases (A to F) for SLIC compactness 0 and 5, using number of superpixels $k = 5000, 11000$ and 17000 , with the associated DSC scores for each case.....	154
3.62	Mean computational processing time for individual cases a) superpixel generation alone b) superpixel generation and application of lazysnapping segmentation.....	155
3.63	Mean boundary recall for superpixel generation as a function of desired number of regions. a) Training run 1; b) Training run 2; c) Training run 3.....	156
3.64	Mean undersegmentation error for superpixel generation as a function of desired number of regions. a) Training run 1; b) Training run 2; c) Training run 3.....	157
3.65	Mean Dice similarity coefficient in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.....	158
3.66	Mean recall scores in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.....	158
3.67	Mean precision scores in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.....	159

3.68	Mean computational time of individual cases for superpixel generation. a) Training run 1; b) Training run 2; c) Training run 3.....	160
3.69	Mean computational time of individual cases for lazysnapping segmentation in relation to number of superpixel regions. a) Training run 1; b) Training run 2; c) Training run 3.....	160
3.70	Performance with variation of edge weight scaling factors for the training runs. (Superpixel generation: $k = 7000$, SLIC $m = 10$) a) Mean recall and precision; b) Mean Dice similarity coefficient.....	161
3.71	Representative images slices of individual cases (A to F) for edge weight scale factors of 10 to 900. (Superpixel generation: $k = 7000$, SLIC $m = 10$).....	163
3.72	Graph-cut segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case).....	165-166
3.73	Mean performance of lazysnapping segmentation (edge weight parameter = 40) on superpixels (number of regions = 7000, SLIC compactness = 10) on each fold of the validation datasets. (error bars represent standard deviation).....	167
4.1	Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for independent test dataset.....	169
4.2	Conformity indices for segmentation of individual cases in comparison to reference contours. a) Dice similarity coefficient; b) Geographical miss index; c) Discordance index.....	170
4.3	Percentage volume difference in relation to the reference contours for individual cases (mean and standard deviation shown in table). Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.....	171
4.4	Segmentation results for representative individual cases. Blue – Watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow – reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour).....	173-174
4.5	Time (minutes in logarithmic scale) for processing individual cases, with the mean and standard deviation shown in the table. (Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC).....	175
5.1	Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for peripheral lung primary disease within training set.....	185
5.2	Percentage volume difference of peripheral primary disease in relation to the reference contours for individual cases (mean and standard deviation shown in table). Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC....	186
5.3	Segmentation results for training dataset (peripheral primary disease; cases p1 – 5). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour).....	187
5.4	Time (minutes in logarithmic scale) for processing individual cases (peripheral primary disease only), with the mean and standard deviation shown in the table. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.....	188
5.5	Segmentation results for testing dataset (peripheral primary disease; test cases p1 – 2). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour).....	188
5.6	Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for peripheral lung primary disease with ISTART dataset.....	190

5.7	Percentage volume difference of peripheral primary disease in relation to the reference contours for individual cases (mean and standard deviation shown in table) of ISTART dataset. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.....	190
5.8	Conformity indices for peripheral primary disease segmentation of individual cases in comparison to reference contours (ISTART dataset). a) Dice similarity coefficient; b) Geographical miss index; c) Discordance index.....	191
5.9	Segmentation results for ISTART dataset (peripheral primary disease; ISTART cases p1 – 10). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour).....	192
5.10	Time (minutes in logarithmic scale) for processing individual cases of ISTART dataset (peripheral primary disease only), with the mean and standard deviation shown in the table. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.....	193
6.1	Design for LDA classification.....	199
6.2	Example of an axial slice of a CT image normalised and discretised into equally spaced bin widths of 25 HUs as displayed in MaZda. The ROIs for analysis is represented by the coloured regions (Blue = GTV, Aqua = 10-pixel width annulus adjacent to GTV, Pink = 10-pixel width annulus at 10-pixel distance away from GTV).....	200
6.3	Design for k-NN classification.....	205
6.4	Results of LDA with respect to MDF, red “1” denotes GTV data points, green “2” denotes data points for region surrounding GTV. A) Training run 1; b) Training run 2; c) Training run 3...212	212
6.5	Visualisation of texture parameters separability between GTV and adjacent tissue, based on the first, second and third principal components.....	213
6.6	QQ plots for assessment of distribution for training data a) Linear discriminants b) Quadratic discriminants.....	213
6.7	Boxplot comparing the values for sum variance of the GTV versus the adjacent tissues at 4-pixel distance and 45° direction.....	214
6.8	Boxplot comparing the mean values of the GTV versus the adjacent tissues.....	215
6.9	Mean ROC curves comparing performance of k-NN classification on training data using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value). Shaded regions represent confidence interval.....	216
6.10	ROC curves comparing performance of k-NN classification on independent test set using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value).....	218
6.11	Mean ROC curves comparing performance of trained k-NN classifier on testing data using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value) in the classification of GTV versus tissue at 10-pixel distance away, where models were trained on features derived from GTV and adjacent tissue. Shaded regions represent confidence interval.....	219
6.12	ROC curves comparing performance of trained k-NN classifier on independent test set using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value) in the classification of GTV versus tissue at 10-pixel distance away, where models were trained on features derived from GTV and adjacent tissue.....	219
7.1	Example of a representative slice for ROI placement (16-pixel square size shown). For illustration, three ROI samples are displayed for each group, though five samples for each group	

	was used in all the experiments. Green: Non-boundary tumour tissue; Blue: Non-boundary non-tumour tissue; Orange: Boundary tumour tissue; Pink: Boundary non-tumour tissue.....	225
7.2	Diagrammatic representation of feature selection workflow.....	226
7.3	Workflow for k-NN classification.....	227
7.4	Mean misclassification errors for k-NN classification of non-boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	229
7.5	Mean sensitivity and specificity plots of nested validation data for k-NN classification of non-boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	230
7.6	Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Non-boundary regions with ROI size of 16-pixel square). Shaded regions represent standard deviation.....	230
7.7	Mean misclassification errors for k-NN classification of non-boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	231
7.8	Mean sensitivity and specificity plots of nested validation data for k-NN classification of non-boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	232
7.9	Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Non-boundary regions with ROI size of 8-pixel square). Shaded regions represent standard deviation.....	232
7.10	Mean misclassification errors for k-NN classification of boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	233
7.11	Mean sensitivity and specificity plots of nested validation data for k-NN classification of boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	234
7.12	Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 16-pixel square). Shaded regions represent standard deviation.....	234
7.13	Mean misclassification errors for k-NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	235
7.14	Mean sensitivity and specificity plots of nested validation data for k-NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	236
7.15	Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 8-pixel square). Shaded regions represent standard deviation.....	236

7.16	Mean misclassification errors for k-NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars), where only sequential feature selection was applied for feature reduction. a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.....	237
7.17	Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 8-pixel square), where only sequential feature selection was applied for feature reduction. Shaded regions represent standard deviation.....	237
7.18	Mean ROC curves displaying classification performance for distinguishing between tumour versus non-tumour ROIs at non-boundary and boundary regions, using 16-pixel and 8-pixel ROIs, using outer validation datasets. Shaded regions represent 95% confidence interval....	238
7.19	Mean number of selected features for optimised classification models (error bars represent standard deviation).....	239
7.20	Proportion of selected feature class from optimised classification models across three cross-validation folds.....	240
7.21	Plots for parameter tuning of final k-NN classification models. a) 16-pixel non-boundary ROIs; b) 8-pixel non-boundary ROIs; c) 16-pixel boundary ROIs; d) 8-pixel boundary ROIs.....	242
7.22	ROC curves displaying classification performance for distinguishing between tumour versus non-tumour ROIs at non-boundary and boundary regions, using 16-pixel and 8-pixel ROIs, on independent test data. Shaded regions represent 95% confidence interval.....	243
7.23	Number of selected features for the different texture classes of the optimised final classification models.....	244
A.1	Mean misclassification errors using k-nearest neighbours classification (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size using multiple texture features for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.....	280
A.2	Mean sensitivity and specificity plots nested validation data using k-nearest neighbours classification (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size with multiple texture features for classification. a, b) Outer cross-validation run 1; c, d) Outer cross-validation run 2; e, f) Outer cross-validation run 3.....	281
A.3	Performance of k-nearest neighbour classifier (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both nested training and nested validation data), as well as the mean sensitivity and specificity of the nested validation data using sum variance (4-pixel distance and 45° direction) for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.....	283
A.4	Performance of k-nearest neighbour classifier (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both nested training and nested validation data), as well as the mean sensitivity and specificity of the nested validation data using mean values for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.....	285
B.1	Performance of k-nearest neighbour classifier (3-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both training and validation data), as well as the mean	

sensitivity and specificity of the validation data. Feature set used for classification: a) Multiple texture features; b) Sum variance (4-pixel distance and 45° direction); c) Mean values.....287

List of tables

1.1	Summary of (semi-) automatic segmentation techniques applied to pulmonary lesions on CT imaging.....	22-30
1.2	Summary of (semi-) automatic segmentation techniques applied to lymph nodes on CT imaging.....	36-37
1.3	List of commercial software products and associated segmentation techniques that are currently available.....	39-40
2.1	Normal tissue outlining definitions and methodology.....	46-47
2.2	Volumes (cm ³ , corrected to 2 decimal places) for GTV contours axial slices (5 slices apart), as well as total GTV volumes for 3 training cases in Pinnacle ³ , VODCA and MATLAB displays (Friedman Test: <i>p</i> -value = 0.49).....	54
2.3	Clinical characteristics of all data (79 cases).....	56
2.4	Scanning parameters of all data (79 cases).....	57
2.5	Mean (and SD) of signal and noise for the muscle and fat in 18 subsample training cases....	57
2.6	Mean (and SD) of SNR and CNR for 18 subsample training cases.....	58
2.7	Hounsfield unit values for GTV of 18 subsample training cases from 90 th to 100 th percentile...59	
2.8	Hounsfield unit statistics for 95 th and 99 th percentiles for all 63 training cases performed with bootstrapping (1000 number of samples).....	60
2.9	Hounsfield unit values for GTV of 18 subsample training cases from 0 to 10 th percentile.....	60
2.10	Descriptive statistics for non-solid versus solid lesions for 18 subsample training cases (Mann-Whitney U test).....	63
2.11	Descriptive statistics for non-solid versus solid lesions for all 63 training cases (Mann-Whitney U test).....	65
3.1	Mean conformity indices for the control synthetic images comparing different gradient computation.....	102
3.2	Mean conformity indices comparing different gradient computation for images degraded by ROI boundary blurring and image noise.....	105
3.3	Performance of watershed segmentation using Prewitt and Sobel operators for gradient computation.....	108
3.4	Comparison of performance of watershed segmentation with and without further exclusion structures (Mann-Whitney test).....	112
3.5	Individual and aggregate performance for semi-automated watershed segmentation.....	114
3.6	Comparison of performance of automatic and semi-automated watershed segmentation....	115
3.7	Mean performance of watershed segmentation on the validation datasets.....	116
3.8	Optimal parameter settings for Chan-Vese active contour segmentation with results of performance.....	122
3.9	Optimal parameter settings for Chan-Vese segmentation with results of performance.....	125
3.10	Performance of Chan-Vese active contour segmentation using contraction bias of 0.75 and smoothing factor of 0.3 on the validation datasets.....	132
3.11	Optimal parameter settings for Edge-based active contour segmentation with results of performance.....	138
3.12	Optimal parameter settings for edge-based active contour segmentation with results of performance.....	141
3.13	Performance of edge-based active contour segmentation with contraction bias of -0.075 and smoothing factor of 0.6 on the validation datasets.....	148

3.14	Performance for the training runs with edge weight scaling factor between the range of 10 and 60. (Superpixel generation: $k = 7000$, SLIC $m = 10$).....	162
3.15	Performance of lazysnapping segmentation (edge weight parameter = 40) on superpixels (number of regions = 7000, SLIC compactness = 10) on the validation datasets.....	166
4.1	Summary of the performance and general behaviours of the different techniques specific to the established workflows developed in this project.....	176
5.1	Absolute and percentage volume difference of peripheral primary disease from reference contours for individual cases, with the respective computational processing time.....	189
5.2	Scanning parameters for the 10 evaluated cases in the ISTART dataset.....	189
5.3	Summary of qualitative measures for segmentation of advanced and isolated peripheral lung tumours. (Computational time displayed for processes using an Intel Core i5-3317U CPU @ 1.70GHz, 4GB RAM on a Windows 10 64-bit environment.).....	194
6.1	Division of data and sample sizes for feature selection and LDA classification.....	199
6.2	Classes and sub-categories of histogram and texture features used for this work.....	201
6.3	Division of data and sample sizes for k-NN classification.....	206
6.4	Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied sequentially for each cross-validation training set, with associated scores. ^ denotes features common in at least one other feature selection method in each run.....	208
6.5	Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied in combination to yield the top thirty ranked texture parameters for each cross-validation training set.....	209
6.6	Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied in combination to yield the top thirty ranked texture parameters for each cross-validation training set (exclusion of co-occurrence matrix at pixel distances 2, 3 and 5). *denotes features present in all three runs; ** denotes features present in two runs.....	210
6.7	Final list of 28 features used for classification analysis.....	211
6.8	Classification results for three training runs using LDA and 1-NN classifier.....	212
6.9	Error rates with 5-nearest neighbour classifier on nested validation data using multiple texture features.....	214
6.10	Error rates with optimum parameters on nested validation data using sum variance (4-pixel distance and 45° direction) for classification.....	215
6.11	Error rates with optimum parameters on nested validation data using mean values for classification.....	215
6.12	Estimate of classification performance with optimised classifiers parameters.....	216
6.13	Optimum error rates from re-training using optimum parameters.....	217
6.14	Error rates of trained data associated with final models using optimum parameters.....	217
6.15	Performance of final k-nearest neighbour classifier on independent test set for multiple texture features, sum variance (0,4) as sole parameter, and mean value as sole feature.....	217
6.16	Comparison of performance of optimised k-nearest neighbour classifier on classifying GTV versus adjacent tissue, and tissue at 10-pixel distance using training data.....	218
6.17	Errors in application of data for GTV versus tissue at 10-pixels on re-trained final models...219	219
6.18	Performance of k-nearest neighbour classifier on classifying GTV versus tissue at 10-pixel distance away, using multiple texture features, sum variance (0,4) as sole parameter, and mean value as sole feature on independent test set. Classification models were trained based on classification of GTV versus adjacent tissue for the respective feature sets.....	220
7.1	Division of data and sample sizes for classification tuning and testing.....	228

7.2	Error rates with optimum parameters on nested validation data (Non-boundary regions with ROI size of 16-pixel square).....	231
7.3	Error rates with optimum parameters on nested validation data (Non-boundary regions with ROI size of 8-pixel square).....	232
7.4	Error rates with optimum parameters on nested validation data (Boundary regions with ROI size of 16-pixel square).....	234
7.5	Error rates with optimum parameters on nested validation data (Boundary regions with ROI size of 8-pixel square).....	237
7.6	Estimated classification performance with optimised classifiers parameters.....	238
7.7	List of selected features from optimised classifier models. *Features present in two runs; **Features present in all three runs.....	241
7.8	Error rates with optimum parameters on validation data after re-training.....	242
7.9	Error rates of final models based on the total trained data using optimum parameters.....	242
7.10	Performance of final k-nearest neighbour classifier on independent testing data.....	243
7.11	List of selected features from optimised final classifier models.....	244

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
4D	Four-dimensional
AAM	Active appearance model
AAPM	American Association of Physicists in Medicine
ACC	Average correlation coefficient
ANN	Artificial neural network
ASM	Active shape model
AUC	Area under curve
BMP	Bitmap
CAD	Computer-aided diagnosis
CB	Contraction bias
CNN	Convolutional neural network
CNR	Contrast to noise ratio
CI	Confidence interval
CT	Computed tomography
CTV	Clinical target volume
DI	Discordance index
DSC	Dice similarity coefficient
ELCAP	Early lung cancer action project
FDG	¹⁸ F-fluorodeoxyglucose
GGO	Ground-glass opacity
GMI	Geographical miss index
GTV	Gross tumour volume
GUI	Graphical User Interface
HU	Hounsfield Unit

ICRU	International Commission on Radiation Units and Measurements
IGRT	Image-guided radiotherapy
IPEM	Institute of Physics and Engineering in Medicine
IM	Internal margin
IMRT	Intensity-modulated radiotherapy
ITV	Internal target volume
IV	Intra-venous
k-NN	k-nearest neighbour
LDA	Linear discriminant analysis
LIDC	Lung Image Database Consortium
MDF	Most discriminatory feature
MEF	Most expressive feature
MI	Mutual information
MRI	Magnetic resonance imaging
MRF	Markov random field
NCRI	National Cancer Research Institute
NSCLC	Non-small cell lung cancer
OAR	Organ at risk
PCA	Principal component analysis
PET	Positron emission tomography
POE	Classification error probability
PTV	Planning target volume
QA	Quality assurance
QDA	Quadratic discriminant analysis
RCC	Royal College of Radiologists
ROC	Receiver operating characteristic
ROI	Region of interest

RTTQA	National Radiotherapy Trials Quality Assurance Group
SBRT	Stereotactic body radiotherapy
SCLC	Small cell lung cancer
SD	Standard deviation
SF	Smoothing factor
SLIC	Simple linear iterative clustering
SM	Set-up margin
SNR	Signal to noise ratio
SUV	Standardised uptake value
SVM	Support vector machine
TMG	Trial management group
TNM	Tumour, node and metastasis

Chapter 1

Introduction

1.1 Advances in radiotherapy

Radiotherapy plays a major role in cancer treatment with about half of all patients requiring radiotherapy during their illness, and it is estimated to contribute to 40% of cases where cancer is cured (1). Over the last few decades, there has been significant changes in radiotherapy treatment planning. The move away from conventional radiotherapy using simple rectangular fields towards more advanced treatment techniques has resulted in an improvement in many aspects of radiotherapy delivery. Approaches such as three-dimensional (3D) conformal radiotherapy and intensity-modulated radiotherapy (IMRT) allow for better conformity of the delivered dose to the target, whilst reducing the dose to normal tissues and sparing organs at risk (OARs) (2). This can lead to a reduction in toxicity to normal organs and thereby improve quality of life (3). The resultant increase in the therapeutic ratio also allows the potential for dose escalation to the target whilst keeping the dose to normal tissues below the maximal tolerated dose, and consequently improve tumour control rates (4-8). Major developments have also been made to improve the precision in radiotherapy delivery such as in image-guided radiotherapy (IGRT) and stereotactic body radiotherapy (SBRT), where good target volume coverage is maintained whilst minimising the dose delivered to normal tissues (2).

1.2 Target definition in radiotherapy

Current practice in radiotherapy uses target volume definitions from the proposed framework set out by the International Commission on Radiation Units and Measurements (ICRU) reports 50 and 62 (9, 10) as illustrated in figure 1.1. Gross tumour volume (GTV) is defined through clinical, radiological and pathological means as the demonstrable extent of the malignant growth. Clinical target volume (CTV) is an anatomical concept which comprises of the GTV and tissues harbouring subclinical microscopic disease below the resolution limits of imaging, typically based on experience and clinical knowledge on the patterns of spread. Planning target volume (PTV) is a geometric concept which includes the CTV as well as the internal margin (IM) (variations in size, shape and position of the CTV relative to the anatomical reference points), and set-up margin (SM) (uncertainties in patient position and treatment beam alignment in treatment delivery) which takes into account tumour movement during treatment delivery. The concept of an internal target volume (ITV) which includes the CTV and IM has also been introduced to account for tumour movement, which has been adopted in the treatment of sites such as lung cancer where tumour motion is estimated using four-dimensional (4D) computed tomography (CT). OARs are normal tissues whose radiation sensitivity may influence the prescribed dose or treatment planning. The shape of the treated volume, which represents the volume enclosed by the specified isodose surface that is intended to be delivered, depends on the conformality of the technique used for treatment. The irradiated volume is the volume that receives a dose considered to be significant in relation to normal tissue tolerance, which is also dependent on the technique used for treatment delivery.

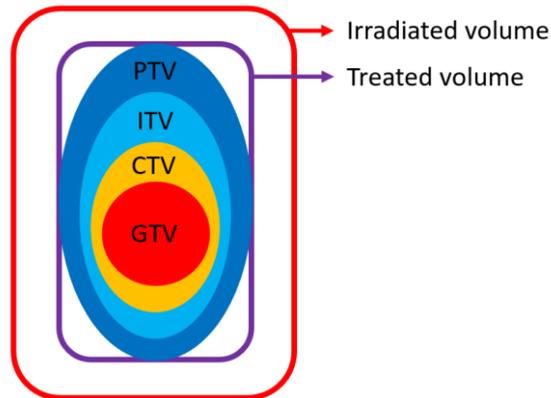


Figure 1.1. Illustration of relationships between gross tumour volume (GTV), clinical target volume (CTV), internal target volume (ITV) and planning target volume (PTV), in relation to the treated and irradiated volumes.

These concepts indicate the many sources of uncertainty that exist in radiotherapy treatment (11), which can occur at any of these points to varying degree, and cause deviations between the intended and received dose in the respective treatment volumes. The integration of methods to improve treatment precision with highly conformal delivery techniques are aimed at minimising this. Errors from set up for example, can be reduced with the appropriate application of immobilisation and fixation devices as well as set up checks. With the implementation of IGRT, not only are set up errors minimised, the impact of organ motion is also reduced. In this framework, target volume definition is generally considered to contribute to the largest source of error (12).

1.3 Need for outlining assessment in radiotherapy

Improvements in precision of radiotherapy treatment are only as good as the accuracy of the defined target. Even with the application of IGRT which can reduce geometric uncertainties, errors from target definition are not mitigated, which can be greater than other sources of errors such as set up inaccuracies (13). It is known that there is significant variation between clinicians in target volume delineation (14), which is a major source of error in the treatment process. It is critical that accurate segmentation of the GTV is performed, as it has direct implications on the treatment volume and the resulting radiotherapy plan. The target definition framework as described above does not account for errors in the delineation of the GTV, which if arises, can lead to a systematic error during the treatment process, and consequently failure in tumour local control and worsened toxicity profile.

Although manual segmentation of target volumes is being used as the “gold-standard” approach to volume definition in the clinical setting, there are many limitations associated with this. Although clinicians are good at recognising tumours on imaging, identifying the tumour boundary is more of a challenge. Clinicians tend to overestimate the boundaries of the lesion to ensure that the entire tumour is identified (15). Human error and mis-identification of involved tumour and lymph nodes can also result in missed targets, which can potentially be avoidable. For many of the tumour sites, both target and OARs delineation can be very time consuming.

There is substantial evidence that there is significant inter- and intra-observer variability for both target volumes and OARs definition across many body sites (16, 17). A review has identified that the widest inter-observer variation of GTV delineation was observed in lung, oesophagus, and head and neck cancers, wherein the size of the largest defined GTV was more than eight times the size of the smallest (18). It has also been reported that a major change was recommended in up to 23% of contours evaluated in radiotherapy planning for primary lung cancer using SBRT, and quality assurance (QA) of target volume delineation is warranted to ensure consistency and quality in treatment planning (19).

In order to achieve high target conformity with IMRT, there is often a high gradient fall-off of the dose at the boundary of the target volume. The advent of IGRT has also led to a move towards reducing the PTV margins in the identification and correction of inter-fractional and patient set-up errors (20, 21). These advances make tumour definition even more critical, as a geographical miss of the target can lead to an inhomogeneous dose being delivered, which can result in a loss of tumour control.

Thus, a number of different approaches has been adopted in the clinical setting to improve the consistency of tumour delineation. For example, educational resources have been developed to provide training to oncologists in volume definition, which includes outlining workshops across a range of body sites (22, 23). Additionally, a range of outlining atlases has been published by expert groups setting out the consensus guidance for both target and OAR delineation (24-36), aimed at helping to improve the consensus of volume definition between clinicians. The use of multi-modality imaging such as positron emission tomography (PET) and magnetic resonance imaging (MRI) scans can also improve the identification and delineation of targets (37, 38). In spite of these interventions, inter-observer outlining variation still exists. For example, although the integration of PET information to planning CT has been shown to reduce variability in non-small cell lung cancer (NSCLC) contouring, persistent variation was felt to have been attributed to differences in clinician judgement (39). Likewise, Senan et al reported significant differences in the size of GTV delineation of NSCLC in spite of the use of a standardised outlining protocol (40). Vorwerk et al also found large inter-observer delineation variability of GTV even with detailed outlining instructions (41).

It is recognised that peer-review of treatment plans by at least another colleague, such as a clinical oncologist or radiologist, can help with detecting inaccuracies in volume delineation (42, 43) and decrease outlining variation in practice (44-46). Recently, a working party through the Royal College of Radiologists (RCR) Clinical Oncology Professional Support and Standards Board has developed recommendations on the implementation of target volume delineation peer review in the UK (47), where the content, structure as well as the benefits of peer review in minimising errors in volume delineation have been described. Because of these reasons, some departments have introduced radiotherapy quality assurance meetings in the clinical setting with an aim of detecting and correcting treatment plans prior to treatment delivery.

1.4 Radiotherapy trial quality assurance

As such, radiotherapy errors which include target delineation inaccuracies makes quality control and QA of the radiotherapy process an integral component in both day-to-day clinical practice and within radiotherapy trials. It has been reported that even in the setting of clinical trials with rigorous trial set ups, there exists a significant number of radiotherapy delivery

variations from the trial protocol (48-53). Poor radiotherapy compliance to a trial protocol has been reported to be associated with adverse clinical outcome. For example, deviations from the trial protocol have been shown to result in an increase in normal tissue toxicity and poorer overall survival in pancreatic cancer (51, 52). Similarly, in head and neck cancer, major radiotherapy protocol deviations are associated with a reduction in both overall survival and loco-regional control (49, 50). In addition to deviations in the planning parameters, these studies have reported non-compliance in volume delineation to the recommended practice. Thus, implementation of the QA process is vital in the setting of a trial not only to ensure that treatment complies with nationally accepted standards, but to ensure adherence to the trial protocol and minimise variations across different recruiting sites. This means that clinical trial outcomes truly reflect differences in the randomised intervention schedules of the trial rather than departures from the protocol.

The National Radiotherapy Trials Quality Assurance Group (RTTQA) has established different QA programmes tailored to the complexity, requirements and treatment delivery used in individual trials. It has been demonstrated that trials where such a credentialing process have been adopted are associated with low major deviation rates from the trial protocol (54). Typically, the credentialing programme consists of a number of different modules, one of which is an outlining assessment. Depending on the needs of a particular trial, both pre-accrual contouring exercises and real-time contouring reviews may be carried out. Pre-accrual contouring exercises come in two forms, a) benchmark case(s), comprising of a standard outlining case(s) undertaken by all relevant investigators in the trial. A set of consensus volumes is typically pre-defined by experts, usually from the trial management group (TMG), to serve as a comparison to the investigator outlines. b) Dummy-run(s), where relevant investigators submit clinical cases from their centres that have been treated according to the trial protocol. Real-time contouring reviews may be performed either prospectively before radiotherapy treatment planning and delivery of the recruited patient, or retrospectively after the patient has been treated but prior to the recruitment of the subsequent patient. Feedback is then provided to the institutions to recommend modification of the contours, should they be non-compliant.

Contouring assessment performed both in the pre-accrual and real-time settings are currently performed manually on a slice-by-slice basis, which is labour intensive and time consuming. This can impact review of benchmark cases, which must be performed in a timely fashion to avoid any delays to opening of new centres to a trial. More particularly, there can be immense time pressure for the assessments to be carried out within a relatively short timeframe for prospective case reviews (e.g. 48 – 72 hours), stipulated to minimise treatment initiation delays by the case review process. Where there is a need for an error to be rectified, re-submission of the case is usually requested, necessitating another review prior to approval, which thereby compounds this problem. To ensure that the reviews are performed promptly, most trials enlist more than one clinician as an assessor for this process.

1.5 The challenge of real time assessment of contouring

Unlike the pre-accrual contouring benchmark cases, the lack of a reference volume can make real-time assessments a more difficult process. In addition to taking up more time, the real-time review process is more subjective and assessor dependent due to the absence of a consensus opinion tailored to the specific case. Consequently, differences in feedback may arise from multiple reviewers if they are not working in tandem. Numerical evaluation through

conformity indices requiring at least a comparator volume also cannot be performed. These issues also apply to pre-accrual dummy-runs, where hold ups to assessments can potentially affect patient recruitment to a trial.

For these reasons, there is a need to improve on the current trial outlining review workflow, in regard to increasing the efficiency and decreasing the subjectivity of the process.

1.6 Lung cancer as tumour site focus for this study

The issues set out in the above sections are encountered across many tumour sites as they are generic to radiotherapy treatments and QA processes. However, rather than to explore a range of different body sites, primarily due to the image processing challenges that would be faced in this project, it was felt that the focus of this study should be narrowed down to a particular tumour site.

Lung cancer was selected as the main subject of study for a number of reasons. Firstly, there is a very high prevalence and incidence of lung cancers, being the third most common cancer in the UK (55) and the most common cancer worldwide (56). Moreover, a high proportion of lung cancer patients receive radiotherapy with either a curative or palliative intent as part of their primary cancer treatment, with rates of 42% and 28% for small cell lung cancer (SCLC) and NSCLC respectively in 2013 – 2014 (57). As the main focus of this work is based on the assessment of tumours on their imaging appearance, the presence of macroscopic lesions (i.e. GTV) is imperative, which would not apply in tumour sites where adjuvant radiotherapy is commonly given in the post-operative setting, such as in breast cancer. Similarly, for sites such as the prostate, the GTV is not typically defined in current practice. Instead, the CTV comprising of at least the whole prostate is outlined, where the segmentation is organ-based. Additionally, there has been a steady number of lung cancer radiotherapy trials in the National Cancer Research Institute (NCRI) portfolio that can potentially benefit from this work, an example of which is ADSCaN, where different dose-escalated accelerated radiotherapy treatment schedules in NSCLC is explored (58). Significant variation in lung cancer delineation has also been reported, necessitating review and quality assessment to ensure inaccuracies are kept to a minimum (18, 19).

1.7 Cross-sectional anatomy in the context of lung cancer

The thorax consists a variety of tissue types (e.g. lung parenchyma, chest wall (bone and musculature) and the mediastinum comprising of vessels, trachea, oesophagus and mediastinal fat), where differences in the radiodensity allows the anatomy to be well visualised on CT imaging. There are a number of resources detailing the anatomy of the chest on cross-sectional CT imaging, and example of which is shown in figure 1.2 (59).

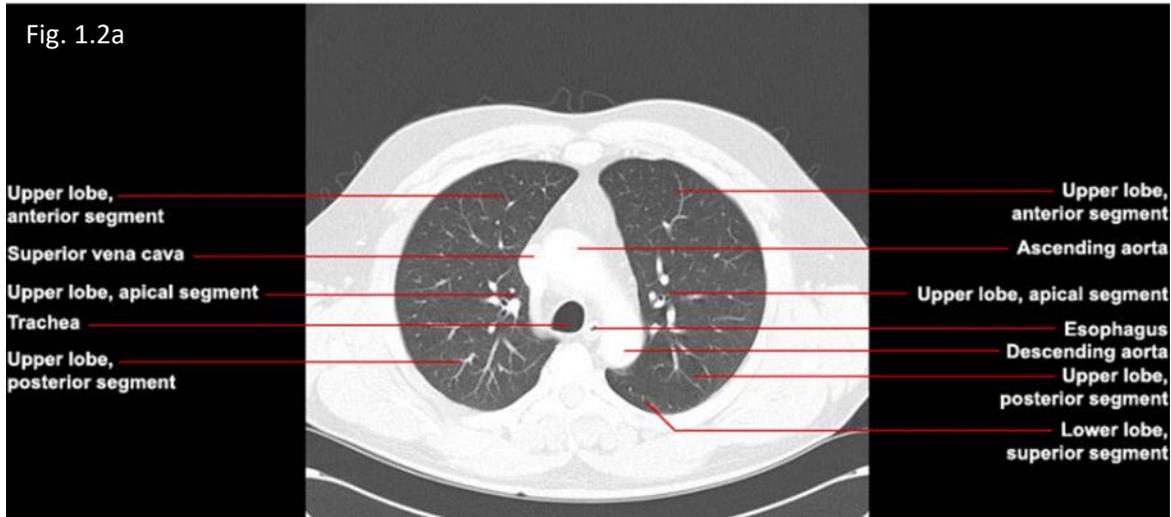


Figure 1.2. Anatomy of cross-sectional imaging of the thorax in at the level of the a) aortic arch, b) carina and c) left pulmonary vein (reproduced from (59)).

In the context of lung cancer, the primary tumour can differ greatly in size and location, which together with the extent of nodal involvement, is associated with different prognosis. The eighth edition of the tumour, node and metastasis (TNM) classification for lung cancer describes the different stages of lung cancer in relation to its spread within and outside of the thorax in association with outcome (60). By size criteria, the primary disease is classified into T1, T2, T3 and T4 tumours for sizes ≤ 3 cm, between 3 and 5cm, between 5 and 7cm, and ≥ 7 cm respectively. T staging is also determined by tumour location and other additional features. For instance, T1 disease is surrounded by lung parenchyma or the visceral pleura, with no involvement of the main bronchus, whereas tumours are classified as T2 if there is involvement of the main bronchus (without carina involvement), visceral pleura invasion or atelectasis/ post obstructive pneumonitis extending to the hilum. Tumours are classified as T3 if there is direct invasion of the chest wall (includes parietal pleura and superior sulcus), parietal pericardium or phrenic nerve, or if there are separate tumours in the same lobe. T4 tumours consist of disease with invasion of the heart/ great vessels/ diaphragm/ mediastinum/ trachea/ carina/ oesophagus/ recurrent laryngeal nerve/ vertebral body, or if there are separate tumours in different lobe of the same lung. Nodal classification follows similar principles, where the presence of ipsilateral peribronchial and/or hilar nodes and intrapulmonary nodes is classified as N1 disease; N2 disease consists of the presence of ipsilateral mediastinal and/or subcarinal nodes, and N3 disease involves contralateral mediastinal and/or hilar nodes or the involvement of scalene/ supraclavicular nodes.

The heterogeneity in presentation of lung cancers can make automated segmentation processes more difficult. Nonetheless, differences in the radiodensity between tumour and their surrounding tissue types can be exploited for this purpose, although this can be challenging where tumours are juxtaposed with tissue types of similar radiodensity.

Aims

This project addresses the hypothesis that computational segmentation techniques, with a given prior, can be used to develop an image-based tumour delineation process. The computer-generated contours can then be used as a reference volume for outlining assessment as part of the radiotherapy trials QA review. Ultimately this will allow better accuracy of tumour volume definition, improve radiotherapy dose delivery to the target, and reduce dose to adjacent normal tissues.

In the trial setting, this can help particularly for real time reviews and pre-accrual dummy-runs, both of which lack a gold-standard reference volume that is required for objective assessments to be performed. Additionally, by reducing the need for clinician input and speeding up the contour evaluation process, timely feedback on the submitted volumes can be provided to the treating clinician and avoid delays to patient treatment.

Overview of project

This thesis focuses on the exploration of the segmentation techniques that can be used to develop an automated image-based method for generating reference delineations, where outlining assessments can be performed without a need for expert-generated reference contours.

The novelty of this project is in the use of the clinician submitted outlines as a prior for image segmentation in generating a reference contour, which can then be used to serve as a comparator to the former to evaluate the manual outlines. For the development and training of the algorithm, anonymised image datasets collected from multi-centre trials will be used, to better reflect the heterogeneity of data seen in practice. This also allows testing of the rigour of the system in its ability to handle images and contours from a wide range of datasets obtained from different CT scanners and planning systems.

Building on the existing evidence in the literature on image analysis and segmentation techniques that is promising, lung cancer segmentation is explored in this project.

The rest of this chapter provides the fundamentals and related work to this study, which includes the literature review on various segmentation techniques. Additionally, the application of segmentation algorithms to lung tumours and lymph nodes is discussed in more detail. A review of the available software products is also described in section 1.12.

Literature review

1.8 General principles of image segmentation techniques

There is great interest in the use of computer-based approaches for evaluation of medical imaging across the facets in a clinical pathway, with increasing confidence in computers generally outperforming humans in areas requiring quantification of information derived from imaging (61, 62). One of the main domains of computer vision is image segmentation, which has been applied extensively across numerous non-medical fields (e.g. facial/fingerprint detection and recognition, video surveillance), and has been widely explored in various aspects of medical imaging.

The aims of image segmentation in the medical setting is to identify and subdivide an image into a number of regions with uniformly homogenous features that distinguishes one region from the next. Depending on the objective of the segmentation, this may be an organ, a particular tissue type, or a lesion within an organ. Segmentation plays a crucial role in image analysis especially in the exploration of (fully or partially) automatic workflows, such as in computer-aided diagnosis (CAD) systems (63), treatment planning (64), as well as in the emerging field of radiomics, where images are converted to minable data through the extraction of quantitative imaging features (65, 66). These processes should not only be accurate, but also efficient.

However, automated segmentation of digital images remains one of the most difficult tasks in digital image processing. Some of the factors that limit the accuracy of segmentation include the inhomogeneity of intensity, partial volume effect, image noise and artifacts, and boundary insufficiencies (e.g. 'missing' edges and/or lack of contrast between regions of interest (ROIs)) which are common in medical images. Nonetheless, various methods of computer-aided segmentation have been applied in a range of settings, in an attempt to improve their efficiency, accuracy and applicability.

The principles behind a range of different image segmentation techniques is described below.

1.8.1 Intensity thresholding

Thresholding is a simple concept where a particular -scale value (i.e. threshold) is chosen, and pixels of the image with values higher than the threshold are assigned to one region, whilst those that are below the threshold are assigned into another region. This creates a binary partitioning from an intensity image (see figure 1.3). The fundamental issue which affects the efficacy of this technique lies in the definition of the optimal -scale value (i.e. threshold) used for segmentation. Selection of the optimum threshold can be executed either globally, where the thresholding relies solely on the characteristics of the individual pixel, or locally, where information from neighbouring pixels is used in the segmentation.

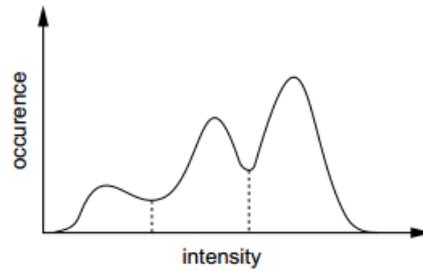


Figure 1.3. Histogram showing three apparent classes where dotted lines represent identification of threshold between the three classes (figure reproduced from (67)).

In the simplest form, global thresholding can be achieved by selecting a single threshold manually which requires *a priori* knowledge about the image, usually evaluated from training data. A way of performing this automatically is through plotting a histogram, fitting a polynomial function to it and selecting the threshold at the minimum turning point of the curve (68). A well-known automatic approach to performing thresholding on a global scale is Otsu's method, which splits the image pixels into classes where there is minimal intra-class variance (i.e. maximal inter-class variance) with the tightest clustering of the pixels of the groups (69). There are many other approaches for threshold selection, which include Kapur's entropy-based method, and Ramesh's shape-based method (70-72). Although these approaches have been traditionally used to partition an image into two classes, further algorithms have been developed to extend this to multi-level thresholding (73-75).

In addition to global threshold approaches, methods for local thresholding have been developed (76). Adaptive thresholding is a means of application of a local threshold in an automatic fashion, where variation in the background -scale intensity is taken into account to partition the object through the comparison of neighbouring pixels (77). This is commonly applied in the presence of uneven background intensities across an image.

The implementation of thresholding techniques can be straight forward, and it often gives a good approximation of the segmentation. Thus, it has been used in many studies as one of the pre-processing steps in image segmentation (78). Because of the impact of noise and artifacts in complex images which can limit its accuracy, thresholding is rarely used as a sole segmentation approach but more commonly in combination with other techniques.

1.8.2 Edge-based segmentation

Edge-based methods are classical tools in image segmentation, where object boundaries are detected and used to separate an image into different entities. These methods are based on locating an edge, which serves as a local image feature, defined as the border between two discrete regions demarcating them into two separate regions. In the identification of an edge, in addition to finding where the abrupt change in grey-level intensity between the two regions lies, the derived boundary should split the two regions into zones with distinctly different properties. Boundary finding algorithms work by detecting the non-homogeneity at an edge through the use of discontinuity measures.

Edge detection can be generally divided into two categories; edge detection techniques based on first derivative operators, and approaches based on derivatives of higher order.

Gradient-based edge detection techniques use first derivative operators on an image to calculate the gradient change between the original pixel values and locates an edge by evaluating the maximum and minimum values of the first derivative image. These include Prewitt (79), Sobel (80) and Roberts cross operators (81) as filters for edge detection. Canny edge detection is a multi-step approach where a Gaussian filter is used to smooth the image prior to constructing the first order gradient intensity map, which is subsequently thinned (82). To remove spurious boundaries, hysteresis thresholding is applied to each of the marked pixel (upper threshold) and its neighbouring edge pixel (lower threshold) to determine if it should be included as part of the edge.

Laplacian-based edge detection works on a similar approach, though instead of using the first derivative values to detect the edge, it uses the second derivative that is performed on the first order derivative (83). In place of detecting the maxima of the gradient magnitude, optimal edges are detected where the second derivative is zero. As the zeros rarely fall exactly on a pixel, zero crossings are isolated in places where one pixel is positive and a neighbour is negative. This does however, tend to produce thicker edges.

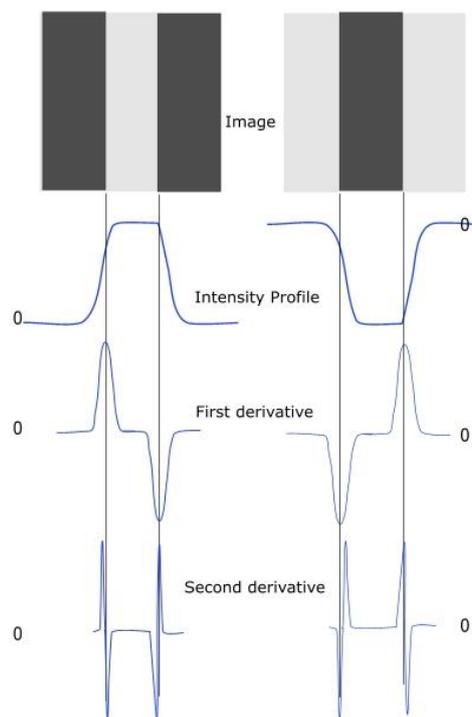


Figure 1.4. Illustration of edge detection by first and second derivative operations for two images (figure reproduced from (84)).

A number of surveys have been performed to compare the various approaches to edge-based detection, and described the advantages and disadvantages seen in the application of the different approaches (85-90). Edge-based approaches generally perform very well in localising boundaries with high contrast. However, the problem with edge detection is that the detected boundaries often do not enclose the object of interest completely, and commonly requires a post-processing step to link the edges identified to create a single boundary contour. This occurs where the boundary between regions is ill-defined, or where an uncertain boundary exists, leading to true edges being missed, and discontinuities in the resultant segmentation. Also, these algorithms tend to be affected by noise, which can cause errors in identifying of an

edge. Conventional edge-based methods that detect changes in grey level rather than absolute value may also be less sensitive to changes in image contrast. Because of these issues, edge-based approaches are often combined with other techniques in medical image segmentation (91-93).

1.8.3 Region-based segmentation

Region-based approaches are based on the identification of homogenous regions i.e. similarities of pixel properties. This can be performed in two main ways, region growing and region splitting.

Region growing is a very popular technique that locates a region of the image by grouping pixels of similar attributes together iteratively. From a particular seed point where the algorithm is initiated, adjacent pixels are recruited into the region, until a set of pre-defined criteria is met (94, 95). The converse can also be adopted i.e. region-splitting, whereby a region that has already been segmented can then be subdivided to segregate the ROI into regions with homogenous pixels. Split and merge methods use a combination of splitting and growing algorithms in the segmentation of images (96).

One advantage of region-based approaches is that they tend to be less affected by noise as compared to edge-based methods. However, they have an inclination to leak into neighbouring regions, resulting in the inclusion regions with similar intensity values.

If a region is homogenous with high contrast compared to the adjacent region from which it is to be portioned from, the detection of the region boundary is a simpler task, where one would expect similar results from either edge or region-based approach. In reality, most images are complex and do not fall into this category, requiring a combination approach to segmentation. To overcome the issue of segments overgrowing into the surrounding regions, region-growing methods typically incorporate some edge-based information in the form of a discontinuity measure(s) as a criterion for the segmentation to stop growing.

There are many different region-growing algorithms that have been developed (94, 95, 97, 98), some of which have been incorporated into software products. Region-growing techniques have been used on a variety of body sites including breast (99, 100), kidney (101, 102), liver (103-105), and lung (106, 107). OncoTREAT is an example of an interactive segmentation system that uses both region growing and mathematical morphology for semi-automated delineation (108).

1.8.4 Edge and region-based hybrid approach (Watershed segmentation)

A popular hybrid approach which uses a region-based technique in addition to gradient information is the watershed algorithm. The watershed approach is based on the topographic concept, whereby regions can be divided by watershed lines at higher elevations, into catchment basins. In topography, a drop of water on one side of the watershed line would flow down into a catchment basin or local minima, with another drop on the other side of the line into an adjacent basin. This concept is easily translatable to image processing, where elevation can be represented by the greyscale levels of each pixel.

Watershed algorithms in image processing have been around for many years, originally described in 1979 (109). Although some of the earliest work is either inefficient or inaccurate,

there has since been numerous watershed transforms developed. The basis of the algorithms is to locate the zone of influence (or catchment basin) of a component within the image, and the boundary of all zones of influence (or watershed lines). In digital images, regions can be partitioned based on the local maximum and minimums of the pixel scale values (see figure 1.5).

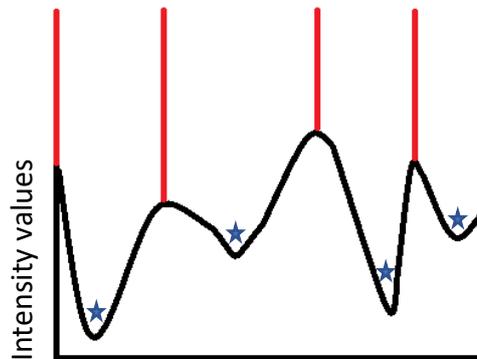


Figure 1.5. Scale profile of image data shown in black lines, with watershed segmentation whereby the local maximum pixel scale values define the watershed lines (denoted by red lines), and local minimum pixel scale values define the catchment basins (denoted by blue stars) (modified diagram based on figure from (110)).

The popularity of watershed segmentation can be attributed to it being computationally fast and its relatively simple and intuitive methodology. Moreover, it has the ability to produce complete partitioning of regions, resulting in no further post-processing to join unlinked contours. The main disadvantage to this approach is in its tendency to result in oversegmentation (see chapter 3 section 3.2), as well as sensitivity to noise.

1.8.5 Model-based segmentation

Another technique that has been employed in medical image segmentation is through deformable or active contour models, which can be thought of as curves evolving within the image to fit the contour of the object boundary, such that it separates foreground from background regions. They work through the use of either surfaces or closed parabolic curves that deform under the influence of internal and external forces, that can vary according to the shape and geometry of the ROI. This is usually performed after the curve/surface has been positioned near the desired boundary of the object, which then undergoes a series of iterative processes that allows the deformation of the surface to fit the boundary of the ROI (90, 111).

Depending on the technique used to track the moving contour, active contour models can be classified into parametric or geometric-based algorithms. Parametric deformable models work by tracking the evolution of the parabolic curve through sampled contour points and solving for the parametric equation determined by the energy functions. The classical active contour model, known as Snakes, was proposed by Kass et al (112), where a contour is attracted to edges of objects in an image by using the idea of energy minimisation. This is based on a spline or curve which is guided by the constraint forces that is also influenced by the image forces. The information within the image can alter the evolution of the spline and cause it to be pulled or pushed towards image features. This means that the evolution of the contour is a dynamic process, which continuously deforms itself from its starting point to conforming to the nearest feature over a period of time.

The principles of parametric active contour models are as follows. Simplistically, propagation of the model is determined by three different forces, external, internal as well as image. The weights of each of these forces can be adjusted to create a range of snake behaviour.

External constraint forces are responsible for placement of the spline near the desired local minimum. This allows information in the form of higher level interpretation to be taken into account in the deformation of the spline, and can be, for example, in the form of a user input. Internal constraint forces govern the regularity of the contour, through geometric properties such as curvature, area or length. With this, the elasticity and rigidity of the spline can be adjusted, thereby allowing control over the smoothness of the resultant contour.

There is a range of different image forces based on the image information that can be used in the energy function. An example is the use of the image intensity itself, where the weight of the function determines whether the contour is pushed towards lines of high or low intensity values. Another example is the use of edge information, where the function can be set up to attract the spline to the location of large image gradients. Because of the energy function of the spline, where it is attracted to the object boundary, neighbouring parts of the curve would also follow, towards a possible continuation of the feature in question. This effect can also be achieved in part by the smoothing effect of the edge- or line-energy function.

The energy of the spline is controlled by the different weights associated with the functions defining the external, internal and image forces, that ultimately determines the location of the curve. The goal of the algorithm is to achieve the state of equilibrium under the push and pull of the various forces at the structure boundary. Adjustment of the weights of the forces thereby exerts different amount of bendiness on the contour and controls the flexibility of the spline. Typically, a termination function is also usually defined, which stops the evolution of the spline once the equilibrium is attained.

In addition to the adaptations of the classical active contour algorithm aimed at improving its performance in the setting of noisy images and being less dependent on the location of the initial curve (113-115), statistical methods have been incorporated and applied in conjunction with the active contour approach, such as in active shape models (ASM) (116) and active appearance models (AAM) (117, 118). In ASM, training data is used to build a statistical boundary shape model of the object of interest, consisting of a mean and the permissible variations pertaining to the shape of the object. After locating to the approximate position of the new image to be segmented, the model is then adjusted to fit to the edge information of the new image. The way in which AAM works is very similar to ASM, which incorporates additional texture information (in the form of mean and permissible range of pixel intensities) across the object to the shape information, which is used in building and fitting the model.

In tracking the contour points explicitly as with the parametric approach, there are situations in which the method would not work, such as at a sharp propagating front where ‘corners’ may be in an unknown state, or in regions of topological changes (e.g. peaks and troughs) where splitting or merging of contours cannot be performed. To better handle these issues during the curve evolution, another technique has been developed which utilises the level set approach (119, 120), of which geometric deformable models are based on. Instead of tracking the contour points explicitly within a curve, for the level set method, a surface is tracked in place of a front, with the front defined at the points in which the height of the surface is zero. In other words, the contour is transformed into a higher dimension level set function, where the contour is

represented as its zero-level set. The main advantage to this approach is in its ability to handle the topological changes and geometric properties, thereby allowing splitting and merging of the implicit contour.

Segmentation of medical images have been performed with Malladi's speed model which uses a level set technique, where the evolution of the curve is inverse to the presence of gradient magnitude in the image (121, 122). Further modifications to the algorithm with less dependence on gradient information have been developed by other groups to overcome the leakage problem seen with Malladi's approach, which affected noisy and blurred images (123, 124). The geodesic active contour derives the energy function (similar to the parametric approach) which is then embedded into the level set equation, the solution to which corresponds to the minimal distance curve in the image (125).

Unlike edge-based methods, model-based segmentation allows the curve for the object boundary to be retained as a closed polygon, and the nature of the algorithms ensures that spurious edges can be avoided and is therefore less sensitive to image noise. These methods have been used in a number of autosegmentation studies on MRIs and CT scans across sites such as the brain, lung and liver (126-130). ITK-SNAP is an example of a software application that utilises active contour methods for its semi-automatic segmentation tool (126).

1.8.6 Graph-cut approach

Image segmentation using graph-cuts has been a popular choice for many years in computer vision, which is based on using maximum-flow/minimum-cut algorithms to minimize certain energy functions and to partition a graph into disjoint sets.

In graph theory, a graph is a structure consisting of a number of objects (vertices or nodes), in which some of the objects are related to another, where each of the related pairs is known as an edge. There are two main graphical models, Bayesian networks (also known as directed graphical models), where there is an orientation to the relationship/edge between the nodes, and Markov random fields (also called undirected graphical models), where there is no directional significance between the nodes. Within an image, the nodes are made up of the image elements, which can be in the form of individual pixels or regions. Each edge has a corresponding weight that specifies the quantity based on the property between the two nodes connected by the edge.

A graph normally also contains some special nodes, called terminals, which correspond to the set of labels that can be assigned to pixels in an image. Figure 1.6 illustrates a 3 x 3-pixel image in the presence of two terminals, known as the source, s , and the sink, t . The edges connecting neighbourhood pixels are known as n -links, where a penalty can be given for any discontinuity between the pixels. The edges connecting the terminals to the pixels are known as t -links, where similarly, a penalty can be given in the assignment of one pixel to one of the labels and not the other.

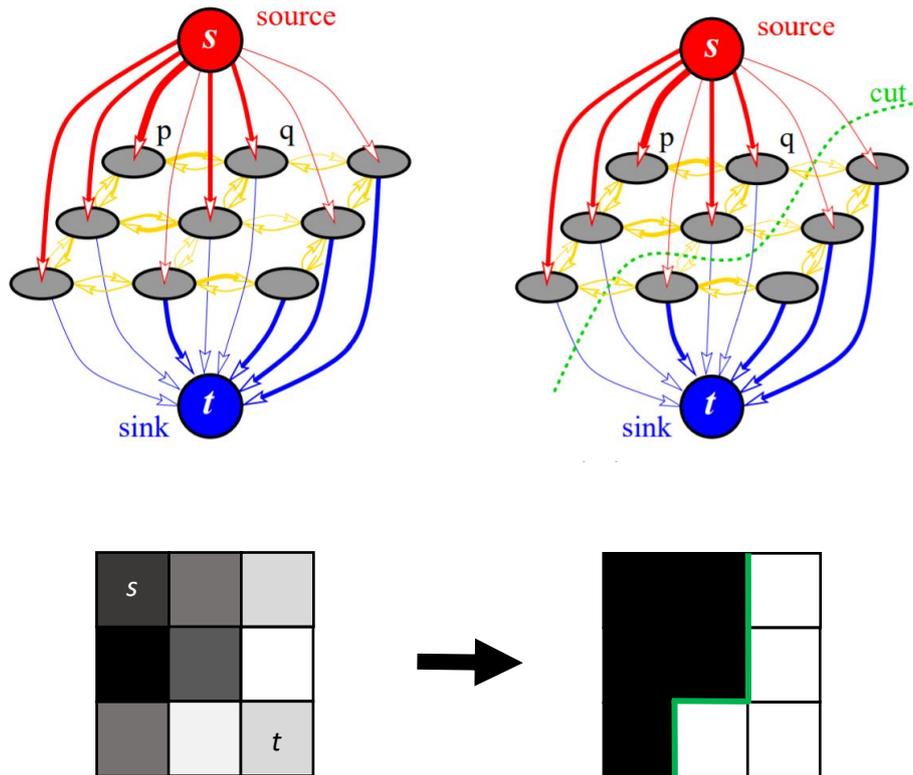


Figure 1.6. Example of a directed weighted graph of a 3 x 3 image, with each pixel represented by grey nodes, and edges represented by yellow (n-links) and red (t-links) arrows (figure reproduced from (131)). In the image, s and t denote the location of the source and sink respectively, with the green line representing the segmentation results.

Simplistically, the algorithm requires input with two user-defined seed points (or groups), which determines the samples of the foreground object (corresponding to the source) and the background (corresponding to the sink) in the image.

For image segmentation, the image is viewed as a graph where the pixels are associated with nodes and the edges are weighted to reflect image gradients. Partitioning to create mutually exclusive regions between the foreground and background (i.e. the two pre-defined seed points) can be performed with divisions known as cuts, where cuts are associated with an energy. Maximum-flow/minimum-cut is one of the ways of minimising the energy function to achieve this, as initially demonstrated by Greig et al in the setting of binary images (132), where the labelling resulting from a minimum cost cut would correspond to the minimum value of the energy, when edge weights are appropriately set based on parameters of an energy. Because of the way in which graph-cut algorithms are solved (i.e. through minimising the objective function), this approach is generally considered as an energy-based/ model-based technique. Unlike the classical active contour approach that utilises boundary information and thus is very sensitive to the initialisation curve, the energy function in graph-cuts is constructed based both on regional and boundary information, allowing the globally optimal result to be achieved. Additionally, some graph-cut algorithms also incorporate clustering methods for pixel labelling, where the partitioning is performed in the spatial domain by the graph-cut solution (133).

Since then, there has been extensive progress to further improve algorithms in the application and processing of cost functions for different graph-cut models (134-148). There has also been development of software products that perform segmentation based on graph-cut algorithms. Examples of these include non-medical applications such as Graph-cut Textures (149) and Photomontage (150), programmes where there has been some use in the medical field such as GrabCut (151) and GridCut (131, 152, 153), and others designed specifically for medical image segmentation, for instance Sim Cut (154).

1.8.7 Atlas-based segmentation

This approach differs from the techniques described above, in that it requires the prerequisite of pre-defined atlases or templates of the anatomy as a reference framework which is used to segment new images. Therefore, atlas-based approaches can be considered as supervised algorithms, requiring labelled training image data which are processed, warped and transformed onto the new image. The basis of these algorithms is the use of registration techniques in order to align the atlas image(s) to the new image, prior to propagation of the volumes.

A review on the types of atlas-based segmentation has been published recently (155). Most state-of-the-art algorithms use a multi-atlas approach, which classically involves a search of (a proportion of) atlases in the library to find the one with the best match to the new image. This is performed through a registration process, of which many different deformable methods have been explored. The labelling of the new image can be based on a single atlas, or on multiple atlases through a voting scheme (majority, or weighted voting, which takes the intensity values of the new image into account). Post-processing to further refine the segmentation can also be carried out.

In radiotherapy planning, the vast majority of atlas-based segmentation has been applied in the auto-segmentation of normal tissues and OARs (156-160) or CTV delineation determined by organ/ anatomical boundaries (161, 162), rather than on tumour segmentation (163, 164). This is due to the inherent nature of most atlas-based algorithms which are based on registration, matching and labelling processes that maintains the contextual information from the original atlas(es). Thus, due to the heterogeneity of tumour location, size and shape, they generally do not subscribe to the classical atlas-based scheme.

1.8.8 Classification and clustering

A different approach to image segmentation is through clustering techniques, which is also popular in the medical domain. These methods work by partitioning the data of particular points within an image into a number of sections. These can be broadly classified into supervised classification and unsupervised clustering algorithms.

1.8.8.1 Supervised

A review by Kotsiantis et al summarises various supervised classification techniques (165). Supervised algorithms work on the premise that information from trained data points with pre-defined class labels are used to classify new data points into the relevant classes. The statistics of the structures of interest in the training dataset is extracted in the first instance, from which the functions of the classification system are derived. These are then applied to the new data points on which the classification, and thereby segmentation, is performed. Examples of such

techniques include statistical methods (e.g. discriminant analysis (166)) and machine learning algorithms such as k-nearest neighbour (k-NN) classifiers (167), random forests (168), support vector machines (SVM) (169) and supervised artificial neural networks (ANN) (170).

Classification of data points through discriminant analysis involves deriving a mixture model to describe the data and finding the best separator of the data into their respective classes. This uses maximum likelihood estimation where the parameter values that best describe the models in their respective classes are estimated, which are then applied to the new data points for classification. This form of classification tends to work well for data with a Gaussian distribution. However, with a non-normal distribution, the results may be unsatisfactory.

On the other hand, in k-NN classification, no assumptions about the underlying distributions of the data are made. In brief, the k-NN algorithms work by assigning the most frequent appearing class label in the selected group of trained data points to the data point in question. Optimisation of the classification is performed through selection of the size of the group of data points (neighbour size). The training phase for k-NN algorithms is minimal and thus fast. Performance of the classifier generally improves with larger sample sizes. The computational cost of k-NN classifier is in the testing phase, both in terms of speed and memory use. Depending on the selected neighbour size, classification accuracy may also be affected in the presence of imbalanced datasets.

Like k-NN algorithms, SVMs are non-parametric classifiers. They work by identifying the hyperplane(s) that best discriminates between the classes of the training samples, where the selection of the hyperplane(s) is such that the distance between the given classes are maximised. New data is then classified based on the defined hyperplanes. One of the main advantages of SVMs is in the use of functions (kernel trick) to map to higher dimensional data without the need of transforming individual data points to obtain the decision boundary between the classes, thereby reducing the computational needs. This is helpful for non-linearly separable data that can be segregated into the respective classes in higher dimensional space.

Random decision forests are also non-parametric classifiers, which are made up of multiple decision trees, each of which work on a rule-based system to classify a variable with its respective class label. The set of rules within the trees are computed during training of the model corresponding to the predictor variables and class labels of samples from the training dataset. During the testing phase with new data, all the output from the terminal nodes are then considered through a weighted average or voting scheme to determine the class label. Random forests are generally fast to compute and tend to be less affected in the presence of data imbalance, although they tend to over-fit with noisy datasets.

ANNs operate via computation through a series of nodes, similar to neurons in a human brain. Each of the nodes are associated with a probabilistic behaviour, and a grid of these nodes act as a bridge between the input and output layer of nodes. The whole complex is trained using the training data where information is fed to the input layer of nodes, which is then transmitted through the whole complex of nodes, until the desired output is obtained in the output layer. This process is performed repeatedly and automatically by the ANN where the weights of the nodes are altered in order to change the bias in which the information is passed through the complex (backpropagation), to fit the input to the output as determined by the training samples, i.e. the system learns and updates itself as training data is processed. Testing data is then passed through the trained complex in order to make classification predictions. To optimise ANNs,

the number of nodes within the network can be selected, where a greater number would increase the complexity of the network. Other variations of neural network-based algorithms include deep neural networks, where there are multiple (hidden) layers of nodes as compared to ordinary ANNs. Convolutional neural networks (CNNs) are similar to ordinary ANNs but instead of connecting all the nodes from one layer to a single node in the next, only a proportion of nodes is connected (i.e. convolving with a filter). The new feature maps are generated as a result of each neuron being connected only to a subset of the input image (i.e. local connectivity), and the computational efficiency is typically improved through pooling within each feature map generated.

One of the main advantages that ANNs offer is its ability to detect complex non-linear relationships between the input and output variables, without requiring restrictive assumptions to be made about the data. However, as better predictions are made with more training data, ANNs tend to work poorly for smaller datasets. Additionally, optimisation of ANNs is difficult and is generally based on trial-and-error, which can take a long time to set up. Moreover, the processing of the data within ANNs is a 'black-box', and generally little information can be extracted as to how a prediction for a particular variable is made.

1.8.8.2 Unsupervised

Clustering methods involve segregation of the image pixel values into groups closer to the respective means of the distributions (171). They have been applied to a number of medical image segmentation processes, although they are more commonly used as an initial step in the sequence of image processing rather than a sole approach (67, 71, 72, 172-174). This is due to their poor performance in the setting of image noise and artifacts. K-means clustering (175), where k represents the number of groups, is a common application that works by assigning each data point to one of the k groups based on feature similarity. This results in disjoint partitions. An extension to this is soft (fuzzy) clustering where a membership function is assigned to each cluster, such that data points can belong to more than one cluster (176). Fuzzy C means is an example of such an approach that combines K-means clustering with fuzzy theory (177), which has been used in the setting of medical image segmentation. Unsupervised neural networks have also been used to perform clustering tasks.

1.8.9 Texture-based segmentation and texture analysis

At the most basic level, the methods described above can be applied on the individual pixel values of an image, without consideration of neighbouring pixel values. Recently, there is great interest in analysing the relationships of pixels with their neighbours, in an attempt to extract more information from their spatial relationships. This forms the basis of texture descriptors that describe the appearance of regions within an image, akin to a surface being considered smooth, rough, fine or coarse. As such, any property that describes the appearance of a region can be considered as a texture feature. Laws identified that these properties are important in the evaluation of texture: uniformity, density, coarseness, roughness, regularity, linearity, directionality, direction, frequency and phase (178). The texture of a tissue is based on the subtle patterns that differs between tissues, which can be used to partition one region from another forming the premise of texture segmentation.

The field of radiomics has vastly increased recently (65), with a growing number of studies investigating textural analysis (TA) in the clinical setting. The methods in which texture can be evaluated can be classified into four main groups.

1.8.9.1 Statistical methods

Statistical methods are commonly used to evaluate texture, which entail the quantitative analysis of the pixels with their neighbouring elements. In two-dimensional (2D) images, texture is related to the grey-level values of the pixels of the image in the plane of interest. There are different ways of evaluating the grey-level intensity in relation to their position (179). The spatial grey-level co-occurrence is a second-order statistic representing the frequency of specific grey-level values in the ROI. Haralick has proposed a number of useful features that can be computed from grey-level co-occurrence matrices that can be used for classification tasks (180). Autocorrelation features describes the amount of regularity and repetition that is present in a region, which allows differentiation of regions with peaks and troughs from those with a smooth texture.

1.8.9.2 Geometrical methods

Instead of analysing texture on a pixel-wise basis, this class of methods define larger texture elements on which TA is performed. One approach is to construct elements using Voronoi tessellation, where the image is partitioned into regions based on distance to points (181). After the computation of features within these regions, elements are then grouped together to result in segmentation based on edges. This method has been used for segmenting cellular histopathological specimens rather than on imaging (182). Structural methods are not generally applicable in the medical setting as they work predominantly on very regular textures.

1.8.9.3 Model-based methods

These approaches include analysing the image in the form of Markov random field (MRF) models that assume each pixel intensity depends on the intensities of only the neighbouring pixels. With these relationships, the texture of the image can be modelled globally by specifying the total energy, or locally by specifying the local interactions of neighbouring pixels through conditional probabilities (181). Another method is through computation of the fractal dimension which indicates roughness as a texture property, which has been found to be useful for stratification of lung cancer aggressiveness as well as prognostication (183, 184).

1.8.9.4 Signal-processing methods

Human visual perception has been shown to involve the analysis of frequency variation (185). Thus, frequency analysis through Fourier filtering which breaks down an image into its frequency and orientation components have been used in texture analysis of images, and similar principles apply to wavelet analysis. Gabor filters have also been used, which are 2D transforms that model on the human cortical receptive field (186, 187). There are many uses for these filters in image processing including denoising, analysis, segmentation, and are heavily applied in pattern recognition processes such as iris, fingerprint and facial recognition (188).

Texture descriptors as discussed above have been used as classification features to partition an image into regions of different texture qualities. Texture-based segmentation has been performed in the setting of head and neck cancer using PET-CT as well as in the brain (189-

192). Not only has texture been used for segmentation of pathological lung tissue, they have also been applied to lung tumours (see sections 1.9.7 and chapter 6 introduction).

1.8.10 Deep learning

Compared to conventional techniques where expertise is required to design feature extractors that generate representative patterns which are then fed into a detector system, deep learning methods process unlabelled data in its raw form, allowing the machine to discover the patterns required for classification by itself (193). There have been breakthroughs across many fields with the use of these methods, including image recognition (194), speech recognition (195), and language translation (196).

In brief, deep learning approaches are made up of multiple processing layers to learn data with information that is challenging to analyse. A typical architecture of such a network comprises of a CNN made up of stacks of convolutional layers where max pooling (maximum value from each cluster of neurons of the prior layer) is performed in order to decrease the computational demands.

To enable the machine to self-generate its own features for use, a further unsupervised neural network is built into the architecture as an autoencoder within the CNN. Unlike supervised learning where there is the need for labelled data, output data is not required for unsupervised neural network algorithms. As such, these neural networks do not distinguish the input data into their respective classes when used on their own. Instead, they are capable of self-learning from the input data and can distil the input information into forms that can be better used for segregation purposes. In deep learning, unsupervised neural networks are incorporated in the early steps within the architecture of CNNs as autoencoders. Not only does this remove the need for feature extraction, autoencoders reduce the representation of the original image, and thus speed up the learning process of the CNN.

A recent review of biomedical applications of deep learning algorithms has shown numerous applications in detection, classification and segmentation tasks (197). At present application of tumour segmentation using deep learning is predominantly within the brain (198-201), though it is being explored in other sites such as head and neck and liver cancers (202-204).

1.9 Segmentation of lung lesions – CT segmentation techniques

The segmentation techniques described above have been extensively applied in the thorax, both in terms of normal tissue segmentation e.g. the lung parenchyma, and nodule or tumour segmentation. This section focuses on the reported (semi-)automatic methods for the segmentation of lung lesions on CT imaging, which is summarised in table 1.1. Most of the studies have been performed in the diagnostic setting as part of the CAD process, aimed at the detection and analysis of pulmonary nodules. Fewer studies have been performed on larger tumours in the setting of more advanced disease.

Study	Year	Data	Lesion type	Segmentation approach	Automation	Description	Performance
Thresholding							
Iqbal et al (205)	2014	60 cases from <i>Lung Image Database Consortium (LIDC)</i> database	Diameter 3 – 30mm	Adaptive thresholding	Fully	Lung region extraction followed by variable multistep thresholding, with false positive reduction based on shape indices.	Sensitivity 92%
Jibi et al (206)	2016	10 cases from LIDC and Vision and Image Analysis Group database	Solid, non-solid; juxtaleural	Adaptive thresholding	Fully	Lung region extraction followed by multistep thresholding, comprising of an intermediate thresholding step, and further filtering on 11 morphological and intensity-based features based on thresholds.	Qualitative and number of nodules reported
Mullally et al (207)	2008	40 nodules from phantom data; 12 nodule pairs from clinical data	Solid; isolated; juxtaleural	Adaptive thresholding	Fully	Automated nodule detection through adaptive thresholding, with final segmentation from threshold based on gradient and shape information.	Root mean squared error for phantom data 0.54; Mean volume difference 13mm ³
Taşcı et al (208)	2014	1269 nodules from LIDC dataset	Juxtaleural	Otsu's thresholding; morphological	Fully	Otsu's thresholding to segment lung fields, followed by morphological operations including alpha hull to segment juxtaleural nodules. Texture features extracted for classification.	Train set: Area under curve (AUC) 0.9679, accuracy 95.88; Test set: AUC 0.887, accuracy 91.49
Zhao et al (209, 210)	1999	9 and 12 nodules for 2D and 3D evaluation respectively	Diameter < 10mm, juxtavascular	Thresholding	Fully	A multi-criterion means of selection of optimal threshold level, based on gradient strength and shape compactness associated with at each threshold.	2D: No statistical difference to manual contours; 3D: acceptable rate 77.4%
Mathematical morphology							
Fan et al (211)	2002	Synthetic phantom (12 nodules); 2 clinical cases (7 nodules)	Diameter 2 – 10mm for phantom, 2.5 – 6mm for clinical cases	Thresholding, morphological and application of 3D template (intensity-based)	Fully	Thresholding applied, then morphological operations to initialise 3D template, followed by propagation through cross correlation to ROI (based on intensity). Refinement through spatial reasoning.	Interscan volume measurement deviation of 2.8% for phantom data and 8.1% for clinical cases
Fetita et al (212)	2003	300 nodules from 10 clinical cases	Diameter 2 – 20mm;	Morphological	Fully	Grey level morphological operations with selective marking and depth constrained connection cost.	Sensitivity 85 – 97%; Specificity 90 – 98%

			Isolated; juxtaleural; juxtavasular				
Kostis et al (213)	2003	105 nodules from clinical cases	Diameter < 10mm; juxtaleural; juxtavasular	Region growing; Morphological	Fully	Fixed thresholding, connected component analysis followed by morphological processes (iterative morphological opening) for vascular subtraction and pleural surface removal.	Acceptable rate 80% and 72% for juxtavasular and juxtaleural cases respectively
Moltz et al (214)	2009	333 nodules from clinical cases	Juxtaleural	Region growing; Morphological	User stroke initiation	Region growing followed by sequential morphological processing to exclude pleural surfaces	Acceptable rate 89%
Region-growing							
Setio et al (215)	2015	238 out of 888 cases from LIDC dataset	Solid nodules only, sizes > 10mm	Region growing	Fully	Thresholding, connected component analysis and region growing. Classification based on intensity, contextual and shape features.	Detection rate 92.2%
Velazquez et al (216)	2012	20 cases from clinical dataset	Stage Ib – IIIb	Region growing	User click for initiation	Single click ensemble method of region growing from user seed point	Overlap fraction 0.925 ± 0.09 for observer intersection
Velazquez et al (217)	2013	20 cases from clinical dataset	Stage Ib – IIIb	Region growing	User defined foreground and background	Grow cut application based on cellular automata for region growing	Overlap fraction 0.943 ± 0.044 for observer intersection
Song et al (218)	2016	850 lesions from LIDC database, 121 lesions from 100 NSCLC patients	Solid, ground-glass opacity (GGO), Cavitating	Region growing	Fully	Extraction of lung parenchyma followed by application of toboggan algorithm to gradient image to generate seed points. Iterative region-growing with distance and growing-degree constraint applied, followed by boundary refinement based on geometry.	True positives = 96.4%, False negatives 95.0% Overall DICE (mean, SD) = 0.82 ± 0.04 Overall Hausdorff distance 3.52
Namin et al (219)	2010	63 cases (134 nodules) from LIDC dataset	Diameter 2 – 20mm	Region growing	Fully	Adaptive thresholding and morphological operators for lung field extraction, followed by gaussian filtering, shape analysis and region growing	Sensitivity 88%, 10.3 false positives per CT scan
Parveen et al (220)	2013	11 clinical cases	General	Region growing	Fully	Lung field extraction through morphological processing followed by region growing for nodule segmentation	Qualitative
Gu et al (221)	2013	15 out of 129 clinical cases	Stage I and II NSCLC	Region growing	User defined seed initiation	From a single user seed point, multiple seed points generated on which region growing technique is applied based on intensity mean, standard deviation,	Mean Jaccard index 0.7829 and 0.7772

						shape and connection status. Voting strategy applied voxel-wise to determine inclusion class.	for user 1 and 2 respectively
Kubota et al (222)	2011	23 nodules from LIDC dataset 1; 82 nodules from LIDC dataset 2; 820 nodules from clinical cases	Juxtapleural; juxtavascular; solid; non-solid	Region growing	Fully	Foreground and background separation with competition-diffusion filtering, followed by nodule core extraction based on Euclidean distance map. Region growing applied to seed points, followed by surface estimation and convex hull.	Mean Jaccard 0.69 ± 0.18 , 0.59 ± 0.19 for LIDC datasets 1 and 2 respectively. 85% of diameter estimates within 30% of manual measurement for clinical cases
Diciotti et al (223)	2008	Synthetic phantom; Development data: Italgung-CT lung cancer screening program; Test data: LIDC	Juxtavascular; mean diameter 5 – 10mm for developmental data	Region growing	Semi-automatic; initial region selection and input for non-nodule marker	Within candidate regions selected by user, semi-automatic marker placement (local maxima of LoG filtered image) requiring user input to distinguish nodule from non-nodule, followed by region-growing (based on intensity values and geodesic distance).	Acceptable rate 86.3% and 83.3% for developmental and test data respectively (Juxtavascular 79.7% and 75% respectively)
Kuhnigk et al (224)	2006	Phantom data for development; 105 nodules from clinical cases	Juxtapleural; juxtavascular	Region growing; morphological	Fully	Region growing applied from fixed threshold, chest wall separation through approximation with convex hull, vascular removal through morphological operations based on distance map.	Acceptable rate 91.4%
Lassen et al (225)	2015	LIDC dataset (19 nodules; 40 nodules)	Non- and part-solid	Region growing; morphological	User initiating stroke	Different thresholds applied to different cases for region growing approach, morphological refinement to remove vessels and chest wall.	Dataset 1: Jaccard = 0.52 ± 0.07 ; Hausdorff distance = 2.79 ± 1.22 Dataset 2: Jaccard = 0.50 ± 0.14 ; Hausdorff distance = 3.37 ± 2.47
Krishnamurthy et al (226)	2016	10 cases from American Association of Physicists in Medicine (AAPM) database; 10 cases from LIDC dataset	Isolated; juxtapleural; juxtavascular	Region-based; morphological	Fully	Lung extraction using thresholding and region-growing. Bridge and fill morphological operations applied to include juxtapleural nodules. Removal of false positive nodules through classification based on 3D centroid shift analysis.	Acceptable rate 100%
Diciotti et al (227)	2011	Development data (256 nodules):	Juxtavascular; mean	Region growing with shape	Fully automatic	Initial segmentation with connectivity based on geodesic distance map, followed by application of local	Acceptable rate 84.8% and 88.5%

		Italung-CT lung cancer screening program; Test data (157 nodules): LIDC	diameter 3.8 – 9.6 mm for developmental data	analysis refinement	(with optional mode for interactive refinement)	shape analysis through evaluation of grown regions and geodesic distance map.	for developmental and test data respectively
Santos et al (228)	2014	140 cases from LIDC database	Diameters 2 – 10mm	Region growing; model fitting; shape analysis	Fully	Thresholding then region growing with morphological process to extract lung fields and lesions. Gaussian model fitting and shape analysis (Hessian matrix) applied to remove vessels and pleura. Classification of texture descriptors.	Sensitivity 90.6%; specificity 85%
Dehmenshki et al (107)	2008	Developmental data: 343 cases (608 nodules) Testing data: 80 cases (207 nodules) from 2 centres	Mixture of solid, non-solid nodules; juxtaleural, juxtavasculature	Region growing; Fuzzy connectivity	User seed placement and selection of final segmentation	Contrast-based region growing from selected seed point within a fuzzy connectivity map; alternative solutions provided should initial results be suboptimal.	85% and 83% acceptable rate for development and testing data respectively
Badura et al (229)	2014	23 and 551 cases from LIDC database for development and assessment respectively	Nodules > 3mm for assessment dataset	Fuzzy connectivity with adaptive thresholding	Single manual click on nodule for initiation	Seed points (manual for nodule; automatic via genetic algorithm for background) applied to binary masks generated using Otsu's thresholding +/- connectivity analysis, morphological operations, followed by fuzzy connectedness with adaptive threshold estimation, and morphological post-processing for vessel removal.	Mean Jaccard = 0.60 for development dataset; Jaccard = 0.69 at 50% probability for assessment dataset
Watershed							
Brown et al (230)	2014	108 cases from LIDC	General	Region/Edge-based (Watershed)	Fully	Intensity thresholding, detection of nodules using watershed on Euclidean distance, followed by region growing. Shape and volume rules for vasculature removal.	Concordance correlation coefficient 0.90 and 0.91 for volume and longest diameter respectively
Tan et al (231)	2013	32 cases from RIDER dataset and 23 cases from LIDC dataset; 22 lesions from anthropomorphic phantom dataset	Solid, part-solid, GGO. Isolated, juxta-pleural Median size: 12100 mm ³ from RIDER dataset, 127 mm ³ from LIDC dataset	Region/Edge-based; +/- Model-based	Manual region initiation	Classification of lesions into solid versus non-solid (part solid or pure GGO) based on mean density. Threshold applied based on Gaussian mixture model, followed by marker-controlled watershed. Further refinement performed with geometric snake model. Non-solid lesions further segmented with MRF model.	Overall overlap ratio (mean) = 69% Solid lesions overlap ratio = 71% Non-solid lesions overlap ratio = 60%

Goodman et al (232)	2006	50 nodules in 29 clinical cases	Mixed (solid and nonsolid, spiculations, juxtaleural, juxtavasculer)	Region/Edge-based	Manual region selection and seed initiation	Automatic reposition of seed point followed by watershed segmentation. Model-based shape analysis to allow nodule separation from chest wall and mediastinum.	Acceptable rate 97% out of 450 observations
Vivanti et al (233)	2015	40 clinical cases	Mean volume 43.8 ± 49.9ml. Isolated; juxtaleural	Region/Edge-based; model fitting	Fully	Segmentation on serial CTs based on maximum likelihood estimation of registered initial outline. Segmentation leaks removal by watershed and modelling parabolic curve.	Improvement of DSC by 30% over fast marching approach
Deformable models							
Kawata et al (234)	1998	62 clinical cases	Diameter 6 – 25mm	Active contour	Fully	Application of active contour segmentation using level sets.	Qualitative with segmentation improvement for ill-defined nodules
Cascio et al (235)	2012	84 cases from LIDC with 148 nodules	Parenchymal and juxtaleural nodules > 3mm	Mass-spring model for nodule segmentation	Fully	CAD system for detection of lung nodules. Whole lung parenchymal extraction with region-growing and morphological operators, whilst nodule segmentation performed based on 3D mass-spring model. Neural network classification based on 3 geometric and 4 intensity features performed for false positive reduction.	Sensitivity of 88% with 2.5 false positives per CT scan
Chen et al	2012	20 local cases (416 nodules) and 20 cases (55 nodules) from LIDC	Solitary pulmonary nodules	Fast marching method	Fully	CAD system allowing of extraction of vessels separate to nodules. Whole lung extraction, application of linear and blob structure filters for vessel and nodule enhancement, followed by front surface propagation through fast marching method.	True positive of 95% and 91.5% with 9.8 and 10.5 false positives per scan for local and LIDC data respectively
Farag et al (236)	2006	29 clinical cases (350 nodules)	Diameters 3 – 30mm; Solid, cavitations, juxtaleural	Deformable model	Fully	Initial rough segmentation with lung extraction, followed by energy-based deformable model (with incorporation of a generic MRF-based prior in energy term)	Error range of 0.4 – 2.35%
Farag et al (237)	2011	115 out of 397 nodules and 50 cases from the Early lung cancer action project (ELCAP)	Juxtaleural	Deformable model	Fully	Segmentation through level set approach with narrow band implementation	Acceptable rate 70%

Farag et al (238)	2013	742 nodules from 4 databases (ELCAP, low dose CT, standard dose CT, LIDC)	Diameter 2 – 20mm; Pre- and post-contrast; juxtaleural, juxtavascular	Deformable model	Fully	Application of adaptive shape model prior to variational level set with adaptive object and background probability density function.	Overall acceptable rate 94.12%
Soltaninejad et al (239)	2012	58 cases from clinical and ANODE09 study (240)	Isolated; juxtaleural; cavitating	Classification for nodule detection and active contour for segmentation	Fully	Lung extraction with adaptive thresholding and morphological processes. 2D intensity based and 3D averaging features used in k-NN classifier for nodule detection, segmentation performed using active contour.	Acceptable detection rate 90%, 5.63 false positives per scan
Way et al (241)	2006	96 nodules from clinical cases and 23 cases from LIDC dataset	General	Deformable model with active contour	Fully	Lung field extraction followed by k-means clustering and morphological opening for initial boundary estimation. Active contour with gradient, curvature and mask energy terms applied. Texture features extraction for classification purposes.	Mean Jaccard over 0.5 at 50% probability
Yip et al (242)	2017	354 nodules (274 cases) from LIDC dataset	Solid, part-solid, non-solid; Juxtavascular, juxtaleural. Median volume 309ml	Deformable model with level set	Fully	Seed points generated from manual contours, search region 30mm from seed point. Geodesic active contour applied with energy terms for chest wall (thresholding and morphological operations), Sato vesselness filter, canny edge detector and sigmoid function.	Median DSC 0.60; acceptable rate 13%
Yoo et al (243)	2006	3 nodules	GGO	Deformable model	Localisation of initialisation not discussed	Asymmetric multi-phase deformable model of 2 level set functions.	Qualitative
Plajer et al (244)	2010	5 clinical cases	Mixed, large, juxtaleural	Active contour	User initialisation	Active contour model with mixed internal-external force based on a cluster function.	Qualitative
Zheng et al (245, 246)	2007 2009	3 clinical cases (245); 12 lesions from 10 cases (246)	Solid; GGO	Graph-cut	Fully	Segmentation of serial CT scans. B-spline nonrigid registration for lung and rigid registration for tumour. Graph-cut segmentation applied for nodules.	Mean of volume variation 0.8%
Lermé et al (247, 248)	2010	10 clinical cases	Atelectasis; Hilar	Graph-cut	Semi-automatic	Interactive segmentation of lung lesions based on graph-cut	Mean DSC 0.7989
Shen et al (249)	2017	10 clinical cases	Large, juxtaleural	Graph-cut	Semi-automatic	Graph-cut segmentation applied in 4D	Mean DSC 0.855 ± 0.048

Classification and clustering							
Brower et al (250)	2007	75 cases from ELCAP; single radiologist as comparator	Non-solid lesions; diameters 5.6 – 17.5mm	Probabilistic classification based on HU	Manual initiation with line through largest nodule slice	Conversion to isotropic voxels and noise filtering, followed by voxel classification with gaussian intensity model into 3 tissue groups (solid; non-solid; parenchyma) and vessel removal filter.	Median growth consistency 1.87 versus 3.12 by radiologist
Zhang et al (251)	2004	23 nodules (8 patients) from clinical dataset	GGO	Probabilistic classification	User pre-defined click	MRF segmentation with spatial constraints from neighbouring pixels with probabilistic density to assign segmentation class voxel-wise.	Acceptable rate 91.3%
van Ginnekan et al (252)	2006	23 cases from LIDC dataset	Solid and non-solid	Discriminative classification within region growing process	Initiation not explored (seed point within nodules provided)	Extraction of lung fields, followed by application of k-NN regression as discriminative classifier using a region growing approach. Features derived from density values, shape and morphological processes. Iterative morphological filtering used for solid lesions as additional feature.	Mean soft-overlap 0.62 ± 0.1
Netto et al (253)	2012	198 nodules (50 cases) from LIDC dataset	General	Clustering and region growing	Fully	Lung field extraction followed by clustering through growing neural gas algorithm to extract high density tissues. Region growing applied to distance transform map to remove vessels, followed by feature extraction and SVM classification.	Sensitivity 85.93 ± 3.98 , specificity 90.79 ± 1.19 (out of 29 cases)
Nie et al (254)	2007	39 nodules	General (solid and sub-solid)	Clustering based on density distribution	Fully	Computation of convergence index features followed by application of unsupervised mean shift clustering to segment nodules.	Mean accuracy 89%
Nithila et al (255, 256)	2016	106 cases from LIDC and AAPM	Solid, part- and non-solid; juxtaleural, juxtavasular	Clustering	Fully	Lung extraction with deformable model, followed by unsupervised nodule extraction with fuzzy C means. Classification within CAD based on texture features.	Spatial overlap 0.584; Classification accuracy 98%, 99.5% and 97.2% for solid, part-solid and non-solid respectively
Zhou et al (257, 258)	2005	10 lesions from clinical dataset	GGO	Supervised classification	Fully	Nonparametric density estimation and likelihood map (based on texture) with k-NN classifier, vessel removal based on shape (Hessian matrix).	Qualitative
Texture analysis							
Kakar et al (259)	2009	42 images from 2 clinical datasets	General	Texture analysis (Gabor filter); clustering;	Fully	Extraction of 20 Gabor features followed by feature selection. Fuzzy C means clustering applied via genetic algorithm to automate clusters, followed by supervised classification with a trained SVM.	Accuracy (True positive fraction) 89.04%; Receiver

				supervised classification			operating characteristic (ROC) 0.9548
Tao et al (260)	2009	100 out of 1100 nodules from clinical dataset	GGO	Texture analysis; supervised classification; shape prior	Fully	Voxel-wise labelling determined by multiphase LDA classification based on 39 texture features. Multiscale blobness filtering applied to obtain shape prior probability map, to which labelled voxels are applied.	GGO: Jaccard 0.68, voxel scale accuracy GGO: 89.87%, overall: 92.28%
Hossain et al (261)	2015	18 clinical datasets	General	Texture analysis; supervised classification	Fully	Lung field extraction, followed by texture analysis (10 features) used for classification by chi-square distance measure.	DSC 88% and 84.4% for GLCM and histogram features respectively
Zinoveva et al (262)	2011	39 nodules from LIDC	General	Texture analysis; supervised classification	Fully	Decision tree used for classification (soft segmentation) of intensity, Gabor and MRF features. Post-processing with VI Trimming to improve segmentation leakage.	Median soft overlap 0.52
Model fitting							
Xu et al (91)	2002	4 nodules	Juxtapleural, juxtavascular; calcifications	Boundary model fitting	User initiated seed	Pre-processing with calcification removal through expectant maximization algorithm. Modelling of lesion with an ellipse and boundary optimisation through dynamic programming.	Qualitative
Jirapatnakul et al (263)	2011	150 nodules from Weill Cornell Medical Centre database	Solid; Juxtapleural	Surface model fitting	Fully	Detection of pleural surface points which are then used for an iterative 3D surface parameter estimation. The estimated pleural surface is then applied to the binary image comprising the thoracic wall and nodule, for nodule separation.	Acceptable rate 98%
Matsumoto et al (264)	2008	66 nodules from clinical cases	Diameter 10 – 30mm; Solid; Part-solid	Surface model fitting	Fully	Lung fields extracted through thresholding and morphological operations. Deformable ellipsoid models fitted to nodule candidates, followed by feature extraction and classification.	Detection sensitivity 91%
Yong et al (265)	2014	10 cases	Juxtapleural	Geometric and morphological fitting	Fully	Seed points generated from Otsu's thresholding. Geometrical model based on variation of incline angle used to locate and segment juxtapleural lesions.	Mean DSC 0.912
Okada et al (266-268)	2005	1312 nodules from 39 clinical cases; 108 juxtapleural cases	Solid; Part-solid; juxtapleural, juxtavascular	Density distribution fitting	Semi-automatic requiring marker initialisation	Anisotropic Gaussian fitting and mean shift-based analysis (266). Anisotropic Gaussian fitting followed by likelihood ratio test segmentation (267). Juxtapleural cases processed with morphological opening and mean shift framework with prior (268).	Acceptable rate 81% (266), 86.9% (267), 88.1% (268)

Others							
Gonçalves et al (269)	2016	569 nodules from LIDC database	Solid and mostly solid nodules, volumes ranging from < 500 to > 1000 mm ³	Hessian-based	Fully	Shape descriptors of lesions calculated through 2 representations of based on calculation of the Hessian matrix (shape index and curvedness, and central adaptive medialness), final results obtained through union of both approaches.	Mean Jaccard index 0.713 ± 0.077
Wang et al (270)	2007	23 cases from LIDC dataset 1; 73 cases from LIDC dataset 2	Diameter 4 – 33.6mm and 3.8 – 30.2mm for datasets 1 and 2	Dynamic programming	User initiation	Transformation of 3D volume of interest to 2D via ‘spiral scanning’ technique, segmentation performed with dynamic programming in 2D prior to transformation of surface back to 3D image space	Mean Jaccard 0.66 and 0.58 in datasets 1 and 2 respectively

Table 1.1. Summary of (semi-) automatic segmentation techniques applied to pulmonary lesions on CT imaging.

1.9.1 Thresholding

As thresholding is one of the most straightforward technique for solving segmentation problems, its use is ubiquitous across many studies, it is often incorporated as one of the first steps in many algorithms. It serves particularly well in the rough segmentation of lung parenchyma because of the low intensity values in relation to soft tissue, which is often performed to allow extraction of the lesions of interest.

Fixed thresholds have been applied to initiate other segmentation techniques. For example, an automatic seeding of a region growing approach developed by Kuhnigk et al was derived from a pre-determined threshold value (224).

However, due to the heterogeneity in intensity values seen with lung lesions which are often not solid in its entirety, fixed threshold approaches may be unsuccessful. In order to detect non- and part-solid lung tumours, Lassen et al developed a way of applying different thresholds to different cases to initiate a region growing approach to account for their variation in appearances compared to solid lesions (225). Taşcı et al used Otsu's thresholding in segmentation of the lung parenchyma to which morphological operations were applied, including an alpha hull (generalisation of the convex hull) application in order to segment juxtapleural nodules (208). The approach taken by Zhao et al and Wiemker et al involve an automatic means of optimal threshold selection determined by the gradient strength and shape compactness that is associated with each threshold level (209, 210, 271). Adaptive thresholding approaches have also been used (205-207, 264).

1.9.2 Morphological operations

Mathematic morphology is a set of tools that allows the extraction of shape information from images. It comprises of four basic operators (erosion, dilation, opening and closing) whereby a structuring element is processed on the original image to obtain an output image. Binary operators are commonly used, although there are grey scale versions.

There are less studies that have based the segmentation process primarily on mathematical morphology. For instance, following an initial rough segmentation through thresholding and connected component analysis, Kostis et al applied a local filtering approach based on binary morphological operations that was successful at refining the segmentation though removing vasculature as well as the pleural surface (213). Similarly, although the method developed by Moltz et al also used region growing in the initial phase, it relied heavily on morphological processing to produce a better fit for juxtapleural disease (214). Fetita et al successfully applied a methodology that is based primarily on a sequence of grey scale morphological operations to segment lesions irrespective of their location (212).

Instead, many image processing algorithms employ these tools to either select or remove specific regions. In most studies, these operators have been applied to improve the performance of other segmentation techniques, rather than being the primary method for ROI extraction. Within the thorax, they have been shown to be helpful in the setting of juxtapleural disease to separate lesions from the chest wall (e.g. through bridging gaps (226) and convex hull approximation (222, 224, 225, 272)). Additionally, they play a role in vasculature removal, which is achieved by exploiting the shape differences between vessels and lesions (224, 225). Morphological operations have also been used for initialisation purposes (211). Grey scale

morphological operators have also been applied to generate image features helpful for classification purposes (252).

1.9.3 Region growing

This is a popular technique that has been adopted by many groups as it fits well with the segmentation problem posed by lesions within the lung. The initiation seed is typically initiated within the lesion followed by the incorporation of neighbouring pixels within a connected-component region, until the pre-defined criteria is met. The GrowCut application in 3D slicer which is a semi-automated tool based on region growing has been applied in the setting of advanced lung cancer (217, 273).

Gu et al used a semi-automatic approach which required user-input with a single click in order to derive the placement of a further 10 seed points, each of which was used to initiate a region-growing algorithm based on intensity, shape and connection status (221). This ensemble approach involved a voting strategy on individual voxels based on the resultant 10 segmentations, where final segmentation comprised of voxels that were segmented in at least half of the occasions. A similar approach was used by Velazquez et al in one of the few studies evaluating lung cancer in a more advanced stage (216).

One limitation often seen with the region-growing approach is the inclusion of pixels with similar intensity as the delineated region, which in the context of pulmonary nodules are the vessels. Thus, further refinement of the segmented results by cropping the vasculature is often applied to improve the accuracy of the algorithm (224-226). For instance, Diciotti et al developed a region-growing approach that assigns voxels to the nearest connected region according to the geodesic distance (223). In order to distinguish nodules from vascular structures, the algorithm was designed to be semi-automatic, which incorporated user knowledge in the decision of the membership for the initiating markers. The same group also developed a means of correcting for the attached vessels based on the local shape analysis (227). Shape index that describe curvature (e.g. the eigenvalue of the Hessian matrix) has also been used to initiate region growing for the same purpose of vessel exclusion (228, 259).

Additionally, following the application of region growing, corrections are also often employed to ensure that juxtapleural regions are also appropriately included in the segmentation (214, 222, 224-226, 228, 272). This also stems from the similarity of pixel values between the lesions and structures of the mediastinum and chest wall. The algorithm developed by Song et al has a geometric-based refinement step following a multi-constraint region growing phase that removes both the chest wall and vasculature (218).

Some groups have adopted the use of a soft computing approach through fuzzy connectedness, which belongs to the region-growing class as a segmentation technique (107, 229). Principally, the method is based on a graphical approach where the relationships between pairs of pixels are described in order to discriminate them into their membership class. Based on the fuzzy affinity relationships of the pixels, the resultant discrimination into the classes becomes probabilistic rather than in binary terms.

1.9.4 Watershed

The watershed approach has been applied successfully in the semi-automated segmentation of nodules within a study on repeated volumetry, within a user pre-defined constraint of a

background region and seed point (232). The algorithm for segmentation of lung lesions developed by Tan et al is also predominantly based on watershed segmentation, with further refinements made through active contour approach (231). Similarly, Li et al describes an automatic approach to segmenting lung nodules using the watershed approach (274).

It has also been used as a tool to improve segmentation performance. In a study on the segmentation of serial CTs by Vivanti et al, the watershed approach was used in conjunction with parabolic curve fitting to remove segmentation leaks following maximum likelihood estimation of the initial segmentation (233). It also plays a role as a technique to aid segmentation initialisation, such as in the technique developed by Brown et al where watershed segmentation was applied on the Euclidean distance transform map to detect nodules, from which region growing was performed (230).

1.9.5 Deformable models

Active contour models have been shown to improve the segmentation quality for lung nodules with ill-defined boundaries (234). Because of the smoothing properties of the curve, active contour techniques can help with correcting regions of deformations (239). They have also been seen to work in the setting of cavitating lesions (239). Plajer et al demonstrated the use of a semi-automated active contour approach in the setting of more advanced lung cancer (244). Physically-based models that better reflects the natural motion seen with flexible objects have also been applied, such as the gradient-based system developed by Cascio et al which utilises the mass-spring model (235).

Some groups have incorporated further energy terms to improve the performance of the active contour model. In addition to a 3D gradient and curvature energy terms, in the work by Way et al an additional mask energy term was also incorporated to improve the performance at the chest wall (241). A geodesic active contour approach was used by Yip et al, where a level set was formulated with energy terms to slow the contour propagations at specific tissue interfaces (242). This included a term for the chest wall (map created from thresholding and morphological operations), vasculature (Sato vesselness filter), lung parenchyma (canny edge detector), and non-nodular regions (sigmoid function). Farag et al used a variational level set approach where instead of using a single Gaussian estimation to model the intensity distribution, at each iteration of the level set solution the probability density estimation of both the object and background is performed and updated (238, 275). This was to allow for better representation of the inhomogeneity within the thorax.

The graph-cut approach has also been applied to lung tumour segmentation. Zheng et al developed an automatic means of segmentation of serial CT images using a combination of registration and graph-cut techniques (245, 246). Lermé improved the efficiency of the graph-cut algorithm of a semi-automated interactive system, demonstrating its capability in the setting of more advanced lung cancer (247, 248). Shen et al used graph-cut segmentation across multiple phases in the 4D setting, demonstrating its superiority over 3D in the presence of increased contextual information (249).

1.9.6 Classification and clustering

Supervised classification approaches applied on a voxel-wise basis has been used for segmentation. In addition to density values, van Ginneken et al, incorporated shape information

and grey level openings on a trained k-NN regression classifier within a region-growing process which computed the probability of individual voxels being within the lesion (252). Browder et al demonstrated that segmentation of non-solid nodules across serial scans using a probabilistic voxel-wise classification approach was more consistent than the radiologist's attempts (250). This was based on a Gaussian intensity model built from pre-determined HU values whereby individual voxels were classified into non-solid, solid and parenchymal regions. Both of these studies necessitated further vessel removing filters as a post-processing step.

Unsupervised clustering such as k-means clustering (276) and its soft computing equivalent (fuzzy C means) (255, 256) have also been explored for lesion segmentation. These approaches required a method of defining the number of clusters. Netto et al employed a growing neural gas algorithm which has the adaptive capability to increase the number of clusters as the algorithm executes to perform this automatically (253). However, further pruning was required to remove vasculature.

1.9.7 Texture analysis

Within the thorax, texture features have been used for a variety of purposes and is especially popular in classification tasks. Examples of its use in CAD schemes include the distinction between nodules and non-nodules (228, 255, 256), the classification of juxtapleural nodules from isolated nodules (208), and in the differentiation of benign and malignant nodules (241).

In terms of its adoption into automatic segmentation tasks specifically, there are comparatively less reports. Nonetheless, some groups have attempted this albeit with different types of texture descriptors. Hossain et al demonstrated that texture features described by Haralick et al, which are based on statistics of cooccurrence matrices, can be used to segment lung lesions through supervised classification (261). Kakar et al presented a multistep clustering and classification approach based on texture features derived from the application of Gabor filters, which are capable of detecting edges and points of texture changes (259). These features were automatically processed through fuzzy C means clustering into distinct samples, which were then separated into their respective classes through a trained SVM. Tao et al developed a voxel-wise labelling approach aimed at segmentation of nodules with ground glass-opacity (GGO) (260). This was based on 39 descriptors consisting of wavelets as the texture descriptor, in addition to first order statistics and shape descriptors. These were trained in a multiphase LDA classifier, where the positive class output from each phase was used in the next phase on three occasions. The output was applied to a shape prior in order to obtain the final segmentation.

1.9.8 Model fitting

Model fitting has also been employed in lung nodule segmentation, which is based on approximating the lesion to a known distribution. As pulmonary nodules approximate well to Gaussian distributions, Okada et al developed a series of algorithms based on this to segment small lung lesions (266-268).

Geometrical model fitting has also been used for pulmonary nodule segmentation. Matsumoto et al applied a deformable ellipsoid model to segment lung nodules as part of a CAD system (264). Jirapatnakul et al used a surface-fitting model through a 3D polynomial function to approximate to the pleural surface points, by exploiting its proximity to the anatomical

structure (263). In this way, juxtapleural lesions can be separated from the pleural surface. However, these approaches are unlikely to be useful for more advanced disease, where tumour irregularity would not approximate well to geometrical shape. Similarly, shape-based segmentation such as the approach by Gonçalves et al suffers from the same limitation (269).

1.10 Segmentation of nodal lesions

In comparison to segmentation of pulmonary lesions, there are far fewer reports in the literature on lymph node (semi-) automatic segmentation (table 1.2). Additionally, the trends in the techniques used for nodal segmentation are different to those used for lung lesion segmentation.

Most of the studies for nodal segmentation used deformable models to delineate the lymph nodes. There were several reports of interactive algorithms based on deformable models which required user seeding (277-282), as well as some which allowed online interaction where the segmentation is iteratively adjusted by the algorithm (278, 281). The deformable snake model was also used by Chen et al in the segmentation of serial CT scans, after registration of an initial prior contour (283).

In contrast to trends for pulmonary lesion segmentation, region-growing based methods are less widely adopted for nodal delineation, an example of which is the region growing based interactive segmentation tool that was developed for pulmonary nodule segmentation and applied in studies by Fabel et al (284, 285) and Buerke et al (286), where a large proportion of delineations required further corrections.

Interesting, Moltz et al designed separate interactive segmentation tools for pulmonary nodule and lymph node segmentation, with the prior based primarily on region-growing and morphological approaches (214). To improve the precision for lymph node segmentation, they developed an interactive watershed-based segmentation on the Euclidean distance map aiming at solving the problem of leakage issues. Marker-based watershed segmentation has also been explored in this setting (287, 288).

There were some studies which based the detection of nodes on shape analysis through the Hessian matrix (289, 290). Clustering (291) and texture classification techniques (292, 293) have also been applied to the segmentation task.

Study	Year	Data	Lesion type	Segmentation approach	Automation	Description	Performance
Fabel et al (284); Buerke et al (286); Fabel et al (285); Höink et al (294); Keil et al (295)	2008; 2010; 2011; 2014; 2009	47 cases (284); 112 cases (742 nodes) (286); 50 cases (285); 63 cases (294); Phantom (295)	Malignant melanoma/ Contrast (284, 285); NSCLC, large B cell non-Hodgkin lymphoma, Hodgkin lymphoma, malignant melanoma/ Contrast (286, 294); Phantom (295)	Region-based; morphological	Semi-automatic (user stroke)	OncoTREAT semi-automatic tool, Region growing applied from fixed threshold, chest wall separation through approximation with convex hull, vascular removal through morphological operations based on distance map. Interactive corrections available.	Acceptable rate 79 – 81% (284); Acceptable rate 64.7% without corrections, 83.3% with 1 correction (286); Acceptable rate 76 – 79% (285); Short axis diameter deviations 5.3% (Manual 6.5%) (294); Mean APE _D 5.18 – 10.12% (295)
Feuerstein et al (290)	2009	5 cases	Contrast; Mediastinal	Region-growing; morphological ; shape analysis	Fully	Sequential segmentation of normal structures (bronchial tree/ aortic arch/ vessels/ bone – region growing from anatomically based seed points; small vessels – thresholding and morphological opening). Mediastinal bounding search region, voxel-wise Hessian matrix, detection of node on blobness measure and size. Min DD filter applied to reduce false positives.	True positive rate 82.1%; Positive predictive value 13.3%
Liu et al (289)	2012	22 cases	Abdominal; Pelvic Contrast (>10mm)	Region-growing; morphological ; shape analysis; Classification	Fully	Extraction of pelvic girdle and spine (thresholding, morphological operations, region growing) and vessel (Hough transform for detection; fuzzy segmentation). Voxel-wise Hessian matrix, false positive reduction via SVM classification (volume, shape, curviness, intensity features).	Sensitivity 83%, 5 false positive/scan
Moltz et al (214)	2009	50 patients (222 nodes)	Melanoma; Mediastinal; Abdominal; Pelvic	Region-/Edge-based (Watershed)	Interactive	Ellipsoid approximation, region-growing followed by interactive watershed segmentation on Euclidean distance map	Acceptable rate 86 – 87%
Yan et al (287)	2006	9 patients (29 nodes)	Lymphoma	Region-/Edge-based (Watershed)	User initialisation	User selection of external region and seed point (automatically generated internal marker based on distance transform). Marker-controlled watershed segmentation.	Mean overlap $83.2 \pm 4.3\%$; Mean Hausdorff distance $3.7 \pm 1.9\text{mm}$
Yan et al (288)	2007	9 patients (29 nodes)	Lymphoma	Adaptive thresholding; Region-/Edge-	Registration; Fully	Serial CT segmentation. Initialisation and internal marker based on adaptive thresholding; dilation of registered contour as external	Mean overlap $73.0 \pm 7.3\%$; Mean Hausdorff distance $3.9 \pm 2.1\text{mm}$

				based (Watershed)		marker. Marker-controlled watershed segmentation.	
Honea et al (279)	1997	Phantom; 1 clinical case	-	Deformable model	User initialisation	Snakes active contour (2D) from 4 user seed points	Phantom: Mean volume error 4.5%
Honea et al (280)	1999	Synthetic image; Phantom	-	Deformable model	User initialisation	3D balloon active contour from 1 user seed point	Error/patient < 5 at noise < 0.1 SD
Lu et al (278, 281)	2011; 2013	50 nodes (281); 20 nodes (2 clinical cases) (278)	Mediastinal; Only 1 case with contrast in (281); No contrast in (278)	Deformable model	User initialisation and supervision	Iterative live wire for segmentation under user guidance in 2D, with extension of segmentation in 3D through projection across orthogonal planar and iterative adjustment.	Mean accuracy $81 \pm 7\%$ (single section seeding); $79 \pm 8\%$ (single click seeding) (281); Overlap $86 - 87\%$ (278)
Yu et al (282)	2015	8 cases (18 nodes)	Mediastinal	Deformable model	User initialisation	Initialisation at 2 points across short axis, initial circle boundary, 3 active contour algorithms (edge-based, region-based and region-based with edge constraint).	Region-based with edge constraint DSC 0.853 ± 0.059 ; Edge-based 0.802 ± 0.125 ; Region-based DSC 0.741 ± 0.199
Chen et al (283)	2013	14 cases (26 nodes), 2 – 3 timepoints	Mediastinal; Abdominal; Lymphoma	Registration; Deformable model	Registration	Serial CT segmentation. Initialisation provided by manual delineation, which was registered to serial CTs and active contour applied.	Mean overlap $74.4 \pm 12.6\%$; Mean Hausdorff distance 3.70 ± 1.56 pixel
Yan et al (296)	204	400 images	General	Deformable model	User initialisation	Seed point and circle initialisation, Fast marching algorithm	Qualitative
Dornheim et al (277)	2007	11 cases (146 nodes)	Neck	Deformable model	User seed point	Mass spring model based on grey value intensity of initialisation sphere.	Hausdorff distance range 1.7 – 3.9 mm
Maleike et al (297)	2008	4 cases (29 nodes)	General	Deformable surface; Statistical shape model	Initialisation not discussed; User correction	Ellipsoid model and deformable surface fitting	Volumetric overlap error 10.1% (6.1% post correction)
Xu et al (291)	2011	70 cases (14 patients, 2 – 3 timepoints)	General	Adaptive region-growing; Clustering	Registration; Fully	Sequential CT segmentation. Registration of manual contour, adaptive thresholding, morphological operations and region growing, k-means clustering	Mean overlap $80.7 \pm 9.7\%$; Mean Hausdorff distance 3.18 ± 1.82 mm
Barbu et al (292, 293)	2010 2012	131 cases (371 nodes) 54 cases (569 nodes)	Mediastinal; Pelvic (> 10mm); Contrast	Texture analysis; classification; model fitting	Bounding box initialisation	Thresholding, detection based on Haar and self-aligning (gradient based) features (Adaboost classifier), segmentation through fitting of sphere triangulation of Gaussian MRF shape prior. Verification through size and feature classification.	Axillary: True positive rate 83.0%; 1 false positive/scan Pelvic: True positive rate 80.0%; 3.2 false positive/scan

Table 1.2. Summary of (semi-) automatic segmentation techniques applied to lymph nodes on CT imaging.

1.11 PET-CT imaging

In lung cancer, ^{18}F -fluorodeoxyglucose (FDG)-PET imaging can help with making treatment decisions (298), and is now routinely performed for cancer staging as a standard of care (299). It has also been shown to decrease variation in target volume delineation of lung cancers when used in conjunction with CT (39, 300). Dose escalation studies have also been performed demonstrating the feasibility and safety of delivering a higher dose to the PET-defined target (301, 302).

With growing applications for PET-CT imaging, many strategies for auto-segmentation of PET scans have been developed, which has recently been reviewed by the AAPM task group (303). However, for this project, the segmentation task was concentrated on CT imaging. PET analysis was not performed for a number of reasons. There was a lack of PET imaging data as all the current lung clinical trials in the RTTQA portfolio do not stipulate planning PET-CT scans, and diagnostic planning scans are not routinely accrued. Even if the diagnostic scans were to be available, the absence of standardising acquisition procedures in a multi-centre setting, in addition to the physiological factors that affect tracer uptake, would contribute to variation in standardised uptake values (SUVs) and affect quantitative analysis (304). Moreover, registration uncertainties in the matching of diagnostic PET-CT scans to the planning CT would render segmentation tasks difficult. This forms the basis of the current clinical practice, where unless the PETs scans have been acquired in the same treatment planning position, diagnostic PET scans are used to help identify lesions, while manual GTV delineation is still being performed based on planning CT appearances (37).

There were two available diagnostic PET-CT scans, one of which is shown in figure 1.7 to illustrate some of the other challenges involved in PET-CT analysis of advanced lung cancer. Within the large primary tumour, necrotic tissue in the centre of the lesion was associated in low FDG uptake. The segmentation task is made more difficult as this region of heterogeneous uptake is in close proximity to mediastinal nodal involvement. This also shows that FDG uptake which corresponds to the biological target volume is different to the GTV that is demonstrated on CT appearances. For these reasons, exploration of the segmentation techniques was performed based on CT imaging.

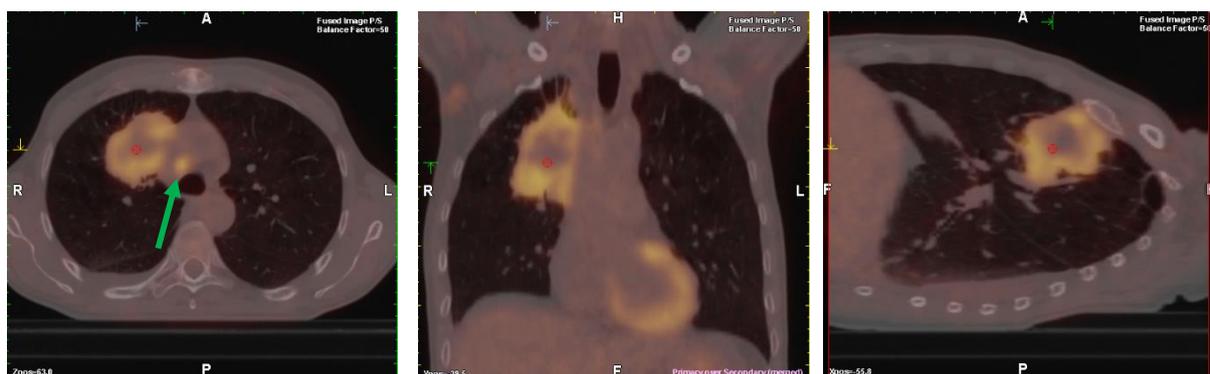


Figure 1.7. Deformable registration of diagnostic PET-CT (orange uptake) with planning CT in three orthogonal planes. Green arrow denotes pathological lymph node.

1.12 Commercial software products

There are many commercial radiotherapy planning software products available for use in radiotherapy planning. At a basic level, these systems enable the use of a thresholding technique to segment entities based purely on the density content, which is typically performed on regions that are very well distinguished between those which exhibit very high- or low-contrast. These include propagation of body contours, lung tissue, as well as bone structures.

Since the review by Sharp et al (64), there has been some developments in the tools offered by commercial systems. All the available commercial products also offer more advanced semi-automatic segmentation tools, as summarised in table 1.3. Atlas-based systems are most commonly offered, although increasing, model-based approaches are available, either as a standalone technique or in combination with atlas-based tools. Recently, Mirada has also developed a deep learning based automatic segmentation software DLCExpert™ (Mirada Medical, Oxford, UK) (305) that has been shown to outperform the company's atlas-based approach for lung and heart contouring, whilst having comparable results for spinal cord, mediastinal and heart contouring (306).

Manufacturer	Software product	Segmentation technique	Application on body sites
Dosisoft	IMAgO (307)	Atlas-based	Brain
			Head and neck
MIM Software	MIM Maestro®	Atlas-based	Brain (308)
			Head and neck (309, 310)
			Thorax (310)
			Pelvis (310) (308, 311)
Velocity	VelocityAI®	Atlas-based	Head and neck (161, 310)
			Thorax (310)
			Pelvis (310)
OSL	OnQ rts®	Atlas-based	Head and neck (312)
			Thorax
			Pelvis (312)
Eclipse	SmartSegmentation® (159)	Atlas-based or model-based	Brain (313)
			Head and neck (314)
			Thorax (315)
			Pelvis (315)
Brainlab	iPlan® (316)	Combined atlas- and model-based	Brain (313)
		Atlas-based	H&N (317)
		Model-based	Pelvis (318, 319)
Accuray	MultiPlan®	Combined atlas- and model-based	Brain (308)
		Model-based	Head and neck (312)
			Pelvis (308, 312)
Philips	SPICE® (320)	Atlas and/or model-based	Head and neck (312, 321)
			Thorax (322)
			Brain (308)
			Pelvis (308, 311, 312)

RaySearch	RayStation® (323)	Atlas and/or model-based	Head and neck (312)
			Brain (308)
			Thorax
			Abdomen
Elekta CMS	Atlas-Based Autosegmentation® (ABAS) (324)	Atlas and/or model-based	Head and neck (156, 310)
			Thorax (310)
			Pelvis (310, 311, 319, 325)
Mirada Medical	RTx® Workflow box® (326)	Atlas-based	Head and neck
			Breast (327)
			Pelvis (311)
	DLCExpert® (306)	Deep learning	Thorax

Table 1.3. List of commercial software products and associated segmentation techniques that are currently available.

Success in using such systems for delineation has been variable. One common theme for atlas-based systems is that the performance seems to be better with more atlases of high quality segmentations. Although there are many reports of the time-saving benefits with the use of these products, there still remains the need for clinician input to correct erroneous segmentation (321). Additionally, although atlas-based systems can be suitably used for contouring of whole organs or constant parts of organs and OARs, they are not suited for automatic tumour volume delineation.

Several commercial CAD software products are available that performs segmentation of pulmonary nodules/tumours as part of the analysis pathway. Examples of these include Lung VCAR (GE Medical Systems LLC) (328), Veolity (MeVis Medical Solutions AG, Bremen, Germany) (329), Syngovia Via (Siemens Medical Solutions, Forchheim, Germany) (330) and ImageChecker CT (R2 Technologies, Sunnyvale, CA) (331). Despite the availability of such products, to my best knowledge, these systems are not currently adopted in routine clinical practice in the UK, although some products have been approved for use in the US as a second reader in addition to the interpretation by radiologists. These systems are not set up for use in radiotherapy.

In summary, there is no suitable automated commercial product that can perform our task.

Challenges of this project

Automatic segmentation of advanced lung cancer is a big challenge in its own right due to the presence of larger, heterogenous lesions often with mediastinal nodal involvement. However, most of the studies on CT segmentation of lung tumours have been performed in early disease, with a general lack of research in the delineation of more advanced disease. With little precedence, much of the consideration in this project was on evaluating suitable segmentation techniques and workflows.

Thesis structure

The main objective of this work is to develop and identify the most appropriate method(s) of generating an automated image-based lung tumour segmentation with which outlining assessments can be based on.

The rest of the thesis is laid out into three main sections. Chapter 2 explores the initial ground work to the project, of which there were four broad areas of evaluation. Firstly, the development of the processes for data import and handling is described. In addition to the clinical and imaging characteristics, imaging quality is explored to determine the need for further pre-processing of the images. Thresholding was investigated through an assessment of the pixel intensities of the tumour in relation to the surrounding tissues, and the use of different threshold levels determined by the presence of non-solid tumour regions was explored.

In chapter 3, the development and training performance for each of the four evaluated segmentation techniques is described. The comparison of the performance for the different approaches is presented subsequently in chapter 4. Chapter 5 pertains to the application of the algorithms in isolated peripheral lung cancer, which is structured in a similar way to chapters 3 and 4 according to the respective datasets (training and testing) for the different techniques.

The analysis of texture features between whole tumour and non-tumour regions is addressed in chapter 6. This work is further developed in chapter 7 evaluating standard ROI sizes at different locations in and around the tumour region.

Conclusions are drawn in chapter 8, together with some proposals for future work.

Chapter 2

Specific Aim A: Processing of data, quantification of image quality, and analysis of GTV with thresholding as an initial segmentation technique

Introduction

The issues in relation to the initial set up and pre-processing of the images are discussed in this chapter.

2.1 Use of MATLAB

MATLAB (The MathWorks, Inc., Massachusetts, US) is being used widely across many fields for its strength for advanced computing algorithms. There are several advantages of using MATLAB for this work in image analysis over other computing languages.

MATLAB has an Image Processing Toolbox extension which provides an extensive set of algorithms and functions for use in image processing and analysis. In addition to its built-in functionality for handling and reading DICOM images, it also has a diverse set of tools that can be used for analysing the properties of regions of interest. Its algorithms for image segmentation can also be readily applied in this project.

In addition, MATLAB allows numerical precision in its use for image processing. Typically, a pixel sample of a CT scan is stored to a 16-bit precision in the DICOM format for medical use, which gives rise to a potential of 65536 pixel integer values per sample. Different image processing algorithms deal with this high-precision data differently; many image processing systems are more attuned to dealing with data with an 8-bit precision and therefore may rescale the data to a range of 0-255 (i.e. up to 256 different grey-scale levels for display) after performing floating point arithmetic, or through truncation of the values. If this is applied to CT numbers scaled in HUs, much of the information will be lost as many structures will be displayed with the same grey-level as their surroundings. MATLAB offers the potential to perform arithmetic computation whilst ensuring the fidelity of high-precision data, to minimise the loss of image information. Maintaining the accuracy of the data will be invaluable for this project, where multiple steps in image segmentation were envisaged to be used.

Thus, the automatic image segmentation algorithms were developed within MATLAB, which served as a single platform for both image processing and assessment of conformity between the generated volumes. Although the reading of DICOM image data is well established in the algorithms within MATLAB, importing and reading of RTSTRUCT files have been less widely applied in MATLAB. One criterion in the design of this project was the need to be in control of the data handling at every step in the process, in order to maintain, check and ensure data integrity. This would provide the flexibility of being able to amend the segmentation workflow and to interrogate the data to suit the purposes of the project. Thus, as opposed to using another software such as CERR (A Computational Environment for Radiological Research (332)) to import data into MATLAB, the initial phase of the work included writing an in-house algorithm specifically for this. A number of checks were then performed to verify that this is in accordance with the performance of other software programmes, before proceeding with the preparation for segmentation.

2.2 Assessment of data heterogeneity

The initial data exploration included an assessment of the heterogeneity of the datasets, both in terms of imaging and clinical factors. This is of particular importance, as unlike many other studies on data from a single institution, there would be greater variability inherent in the setting of data from a multi-institutional trial.

2.2.1 Imaging acquisition heterogeneity

One consideration is the differences in imaging factors, of which there are 3 main sources contributing to the noise in a CT image. Quantum noise is determined by the number of X-ray photons that is detected. This is influenced by the scanning technique (tube voltage, tube current, slice thickness, scan speed, helical pitch etc) which affects the number of photons delivered to the patient, and the scanner efficiency, which affects the percentage of photons that gets converted to useful signal (84). The second source is the inherent physical limitations of the system (electronic noise in detector and data acquisition system) which is not usually amenable to adjusting (84). The third source is in the image generation process, which includes reconstruction algorithms, of which there can be different reconstruction parameters, such as reconstruction filter kernels, field of view, image matrix size and post-processing techniques etc (84). Selection of different reconstruction algorithms and their parameters, where different reconstruction kernels are designed for specific clinical applications, can result in variation in CT numbers. For example bone or lung algorithms can enhance the visibility of bony objects and falsely elevate the CT number of small lung nodules, as compared to a standard algorithm (333). The choice of reconstruction is manufacturer and centre specific.

There are different CT manufacturers, models, and scanning protocols in use across the UK, reflecting a range in practice with different selections of parameters in keeping delivered dose to a patient as low as possible, whilst optimising the images in balance with the associated trade-offs. Two surveys on diagnostic CT imaging in the UK have shown not just wide variation in scanning techniques between centres for similar body sites, but also changes in practices over time as the technology of CT continue to advance (110, 334). An audit by an Institute of Physics and Engineering in Medicine (IPEM) working party has recently highlighted the variation in the radiotherapy planning CT scanning procedures, where dose indices were seen to differ between centres for lung planning CTs, albeit with greater variation seen in 4D than 3D scans (335). Variation in radiotherapy planning CT acquisition could contribute to HU differences and thereby add to heterogeneity in images obtained across centres (336), although there remains a paucity of data on how this impacts on image quality in the radiotherapy setting.

2.3 Considerations for image enhancement

One other consideration is the potential need for any enhancement of image data. These procedures are commonly applied in the initial step of image processing in order to improve the quality of the image. Different filters can be applied for noise reduction, smoothing, contrast stretching, and edge enhancement (337), which depend on the segmentation approach as well as the image quality. However, the selection of a particular denoising filter is seemingly empirical for many segmentation studies. Additionally, the application of noise reduction filters usually comes at the expense of increasing image blur and edge information loss. Thus, an assessment into the image quality had to be performed to ascertain if noise reduction filters would be required in this work. This was assessed quantitatively through computation of the signal, noise, signal to noise ratio (SNR) and contrast to noise ratio (CNR).

2.4 GTV and normal tissue density distribution

The density differences between the GTV and other organs was also analysed. This is to allow thresholding to be applied as a simple and effective means of removing the surrounding normal tissue structures. Due to its simplicity and ease of implementation, this was in preference to region-growing methods. Moreover, region-growing methods often required further post-processing steps to improve its precision, adding further complexity to the workflow.

It is vital to ascertain appropriate thresholding values in the presence of part- and non-solid tumours, in order to avoid removing tumour regions such as GGOs. Thus, in addition to determining suitable threshold values, an evaluation was performed to find a means of distinguishing solid from non-solid lesions.

2.5 Summary of tasks

These topics are explored in this chapter.

Task A.1 Data import and handling

Task A.2 Determination of heterogeneity and imaging quality of cases to assess need for image enhancement

- A) Clinical characteristics of cases
- B) Scanning parameters of cases
- C) Assessment on image quality

Task A.3 Exploration of the descriptive statistics for the GTV and surrounding tissues

Task A.4 Determination of thresholding procedures

Methods

2.6 Clinical datasets

A search of the RTTQA database was performed to select lung cancer trials and datasets that were suitable for this work. All patients included in this study would have received external beam radiotherapy for treatment of lung cancer, with target volume delineation approved by the trials QA team. Complete thoracic CT datasets and RTSTRUCT files were retrieved for these cases, which was assessed for their suitability to be used in the different phases of the study.

2.7 Study design

The design of the studies in this section is shown in figure 2.1. Description of the clinical and scanning parameters was evaluated for all the data to establish the heterogeneity of the case mix.

A subsample of the total training data was used to develop the workflow for importing and viewing the DICOM files. The assessment of image quality, descriptive statistics and thresholding was also performed on a subsample of training cases. The evaluation of thresholding was further extended to include the whole training set to obtain a better estimate of the threshold level.

Apart from its description, the independent test data was not used for any of the work in this chapter.

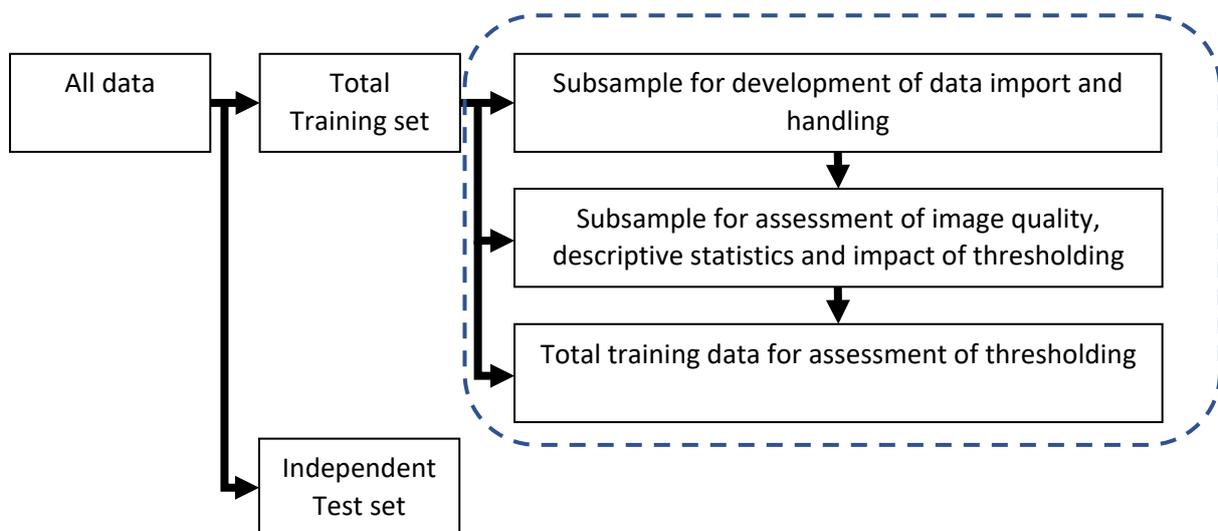


Figure 2.1. Design of pre-segmentation projects.

2.8 Clinical and imaging parameters

A visual assessment of the clinical characteristics of the cases was performed in VODCA v4.3 (Medical Software Solutions, Hagendorn, Switzerland), to ascertain the use of contrast-enhancement, and to detect the presence of cavities within tumours, or atelectasis and effusions in the affected lung. Lesions were divided into two groups, solid lesions versus non-solid lesions, defined by the presence of cavities, ground-glass opacities (GGO), and airways that are not practically distinct from the GTV. The latter is a common occurrence in the region of the hilum where small areas of airways are not excluded from the GTV contour, due to the involvement of hilar nodes. The location of the GTVs were also assessed and classified into

peripheral tumours (surrounding by lung tissue in its entirety), and where it is adjacent to or involving the mediastinum, hilum and the chest wall. Additionally, the presence of nodal disease (distinct or indistinct from primary tumour) was also noted.

Imaging parameters were retrieved from the DICOM metafile using the combination of an in-house DICOM tag viewer, and the mean tube current for each scan was extracted in MATLAB.

Based on the clinical and imaging parameters, the cases were split into training and an independent testing dataset.

2.9 Gold-standard reference ROIs

All the CT and RTSTRUCT datasets were imported into Pinnacle³ (Philips, Eindhoven, Netherlands) to assess the segmentation of the GTV contours. Although these structures have been QA approved within the trial, a check was performed to ensure that there were no erroneous contours, which were manually corrected if present. A further check of the delineation was performed in VODCA, where empty structures and structures associated with single data points were removed.

For the training dataset, additional normal tissue ROIs were processed in a similar way to the GTVs. A description of the outlining definitions and methodology for each of these ROIs is shown in table 2.1 referenced to their respective atlases where available. For the lung, bone, vessel and mediastinal soft tissue contours, further edits were performed to ensure that exclusion areas within the contours, where present in the axial plane, were linked to the contour boundary, to avoid any errors in generation of binary masks in subsequent processing.

Normal tissue	Window levels (Pre-set in Pinnacle ³)	Outlining description
Bone	Bone	To speed up the outlining process, thresholding was applied to include pixel raw values of above 1200, then manually edited for all axial slices. The cortex and spongy bone for the sternum, ribs, vertebral body and transverse/spinal processes were included; scapula was excluded. The spinal canal was also excluded from the ROI.
Chest wall	Mediastinal	A 1-cm rim at the chest wall from the superior to inferior border of the lungs was manually placed to include the ribs, chest wall musculature, and vertebrae. The sternum was not included. (Modified from RTOG atlas (30) to minimise inclusion of chest wall subcutaneous fat, with the inclusion of vertebrae)
Heart	Mediastinal	The entire outline of the heart was outlined from the axial slice at the separation of the pulmonary trunk and arteries superiorly, to the slice containing the most inferior wall of the left ventricle (338). The inferior vena cava was excluded from the ROI.
Ipsilateral Lung	Lung	The submitted contours for the bilateral lung (338) were edited to isolate the ipsilateral lung. Further manual edits were performed to ensure the whole of the ipsilateral lung was outlined, and GTV was excluded.

Airways	Mediastinal	The trachea (superior border at the cricoid) and proximal bronchial tree were outlined as a single structure, ensuring that the walls of the airways were included (Modified from RTOG atlas (30) to include trachea).
Mediastinal Soft Tissue	Mediastinal	The soft tissue of the mediastinum was included in this structure, defined superiorly at the cranial border of the ipsilateral lung, to the caudal border of the heart. Boolean operations were performed to exclude vessels, heart, oesophagus and airways.
Oesophagus	Mediastinal	The outer wall of the oesophagus was outlined, from the level just inferior to the cricoid, to the gastro-oesophageal junction (30).
Vessels	Mediastinal	Large vessels for the length of the ipsilateral lung in the Z-axis were outlined, which included the mediastinal vessels and the descending aorta. Vessels within the mediastinum were contoured from the superior border of the ipsilateral lung to the slice above the heart. Inferior to this level, the descending aorta was outlined, to the most caudal slice of the ipsilateral lung. Fat surrounding the vessels were excluded.

Table 2.1. Normal tissue outlining definitions and methodology.

The final DICOM dataset with the respective ROIs was then exported from VODCA to be used in the analysis.

2.10 Data handling

2.10.1 MATLAB scripts

MATLAB and Image Processing Toolbox Release 2014a was used initially for scripting of the code, which was subsequently updated to 2015b, 2016a and finally 2017b, which was used to run all the analysis. For clarity, the built-in MATLAB functions are specified in the text below in this font: `MATLAB function`.

2.10.2 Importing image datasets

The DICOM metadata was extracted using `dicominfo` where the information on the data elements of interest from the image files were retrieved. Information on CT orientation (patient position), pixel spacing, slice spacing, as well as rescale intercept and rescale slope were obtained from a single axial slice, whilst information on image position patient was extracted from all the files of the image dataset.

Information of the Z-coordinates for all the CT axial slices were extracted from the data retrieved from image position patient. CT orientation was checked and used to sort the images in descending order according to the Z-coordinates.

`Dicomread` was used to read the image data from the DICOM file. In volumetric digital CT scans, the radiographic density is represented by the attenuation coefficient of each voxel called the CT number. Each of these values correspond to an absolute physical radiodensity in the Hounsfield unit (HU) scale. The HU is a linear transformation of the linear attenuation coefficient measurement where the radiodensity is defined as 0 HU for distilled water (at standard pressure and temperature). It ranges from -1000 HU for that of air, to 1000 HU for

that of bone (3000 HU for dense bone). For a given material X with a linear attenuation coefficient μ_x , the HU value is defined by the equation:

$$\text{Hounsfield Unit} = 1000 \times \left(\frac{\mu_x - \mu_{\text{water}}}{\mu_{\text{water}}} \right) \quad [2.1]$$

where μ_{water} and μ_{air} are linear attenuation coefficients of water and air respectively (339).

The individual pixel values of the CT images were corrected into HU by applying the rescale intercept and rescale slope for each of the images based on the equation:

$$HU = m \times SV + b \quad [2.2]$$

where HU is the Hounsfield unit, m is the rescale slope, SV is the stored pixel value, and b is the rescale intercept.

2.10.3 Normalisation of image datasets

As the cases were obtained from different sources, to ensure that all images were normalised to the same range, this was selected to be from -1000 to 1000 HU. Values higher than 1000 is unlikely to be useful, especially as the values of the GTV, the main area of interest, is expected to be well within this range. Analyses of the HU was performed with this normalised image data. For other calculations within MATLAB, to ensure the compatibility of the data with the in-built functions, the images were then converted into double precision, within a range of 0 to 1 using `mat2gray`. This ensured the fidelity of 16-bit data and circumvented the issue of processing signed integers.

2.10.4 Importing RTSTRUCT files

As there is no inbuilt functionality in MATLAB to read RTSTRUCT files directly, an in-house algorithm had to be written specifically for this. This was carried out in two ways, with the first approach being an indirect import through the creation of a separate STRUCT file. A second approach with an in-house algorithm that performed a direct import into MATLAB was also explored.

2.10.4.1 First approach: Indirect import through generation of STRUCT file

In order to import the information on the structure sets from the RTSTRUCT file into MATLAB, an in-house software programme was developed (by Jena R) and used to read each RTSTRUCT file. The names of all structures present within the RTSTRUCT object and the 3D coordinates of individual pixels for each of the structures were extracted. This data was then exported as a single STRUCT file for each case, which was compatible with import into MATLAB.

To speed up the computing process, data truncation was built into the algorithms of the in-house software whereby contours in excess of 350 data points were not processed. Affected volumes such as the body contour were removed from the STRUCT file as a post-processing step in preparation for importing the file into MATLAB.

2.10.4.2 Second approach: Direct import through retrieval from DICOM metadata

In a similar way to the image files, the DICOM metadata of the RTSTRUCT file was extracted using `dicominfo`, from which a list of all the ROIs were retrieved, together with the associated contour points in each axial slice. Unlike the previous approach, no limit was set to the number of data points with this methodology.

2.10.5 Matching of structures to image data

To avoid the processing of irrelevant ROIs for the purposes of this work, the ROI(s) to be evaluated was specified manually. The Z-coordinates from the image slices were used to match up with those from the ROI contours in order to pair up the ROI coordinates with its associated image axial slice. Due to the way in which different planning systems generate and export structure set data, there can be a small discrepancy in the Z-coordinate values between the ROI and image files. As MATLAB handles the numeric data with exact precision, the potential Z-coordinate discrepancy was taken into account using the information from the image slice thickness in the pairing.

The X- and Y- coordinates from the data retrieved from the CT DICOM data element *image position patient* was used to transform the X- and Y- coordinates of the individual pixels of the ROI contours to the same coordinate system for that of the image, adjusted for the image pixel spacing. To ensure that closed rather than open polygons would be created, the first X- and Y-coordinates of each contour was appended to the bottom of the list of coordinates. Binary masks for the ROIs were generated using the built-in function `poly2mask` from the contour polygons.

To avoid potential mix-up of pixel coordinate values during the creation of binary masks when there were multiple contours in a single axial plane, these slices were identified and processed separately to generate multiple mutually exclusive contours.

The initial scripts for data import of the raw data in the first approach was assisted by Georgantzoglou A, and the second approach by Al Sa'd M. These scripts were further modified to match the structure masks to the imaging data transformed to HU values and with the appropriate normalisation.

2.10.6 Assessment of data import

To verify that the GTV contours were correctly plotted on the CT images in MATLAB, they were visually inspected for every slice against their placement using VODCA. Additionally, for 3 cases, the total volume of the MATLAB contours, and axial sections every 5 slices apart were calculated. These were evaluated against the volumes computed in VODCA and Pinnacle³. As the information on the number of voxels was not available in both VODCA and Pinnacle³, volumes of the ROIs were compared instead. For the ROIs processed in MATLAB, the volumes were determined by multiplying the area of the GTV binary masks (`bwarea`) with the voxel size of the CT images.

2.11 Image quality assessment

A subsample of training cases was used to quantify and assess image quality on the normalised CT data. The cardiac contours were imported to locate the Z-coordinate of the cranial heart border, which was then used to identify the 5 adjacent CT images superior to this slice. This location was chosen to avoid image slices where there may be greater variation in beam attenuation due to large changes in the anatomy. This is more likely to occur at the borders of large organs e.g. top of the liver, which would result in larger variation in the detected signal. Additionally, as the chosen location corresponds to the position of the hilum near the centre of the thorax, there is a higher likelihood that these slices would be sited near or at the location of the tumour.

A standardised ROI corresponding to a circle with a radius of 4 pixels (49-pixel area) was used. Although it would be preferable to use a larger ROI, this was limited by the anatomy. A windowing level of -150 to 150 HU was used in the manual positioning of the ROIs. Although the signal of the GTV is the main subject of interest, it was decided not to evaluate the signal at the tumour directly due to the potential heterogeneity of the HU values within the disease. Additionally, different locations of the tumour between cases could potentially affect the assessment of noise. Vessels and the heart were also not used due to the differences in intra-venous (IV) contrast use between cases. Thus, for this assessment, the thoracic muscle was chosen, as a surrogate, with the subcutaneous tissue as the background.

Within each axial image, four ROIs were placed in the skeletal thoracic musculature, two on each side, ensuring that the ROI selection was well within the musculature and not encroaching into surrounding tissue. ROIs were placed in the infraspinatus and latissimus dorsi muscles where possible. In the presence of streak artifacts, they were placed in the subscapularis or the serratus anterior muscle instead. A further four separate ROIs were positioned in the subcutaneous fat to serve as background tissue for comparison, similarly with two on each side of the body. The region just anterior to the latissimus dorsi was chosen, as well as a region in the anterior chest wall. There were no areas of overlap for any of the ROIs. An example of the ROI selection is shown in figure 2.2.



Figure 2.2. Axial image illustrating placement of ROIs for evaluation of image quality. Red ROIs denote selection in soft tissue musculature, blue ROIs denote selection in background subcutaneous fat.

For each case, the mean and the standard deviation (SD) for each ROI were calculated, which corresponds to the signal and noise respectively. The SNR and CNR values were computed with the formulas:

$$SNR_{st} = \frac{\hat{S}_{st}}{\hat{\sigma}_{st}} \quad [2.3]$$

$$CNR = \frac{\hat{S}_{st} - \hat{S}_{bg}}{\hat{\sigma}_{bg}} \quad [2.4]$$

where \hat{s} is the mean of the signal, and $\hat{\sigma}$ is the mean of the noise. The subscripts ‘st’ and ‘bg’ denote the soft tissue and the background respectively. Further statistical analysis was performed using the IBM SPSS Statistics v20 software programme (IBM, New York, US).

2.12 Statistics of GTV and surrounding tissues

Descriptive statistics for the GTV and normal tissue were acquired to explore their relationship to determine the use of thresholding in the segmentation process. The difference in HU between solid and non-solid lesions was also evaluated for the subsample of cases, and further extended to the rest of the training cohort. This work was performed using both MATLAB and SPSS. For the boxplots, the mid-line bar represents the median, the top and bottom of the box represents the 75th and 25th percentiles respectively, whiskers represent 1.5 times the interquartile range and the dots represent outliers.

Results of data assessment exercise

Within the RTTQA database, two clinical trials, IDEAL-CRT and ISTART, were identified as potential sources of data for this work. Both of these trials were conducted with approval from the respective research ethics committee. Permission from both TMGs was obtained for use of the trial datasets in this work, which have been suitably anonymised in the RTTQA database in keeping with the Data Protection Act 1998. It was decided that data from the IDEAL-CRT trial would be more appropriate, as the ISTART dataset had a mix of 3D and 4D CT planning scans, with ITV contours rather than GTV.

The IDEAL-CRT trial is a multi-centre UK phase I/II trial of concurrent chemoradiation with dose-escalated radiotherapy in patients with stage II or stage III NSCLC (340). A total of 84 patients were recruited into the IDEAL-CRT trial. 4 cases were excluded from this study, due to missing data (1 patient withdrawn from study; 2 missing structure files; 1 truncated CT scan). A further case was excluded as a GTV structure was not present, leaving a total of 79 cases for this study. Out of the remaining 79 cases, 63 cases were used as the training data, of which 18 cases were used to develop the data import workflow, as well as the studies on image quality, descriptive statistics and thresholding.

2.13 Task A.1 Data import and handling

Data import and ROI display: First approach - Indirect import through generation of STRUCT file

Figure 2.3 show an example of the images generated from MATLAB for an axial slice for one of the datasets, which appears to be consistent with the GTV outline. Despite the congruency of data import using this method, it became apparent that this multi-step import of structure sets is time consuming and not feasible pragmatically with larger datasets. This pushed the need for re-evaluation of the methodology and an attempt to importing the structure files directly into MATLAB, with less user input.

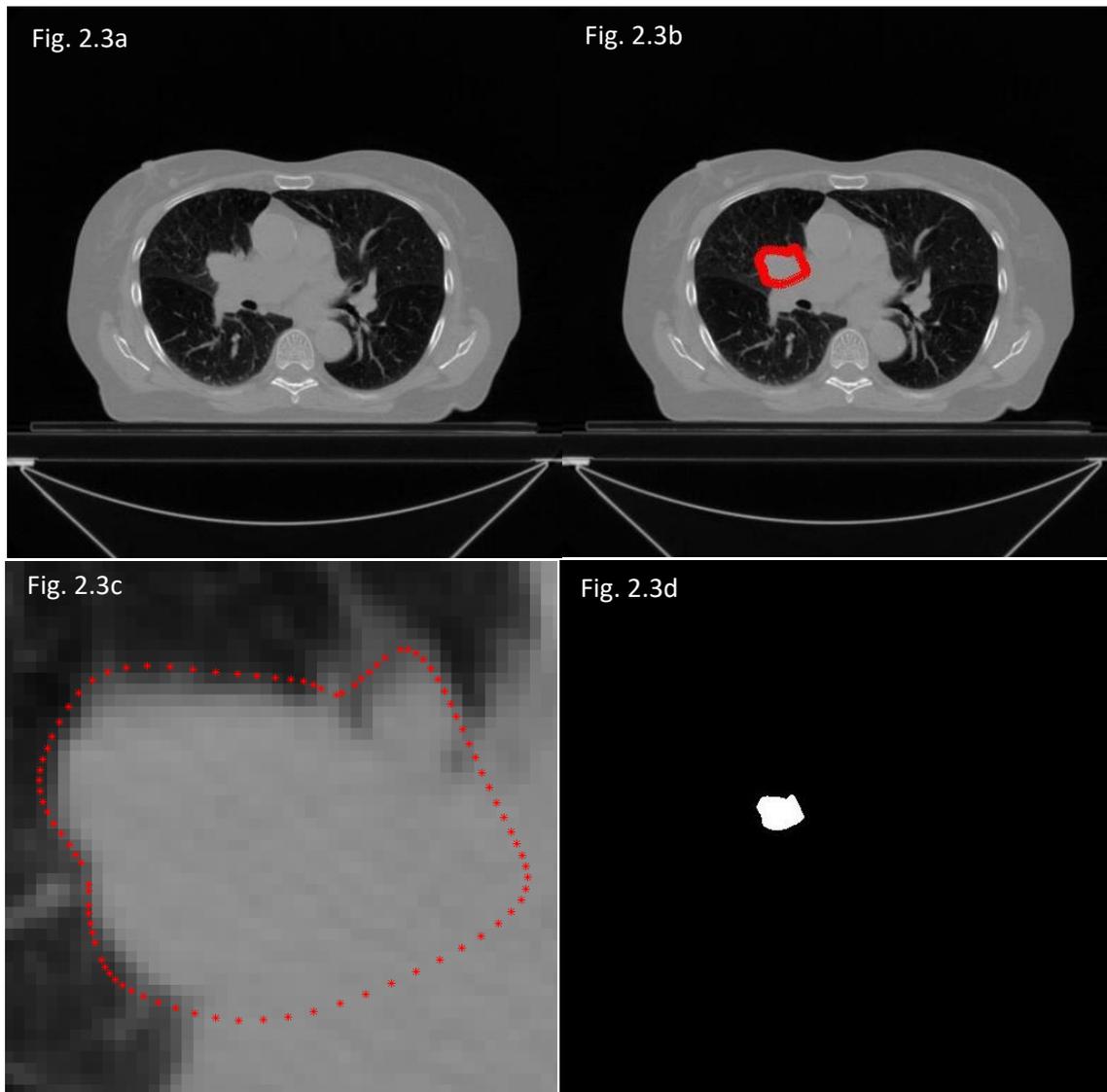


Figure 2.3. Example of images generated for an axial slice for one of the datasets with indirect import through generation of STRUCT file. a) CT axial slice b) Contour superimposed on CT axial slice c) Close up of contour points on CT d) Binary mask generated for ROI.

Data import and ROI display: Second approach – Direct import through retrieval from DICOM metadata

An example of the MATLAB display of a GTV contour in the axial plane is shown in figure 2.4a, where the displayed contour boundary marks the periphery of the GTV binary mask that was generated from the structure file. The corresponding slice using VODCA is shown in figure 2.4b. As compared to the images from MATLAB, the GTV contour and background image in VODCA appears smoother and less pixelated due to interpolation of both the CT image and ROI contours. With this taken into account, visually, the contours in MATLAB appears to be congruent to the display in VODCA. Similarly, the GTV volumes calculated in MATLAB for the 3 training cases are consistent across the 3 software packages (Friedman Test: p -value = 0.49), as shown in table 2.2. No discrepancy in the matching of contours in the CT slices in the superior-inferior plane was seen, even in the presence of non-contiguous slices.

Case	Slice number	GTV Volumes (cm ³)		
		Pinnacle ³	VODCA	MATLAB
Case A	Slice 1	0.69	0.66	0.66
	Slice 6	4.50	4.44	4.47
	Slice 11	7.08	6.99	6.98
	Slice 16	9.58	9.45	9.47
	Slice 21	7.25	7.12	7.11
	Slice 26	1.42	1.41	1.42
	Total Volume	145.39	143.56	143.58
Case B	Slice 1	0.17	0.17	0.18
	Slice 6	0.65	0.65	0.65
	Slice 11	4.28	4.30	4.22
	Slice 16	8.63	8.69	8.58
	Slice 21	6.43	6.43	6.43
	Slice 26	2.80	2.81	2.79
	Total Volume	113.50	114.10	112.95
Case C	Slice 1	0.80	0.80	0.79
	Slice 6	5.47	5.49	5.49
	Slice 11	10.76	10.81	10.82
	Slice 16	12.28	12.30	12.36
	Slice 21	9.72	9.76	9.76
	Slice 26	5.17	5.20	5.17
	Total Volume	230.42	231.30	231.49

Table 2.2. Volumes (cm³, corrected to 2 decimal places) for GTV contours axial slices (5 slices apart), as well as total GTV volumes for 3 training cases in Pinnacle³, VODCA and MATLAB displays (Friedman Test: p -value = 0.49).

Although the visual display in MATLAB could have been improved with interpolation of the CT images, it was not performed as it would not contribute a great deal to the assessment of the contours.

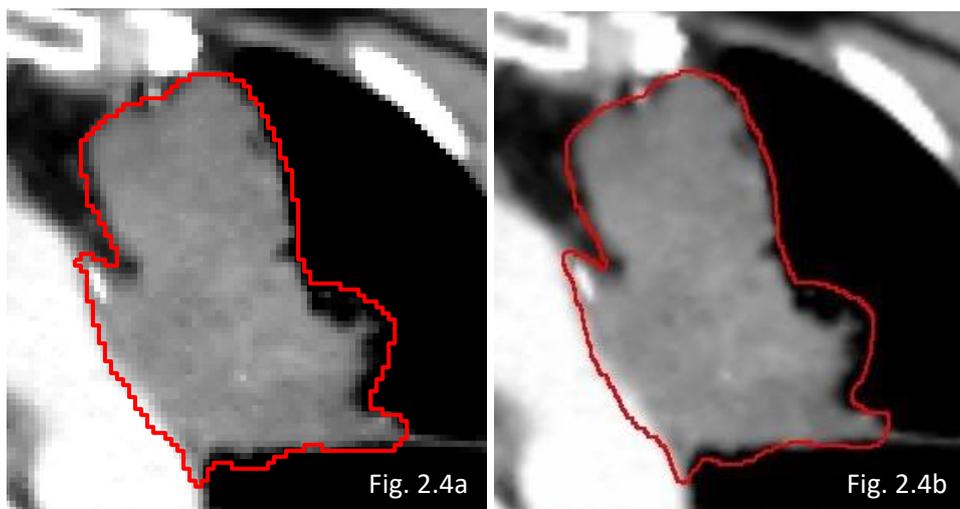


Figure 2.4. Example of display of GTV contour for the same axial slice in a) MATLAB and b) VODCA.

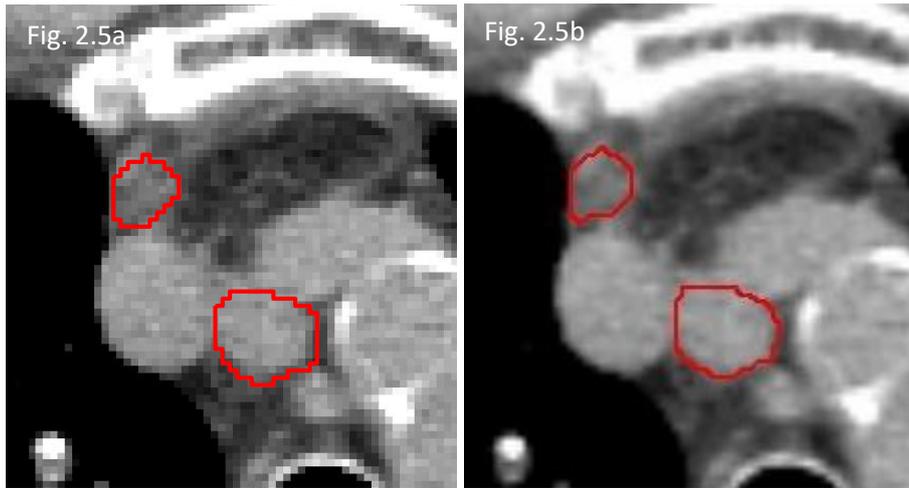


Figure 2.5. Example of display of GTV contour for the same axial slice in a) MATLAB and b) VODCA, showing multiple contours on a single axial slice.

Some of the cases contained slices where multiple GTV contours were present on the same axial slice, an example of which is shown in figure 2.5. As the ROI polygons on these slices were processed separately, the binary masks were generated appropriately to reflect the location of the corresponding polygons, i.e. there did not appear to be a mix up of coordinate points between the ROIs on the same slice.

2.14 Task A.2 Determination of heterogeneity and imaging quality of cases to assess need for image enhancement

The characteristics of all 79 cases (training and testing) are described here.

2.14.1 Clinical characteristics of total data

There were 6, 55 and 18 patients with stage II, IIIa and IIIb disease respectively. There was a wide variation in size of disease across the cases, with a mean volume of $121.75 \pm 101.49 \text{ cm}^3$, ranging from 10.89 cm^3 to 609.10 cm^3 , reflecting heterogeneity of the cases. Majority of the cases had either primary or nodal disease adjacent to or involving the hilar or mediastinum, with nearly half of the cases with disease adjacent to the chest wall. There were 12 cases where the primary disease was surrounded by the lung parenchyma in its entirety and there was nodal disease in all but one case. About a fifth of the cases had atelectasis or collapse of the adjacent lung. 44 cases were performed with IV contrast, leaving 35 cases where contrast was not administered.

	Number of cases
Adjacent to/Involvement of hilum or mediastinal	78
Adjacent to chest wall	34
Peripheral primary	12
Presence of adjacent lung atelectasis/collapse	18
Presence of cavitation	9
Presence of GGO	21
Small airways included as part of GTV	12

Table 2.3. Clinical characteristics of all data (79 cases).

The heterogenous case mix was apparent from the clinical characteristics of the dataset, with disease present in a range of locations within the thorax. A large proportion of cases had disease adjacent to the soft tissue in the chest. This case mix is very different to the majority of studies on lung segmentation, performed on cases with disease of earlier stages.

2.14.2 Scanning parameters of total data

The patients were recruited from eight different UK centres, with the scanning parameters shown in table 2.4. All 79 cases had associated GTV structures. Despite having assessed the DICOM meta file and CT imaging appearances, it was not possible to confirm if any of these scans were obtained as part of 4D CT planning, as much of the data in the DICOM metafile had been removed through the export and anonymization process, including information on dose and pitch. For the same reason, some of the metadata on the scan parameters was also not present for two centres.

Nonetheless, it can be seen that there was a spectrum of different machines and protocols used to acquire the planning CT images. This raised the concern that the variation in scanning practices could result in differences in noise across the scans acquired across the centres. Moreover, the absence of information on some of the scanning parameters also made the assessment of image quality imperative.

Centre	No. of cases	CT Manufacturer	Slice Thickness (mm)	Pixel Spacing (mm)	Tube Voltage (kV)	Modulating or fixed current	Tube current (mA) Mean (range)
A	19	GE Medical Systems	2.5	0.977 or 1.270	120	Modulating	179.1 (70 – 586)
B	17	Toshiba	3	0.931 to 1.404	120	Modulating	Unknown
C	14	Unknown	3	0.977 or 1.094	Unknown	Unknown	Unknown
D	11	GE Medical Systems	3	0.977	120	Modulating	173.4 (135 – 212)
	1	GE Medical Systems	3	0.977	120	Fixed	150
	1	Toshiba	3	1.074	120	Fixed	100
E	4	GE Medical Systems	2.5	0.977	120	Modulating	319.3 (309 – 331)
	2	GE Medical Systems	2.5	0.977	120	Fixed	250
F	4	Philips	3	1.172	120	Modulating	139.7 (109 – 169)
	1	Unknown	3	1.172	120	Fixed	45
G	4	Philips	3	0.965 to 1.065	120	Modulating	156.2 (133 – 174)
H	1	Philips	3	1.061	120	Modulating	203.2

Table 2.4. Scanning parameters of all data (79 cases).

2.14.3 Image quality assessment

The image quality assessment was performed on the 18 subsample cases. The discussion points are made here as decisions will impact the analysis in the following sections.

Table 2.5 shows the signal and noise associated with the muscles and the subcutaneous fat. As the measured noise for both muscle and fat was normally distributed (Shapiro-Wilk test, p -value = 0.688 and 0.101 respectively) and had equal variance (Levene's test p -value = 0.186), a 2-tailed T-test was used which showed no statistical difference between the measured noise of the muscles and the subcutaneous fat (p -value = 0.244).

	Musculature	Subcutaneous fat	p -value
Signal	44.9 ± 11.9	-117.5 ± 12.6	Not applicable
Noise	9.56 ± 3.6	8.33 ± 2.5	0.244

Table 2.5. Mean (and SD) of signal and noise for the muscle and fat in 18 subsample training cases.

There was some variation in the noise between scans, as reflected by the SD of the noise and the range (12.4 and 8.6 for muscle and fat respectively). This variation was expected of the differences in imaging protocols, which was felt to be within acceptable limits. With the 95% confidence interval (CI) for the mean noise measured in the muscle of 7.77 and 11.35, and 7.08 and 9.58 for the subcutaneous fat, overall the levels of noise across the cases is low.

The measured noise in this dataset is comparable to reported studies of diagnostic CT imaging of the thorax using a fixed voltage of 120kV with current modulation (341-343). Image noise

was reported at 9.01 ± 1.63 for the muscle by Hu et al (343), 10.8 by Mayer et al (341) and 11.4 by Schimmöller et al (342).

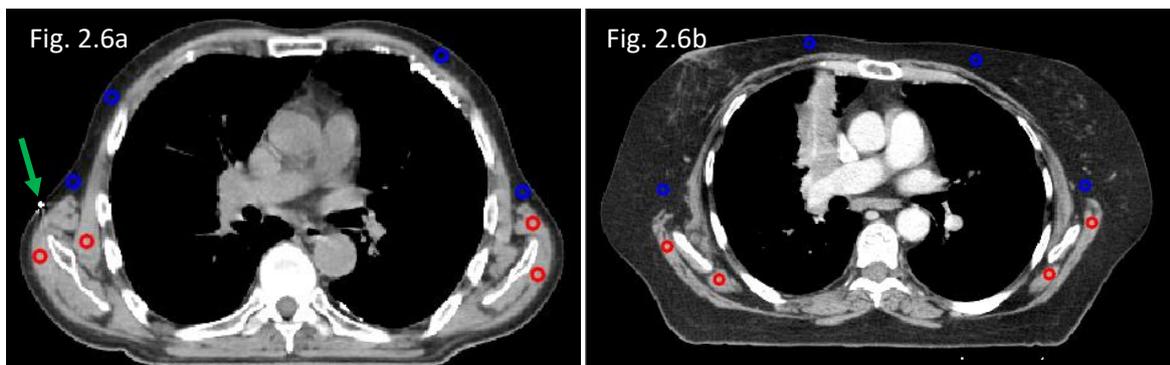
The SNR and CNR is displayed in table 2.6.

SNR	5.1 ± 1.6
CNR	21.8 ± 8.3

Table 2.6. Mean (and SD) of SNR and CNR for 18 subsample training cases.

SNR of the chest muscle to the background fat was reported at 5.83 ± 1.22 by Hu et al (343), which is within the error estimate of these results. A slightly higher SNR of 6.36 was observed by Peng et al (344) as compared to these results. On the other hand the CNR seen here was higher than that reported by Peng et al at 13.35 (344). Higher CNR was also obtained in this work as compared to a phantom study on pulmonary nodules (345).

A limitation of this work is in the selection of the ROIs. There were varying sizes of the musculature as well as the subcutaneous tissue between patients. In subjects with thin musculature or subcutaneous tissue, there were limited regions in which the ROIs could have been placed, in order to avoid tissue boundaries and artifacts.



Figures 2.6a – b. Display of ROI selection for two subjects with differences in anatomy, as well as presence of artifacts from localisation markers (green arrow). Red ROIs denote selection in soft tissue musculature, blue ROIs denote selection in background subcutaneous fat.

Despite the limitations, the image noise seen here is largely comparable to other studies. Additionally, there is suggestion that it may not be feasible to select a global denoising technique that would be applicable for all CT images (346). Judging by the relatively low noise for this dataset, it was decided not to apply any denoising filters in the pre-processing step for this work.

2.15 Task A.3 Determination of the descriptive statistics of the GTV and surrounding tissues

Figure 2.7 shows the distributions of the GTV, ipsilateral lung, bone, and vessels for all 18 subsample training cases collectively. There is least amount of overlap of the distribution of GTV with the ipsilateral lung, compared to bone and vessels. In spite of this, the GTV distribution appears to be negatively skewed with small numbers of pixels at low HU values, which is likely to be attributed by small airways, cavities or GGOs present as part of the GTV.

There is a bimodal distribution for the vessels, which would be explained by the difference in the use of IV contrast between cases. Despite the overlapping distributions of the GTV versus the bones and vessels, judging by the plots, thresholding based on HU is likely to be helpful for separating the GTV from these organs especially in the absence of a positive skew.

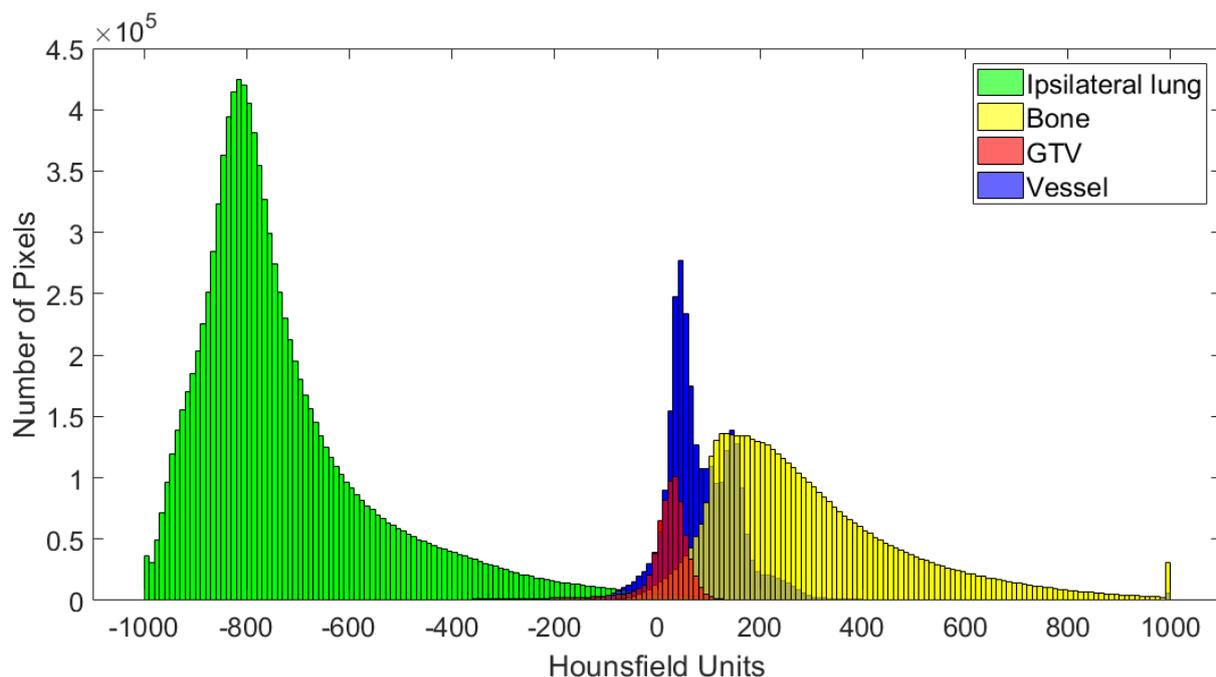


Figure 2.7. Histogram plots displaying distribution of Hounsfield Units for the GTV, bone, ipsilateral lung and vessels across 18 subsample training cases collectively.

2.15 Task A.4 Determination of thresholding procedures

The determination of an absolute threshold should be based on the GTV rather than the other organs to minimise regions of the GTV from being excluded. Two threshold levels, an upper and a lower, would be helpful for excluding high and low HU values respectively. To set this level, the summary statistics of the percentiles of the GTV distribution for the 18 subsample training cases was produced as shown in tables 2.7 (upper threshold) and 2.9 (lower threshold).

	90 th Percentile	95 th Percentile	99 th Percentile	100 th Percentile
Mean \pm SD	57 \pm 20	68 \pm 23	94 \pm 32	288 \pm 194
Maximum	96	114	149	735
Mean + 1 SD	78	91	126	482
Mean + 2 SDs	98	115	158	675

Table 2.7. Hounsfield unit values for GTV of 18 subsample training cases from 90th to 100th percentile.

At the upper percentiles of the GTV, the SDs for the means are low, except at the 100th percentile, which is likely due to delineation errors e.g. inclusion of small regions of bone. Both the 95th and 99th percentiles were considered for the upper threshold level. Although basing the threshold on the 95th percentile would decrease the risk of such delineation errors, this would be at the expense of increasing the false negative regions. To allow for a margin of error, it was decided that the mean + 2 SDs of the 99th percentile was to be used for the absolute upper threshold value, i.e. 158 HU.

To measure the accuracy of this estimate, the 95th and 99th percentile values for all the training cases were evaluated with bootstrapping (1000 number of samples) (table 2.8). The mean + 3SD for the 99th percentile was not evaluated as this was felt to be too high a threshold for it to be useful.

	CI for 95 th percentile		CI for 99 th percentile	
	Lower bound	Upper bound	Lower bound	Upper bound
Mean	67	80	96	117
Mean + 1SD	90	109	129	179
Mean + 2SD	112	138	157	237
Mean + 3SD	136	169	-	-

Table 2.8. Hounsfield unit statistics for 95th and 99th percentiles for all 63 training cases performed with bootstrapping (1000 number of samples).

From table 2.8, it can be seen that the threshold of 158 HU is within the 95% CI for the mean + 1SD, mean + 2SD for the 99th percentile, and mean + 3SD for the 95th percentile. As it is higher than the 95% confidence estimate for the mean + 2SD for the 95th percentile as well as the mean at the 99th percentile, this was felt to be an adequate cut off. Therefore, an upper threshold of 158 HU was fixed for all the procedures.

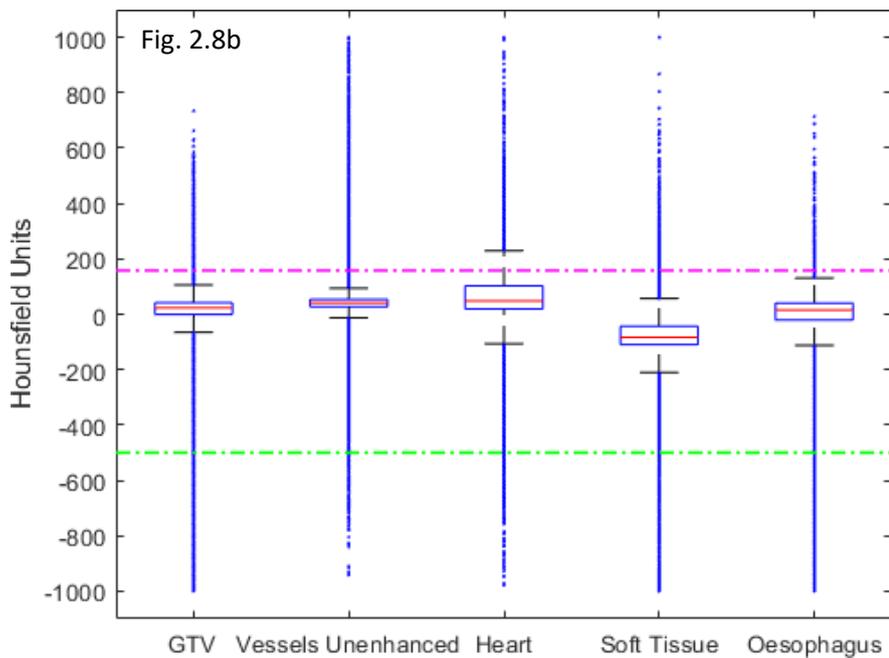
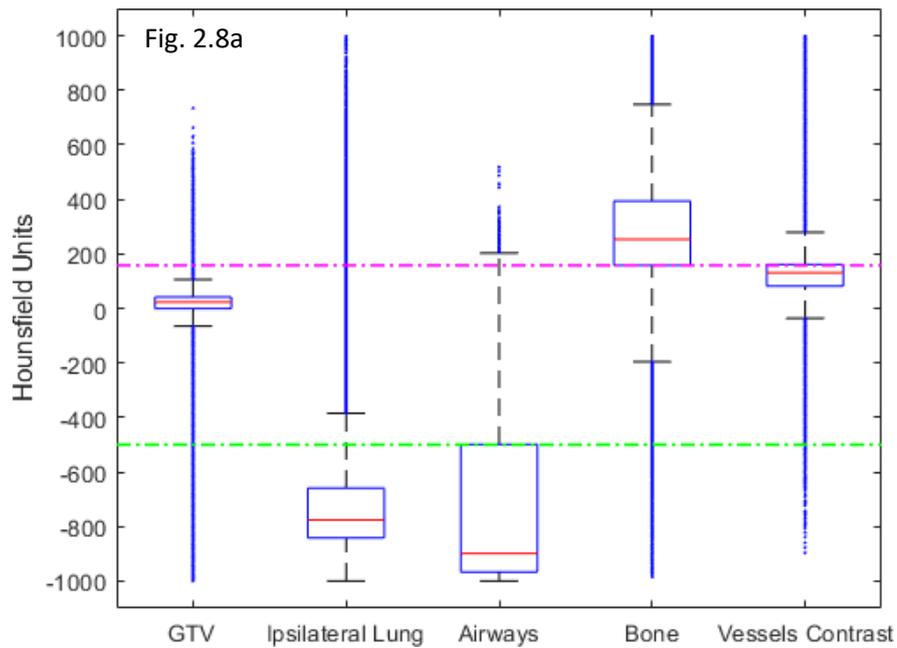
Conversely, there was a wider distribution of Hounsfield units at the lower percentiles of the GTV (table 2.9). This suggested that there were differences between the GTVs in containing regions of low HUs, such as cavities, GGOs and small airways.

	10 th Percentile	5 th Percentile	1 st Percentile	0 Percentile
Mean ± SD	-140 ± 183	-277 ± 222	-510 ± 240	-850 ± 153
Minimum	-779	-848	-940	-1000
Mean - 1 SD	-323	-499	-749	-1000
Mean - 2 SDs	-506	-721	-989	-1000

Table 2.9. Hounsfield unit values for GTV of 18 subsample training cases from 0 to 10th percentile.

An absolute HU threshold of -500 has been used by others to threshold the lung parenchyma (347), which was considered for the lower threshold level.

The boxplots in figure 2.8 show the impact of using -500 and 158 HU as the absolute threshold levels. This would permit the exclusion of a large proportion of the lung parenchyma, airways and bone. Some regions of contrast-enhanced vessels would also be excluded with the upper threshold, but this process would not be helpful in the exclusion of other organs in the mediastinum due to their overlapping HUs with the GTV.



Figures 2.8a – b. Boxplots showing HU distribution for the GTV and other normal tissue across 18 subsample training cases collectively. Pink dash-dot line denotes the use of 158 HU as the absolute upper threshold; green dash-dot line denotes the use of -500 HU as the absolute lower threshold.

However, if -500 HU were to be used as the lower threshold level for all cases, a high proportion of GGOs and cavities would be excluded from the contour. Consequently, it was decided to base the lower threshold on the lowest 5th centile of the GTV. If this value was greater than -500 HU, -500 HU was used as the lower cut-off. Otherwise, the HU value at the 5th centile was chosen as the lower threshold value instead.

2.15.1 Solid versus non-solid lesions

Even with the adaptation of the lower threshold value, there is still a potential risk of excluding true tumour regions for non-solid lesions. Thus, a further evaluation was performed to assess if an automatic process based on tumour statistics could be incorporated to distinguish solid and non-solid cases. Out of the 18 cases, there were 10 cases with solid lesions and 8 cases with non-solid lesions (2 cavitating lesions, 3 GGOs and 3 cases with small airways). Figure 2.9 shows the mean as well as the error bars (95% CI) of the lower percentiles for these two groups, and the descriptive statistics is summarised in table 2.10.

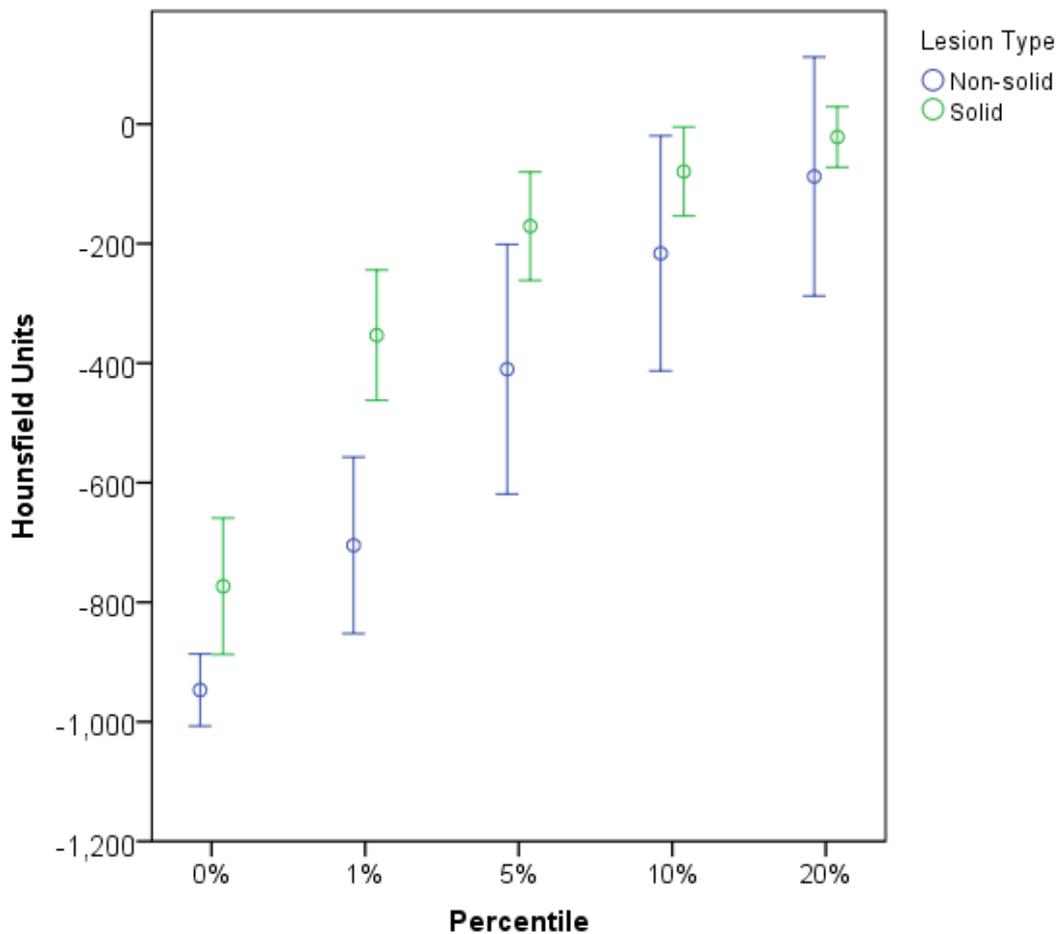


Figure 2.9. Mean and 95% confidence interval of lower percentiles for GTV between non-solid and solid lesions for 18 subsample training cases.

	Non-solid lesions	Solid lesions	<i>p</i> -value
0 Percentile			
Mean ± SD	-947 ± 72	-773 ± 159	0.006
95% CI of mean lower bound	-1000	-887	
95% CI of mean upper bound	-887	-659	
1st Percentile			
Mean ± SD	-705 ± 177	-353 ± 152	0.000
95% CI of mean lower bound	-852	-462	
95% CI of mean upper bound	-557	-244	
5th Percentile			
Mean ± SD	-410 ± 250	-171 ± 126	0.009
95% CI of mean lower bound	-619	-261	
95% CI of mean upper bound	-201	-80	
10th Percentile			
Mean ± SD	-216 ± 235	-79 ± 104	0.021
95% CI of mean lower bound	-413	-154	
95% CI of mean lower bound	-20	-5	
20th Percentile			
Mean ± SD	-88 ± 123	-22 ± 71	0.762
95% CI of mean lower bound	-287	-72	
95% CI of mean upper bound	112	29	

Table 2.10. Descriptive statistics for non-solid versus solid lesions for 18 subsample training cases (Mann-Whitney U test).

The difference between the means of the solid versus non-solid lesions at the 10th, 5th, 1st and 0 percentile for the GTV were statistically significant, with the largest difference for the means at the 1st percentile. Similar trends were observed when this was performed across all 63 training cases with bootstrapping (1000 number of samples) to obtain a better accuracy of the estimate, as shown in figure 2.10 and table 2.11. There were 28 cases with solid lesions, and 35 cases which were considered non-solid (6 cavitating lesions, 18 GGOs and 11 cases with small airways).

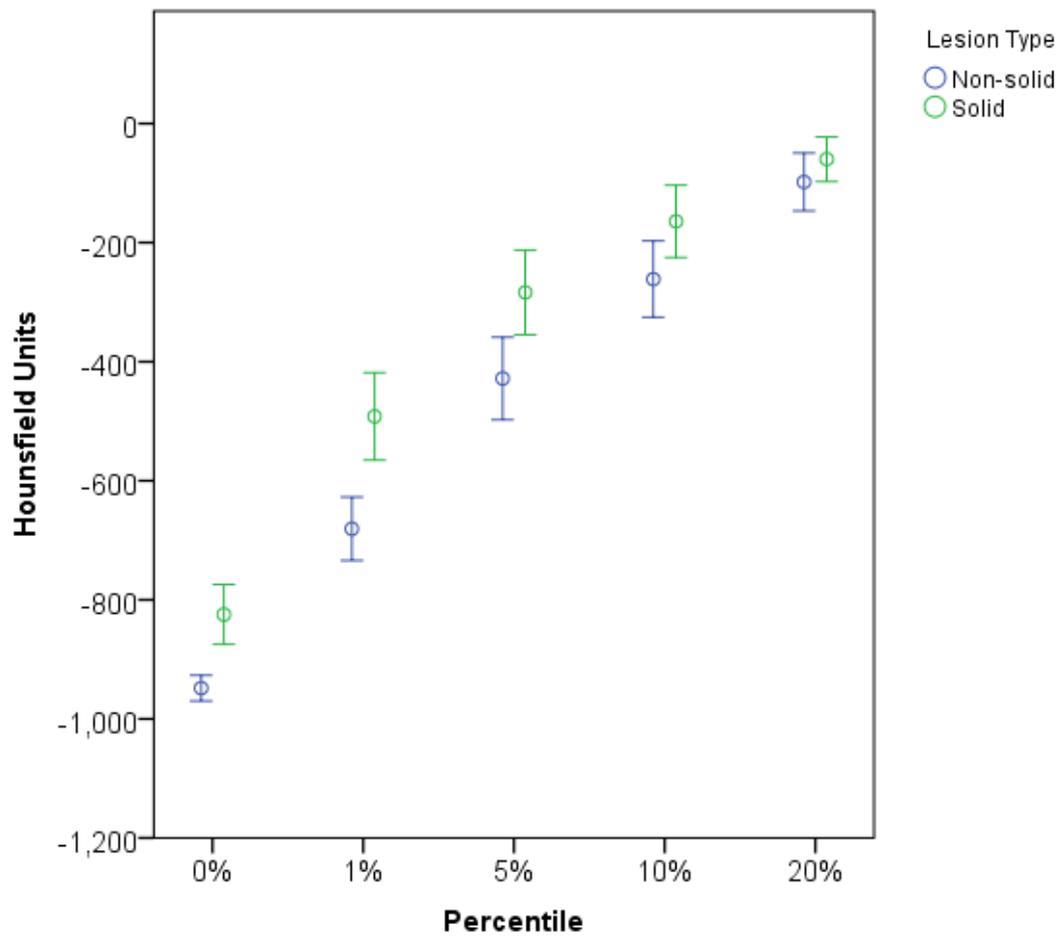


Figure 2.10. Mean and 95% confidence interval of lower percentiles for GTV between non-solid and solid lesions for total training dataset (63 cases).

	Non-solid lesions	Solid lesions	<i>p</i> -value
0 Percentile			
Mean ± SD	-948 ± 63	-825 ± 130	0.000
95% CI of mean lower bound	-970	-875	
95% CI of mean upper bound	-927	-774	
1st Percentile			
Mean ± SD	-680 ± 155	-492 ± 189	0.000
95% CI of mean lower bound	-734	-565	
95% CI of mean upper bound	-627	-419	
5th Percentile			
Mean ± SD	-428 ± 202	-283 ± 184	0.008
95% CI of mean lower bound	-498	-355	
95% CI of mean upper bound	-359	-212	
10th Percentile			
Mean ± SD	-261 ± 187	-164 ± 156	0.030
95% CI of mean lower bound	- 325	-225	
95% CI of mean lower bound	- 197	-104	
20th Percentile			
Mean ± SD	-98 ± 142	-60 ± 97	0.226
95% CI of mean lower bound	-147	-97	
95% CI of mean upper bound	-49	-22	

Table 2.11. Descriptive statistics for non-solid versus solid lesions for all 63 training cases (Mann-Whitney U test).

As the largest difference was observed at the 1st percentile, this was used as the discriminatory factor to distinguish cases with solid and non-solid lesions. To allow a generous margin of error in the estimate, a threshold of -500 HU was selected based on the plots above.

Thus, in the application of the lower thresholding, cases with a GTV 1st percentile value greater than -500 HU (solid lesions) did not require further operations. Conversely, for cases with a GTV 1st percentile value less than -500 HU (non-solid lesions), a selection for the largest connected region segmented by the lower threshold was performed, and other smaller regions were removed. This would help prevent regions of the GTV from being excluded by the lower threshold for non-solid lesions.

Although tighter thresholding limits could be set, it was decided not to do so, to avoid increasing the number of false negative regions from the segmentation. During the preliminary work in setting up the segmentation schemes, it was seen that this process was helpful in the segmentation process.

Discussion

The developments for data import lead to a more time-saving and efficient process, requiring much less manual input. This is imperative to the subsequent segmentation workflows, where the functionality would be extended to a larger dataset. A major benefit with this set up was the flexibility and accessibility of the data to suit the purposes of the segmentation work, rather than the use of an existing software. For example, some segmentation techniques require multiple contours on the same axial slice to be processed separately before amalgamating the results for the final evaluation, which was made possible with this work.

It was apparent from the clinical characteristics of the dataset that the case mix was very heterogenous in terms of disease size, location and juxtaposition to adjacent tissue. Moreover, there were also differences in contrast use. In contrast, many of the segmentation studies have been performed on collections from available databases comprising of cases different to this cohort. This includes the Lung Image Database Consortium (LIDC) which is a large collection of cases with lung nodules stratified into either less than 3mm or between 3 to 30mm in diameter (348). As compared to the RIDER Lung CT collection, the tumour sizes were larger in this dataset (349). Also, the majority of the cases here had disease with soft tissue attachment, which is present in only just over half of the cases in the RIDER. Although the LungCT-Diagnosis Collection comprises of cases at stage II and III, up to 41% of the cases had stage I disease, with mean tumour sizes smaller than our dataset (350). These would have to be taken into context when evaluating the performance of the segmentation techniques against other studies. In the subsequent application of the segmentation techniques, the diversity of this data would allow an assessment of the performance of the different approaches at different tissue boundaries.

Despite the differences in the scanning parameters, acceptable noise, SNR and CNR were observed, which formed the basis of the decision to withhold a noise reduction pre-processing step. Although this may potentially worsen the segmentation results at the extreme of the case spectrum with more noise present, there is a greater need for preserving the image data integrity at the tumour edge borders, especially in consideration of our case mix. This also tested the ability of the segmentation in the presence of noise, and in its general applicability to images from multiple sources.

The threshold level selection was performed with a generous margin to avoid exclusion of tumours in non-solid cases. There are other approaches that can achieve this. For example, filling of tumour cavities could be carried out using region-based methods, due to their relatively homogenous appearance and high contrast at their borders (351). However, the filling of airways (352, 353) and excluding GGOs (258, 354) is a more challenging task, often requiring more complex techniques. Thus, thresholding with some adaptation on a case-by-case basis was applied as a simple means of fulfilling the task, which is similar to the technique adopted by Tan et al (231).

Conclusions

A robust system was set up for processing the data efficiently in preparation for the project. There was heterogeneity in the dataset both in terms of clinical characteristics and scanning parameters. However, image quality was deemed to be sufficient for the segmentation task. Appropriate bi-level thresholding values were determined, with further adaptation for non-solid cases. These form the basis of the segmentation work that is explored in the next chapter.

Chapter 3

Specific Aim B: Development of different approaches to tumour segmentation

Introduction

There are many different approaches to segmentation of an image, as discussed in chapter 1 with their associated advantages and disadvantages. Four segmentation techniques were investigated in this work, including edge-based approaches, watershed segmentation, active contour models, as well as graph-cut segmentation. They represent state-of-the-art in conventional image processing-based segmentation algorithms. Deep learning segmentation approaches were not explored, primarily due to the relatively small number of cases available in this study.

The background to these various methods was described in chapter 1. Here, further details on the application of these techniques in MATLAB is discussed, and the rationale for the selection of some of these approaches for development is explained.

3.1 Edge-based segmentation techniques

The implementation of edge-based segmentation techniques on the CT images was straightforward in MATLAB using the `edge` function. An example of the segmentation results is shown in figure 3.1, using the Sobel, Prewitt, Roberts, Laplacian of Gaussian (default sigma 2) and Canny (default sigma $\sqrt{2}$) edge-detection techniques (79-82, 355, 356), where the threshold selection was set as the default and performed automatically.

It was seen that all 5 approaches appeared to be good at detecting strong edges, such as the body contour, bones, trachea, the boundary between the lungs and the chest wall, as well as vessels in the presence of strong contrast enhancement. However, the Laplacian of Gaussian and Canny methods out performed Sobel, Prewitt and Roberts in the detection of more subtle edges, such as the muscles in the chest wall, and vessels where the contrast enhancement was not as strong.

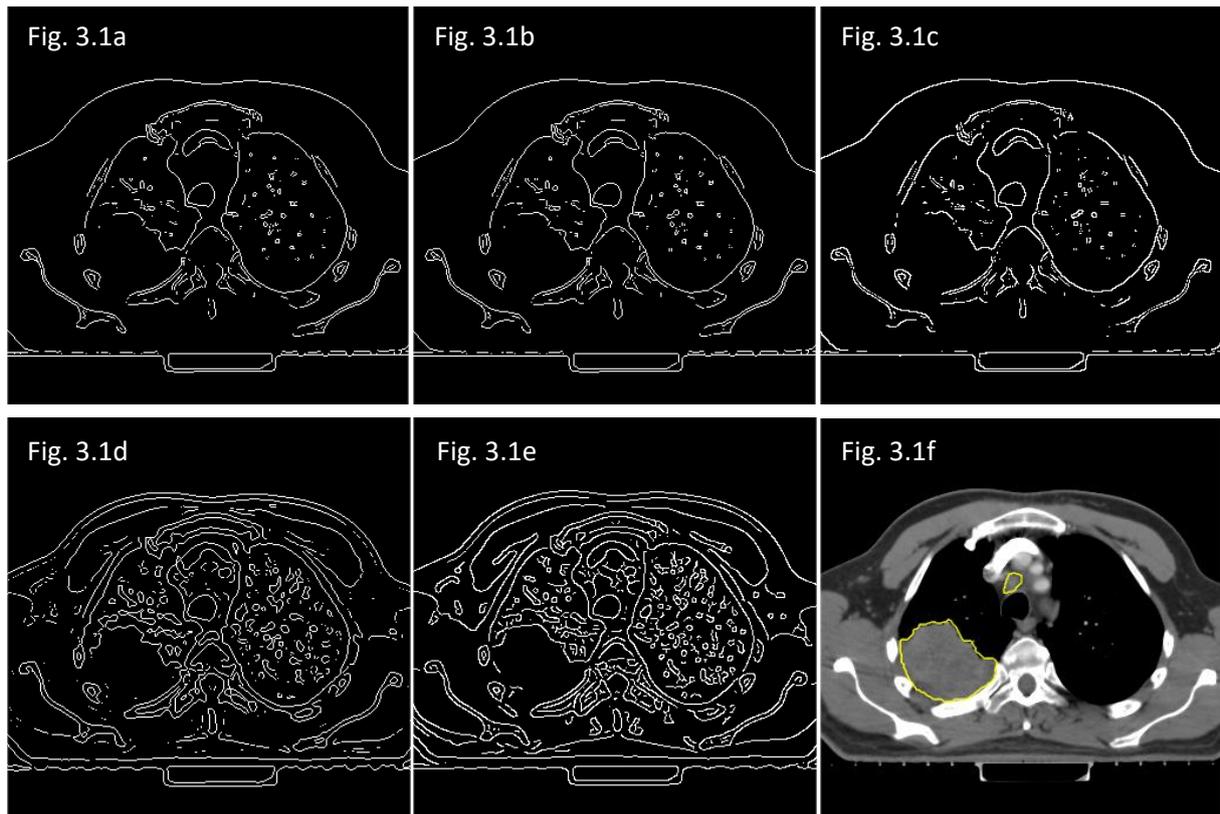


Figure 3.1. Edge-based detection techniques applied on CT image is shown in figure 3.1f, with the reference contours of the tumour indicated in yellow outline. a) Sobel approach (threshold 0.0567) b) Prewitt approach (threshold 0.0560) c) Roberts approach (threshold 0.0546) d) Laplacian of Gaussian approach (sigma 2, threshold 0.0019) e) Canny approach (sigma $\sqrt{2}$, threshold – upper: 0.0313, lower: 0.0125).

The striking feature across the five edge-based techniques was the creation of disjointed lines in the segmentation result, which is a known limitation for edge-detection methods. For the Sobel, Prewitt and Roberts methods, this was apparent even at boundaries where one would expect higher contrast, such as the bones and around the lungs. This effect was less prominent for the Laplacian of Gaussian and Canny methods, where an additional processing step for edge linkage was applied after the detection of the boundary. Although one would expect the influence of this to lessen with the lowering of the threshold, it was seen to remain an issue, as shown in figure 3.2, where Canny edge detection was applied with lowering of the threshold to an upper limit of 0.01, and a lower limit of 0.009. This change resulted in an increase in the detection of weaker boundaries such as the subtler pulmonary architecture and the tumour boundary, but also caused an increase in the detection of soft tissue and intra-tumoural inhomogeneity, some of which would have been created because of the inherent image noise. Significant post-processing would therefore be needed to select the contours of interest. Because of this, as well as the persistent issue with edge linkage, edge-based techniques were not explored further.

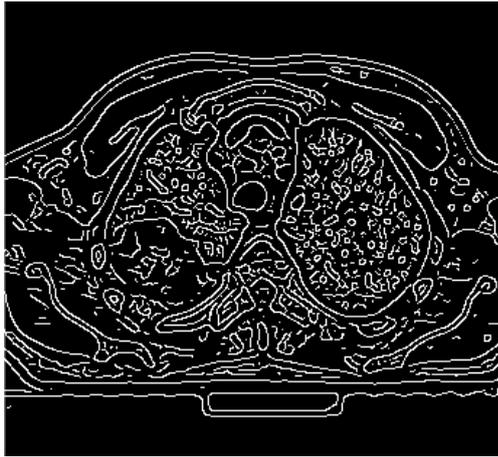


Figure 3.2. Results of segmentation with lowering of threshold using Canny edge detection technique (sigma $\sqrt{2}$, threshold – upper: 0.01, lower: 0.009), displaying increase detection of weaker edge boundaries and persistent limitation of incomplete edge linkage.

3.2 Watershed segmentation

There are many watershed algorithms that have been developed, which can be broadly classified into two main groups. One class is based on a recursive algorithm such as in Vincent and Soille's work, where a histogram is created based on the pixel values of the image, with its bins sorted according to the grey values. A hierarchical approach is then applied in the selection of the pixels fulfilling the criteria to be used in the progressive flooding of the catchment basins (i.e. by immersion), where pixels of lower grey values are flooded first (357). The other approach is based on Meyer's algorithm on distance functions; this is the method implemented in MATLAB (358). For this approach, the catchment basin for a particular minimum is defined as the set of points which are topographically closer to that minimum than another. Pixels are classified into one basin rather than another based on the geodesic lines of the topographic surface that follows the path of the steepest descent. With Meyer's implementation, issues with plateaus (i.e. regions of pixels with the same values) could also be overcome by reconstruction of the image (through the computation of the geodesic distances to the lower boundary of the plateau) before application of the algorithm. Thus, this approach could be useful in the separation of connected regions, as well as partitioning of non-connected regions of an image.

It was apparent early in this work that watershed segmentation cannot be directly applied to the images, the outcome of which is shown in figure 3.3 below, where it was implemented on an image slice in a training case.

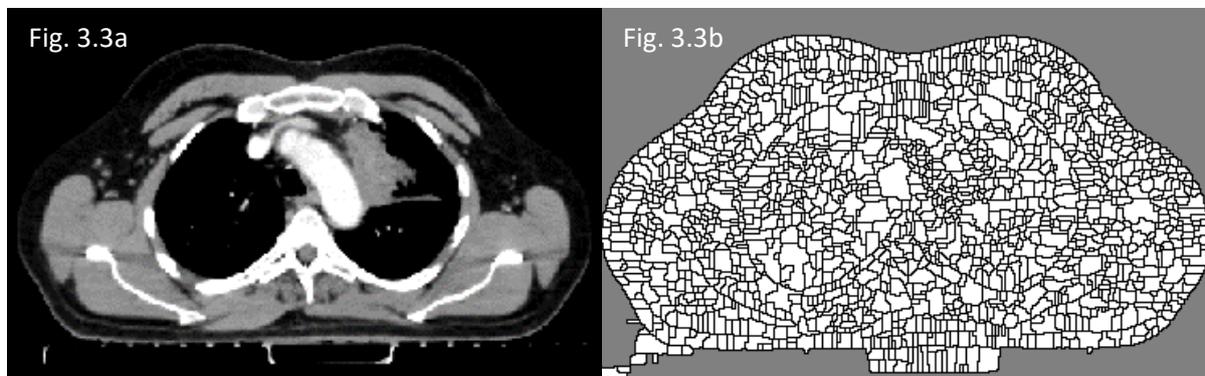
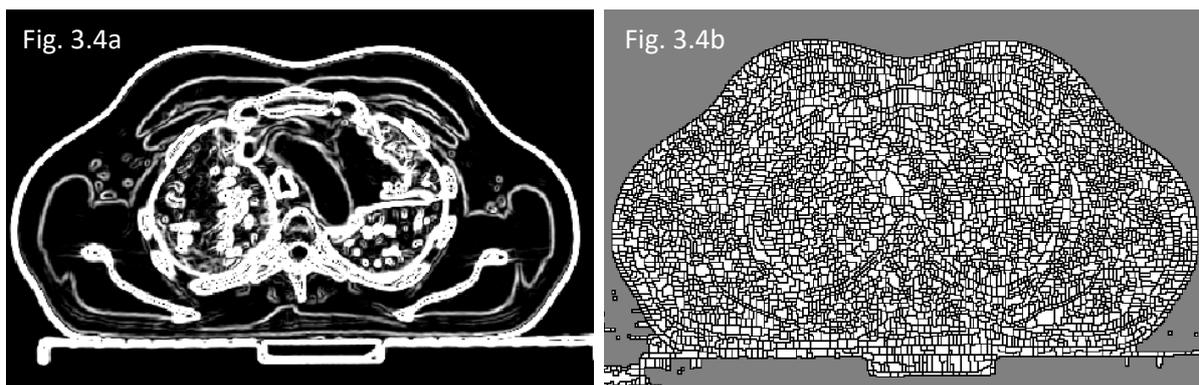


Figure 3.3. a) Axial slice of CT image b) Application of watershed segmentation, with result of watershed lines denoted by black boundary, and their corresponding white catchment basins.

One of the main issues with watershed segmentation is that often, the local minima are far too numerous, many of which are irrelevant to the object to be partitioned, which can result in oversegmentation of the image, as seen in the example above. Secondly, direct application of the watershed algorithm to the images is unlikely to produce any useful results for the GTV, on the premise that the HU of GTVs, similar to other soft tissue, is usually around the middle of the HU range of interest (-1000 to 1000). Conversely in the presence of higher HUs along the true object boundary, the algorithm seems to work well. For example, the segmented contours appeared to be highly congruent to the true body contour, which would be explained by the relatively higher HU of the skin, as compared to the subcutaneous tissue, or the surrounding air. This also seemed to be the case at the boundary between the lung and the chest wall, as well as the trachea and surrounding tissues.

Thus, to be able to apply the watershed algorithm on the CT images in a useful way, an image of the gradient would have to be computed, which had the effect of increasing the pixel values at the boundary of the GTV, as shown in figure 3.4a below. Conversely, the regions inside the object boundaries would have lower intensity values. However, direct application of the watershed segmentation without further processing had the same issue of oversegmentation due to the numerous local minima present in the gradient image (figure 3.4b).



Figures 3.4. a) Gradient magnitude (Sobel) of CT image b) Application of watershed segmentation, with result of watershed lines denoted by the black boundary, and their corresponding white catchment basins.

There are two main approaches in which the oversegmentation effect could be overcome. A separate region-growing algorithm could be applied to over-segmented images to merge the pixels. An alternative approach is to reduce the number of local minima.

This workflow was developed based on the latter through the application of marker-controlled watershed segmentation, which involved regions of local minima to be defined through the placement of markers, and hence result in a significant reduction in the number of resultant delineated regions. Another benefit with this technique is the introduction of *a priori* knowledge through the use of the markers, which can help in localising the segmentation to the section of the image close to the ROI.

The marker-controlled watershed approach has been explored in the medical setting in the segmentation of brain, lung, liver and breast lesions as well as lymph nodes (231, 274, 287, 359-363). In addition to automated methods of marker placement (e.g. thresholding), this was performed with manual input in a number of the studies, especially with regard to the external

marker, with an aim to limit the external boundary in which the segmentation occurred. Examples of this includes the work by Yan et al in the segmentation of lymphoma (287), Tan et al in the segmentation of lung tumours (231), Cui et al in the segmentation of breast lesions (363), and Bellon et al (364) as well as Yan et al (365) in the segmentation of liver lesions, where a semi-automated workflow was developed in the generation of the external and internal markers.

A fully-automated segmentation process was preferred for this work. To enable this, the submitted contours were modified and used to generate these markers. Additionally, a semi-automated process was also set up to assess how the segmentation results would differ to the fully automated workflow.

3.3 Active contour segmentation

Another segmentation technique that was explored is the active contour models. There are two variants of active contour algorithms available in MATLAB, the Chan-Vese (366) and geodesic (edge-based) active contour models (125).

Geodesics are the minimal length space curves lying on a surface that connects two points on that given surface. The geodesic active contour approach is based on finding the solution for the length minimisation problem between points on a curve, whilst taking the image characteristics i.e. the edge, into account. The principles of this algorithm are similar to the classic snakes method described in chapter 1 section 1.8.5, where it acts as an edge detector on the image. However, this algorithm allows greater flexibility to the evolution of the curve and allows for splitting and merging of the contour, a behaviour which the classic snakes approach is not capable of. This means that the contour can conform to the image characteristics to create multiple segmented regions on an image slice, despite starting off as a single contour, and vice versa. For clarity, this technique is referred to as the edge-based active contour approach in this report.

The Chan-Vese model is a variant of the active contour segmentation approach, which is based on the Mumford-Shah model where the image is analysed through regions, and each region is represented by a piecewise (a function which is defined by multiple sub-functions) constant. What sets this algorithm apart from other segmentation methods is in its implementation where edge information is not computed. Instead, the segmentation operates by fitting a two-phase piecewise constant model to the given image (also known as the minimal partition problem), such that the foreground and background are separated based on the intensity values of the two regions. This is akin to a region-based approach, where the curve is set at the boundary of the regions in which it is partitioning, whereby there is minimal difference in the intensity values within each of the regions. Therefore, for the curve to fit an object, there must be minimal variation of the intensities inside the curve as well as outside.

Figures 3.5 and 3.6a–d illustrates how the model works, where the red outline represents the curve (C), and the grey triangle represents the object within the image. The points within the image are represented by (x_n, y_n) . The location of the curve in figure 3.5 has not solved the minimisation problem as yet, as the minimum average intensity values for both within the curve (fitting term one, F_1) and outside (fitting term two, F_2) have not been achieved. To do this, the points at (x_2, y_2) and (x_3, y_3) should be on one side of the curve, whilst (x_1, y_1) and (x_4, y_4) on the other.

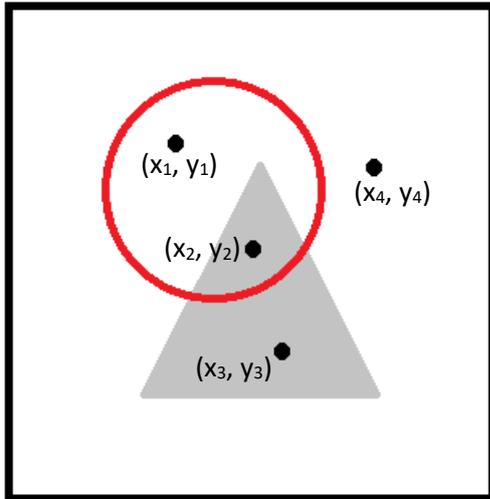


Figure 3.5. Representation of grey triangle region to be segmented from the white background, and the initiation of the Chan-Vese active contour curve as the red circular outline.

Figure 3.6 shows the possible evolutions for the curve. In figure 3.6a, the minimum average intensity within the curve has been achieved with the fitting function. However, as this has not been achieved for the region outside the curve, this would not be a good solution. Figure 3.6b shows the converse, where the region exterior to the curve has been well fitted, but not for the region within the curve. Figure 3.6c is akin to figure 3.5 above, which shows that the fit has not been achieved for both regions inside and outside the curve, whereas figure 3.6d shows the optimum solution for the partitioning.

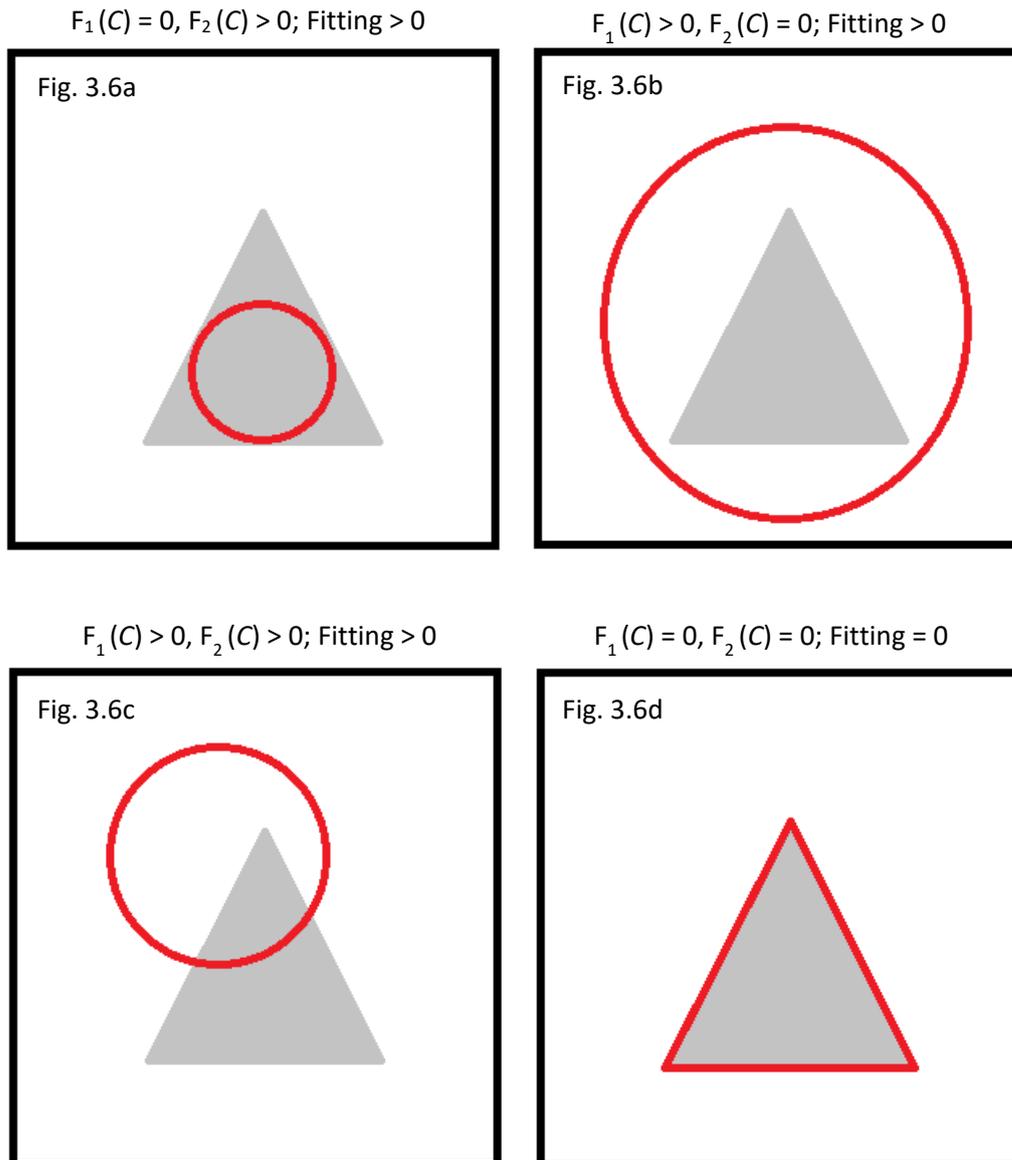


Figure 3.6. Examples of possible evolutions for the segmentation curve based on the Chan-Vese active contour model.

Because of the independence of this approach from edge information, one advantage of this approach is that it is less sensitive to image noise and blur.

The minimisation problem is solved by minimizing over all the boundaries of C . Classically, the propagation of the active contour spline is performed on a parametric curve. However, instead of applying the minimisation directly on the spline and manipulating the curve, like many other active contour models that have been developed, both the edge-based active contour and the Chan-Vese model implemented in MATLAB use the level set technique (Spare-Field level-set) to accomplish this (367), where the curve is represented as the zero-crossing of a level set function.

Although both of these approaches use the same technique (i.e. level set) to generate the contours, as they utilise different aspects of the image information (edge- versus region-based), both active control methods were investigated.

3.4 Graph-cut segmentation

Graphical model-based partitioning is another well-established technique used for image segmentation. As classification of the pixel of interest for this approach is based on the characteristics of neighbouring pixels, considering that tumour characteristics differs from the surrounding tissues, the use of graph-cut technique for tumour segmentation was explored.

The principles of graph-cut segmentation are explained in chapter 1. Conventionally, graph-cut algorithms evaluate an image on an individual pixel basis. However, this approach is computationally expensive and slow, which thereby decreases the efficiency of the workflow.

To decrease the computational time, instead of processing the image on an individual pixel basis, this workflow was developed to include a pre-processing step to group similar pixels together prior to applying the maximum-cut/minimum-flow algorithm. This would result in an over-segmented image comprising of pixel regions. With a far smaller number of nodes to be computed, efficiency and speed is therefore increased.

3.4.1 Oversegmentation and superpixel generation

Because of this advantage, oversegmentation is an increasingly popular approach as part of the pre-processing step in segmentation of images. The aim of this is to categorise pixels into small groups with similar characteristics, such as colour, brightness and texture, which can then be used as sub-regions in the segmentation process. Each group of pixels is known as a superpixel which contains rich local characteristics, where the information on the image structure is preserved. Compared to a single pixel, not only is the resulting smaller number of regions helpful in increasing the speed of the segmentation, the presence of collective information within the pixel regions also reduces the complexity of the partitioning algorithm and improves its efficiency.

The idea of superpixel was first introduced by Ren and Malik (368), where the Normalised Cuts algorithm (369) which globally minimizes an objective function by recursively finding the optimal partition in the normalised Laplacian graph, was applied together with contour and texture cues (370) to generate the superpixel map. Since then, a number of algorithms have been developed, where their performance and parameter selection have been compared in some reviews (371-375). Generally, the means for superpixel generation can be grouped into either a variant of graph-based approaches such as in Normalised Cuts, or region-growing methods, which are initiated from a set of pre-defined seed points. Much of this work has been performed on non-medical coloured images, although some have looked at its value in electron microscopy micrographs, as well as histopathological images (376-379).

Nonetheless, the use of superpixels has been extended to medical imaging. In the setting of MR images, it has been applied to other segmentation algorithms such as graph-cut and active contour based approaches for the prostate gland (380), graph-cut for brain tumours (381), fuzzy clustering in the brain (382), as well other classification methods for colorectal tumours (383). It has also been used in segmenting lung parenchyma in CT images (384), and applied to segmentation of the pancreas in conjunction with deep convolutional networks (385). Other areas in which superpixels have been used include the detection of breast lesions through support vector machines (386).

Despite the introduction of additional pre-processing steps to the segmentation process, the overall efficiency of the workflow can be improved if the partitioning step (i.e. the graph-cut

algorithm) is the main limiting factor in terms of computational speed. The generated superpixels also has to be reflective of the underlying image, to allow the subsequent partitioning to be accurate. Therefore, the accuracy of the superpixel generation is evaluated as part of this work, in addition to assessing the computational speed of the superpixel generation against the whole segmentation process.

3.5 Summary of tasks

Four different segmentation techniques were therefore developed and used to evaluate the tumour segmentation, as summarised below. The same processes for evaluation was used across the different approaches (see chapter 4).

Task B.1 Development and tuning of segmentation techniques on training dataset

- A) Marker-controlled watershed segmentation
 - a. Evaluation on synthetic geometric shapes
 - b. Analysis of gradient magnitude and gradient computation
 - c. Performance with further exclusion structures
 - d. Performance with manual input
 - e. Performance of fully automated process applying cross-validation folds
- B) Chan-Vese active contour segmentation
 - a. Determination of best process for segmentation initialisation
 - b. Parameter tuning on training cohort
 - c. Performance of fully automated process applying cross-validation folds
- C) Edge-based active contour segmentation
 - a. Analysis of initialisation of segmentation
 - b. Parameter tuning on training dataset
 - c. Performance of fully automated process applying cross-validation folds
- D) Graph-cut segmentation
 - a. Evaluation of superpixel generation and associated parameters
 - b. Parameter tuning for graph-cut segmentation on training cohort
 - c. Performance of fully automated process applying cross-validation folds

Methods

3.6 Pre-processing for watershed, active contour and graph-cut segmentation

The pre-processing steps were similar for all the techniques (figure 3.7) to allow for a fair comparison of the different segmentation approaches, whilst further development for each algorithm was performed separately. The generic pre-processing procedure is described here, and the additional processing and optimisation of each of the different techniques is described in the later sections.

Assessment of the segmentation performance for the different approaches was also carried out in a similar way, as described in section 3.11.

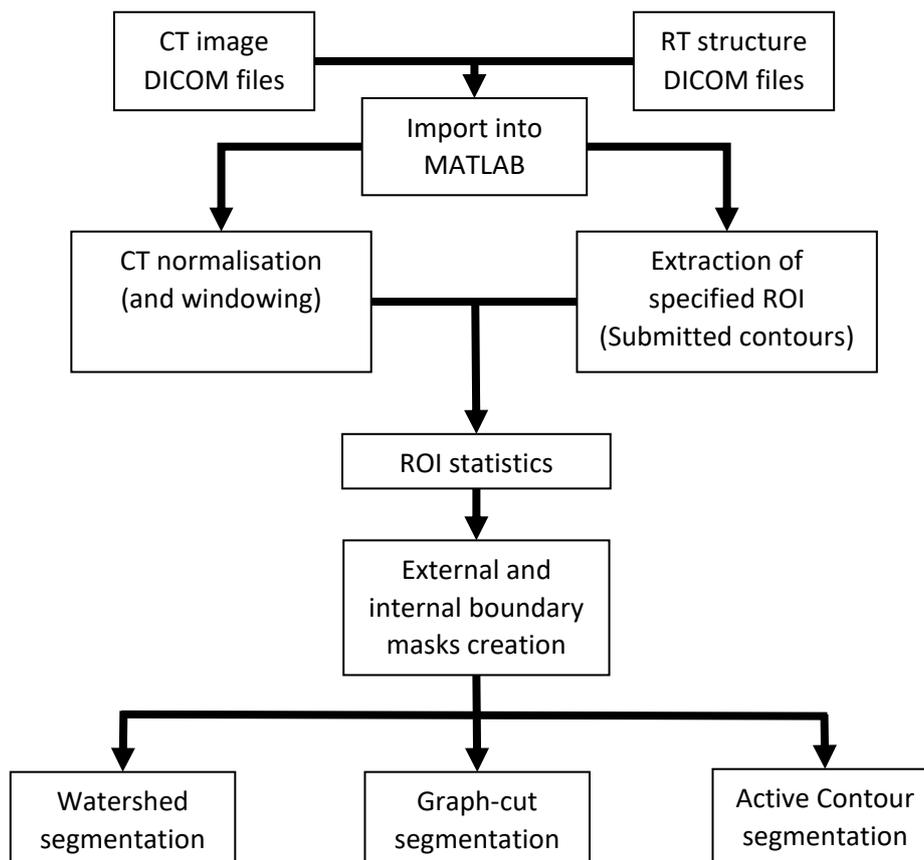


Figure 3.7. Summary of initial pre-processing workflow for different segmentation techniques.

3.6.1 Windowing levels

Consideration was given to adjustment of the window levels, to reflect the clinical practice in the delineation of structures within the thorax. Although there is some variation in the actual values used between clinicians, a level of -600 HU with a width of 1500 HU is typical of the settings for lung windowing, which is used in delineation of regions juxtaposed to lung parenchyma (387). For delineation of regions with adjacent soft tissue, a level of 40 HU and width of 350 HU is commonly used (387). The image data was adjusted to reflect this range to match the visual information as seen by a human observer, which translated to an upper limit equivalent to 215 HU. The lower HU limit was set at -1000, and the images were normalised between the range of 0 and 1. Where a HU value was applied the subsequent processing of the data (i.e. in thresholding), the equivalent value within the range of 0 to 1 was used.

3.6.2 External boundary limits

Bounding regions were defined to set a limit to the search area in which the segmentation was applied to improve the efficiency and accuracy of the techniques. The extent of the bounding region was decided based on the reported interobserver variation seen in the manual delineation of lung tumours (388), where a range of up to 4cm was seen in the axial plane.

Therefore, an external boundary of 20 pixels from the submitted contours was applied, which constituted a region with an approximately 2cm width from the submitted contours. This was felt to be a generous search region. To create this boundary mask, a morphological dilation was performed using a rectangle structuring element with a 41-pixel width, which was applied to the contour masks in the axial plane. This was chosen in preference to a disk-shaped structuring element, to ensure that the dilation was performed not just in the vertical and horizontal directions but also diagonally, which ensured that a larger search region was created.

3.6.3 Thresholding

In addition to the external boundary masks, thresholds were applied as additional limits for the segmentation.

Based on the work described in chapter 2, an upper threshold was set at the equivalent of 158 HU for all cases. This would help with excluding areas of high intensity values, such as bone and contrast-enhanced vessels. After applying the threshold, further morphological operations were performed to fill in any holes created in the process, and small regions of less than 6 pixels in size were removed.

A lower threshold was also incorporated, with an aim to exclude most of the lung parenchyma. Due to the tumour heterogeneity between cases at the lower intensity values, this was adapted for each individual case. Using the normalised CT images without application of the window levels, the statistics of the contoured GTV region was computed for all the slices to determine the intensity value at the 5th percentile, which was chosen as the lower threshold if it was less than -500 HU, otherwise, -500 HU was used. A value of -500 HU was selected as it is in keeping with absolute thresholding techniques in other work (347). Morphological operations were then performed to exclude the body and the chest wall.

It was also determined from the earlier work that non-solid lesions were likely to have a value of less than -500 HU at the 1st percentile of the GTV density values. Thus, to avoid the placement of lower threshold masks in regions of low density within the tumour region, cases with GTV values of less than -500 HU at the 1st percentile were further processed to remove masked regions that were not connected to the rest of the masked lung parenchyma. To avoid the low threshold exclusion regions from encroaching onto the tumour boundary at the tumour edge, a further morphological erosion of 2-pixels was applied.

3.6.4 Internal boundary limits

For the watershed and graph-cut algorithms, an internal marker was placed through the morphological erosion of the submitted contours by 20 pixels. A 10-pixel radius disk-shaped structuring element was used for this process, to decrease the likelihood of the whole mask from being eliminated during the erosion process, which can occur for small ROIs. Additionally, a skeletonised erosion was also performed and superimposed onto the interior

mask. This gave the added benefit of a shape prior to assist the segmentation. This extended the search region by another approximate 2cm width from the submitted contours.

It was observed in the development of this process that there were occasions where the search region was obliterated in certain regions, due to an overlap between the external and internal boundary limits after thresholding was applied. To avoid this, it was ensured that there would be an unmasked region of at least 1-pixel width exterior to the internal boundary limit.

3.7 Division of data

The 79 cases were divided into the same folds as previously described (chapter 2 section 2.7) for the development and evaluation of the segmentation techniques. There were 63 training cases (mean volume $116.24 \pm 87.35\text{cm}^3$) and a separate independent testing dataset comprising of 16 cases (mean volume $124.87 \pm 139.26\text{cm}^3$). Of the 63 training cases, the initial set up of the segmentation algorithms were performed on the 18 subsample cases (mean volume $117.72 \pm 68.63\text{cm}^3$).

After the segmentation workflows were established, a 3-fold cross validation training method was applied to the whole training data, to allow a better estimate of the performance of the segmentation and to avoid overfitting, especially for the segmentation methods where parameter selection was involved. The same cross-validation folds were used for the different segmentation techniques to allow comparison between the various methods. After parameter selection based on the training folds, the performance of the segmentation was assessed on the validation folds.

For each of the segmentation techniques, the optimal workflow and chosen parameter settings were then applied to the independent testing dataset.

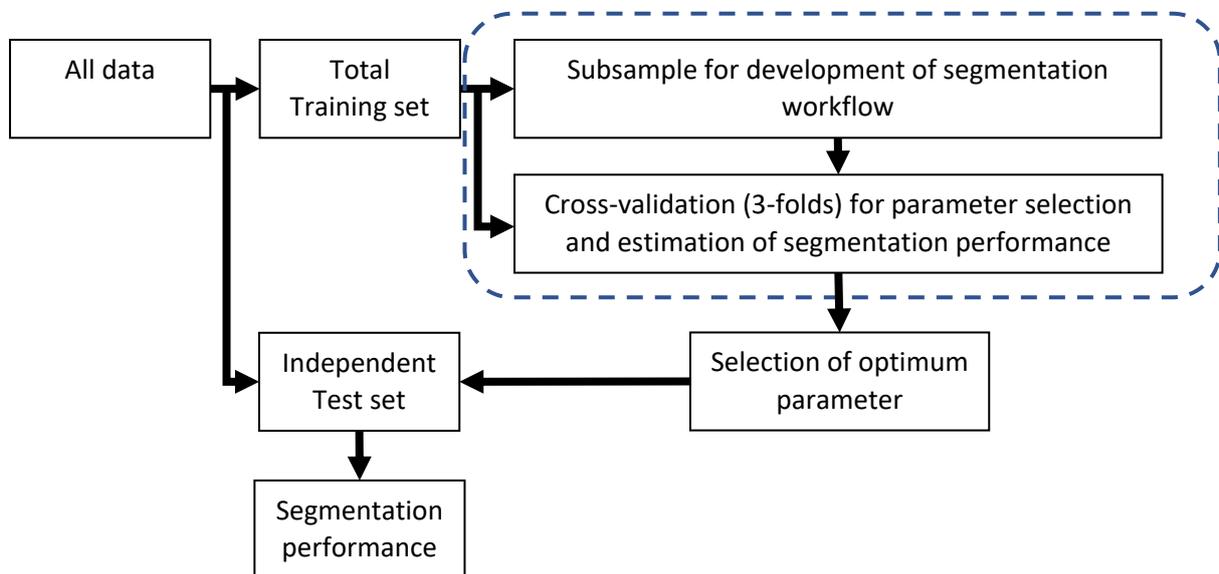


Figure 3.8. Diagram on data division for segmentation processes.

3.8 Watershed segmentation workflow

A summary of the workflow designed for the fully-automated marker-controlled watershed segmentation is shown in figure 3.9.

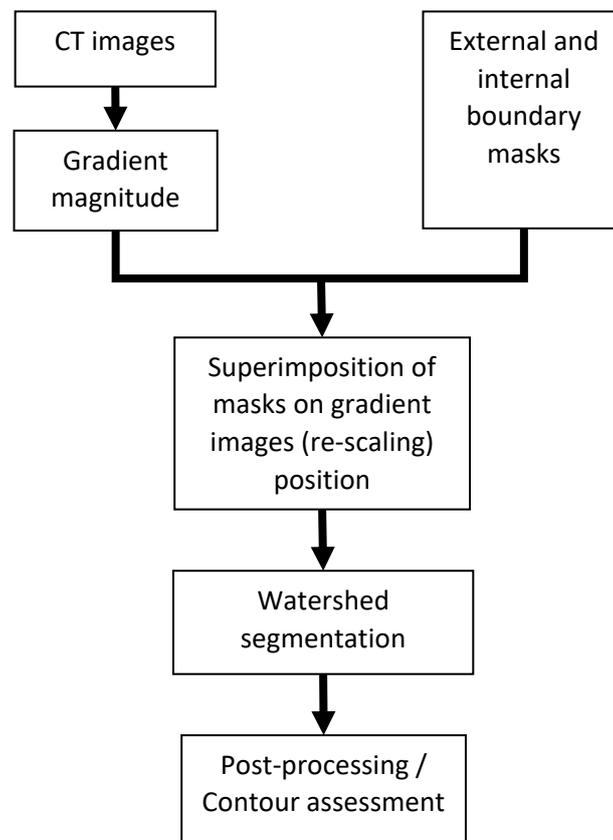


Figure 3.9. Workflow for watershed segmentation.

3.8.1 Image gradient computation

The magnitude of the image gradient was calculated using `imgradient`, which is a reflection of the directional change in the intensity of the image. There are a number of operators that can be used for this computation, some of which also form the basis of various edge detection techniques as described in chapter 1. Five classical first-derivative gradient operators were explored in this work.

The most basic gradient calculations can be computed through one-dimension filters, such as the central difference method. For each point in the image, the difference of the intensity between the pixel and its neighbours is averaged, given by the equation:

$$\frac{dI}{dx} = \frac{I(x+1) - I(x-1)}{2}$$

[3.1]

where I represents the pixel intensity at location x . As a one-dimensional filter, it is standardly computed in one axis, though it can also be applied in the orthogonal direction(s) to yield the gradient magnitude in two (or three) dimensions. Another similar way of calculating the gradient is through the intermediate difference operator, given by the equation:

$$\frac{dI}{dx} = I(x + 1) - I(x) \quad [3.2]$$

where only one single adjacent pixel is used in the calculation. Theoretically, because only two pixels are being evaluated at a time, the intermediate difference method can achieve better accuracy than the central difference method, as it preserves more detail of the original image. However, as the central difference computation averages over more pixels, it may be more advantageous at handling noise.

The downside of one-dimensional filters is that pixels connected diagonally are not taken into account in the computation. This is mitigated by 2-dimensional filters, several of which exist. The three 2-dimensional gradient estimators explored in this work operate on a similar basis, where the original image is convolved with the respective kernels.

The Roberts cross operator (81) performs discrete differentiation through computing the sum of the squares of the differences between diagonally adjacent pixels. This is achieved by convolving the image with the two 2 x 2 kernels:

$$\begin{matrix} \begin{bmatrix} +1 & 0 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & +1 \\ -1 & 0 \end{bmatrix} \\ G_x & G_y \end{matrix} \quad [3.3]$$

In other words, the gradient edges are computed at 45° and at -45°, and not across pixels horizontally or vertically.

Instead, the Prewitt gradient operator (79) is based on convoluting the image with two 3 x 3 kernels, which takes into account not only the diagonally adjacent pixels, but also the pixels in the horizontal and vertical planes. This is given by the matrices below:

$$\begin{matrix} \begin{bmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix} & \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{bmatrix} \\ G_x & G_y \end{matrix} \quad [3.4]$$

Unlike the operators described above, this convolution is independent of the pixel at the centre of the matrix where the kernels are applied.

Another classical gradient estimator is the Sobel operator (80). Similar to the Prewitt operator, it uses two 3 x 3 kernels as shown below, where one kernel is the equivalent of the other with a rotation of 90°.

$$\begin{matrix} \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} & \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \\ G_x & G_y \end{matrix} \quad [3.5]$$

The feature of the Sobel kernels is that a different weight is given to each of the surrounding pixels, which decreases the further away the neighbouring pixel is. This is given by an inverse order of the distance between the centre of the matrix to the surrounding pixels. In a 3 x 3 convolution, this results in a higher weight equivalent to $\sqrt{2}$ for the 4-connected pixels, whilst the other 4 in the diagonal directions have weight of 1.

From the computation of G_x and G_y through the different operators described above, the gradient magnitude G was then computed through:

$$G = \sqrt{G_x^2 + G_y^2} \quad [3.6]$$

The gradient operators can also be extended to larger matrices, but for this work, 3 x 3 kernels were used, to preserve as much of the local features as possible.

In the computation of the gradient operators, a normalisation factor was applied to Sobel, Prewitt and Roberts gradient calculations with a constant of $\frac{1}{8}$, $\frac{1}{6}$ and $\frac{1}{2}$ respectively, to allow cross comparison of the gradient magnitude.

3.8.2 Watershed segmentation

Both the external and internal masks as described in sections 3.6.2 and 3.6.4 were applied onto the gradient image. This was performed using `imimposemin`, where a morphological reconstruction was performed to re-scale the intensity values of the image. In the reconstructed image, regions of exclusion would have values corresponding to the minimum points in relation to the rest of the image.

The watershed transform was then applied to the reconstructed image. As a result of the reconstruction, the pixels values at the local minima of the original gradient image that were not within the excluded regions would be transformed to values higher than the regions of exclusion, thereby resolving the issue of oversegmentation. However, multiple regions could still be generated from the segmentation, as regions of very low local minima in the original gradient image would not transform to values very much higher than the regions of exclusion during the re-scaling process. These regions would therefore also be considered as the local minima in the transformed image, in addition to the excluded regions. Also, isolated exclusion regions that were generated following the application of thresholding would result in that particular region being partitioned as a separate object from the tumour.

3.8.3 Post-processing

Thus, further post-processing steps were undertaken to extract the segmented regions, where a selection process was performed to choose the most appropriate segmentation. The initial workflow for this was based on comparison of the centroid of the segmented regions with that of the original contour, followed by the selection of the region that had a centroid closest to the original contour. However, initial experiments showed that this was not a robust means of selection in the presence of multiple small segmented regions. Also, for slices with multiple contours in close proximity where the centroids were a small distance apart, errors in the selection were observed.

Therefore, the methodology was changed to an analysis of the area of overlap instead. For each of the segmented regions, the area of overlap between the region and the original GTV contour was computed. The region with the greatest amount of overlap was selected as the final segmentation. This process also ensured that the appropriate number of contours were selected, in the case of multiple GTVs in the same axial plane.

3.8.4 Experiments

3.8.4.1 Synthetic geometrical shapes

An initial evaluation on synthetic images was conducted to allow a better understanding of how the watershed segmentation performed, after observing promising results on preliminary clinical tests.

A synthetic image was created, comprising of 9 circles (2000-, 500- and 100-pixel sizes) with the intensity value equivalent to HU of 0 to mimic tumour (see figure 3.10). The background was split into three regions with different intensity values equivalent to HU of -800, -50 and 215 to represent lung parenchyma, mediastinum and areas of high contrast such as bone.

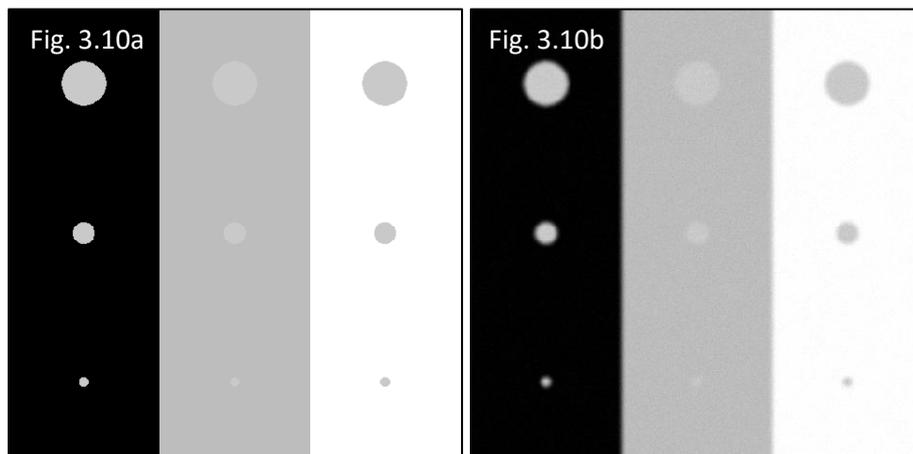


Figure 3.10. Synthetic images for evaluation of watershed segmentation process. a) Original image (also used as assessment reference); b) Degraded image (Gaussian edge smoothing with $SD = 2.5$, and gaussian noise with SD equivalent to 30 HU, mean = 0).

Gaussian filtering (standard deviation between 1 – 2.5) was then applied to blur the edges, to model the indistinct boundary that is often seen between tumours and the surrounding tissue. To evaluate the impact of noise, gaussian white noise was also added to the images, with a mean of 0. Standard deviations equivalent to 10, 20 and 30 HUs were used, to be reflective of the degree of noise seen in clinical scans (see chapter 2 section 2.14.3).

The original image and all the degraded images were then taken through the whole watershed segmentation process, using the five means of calculating the image gradient. Gradient operators displaying better performance was then tested using the clinical dataset.

3.8.4.2 Clinical cases

Gradient magnitude analysis

Further evaluation of the HU values and the gradient magnitude in the vicinity of the tumour boundary was undertaken for the 18 subsample training cases. For each of the reference contours, individual annular masks were created that were between 1 to 4 pixels distance from the GTV boundary, for both directions towards and away from the GTV edge. This resulted in eight masks of varying distance from the GTV contour for each axial slice. These were subsequently applied to the normalised CT images to extract the HU values associated with each distance. The gradient magnitude for each distance was also computed using the better performing gradient operators identified from the evaluation on synthetic objects.

Development of watershed segmentation

The segmentation process described above was developed and tested on the 18 subsample cases using the selected means for gradient computation. Computational time was also recorded.

Cross-validation with training dataset

After developing of the watershed algorithm with the 18 subsample cases, the final algorithm was implemented on the whole training data set in their respective cross-validation runs, to allow an estimation of the segmentation performance, as well as a fair comparison of the performance with the other automated segmentation techniques.

Further exploration and development of watershed technique

An evaluation into the improvement of the segmentation process using other techniques was performed using the 18 subsample cases.

3.8.5 Exclusion structures

The performance of watershed segmentation with additional exclusion structures was assessed. Other normal structures (airways, ipsilateral lung, oesophagus, mediastinal soft tissue, vessels, heart and the chest wall) were contoured and imported with the GTV and CT data. The z-coordinates for all the slices of each of the normal structure was extracted and used to match to those of the GTV. Subsequently, the masks for the normal structures at the corresponding GTV slices were skeletonised morphologically. These skeletonised masks were then merged with the exterior exclusion masks that were generated as described in section 3.6.2, prior to fusing with the interior exclusion limits.

The rest of the watershed algorithm was then applied as described.

3.8.5.1 Atlas-based generation of exclusion structures

To find a means for automating the generation of these structures, the atlas-based normal tissue auto-segmentation tool in OnQ rts (Oncology Systems Limited, UK) was explored. Some pre-defined atlases were already available within the software, for the thorax these structures comprised the lungs, bones and trachea. However, this list was not sufficient to complement the tumour segmentation using the watershed approach, necessitating the creation of new atlases which encompassed other normal tissues too. Thus, nine cases with contours on other normal structures (airways, oesophagus, mediastinal soft tissue, vessels, heart, chest wall and

spinal cord) were imported into OnQ rts to generate more atlases. The atlas-based segmentation was then applied to the remaining nine cases of the subsample dataset.

3.8.5.2 Semi-automated approach

An interactive interface was also developed to allow manual input of additional exclusion structures which ran as part of the algorithm. This was performed to assess if the segmentation results could be improved with additional exclusion limits. To minimise the extent of manual input, the workflow was designed to preserve the automated method of internal marker placement that had already been set up.

After the generation of gradient image, an addition code was implemented before the superimposition of the masks on the gradient image. This involved a visual display of the CT image in the axial plane, where the window levels could be specified and adjusted. There was also scope to magnify the image allowing better visualisation. As it was sometimes difficult to locate the GTV especially the nodal disease, an additional marker denoting the centroid of the GTV was also displayed on the image. The submitted contours were not displayed and not made available during this process to avoid bias and overfitting.

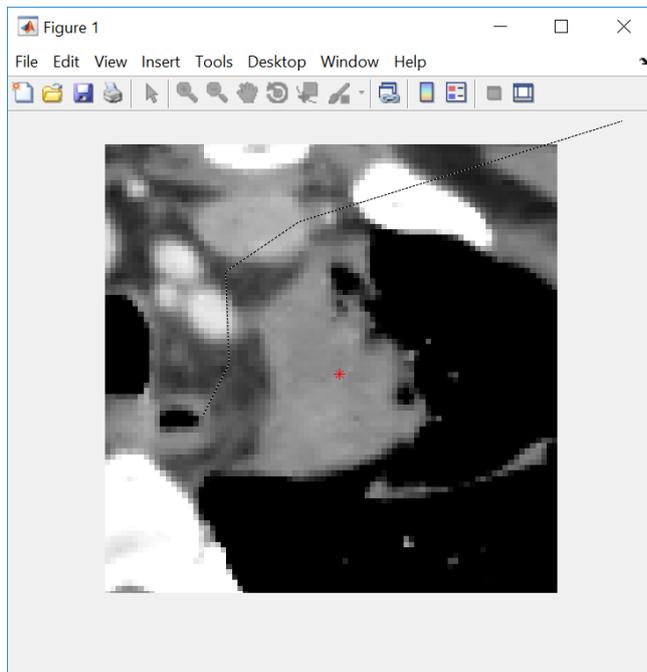


Figure 3.11. User interface for semi-automated approach. A magnified axial slice is shown with mediastinal windowing. The red marker denotes the centroid of the GTV for the slice. The dashed line depicts the point placement and the linear connection between the points, to generate a polygon. Note that the point placement was performed at a distance away from the GTV boundary.

On each slice, a series of points were placed exterior to the GTV by a series of clicks. It was ensured during placement of the points that this was performed at a distance away from the GTV boundary, as depicted in figure 3.11. The points were connected linearly, which was made visible during their placement. The last point placed on the image was also linearly connected to the first point to create a polygon. The polygon was subsequently converted to an exclusion mask and applied to the watershed workflow. This was developed in preference to setting the individual points as the limits to improve the computational efficiency when the watershed segmentation and post-processing is run, as this approach decreased the number of segmented regions created by the watershed process.

All the training cases were processed in one sitting. However, because of the potential bias inherent in manual selection, this procedure was repeated on three separate occasions at least three days apart, in the absence of reference to the reference contours.

3.9 Active contour segmentation workflow

Figure 3.12 summarises the workflow for both active contour approaches.

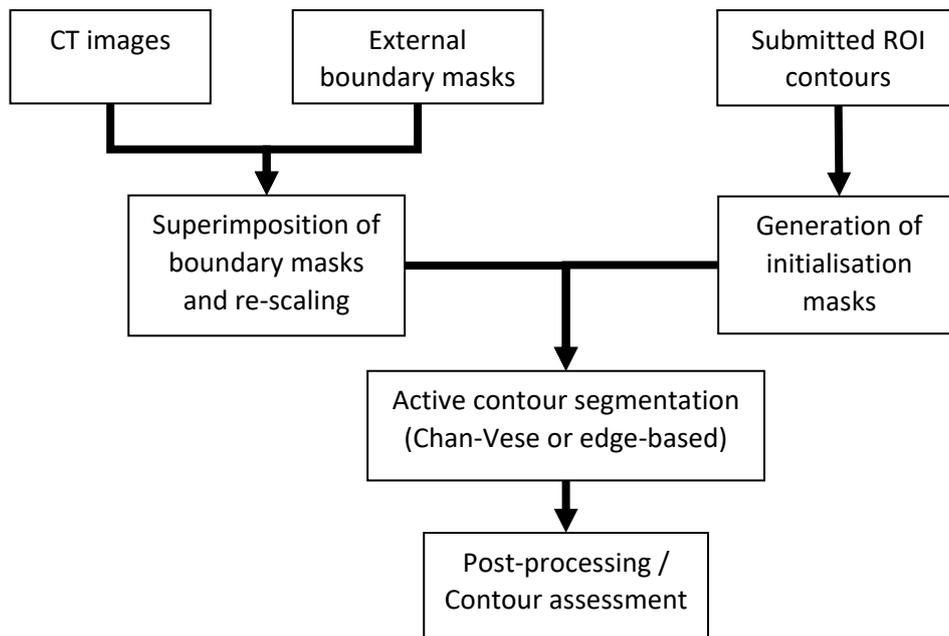


Figure 3.12. Workflow for active contour segmentation.

Unlike the watershed and graph-cut workflows, the internal masks were not used for both methods of active contour segmentation. Instead, these algorithms required further input in the form of initialisation masks which act as the initial splines from which the algorithms evolve.

3.9.1 Pre-processing for active contour

Further steps were undertaken in the pre-processing of the CT images for the active contour technique. The external boundary masks described in section 3.6.2 were used to set the limits of the segmentation. In addition to this, the upper threshold equivalent to 158 HU was also applied. In order to do this, pixels of the CT image covered by the external boundary mask were set to 0 intensity, and a further re-scaling was performed to stretch the intensity between 0 and 100. This was chosen rather than between 0 and 1, as the active contour function did not work well with small decimal numerals. To allow the re-scaling to be reflective of the HUs, it was decided not to use the lower threshold as a bounding limit, as the adaptive lower threshold levels resulted in different lower thresholds between cases.

This method of the image pre-processing was decided after multiple unsuccessful attempts at the application of the active contour algorithm on the CT images de novo. Without a boundary region and without further re-scaling, the evolution of the contours was very unstable, where a slight change in parameter settings (see number of iterations, contraction bias (CB) and smoothing factor (SF) below) resulted in a drastic change of the generated contour e.g. contours were seen to conform to a tissue surface at a particular setting, but a minor change in contraction bias of 0.01 resulted in the shrinking and collapsing of the contour to a point.

3.9.2 Generation of initialisation masks

A preliminary evaluation was performed using three cases within the 18 subsample cases to help decide on the most appropriate way of generating the initial masks. This was important as

the active contour models are sensitive to the placement of the initial mask, and they work better when the initial mask is placed close to the object boundary.

It was decided that the submitted contours should not be used directly for initialisation, to minimise overfitting to the submitted delineation, where differences in delineation would potentially affect the segmentation results. A preliminary evaluation of three different means of initial mask placement was performed. These included a) the smallest convex polygon containing the contour, b) the smallest rectangle encompassing the contour, c) a circle with an equivalent area to the contour centred on the centroid of the contour. The initial masks were generated as follows.

i. Smallest convex polygon encompassing the contour

The convex hull property of `regionprops` were used to derive the coordinates of the smallest convex hull encompassing the contour, which were used to generate the initial mask.

ii. Smallest rectangle bounding the contour

Similar to the creation of the smallest convex polygon, the coordinates of the bounding region of the contours were obtained using the bounding box property of `regionprops` to construct the binary mask.

iii. Circle with an equivalent area to the contour centred on the centroid of the contour.

The number of pixels for each GTV contour in the axial plane was calculated, which was then used to find the radius of the equivalent circle using the formula

$$\text{Area of circle} = \pi r^2 \quad [3.7]$$

where r denotes the radius of the circle. The centroid of the GTV was located through the coordinates obtained from the centroid property of `regionprops`. The coordinates were also used to generate a mask with a labelled point at the centre of the GTV.

At 300 iterations, it was noticed that the second method i.e. using the smallest rectangle bounding the contour performed the worse, as depicted in figure 3.13b. This was because of the greater distance between the actual GTV boundary and the edge of the initial mask, which resulted in leakage of the segmentation at the same parameter settings compared to the other initialisations.

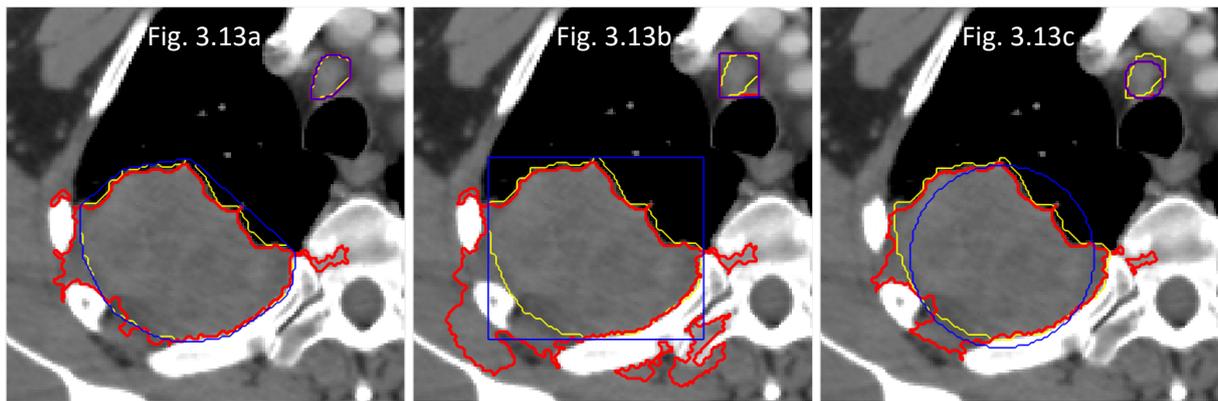


Figure 3.13. Comparison of Chan-Vese active contour segmentation in relation to different initial masks a) Smallest convex hull bounding the submitted contour; b) Smallest rectangle bounding the submitted contour c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask. (Parameter settings – Iteration number: 300, Contraction bias: 0.7, Smoothing factor: 0)

In most parts, the convex polygon and the circular initial masks created similar contours, although the circle masks seemed to result in greater leakage than the convex polygon. On the other hand, due to the way in which the convex polygons were constructed, there were occasions where the initial masks were of the same shape and size as that of the submitted contours, which also raised the concern of overfitting to the shape of the submitted delineation. This effect can be seen for the mediastinal lymph node in figure 3.13a, where there was an overlap of the submitted contour, the initial mask, as well as the resultant segmentation.

Because of the poorer performance of the bounding rectangle initiation mask in relation to the other two initialisations, the bounding rectangle method was not explored further. Additionally, the use of the convex polygon de novo was also not brought forward due to the possibility of overfitting.

Conventionally, the initial mask used in active contour is usually placed exterior to the object boundary, which then shrink during its evolution to conform to the object. This would normally work in the case of a homogenous background with a foreground object of different intensity values. In the case of thoracic CTs, the background is heterogenous as it is made up of different tissue types. Thus, it would be more appropriate to define the initial mask within the tumour rather than exterior to it, to prevent the evolution of the contour from conforming to the other tissue types rather than the tumour. This effect was noticed in the preliminary studies where an exterior initial mask resulted in the segmentation of the normal tissues around the tumour instead of the lesion. Although it was decided that it would not be appropriate to use the convex polygon de novo, it was felt that overfitting would be less of an issue if an erosion of the convex polygon was used.

Based on this work, it was decided that the initial masks to be explored included

- i. Smallest convex polygon encompassing the contour eroded isotropically by an empirical 4 pixels
- ii. Circle with an equivalent area to the contour centred on the centroid of the contour eroded isotropically by an empirical 4 pixels

- iii. Circle with an equivalent area to the contour centred on the centroid of the contour

For all of the mask creations, an additional check was implemented to ensure that the construction of the mask did result in a mask being created. This may not be the case for the first two methods of mask creation where morphological erosion was applied. If so, the centroid was used as the initiation point as a surrogate.

3.9.3 Multiple contours on same axial slice

As it was uncertain as to how the evolution of the active contour models would be influenced by processing multiple contours on the same axial slice, these regions were segregated and handled separately by creating separate initial masks for each of the region. This was also taken into account for the pre-processing of the CT images, which were prepared ensuring that other regions exterior to the region in question was masked out during the computation.

3.9.4 Post-processing of active contour segmentation

Post-processing steps were undertaken due to the potential creation of discrete segmented regions from the splitting of the contour during the evolution. As small regions were deemed not to be significant, areas with pixel sizes of less than 8 were removed. Additionally, 'holes' created in the evolution were also filled morphologically. The ensuing segmentation was defined as the final active contour segmentation result.

For the assessment of the segmentation, the delineation results for the multiple contours on the same axial plane were merged onto the same slice. This not only allowed ease and better judgement of the segmentation visually, but also enabled fair comparison of segmentation results across the different techniques.

3.9.5 Parameter settings

There were a number of parameters that were to be specified to optimise the segmentation for both the Chan-Vese and edge-based active contour algorithms.

3.9.5.1 Number of iterations

This determined the maximum number of iterations for the progression of the segmentation. The evolution would be stopped either when this maximum number of iterations is attained, or when the contour position in the latest evolution is the same as the contour position in the previous five iterations.

During the preliminary evaluation, it was noted that for some runs, the number of iterations required to reach a steady contour was as low as 20. However, it was also seen that for other slices, this was as high as several hundred. Considering that the default number of iterations in MATLAB is set at 100, and that it is recommended that the number of iterations should be increased with greater distance of the boundary object from the initial contour position, it was decided empirically that the maximum iteration number would be set at 300 for all experiments.

3.9.5.2 Contraction bias

This parameter adjusted the tendency of the contour to either grow in the exterior direction, or to contract and shrink inwards, where negative values biased the contour to expand and positive values biased the contour to contract. In spite of the specification, the growth or contraction of the contour was not guaranteed. However, the selection allowed a propensity for the contour

to evolve towards the specified direction, and if the tendency was in the opposite direction to what was specified, it would slow the progression of the contour in the opposite direction. This meant that the contour would evolve based largely on the image intensity values, but yet allowed for some influence in the nature of the evolution, which was helpful especially in regard to the different locations of the initial mask.

3.9.5.3 Smoothing factor

The smoothing parameter determined the regularity of the resultant segmentation boundary, which was specified with a minimum of 0, whilst a higher value produced smoother region boundaries. Although limiting the regularity of the evolution can be helpful in reducing the tendency of the contour to leak out into regions beyond the GTV in an irregular fashion, if the chosen value is too high, it may smooth out the curve based on the finer details of the intensity values.

3.9.6 Experiments

The active contour segmentation process was conducted separately for Chan-Vese and edge-based approaches. The training data was evaluated to select the most appropriate initialisation mask and parameter settings.

3.9.6.1 Selection of initialisation mask

The evaluation of the initialisation mask was conducted using the 18 subsample cohort, where the behaviour of how the initialisation masks affected the segmentation was explored. From this, the most appropriate initialisation method was selected. The impact of parameter tuning was also analysed with this dataset, prior to the application of the segmentation to the rest of the training data.

3.9.6.2 Parameter optimisation

The optimum contraction bias and smoothing factor were determined using a two-phase parameter sweep for the different initialisations. The first phase was performed on three cases within the 18 subsample cases, which comprised of a manual systematic search over a wider range of values, with an aim to narrowing the range for the parameter sweep in the subsequent analyses. This was determined separately for each of the parameters, for both Chan-Vese and edge-based approaches.

After evaluating the best range of parameters in phase one, subsequent parameter sweeps were conducted using the narrower range identified. This involved looping the segmentation through the parameters with an increase in values with a step-wise increment.

Chan-Vese active contour

i) Contraction bias

For the Chan-Vese model, phase one was performed with contraction bias ranging between -1 and 1.5, with the number of iterations fixed at 300 and the smoothing factor fixed at the default value of 0. Good segmentation results were seen around the value of 0.7 for both the circle and convex hull initiation masks. As a result, subsequent parameter sweeps were performed at a narrower range between 0.4 and 1.1 at 0.05 step increments, resulting in a total of 15 different parameters that were analysed.

ii) Smoothing factor

The manual parameter sweep for the smoothing factor in phase one for the Chan-Vese model was carried out using 300 number of iterations, and a contraction bias fixed at 0.7. A range between 0 and 2 was evaluated, and a narrower range of 0 and 0.7 at 0.1 step increments was chosen for the rest of the parameter sweeps, resulting in a total of 7 different parameters that were analysed.

Edge-based active contour

i) Contraction bias

The edge-based active contour segmentation was initially evaluated with a contraction bias between -1 and 1, at 300 iterations with a smoothing factor of 0. Better segmentation results were observed around a contraction bias of 0. Thus, a final range between -0.2 and 0.075 at 0.025 step increments was chosen, where a total of 12 different parameters were evaluated.

ii) Smoothing factor

After the initial parameter sweep using a contraction bias of 0 and at 300 iterations, smoothing factors between a range of 0 and 0.8 at 0.2 step increments was selected for the subsequent analyses, i.e. a total of 5 different smoothing factors were evaluated.

3.9.6.3 Selection of optimal parameters

Using the three folds of the training data, the performance of the segmentation was assessed. The optimal parameters were selected for the respective algorithms and applied on the validation data to estimate the performance of the segmentation.

3.10 Graph-cut segmentation workflow

The segmentation process for graph-cut technique is summarised in figure 3.14.

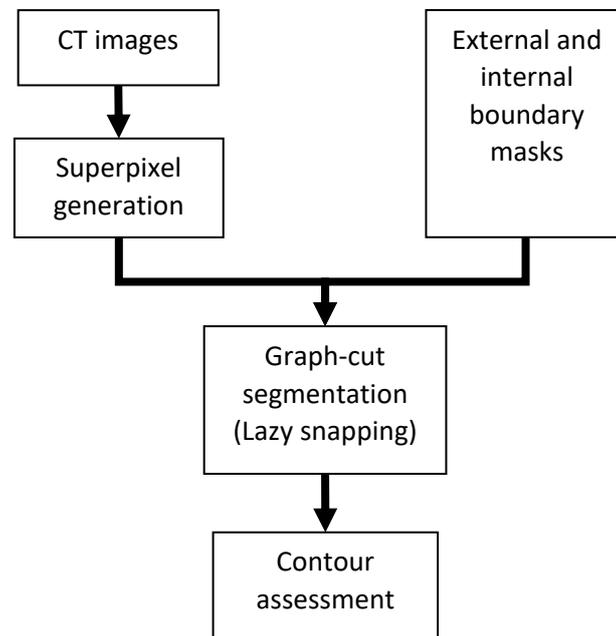


Figure 3.14. Workflow for graph-cut segmentation.

3.10.1 Graph-cut segmentation

Suppose an image is a graph $G = (V, E)$ consisting of a set of nodes (V) and a set of edges (E) that connects each node. The image set V is made up of individual pixel elements v_i . To segment the image, the aim of the algorithm is to assign a unique label to each node into foreground or background. To find the solution, the minimisation of a Gibbs energy $E(f)$ (the energy cost for neighbouring pixels) can be obtained (389):

$$E(f) = \sum_{i \in V} E_1(x_i) + \lambda \sum_{(i,j) \in E} E_2(x_i, x_j) \quad [3.8]$$

where E_1 is the likelihood energy of encoding the cost when the label of the node i is x_i , and $E_2(x_i, x_j)$ is the prior energy, representing the cost when the labels of adjacent nodes, i and j , are x_i and x_j respectively (390).

In the original proposal by Li et al, k -means clustering on the values of the foreground and background seeds is performed and initialised into 64 clusters. In the calculation of the likelihood energy, the minimum distance of the value of each node i to the mean value of the foreground and background clusters is computed, with all nodes in the image used in the optimisation process (390, 391). The difference in the MATLAB implementation is that this step is performed without the clustering into similar foreground or background pixels, which may thereby affect the computational performance.

The prior energy E_2 is the energy of the gradient along the object boundary, which ensures a penalty cost if the adjacent nodes are assigned with different labels. The closer the values of adjacent nodes are, the larger the E_2 , and the less likely the edge is on the object boundary.

Additionally, the likelihood of labelling of a pixel into the foreground or background can be adjusted through specification of the edge weight scale factor.

3.10.2 Graph-cut application using lazysnapping

The implementation of graph-based segmentation in MATLAB, `lazysnapping`, segments an image into binary labels (390). Like many of the other interactive graph-cut algorithms, user input is required for the initial definition of foreground and background. To do this automatically, the same internal and external masks (described in sections 3.6.2 and 3.6.4) were applied as the foreground and background markers respectively, which also permitted the comparison of how the different techniques performed.

3.10.3 SLIC algorithm

To improve the computational efficiency of the segmentation process, oversegmentation of the image was achieved with the generation of superpixels.

The MATLAB implementation, `superpixel`, is based on the simple linear iterative clustering (SLIC) algorithm, which uses an adaptation of k -means clustering for generating the superpixels (392). SLIC has been shown to be efficient with good performance as compared to other algorithms (371-374). In its optimisation process, the number of distance calculations is reduced significantly by limiting the search space to a region proportional to the superpixel size. It also has a weighted distance measure that provides control over the size and compactness of superpixels, whilst combining information on spatial proximity, and grayscale level or if relevant, colour.

The main parameter which needs specifying is k , the desired number of superpixels in the image. From this parameter, cluster centres are initiated in a regular grid, S pixels apart. This is given by the equation:

$$S = \sqrt{\frac{N}{k}}$$

[3.9]

where N is the total number of pixels in the image, where approximately equally sized superpixels are produced on the image. Next, within a 3×3 -pixel neighbourhood, the cluster centres are moved to seed locations corresponding to the lowest gradient position, to avoid seeding on an edge boundary as well as to reduce the impact of noise on the clustering initiation.

Subsequently, within a limited search space of $2S \times 2S$ around the seed points, each pixel i is evaluated against the superpixel centre determined through a distance measure, which combines both spatial as well as grayscale information. The spatial proximity to the cluster centre is computed through:

$$d_s = \sqrt{(x_j - x_i)^2 + (y_i - y_j)^2}$$

[3.10]

where $[x, y]$ indicates the pixel location within the image. The grayscale proximity to the cluster centre is similarly computed by:

$$d_c = \sqrt{(l_j - l_i)^2}$$

[3.11]

where l represents the grayscale intensity of the pixel. Both of these calculations are combined using the equation below:

$$d_s = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2}$$

[3.12]

where m , the compactness factor, is a constant that can be adjusted to increase or decrease the weight of the spatial proximity in relation to the grayscale calculation. When m is large, there is greater importance on the spatial proximity, resulting in superpixels that are more compact. By decreasing m , the superpixels are less compact, with greater irregularity to the size and shape, the boundaries of which adhere more to the edges within the image.

As each pixel has been associated to the nearest cluster centre, a calculation for the mean vector of all the pixels assigned to that particular cluster centre is performed, as well as the residual error between the new and previous cluster centre locations. The process is repeated iteratively until the error converges. A further post-processing step is then executed to connect disjoint pixels to neighbouring superpixels.

3.10.4 Adaptive SLIC

The SLIC algorithm described above has a relatively simple distance measure, which uses a constant to modify the impact of the spatial information for all iterations. SLIC0 is a variation of the SLIC algorithm which has an adaptive component for using the spatial information as it performs its iterations. This means that it has the ability to alter the compactness with each superpixel computation according to the maximum observed spatial and grayscale distances from the previous run. Because of the ability of the algorithm to work out the most appropriate compactness dynamically, there is no requirement for selection of the compactness parameter upfront for SLIC0.

3.10.5 Superpixel parameter optimisation

3.10.5.1 SLIC versus SLIC0

The SLIC algorithm allows a choice in specifying the shape of the superpixel. This is the compactness parameter, where a higher value would make the superpixels more regularly spaced, i.e. a square. Conversely, a lower value would produce superpixels that are more

irregularly shaped, which can adhere to boundaries better. For SLIC, this compactness parameter is kept constant for the clustering process.

On the other hand, parameter selection is not required for SLIC0 as the compactness is refined adaptively after the first iteration.

Both of these algorithms were explored in the training dataset.

3.10.5.2 Number of iterations

For both SLIC and SLIC0 algorithms, the number of iterations used in the clustering phase of the algorithm can also be modified. The authors of the algorithm have found that 10 iterations is sufficient for most purposes (371, 392). This was the default setting in MATLAB, where it was recommended that this factor need not be adjusted in most cases, and thus, the default value of 10 was used for this work.

3.10.5.3 Desired number of superpixels

The desired number of superpixels had to be specified to initiate the algorithm. Reports in the literature suggest that the quality of superpixel generation improves with increasing number of specified regions (368, 371-374, 393, 394). However, it would be preferable to choose the smallest number of superpixels that is required for the segmentation, in order to derive the computational benefits of working with a smaller number of regions. For non-medical imaging, typical values for this in the region of several hundreds.

In order to establish the optimal number of superpixels to use, a parameter sweep was performed.

3.10.6 Assessment of superpixel generation

A number of different metrics have been used to assess the quality of superpixel generation. For this work, boundary recall was calculated, which gave an assessment of how accurate the superpixel boundaries were, in relation to the ground truth. The fraction of ground truth edges that fell within a certain number of pixels, d_{br} , of the superpixel boundary was computed through the equation:

$$\text{Boundary recall} = \frac{TP}{TP + FN} \quad [3.13]$$

where TP represents true positives i.e. the number of boundary pixels in the ground truth image for which a boundary pixel in the superpixel image exists within a range of d_{br} , and FN represents false negatives i.e. the number of boundary pixels in the ground truth image for which there is no boundary pixel in the superpixel image within a range of d_{br} . Values of the boundary recall range from 0 to 1, with 1 being the best score.

Additionally, the undersegmentation error was also computed, which provided another measure to assess the ‘bleeding out’ of the superpixels in relation to the GTV (395). In other words, this penalised superpixels that did not conform to the reference boundaries. This measure has been adopted in a number of other studies evaluating superpixels (374, 392, 394). The undersegmentation error was calculated with

$$\text{Undersegmentation error} = \sum_{V_{GT}} \frac{\sum_{P:P \cap V_{GT} \neq \emptyset} |P_{out}|}{|V_{GT}|} \quad [3.14]$$

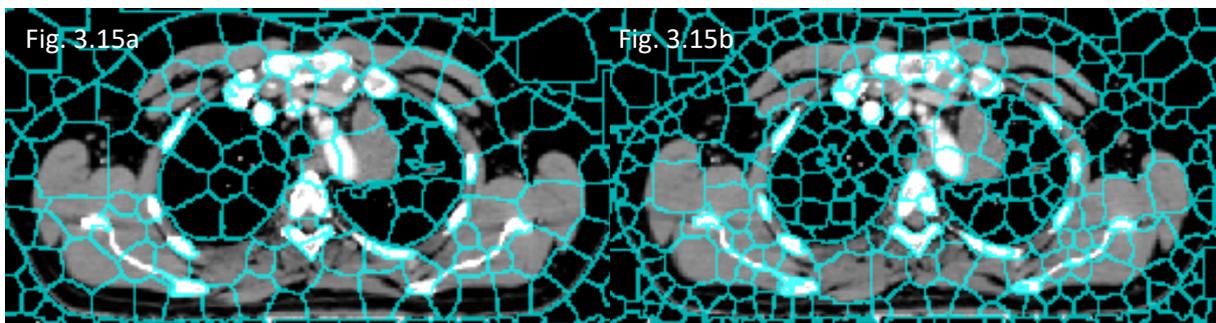
where V_{GT} represents the ground truth region and P represents the individual superpixels. The expression $|V_{GT}|$ is the size of the ground truth in number of pixels, and $|P_{out}|$ is the size of the region of the individual superpixel that is not overlapping with the ground truth in number of pixels. With lower undersegmentation error, there is less leakage of the superpixels into the surrounding regions of the superpixels that is overlapping with ground truth.

3.10.7 Experiments

Many of the studies using SLIC to generate superpixels have a similar design, which include a training phase to select the use of suitable parameters (383, 396-403). Fewer studies use more sophisticated approaches to help with the optimisation process (384). To proceed with this work, there was a need to evaluate the factors that could affect the generation of superpixels and assess the impact of the parameters k and m on the desired outcome, prior to application of graph-cut segmentation.

3.10.7.1 Preliminary development phase

Exploration of the graph-cut application and the evaluation process was performed using the same subsample of 18 cases. In the preliminary development phase, the same three cases were selected within the subsample to initiate the graph-cut workflow, where it was apparent that less than a thousand superpixels was not sufficient for this work, which suffered from poor adherence to the boundaries in thoracic images (figure 3.15). At this range, there was low boundary recall and high undersegmentation error. Not only was this seen in the mediastinum in the presence of small vascular structures and nodal disease, it was also observed for the larger tumour in the lung parenchyma. Thus, in the subsequent work, the range for the number of desired regions was evaluated between 1000 to 19000, at increments of 2000. The upper limit was initially set at 30000 but this was decreased for two reasons; a) there was significant computational expense, where the processing time for the evaluation of each case took several days and b) the greater the number of regions, the closer it would be towards computing on a pixel-wise basis, which defeats the purpose of superpixels.



Figures 3.15a – b. Application of SLIC0 algorithm for $k = 500$ and $k = 1000$ respectively (10 iterations). Aqua lines denote superpixel boundaries. Both examples show poor adherence of the superpixels to the medial boundary of the tumour, as well as the mediastinal structures.

Both SLIC and SLIC0 were also evaluated. In keeping with the reported literature, the increasing regularity of the superpixels was also observed with increasing compactness for the SLIC algorithm (figure 3.16). This effect was more apparent for compactness at 15 and beyond, which was felt to be undesirable for this work. Thus, the compactness for SLIC was evaluated at 5, 10 and 12 for the proceeding evaluation.

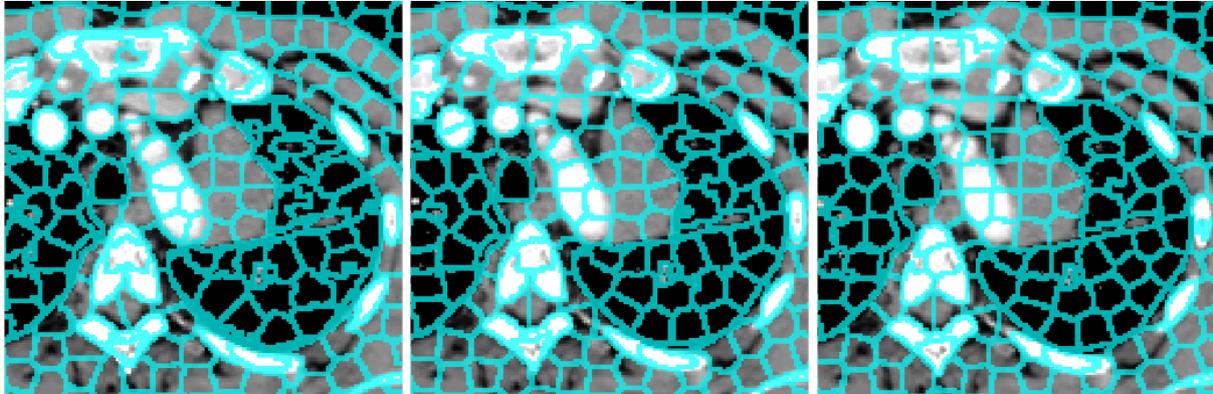


Fig. 3.16a

Fig. 3.16b

Fig. 3.16c

Figures 3.16a – c. Application of SLIC algorithm for $m = 10$, $m = 15$ and $m = 20$ respectively ($k = 2000$, 10 iterations). Aqua lines denote superpixel boundaries. The regularity of the superpixels increased with increasing m , which can be seen in the lung parenchyma as well as in the chest wall.

The boundary recall was calculated for $d_{br} = 1$ pixel, to account for some variation in the ground truth segmentation whilst maintaining a score that is reflective of the accuracy to the ground truth. Although $d_{br} = 2$ pixels have been used in the literature for a number of studies on superpixels (371, 373, 393), the evaluation of this on the 18 subsample cases revealed that a 2-pixel boundary was not sensitive enough to detect acceptable boundary limits. $d_{br} = 0$ pixel was also not used, in order to allow for a margin of error in the placement of the reference volumes.

3.10.7.2 Training: Parameter selection

The training phase was divided into two sequential parts to decide on the parameter settings for the segmentation. Firstly, the parameter settings for generating the superpixels was evaluated, which was performed initially using the 18 subsample cases, and then extended to the three cross-validation training runs. This was followed by assessment of the optimal settings for the lazysnapping algorithm, which was applied to the cross-validation training runs.

3.10.7.3 Superpixel parameter selection

The training data was evaluated to decide on the desired number of superpixels, as well as the compactness for SLIC algorithm. The number of iterations was set at 10 for all evaluation. The boundary recall as well as the undersegmentation error were computed. Additionally, to assess the impact on how the different superpixel parameter settings affected the subsequent graph-cut segmentation, it was applied to each setting and the Dice similarity coefficient (DSC), recall and precision were also computed (described further in section 3.11). Furthermore, as the computational time was also an important consideration in the selection of the optimum number of superpixels, this was also evaluated at each parameter setting. For this work, the edge weight scale factor for the lazysnapping algorithm was set at the default of 500, with connectivity of connected components set at 8.

3.10.7.4 Lazysnapping parameter selection

After the parameters for superpixel generation were established, these settings were used in the evaluation of the graph-cut segmentation. The connectivity of the connected components for the lazysnapping algorithm was set at the default of 8, which was appropriate in the setting of 2D images where the pixels were evaluated against their neighbours in orthogonal and diagonal directions.

The other parameter to be adjusted for the algorithm is the scale factor that affected the edge weights between the regions to be segmented, typically ranging between 10 and 1000. Preliminary observations on the 18 subsample cases showed greater differences in segmentation at lower values. Thus, a range of values at irregularly spaced intervals were evaluated, with a focus on the lower range. The selected values were 1, 5, 10, 20 to 100 at intervals of 20, and 100 to 900 at intervals of 100, giving a total of 16 values.

3.10.7.5 Selection of optimal parameters

Similar to the approach for the active contour workflow, the performance of the segmentation was assessed with the three training folds. The optimal scale factor was selected and applied on the validation data to estimate the performance of the segmentation.

3.11 Assessment of segmentation

The segmentation results for the different techniques were conducted quantitatively with the computation of conformity indices, and qualitatively with the visual inspection of the contours.

3.11.1 Quantitative assessment

There are many different means for assessing the segmentation results in the literature. It was decided that conformity indices would be most appropriate for this work, as it is extensively used in the literature as an assessment technique and was easy to compute. The calculations were performed for all slices in each case and aggregated across all the cases from which the mean and SD were derived.

3.11.1.1 Indices for algorithm development and parameter optimisation (training dataset)

For the assessment of contours in the algorithm development and parameter optimisation phases, recall, precision as well as the DSC were used. These measures were chosen as they are commonly reported conformity indices in image segmentation.

Recall, also known as sensitivity or the true positive rate, measures the proportion of positives that are correctly identified as such. This was defined by the equation

$$Recall = \frac{TP}{TP + FN} = \frac{|V_{seg} \cap V_{GT}|}{|V_{GT}|} \quad [3.15]$$

where TP and FN denote true positives and false negatives respectively, and V_{seg} and V_{GT} represent the segmented and ground truth regions respectively. For good recall, the score should be close to 1.

Precision, also known as the positive predictive value, measures the ability of the segmentation to identify only the true positive region, defined by the equation

$$Precision = \frac{TP}{TP + FP} = \frac{|V_{seg} \cap V_{GT}|}{|V_{seg}|} \quad [3.16]$$

where FP denotes false positives. For good precision, the score should be close to 1.

Typically, recall and precision are often in conflict, thereby raising the need to assess based on another measure that takes both into account. This can be computed through the F-measure, which is the harmonic mean of precision and recall (404). In the special case where the weights of precision and recall are balanced, the F-measure is equivalent to the Dice similarity coefficient, given by the equation

$$DSC = \frac{2|V_{seg} \cap V_{GT}|}{|V_{seg}| + |V_{GT}|} \quad [3.17]$$

where the extent of spatial overlap can be estimated. For a good DSC score, the value should be close to 1. This is also known as the Kappa index in some studies.

Although recall and precision are helpful in the assessment of the segmentation, the DSC is a single measure that takes both recall and precision into account. Therefore, the best DSC score achieved was used to select the optimal parameter setting where appropriate (active contour and graph-cut segmentation), whilst taking into account both recall and precision.

3.11.1.2 Indices for evaluation of testing dataset

DSC, as described above, was also used as a metric to assess the final segmentation in the testing dataset. However, although precision and recall are widely used in the literature in the assessment of image segmentation, they are less commonly used in the reports in the clinical literature (405-408). Therefore, in place of recall and precision, other more commonly used conformity indices were also used to assess the performance of the segmentation for the independent testing dataset in addition to DSC. These indices are analogous to recall and precision through a very simple relationship as described below.

The extent of segmentation that missed regions of true disease was represented by the geographical miss index (GMI) defined as

$$GMI = \frac{|V_{GT}| - |V_{seg} \cap V_{GT}|}{V_{GT}} \quad [3.18]$$

. The GMI is in fact 1 – the recall score.

The extent of spillage of segmentation, the discordance index (DI) was also calculated as follows.

$$DI = 1 - \frac{|V_{seg} \cap V_{GT}|}{|V_{seg}|} \quad [3.19]$$

Similarly, DI is the equivalent of 1 – the precision score.

Thus, good GMI and DI should be as close to 0 as possible.

These metrics would be helpful adjuncts to DSC which assesses how conformal the segmentation results are in relation to the reference volume, through the provision of further information on where the segmentation discrepancies lie.

3.11.1.3 Volume assessment

Additionally, the volumes of the final segmentation for each of the techniques were obtained in the independent testing phase. This was performed by multiplying the number of pixels within the contour in the axial plane with the individual voxel size, and then summing the volume for each slice to obtain the total volume for the case. The reference contours were processed in a similar way.

The absolute volume difference between the segmented results and the reference outlines were computed. However, due to the variation in sizes of tumour across the cases, the percentage volume difference was also calculated, to allow the assessment to be independent of the tumour

size. This was performed by computing the ratio of the segmented results and reference outlines subtracted by the reference volumes. Increasing positive values indicate larger segmentation in relation to the reference volumes, whilst greater negative values would be obtained for smaller segmentation volumes compared to the reference outlines.

3.11.2 Qualitative assessment

MATLAB has an inbuilt function `imshow` that can be used to display the CT image, and boundary plots can be plotted onto the image for visualisation. However, this proved to be an inefficient means of assessing the quality of the segmentation across different slices within a case. Therefore, a graphical user interface (GUI) was developed in order to allow the visualisation of the segmentation results to be performed efficiently (figure 3.17). It was designed to display the segmentation results adjacent to the reference contours in the same axial plane with information on the slice number. A slider function was implemented to enable scrolling capability through the slices of the images. In addition to the magnification and pan functions, the window levels for the images could also be adjusted appropriately for both lung and mediastinal window levels.

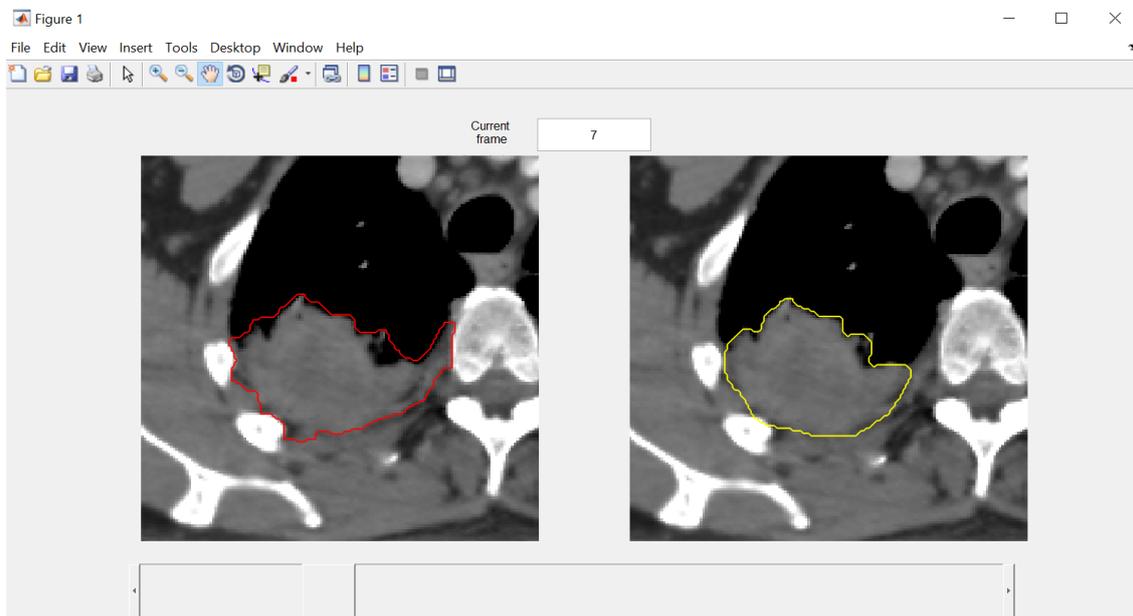


Figure 3.17. GUI used for the qualitative assessment of the segmentation results, showing a segmentation result on the left, and the reference contours on the right. In addition to inbuilt magnification and pan functionalities, windowing levels could also be adjusted. The slider function enabled efficient visualisation of serial CT slices.

3.11.3 Computational time

Where relevant in the algorithm development stage, the computational time for the techniques was also recorded. The development of the watershed segmentation was carried out using an Intel Core i5-3317U CPU @ 1.70GHz, 4GB RAM on a Windows 10 64-bit environment. For the development of the graph-cut segmentation, the processing was performed using an Intel Xeon CPU E3-1226 v3 @ 3.30GHz, 8GB RAM on a Windows 7 64-bit environment.

To allow for fair comparison of the computational time for the various segmentation techniques, the former machine was used to process all the cases in the independent testing dataset.

Results

The results of the algorithm development are presented for the subsample of 18 cases for each of the different segmentation techniques. This is followed by the description of parameter selection for each of the cross-validation runs of the training dataset and a display of representative cases. The estimated performance from each of the validation datasets in the cross-validation runs is then reported.

The combined results of the independent test data with the different segmentation approaches is presented in chapter 4.

Mean \pm SD are shown in the numeric results for the tables unless otherwise stated.

3.12 Task B.1 Development and tuning of segmentation techniques on training dataset

3.12.1 Watershed segmentation

3.12.1.1 Synthetic geometrical shapes

3.12.1.1.1 Control experiments

Table 3.1 shows the mean conformity indices of the 9 circles of the control experiment performed with the original synthetic image using the different gradient computation. It can be seen that the performance for the algorithm using the intermediate gradient operator was poor as compared to the other methods.

	Roberts	Sobel	Prewitt	Central	Intermediate
Dice Similarity Coefficient	0.94	0.94	0.94	0.92	0.54
Recall	1.00	1.00	1.00	1.00	1.00
Precision	0.89	0.88	0.88	0.85	0.41

Table 3.1. Mean conformity indices for the control synthetic images comparing different gradient computation.

To understand why this occurred, the segmentation for one of the circles using the intermediate gradient operator is shown in figure 3.18a, with the corresponding gradient magnitude in figure 3.18b. It can be seen that the failure of the segmentation occurred in the upper left and lower right directions, while the partitioning in the upper right and lower left directions appeared to correspond well to the gradient magnitude. On closer inspection of the gradient magnitude in the upper right and lower left directions, the pixels for the computed gradient did not have 4-pixel connectivity in the horizontal and vertical directions, whereas this was preserved in the upper right and lower left directions. This resulted in leakage of the watershed segmentation in the affected directions, which was not observed for the other methods of gradient computation.

To test if a 4-pixel connectivity correction would rectify the leakage, the gradient image was dilated using a 1-pixel radius disk-shaped structuring element, as shown in figure 3.18d. The corresponding final segmentation is shown in figure 3.18c, which corresponds to a DSC of 0.92, recall of 1.00 and precision of 0.85, comparable to the results obtained using the central gradient operator. This suggests that the watershed segmentation would not be accurate in the absence of pixel connectivity in the horizontal and vertical directions.

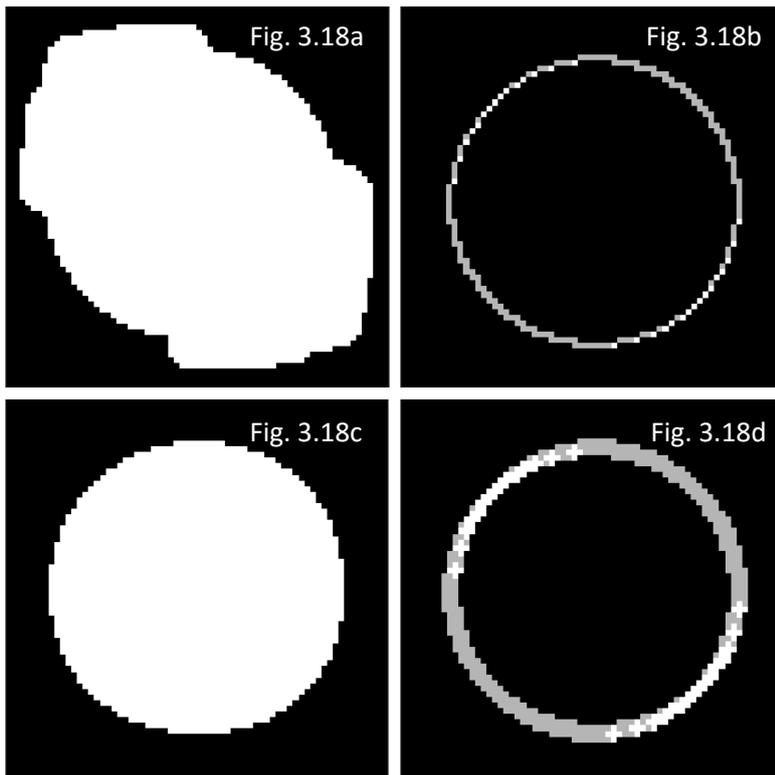


Figure 3.18. Watershed segmentation using intermediate gradient operator for 2000-pixel circle against background equivalent to lung parenchyma; a) Final results of watershed segmentation on uncorrected gradient magnitude image; b) Uncorrected gradient magnitude image; c) Final results of watershed segmentation on gradient magnitude image corrected for 4-pixel connectivity; d) Corrected gradient magnitude image.

3.12.1.1.2 Evaluation of ROI blurring and image noise

Overall, the mean DSC scores decrease with increasing ROI boundary blurring (figure 3.19a) and image noise (figure 3.19b) across the different types of gradient computation.

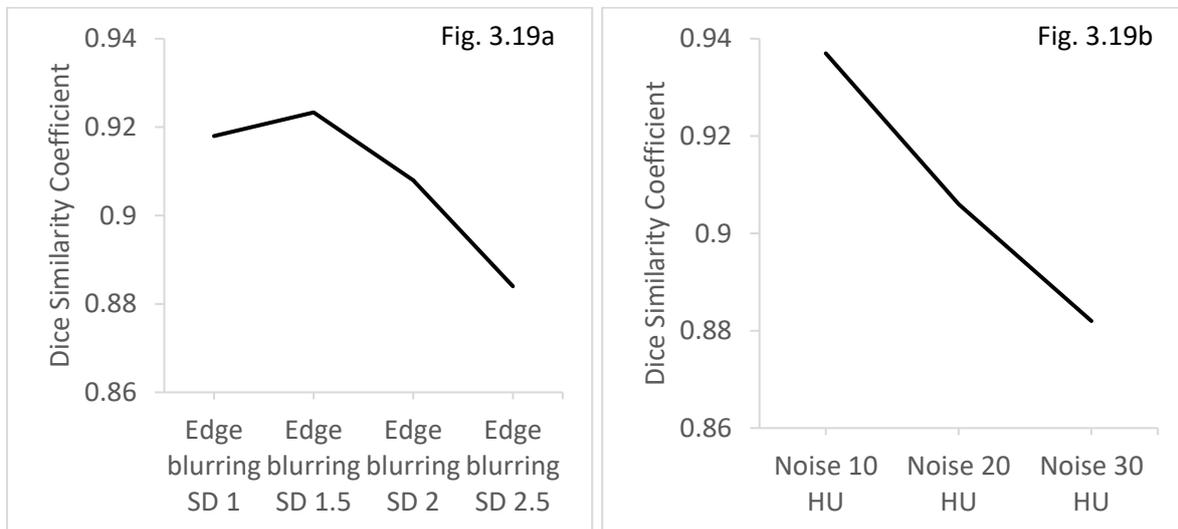


Figure 3.19. Effect of a) increasing ROI boundary blurring and b) increasing image noise, on mean Dice similarity coefficient using watershed segmentation.

The extent of this effect was different between different gradient computation, as seen in figures 3.20a – c. The Roberts, central and intermediate gradient operators appeared to be more influenced by variation of the ROI boundary blurring and image noise, whilst the Sobel and Prewitt operators are less affected by this.

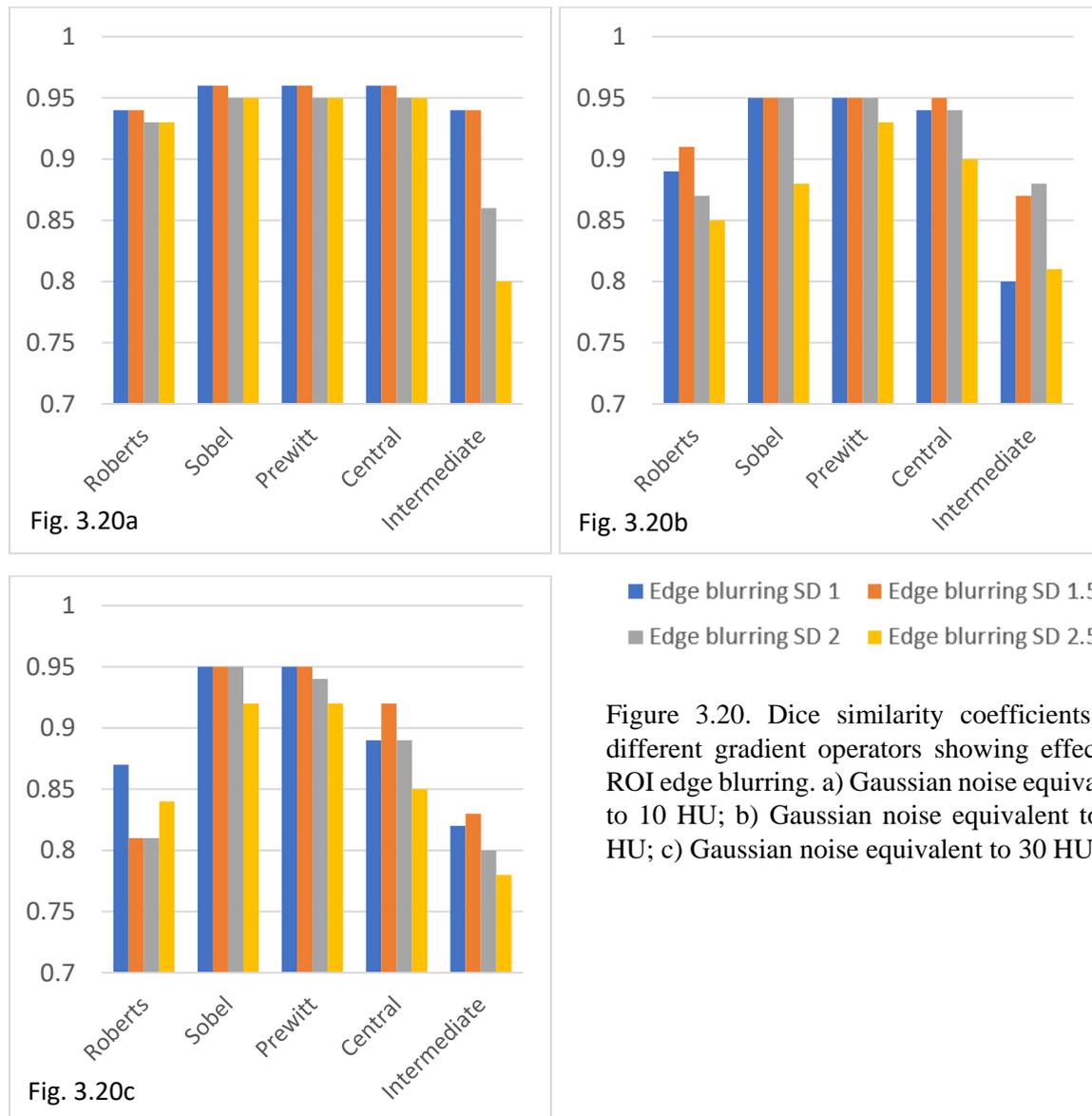


Figure 3.20. Dice similarity coefficients for different gradient operators showing effect of ROI edge blurring. a) Gaussian noise equivalent to 10 HU; b) Gaussian noise equivalent to 20 HU; c) Gaussian noise equivalent to 30 HU.

Of note, in the presence of image noise and the blurring of the ROI edge, the lack of 4-pixel connectivity was not observed when the intermediate gradient operator was used. In the absence of an abrupt change in HU between adjacent pixels where the change in gradient was spread over a number of pixels, 4-pixel connectivity was preserved for the gradient magnitude image.

Table 3.2 shows the mean and SD of the conformity indices for the different gradient computation, where it can be seen that the intermediate and Roberts operators are associated with a lower DSC as compared to the other gradient operators. There was a statistically significant difference in DSC between the gradient operators (p -value = 0.000; Friedman test). Post hoc analysis with Wilcoxon signed-rank tests was performed with a Bonferroni correction (significance level set at $p < 0.005$) revealed that there was no statistically significant difference

between the Sobel, Prewitt, and Central operators (Sobel vs Prewitt p -value = 0.655; Prewitt vs Central p -value = 0.018; Sobel vs Central p -value = 0.062). However, the DSC for the Roberts and Intermediate operators were found to be statistically significantly different to the Sobel, Prewitt and Central operators when compared individually (p -value = 0.002 for all six individual test comparisons).

Judging by the lower precision scores as compared to the recall values, it can be seen that the segmentation tended to produce larger errors in leakage rather than in missing the target, which was more apparent in the Roberts and intermediate operators. Evaluation of the precision also revealed a statistically significant difference in scores between the gradient operators (p -value = 0.000; Friedman test). Post hoc analysis with Wilcoxon signed-rank tests (Bonferroni correction applied using a significance level of $p < 0.005$) also showed no statistically significant difference between the Sobel and Prewitt operators (p -value = 1.000). Similar to the DSC evaluation, Roberts and Intermediate operators were found to be statistically significantly different to the Sobel, Prewitt and Central operators when compared individually (p -value = 0.002 for all six individual test comparisons). However, although the precision scores between the Sobel and central operators were found not to be statistically significant (p -value = 0.028), a statistically significant difference was detected in the comparison between the Prewitt and central computation (p -value = 0.004).

	Roberts	Sobel	Prewitt	Central	Intermediate
Dice Similarity Coefficient	0.88 ± 0.05	0.94 ± 0.02	0.95 ± 0.01	0.93 ± 0.04	0.84 ± 0.05
Recall	0.97 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	0.97 ± 0.02
Precision	0.83 ± 0.06	0.91 ± 0.02	0.91 ± 0.01	0.88 ± 0.04	0.78 ± 0.06

Table 3.2. Mean conformity indices comparing different gradient computation for images degraded by ROI boundary blurring and image noise.

The qualitative segmentation results using a noise level of 10 HU and edge blurring with an SD of 1.5 is shown in figure 3.21. It was observed that there were more irregularities in the segmentation against the mediastinal background as compared to lung and bone background, which was apparent across the different ROI sizes.

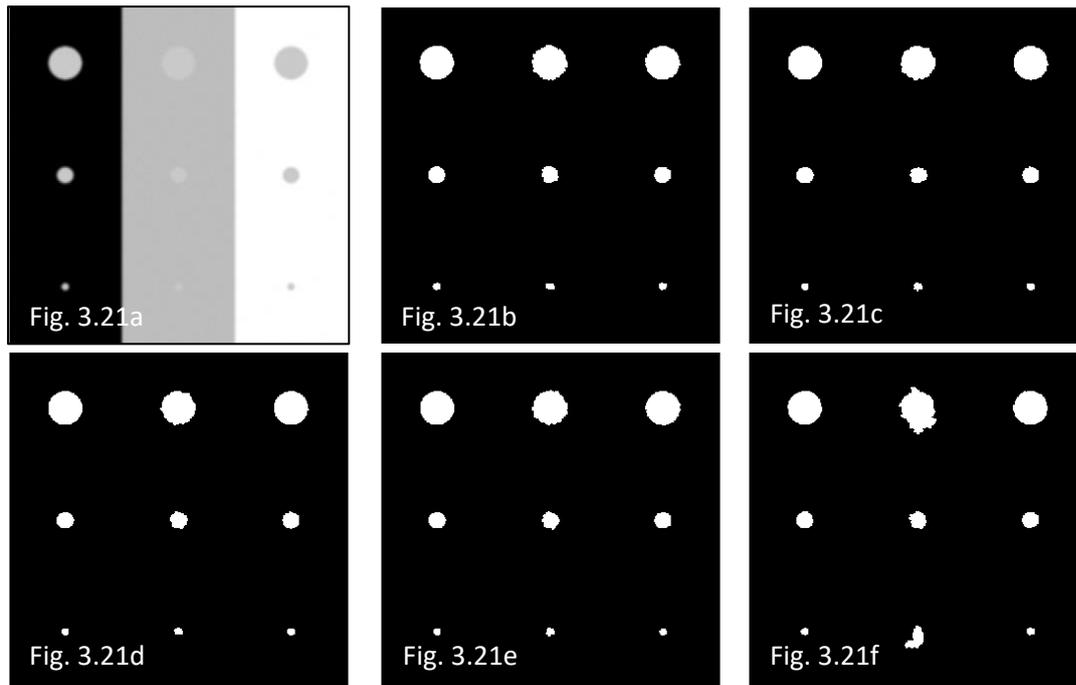


Figure 3.21. Segmentation results using degraded image with noise equivalent to 10 HU and ROI edge blurring with a SD of 1.5 (mean = 0). a) Degraded image; b) Roberts gradient operator; c) Sobel gradient operator; d) Prewitt gradient operator; e) Central difference gradient operator; f) Intermediate difference gradient operator.

From this results, it was concluded that the Roberts and intermediate operators had poorer performance than the Sobel and Prewitt operators when applied with the watershed segmentation in the presence of image degradation, whereby there was an increase in the leakage of the segmentation into the surrounding region. The central difference computation also had poorer performance than Sobel and Prewitt operators in terms of precision. Although the results may be an overestimate as a noise level of up to 30 HU was evaluated, the Sobel and Prewitt operators were favoured over the others as they were less influenced by noise. These two operators were therefore selected to be used in the subsequent evaluation.

Additionally, blurring of the ROI edges was shown to contribute to poorer segmentation performance, even in the presence of modest image noise equivalent to 10 HU. This observation supported the decision for not applying de-noising smoothing filters, which although can decrease the image noise, can result in increased blurring of the ROI edge.

3.12.1.2 Clinical experiments

3.12.1.2.1 Gradient computation for ROI boundary

The relationship of the HUs at the vicinity of the GTV boundary is shown in figure 3.22. It was observed that there was a greater spread in HU exterior to the tumour boundary in comparison to the HU within the tumour. There was one case where the HUs were significantly different to the rest of the cases, due to the presence of a large amount of GGO. However, all the cases appear to have the greatest change in HU values near the GTV boundary.

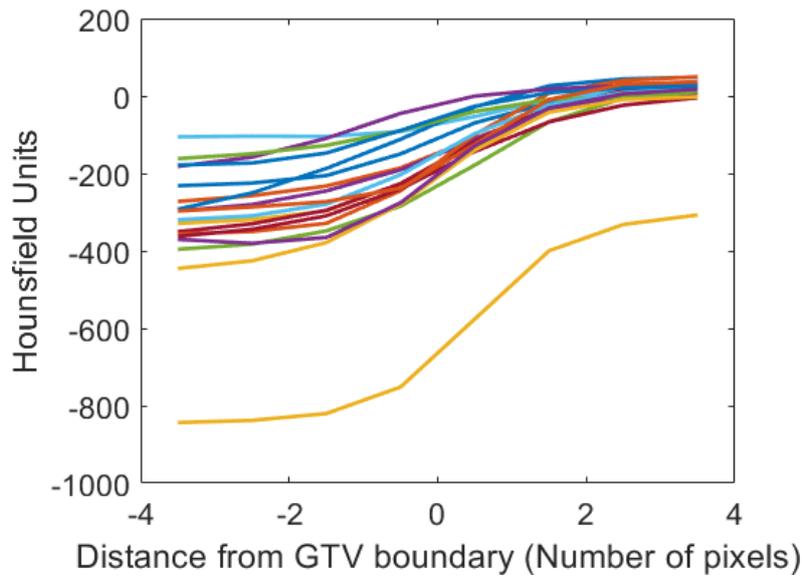


Figure 3.22. Hounsfield units at the vicinity of the GTV boundary. Each of the 18 subsample training cases is denoted by each individual plot. Positive values for the distance from GTV boundary represent the direction towards the centre of the GTV, negative values represent the direction away from the centre of the tumour.

The gradient magnitude computation was performed with the Prewitt and Sobel operators, where the mean gradient magnitude in relation to the distance from the GTV boundary for the 18 subsample training cases is shown in figure 3.23.

Both gradient operators showed very similar trends for the gradient magnitude, with higher gradient magnitude towards the GTV boundary. This supported the use of the watershed segmentation approach which locates the gradient magnitude at its maximum, within the search region.

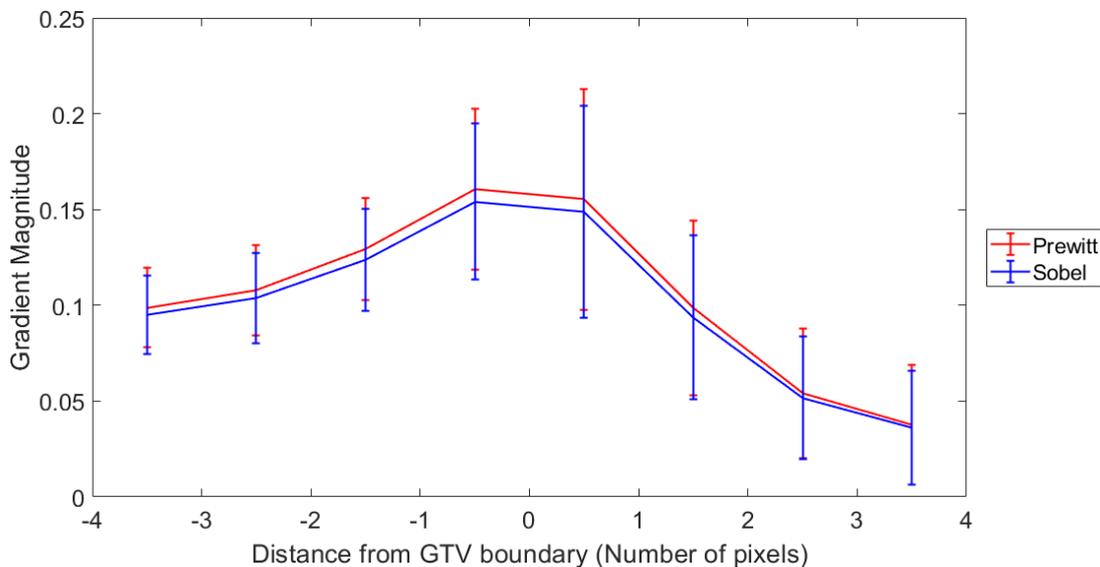


Figure 3.23. Mean gradient magnitude in relation to the distance from the GTV boundary as computed through the Prewitt and Sobel gradient operators, showing the mean of the 18 subsample training cases (error bars denote the SD).

Although the Prewitt operator produced values of gradient magnitude that were slightly higher than that of the Sobel operator, due to the similar trends and values obtained for both methods, it was unclear as to whether there would be an advantage in using one approach over the other. Results from the synthetic dataset suggested that there is no statistical differences between the two types of computation. This clinical dataset was therefore analysed to ascertain if there were any differences in the two methods.

3.12.1.3 Clinical experiments of watershed segmentation

The results of the training cohort using the Prewitt and Sobel operators is shown in table 3.3.

	Dice Similarity Coefficient	Recall	Precision	Time (seconds)
Prewitt	0.75 ± 0.09	0.96 ± 0.04	0.65 ± 0.12	25.5 ± 5.9
Sobel	0.76 ± 0.09	0.96 ± 0.04	0.65 ± 0.12	27.0 ± 7.2

Table 3.3. Performance of watershed segmentation using Prewitt and Sobel operators for gradient computation.

The performance of the algorithm using both Sobel and Prewitt operators were mixed. Overall, watershed segmentation had good recall, with a mean value close to 1, which implying that the segmentation resulted in good coverage of the GTV. There was only one case where the score was less than 0.9. On the other hand, the segmentation did not perform as well according to the precision, which indicated that the segmented regions tended to leak out of the true boundaries. The performance was poor for a number of cases, and there were three cases where the score was less than 0.5, denoting that a large false positive section that was segmented by the algorithm.

It can be seen that there was little difference in the scores between the Sobel and Prewitt operators, which was confirmed not to be statistically significant (Wilcoxon Signed Ranks test; DSC p -value = 0.983; recall p -value = 0.446; precision p -value = 0.777). The total time for the watershed segmentation process was also not statistically significant between the two groups (Wilcoxon Signed Ranks test; p -value = 0.112).

There was little difference between the two gradient computation techniques based on these findings. Sobel was chosen as the gradient operator as it is more commonly used in the literature.

3.12.1.4 Qualitative assessment of segmentation performance

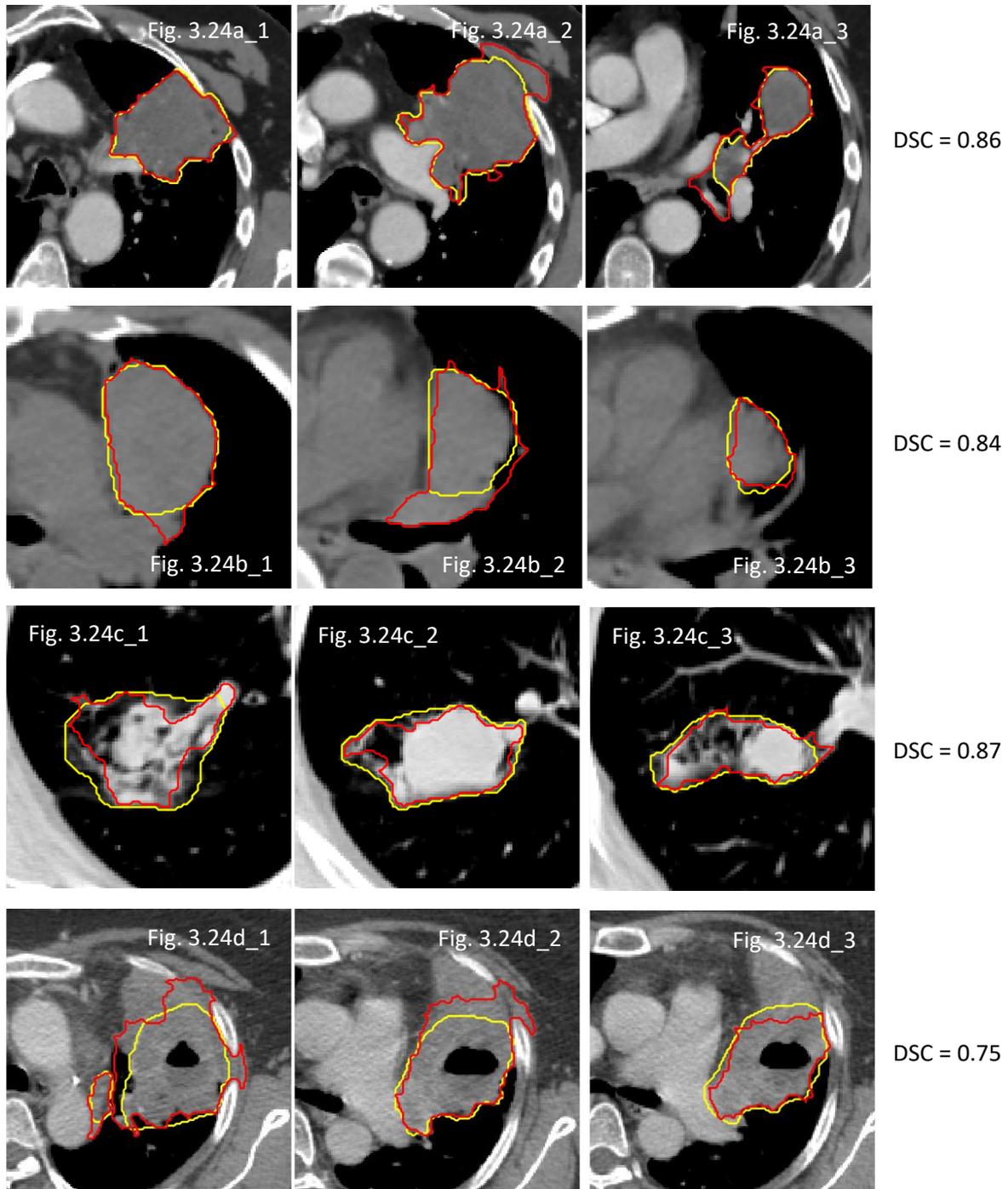
Some examples of the behaviour for the watershed segmentation is shown in figure 3.24. There were cases which exhibited good DSC scores of more than 0.8, where watershed segmentation was seen to work well in the presence of sharp contrast between regions. This can be observed in figures 3.24a, where the generated segmentation was congruent to the the reference contours in the vicinity of the contrast-enhanced vessels, as well as between the tumour and the lung parenchyma. Similar patterns were seen even in the absense of contrast administration (figures 3.24b), where the good distinction between tumour and mediastinal soft tissue was acheived.

Another observation is its acceptable performance in the presence of GGOs. Patchy GGO was present in one of the cases (figure 3.24c) surrouding a region of solid disease. Here, the segmentation included the area of GGO, in addition to solid disease. In figures 3.24d, regions of cavitation within the tumour were also included in as part of the segmentation, and the algorithm appeared to be able to segment the tumour from the region of collapse anterior to the GTV.

However, poorer performance of the segmentation was observed in other cases, and this was seen especially in the presence of disease adjacent to or within the mediastinum. In figure 3.24e, the primary disease wthihn the lung parenchyma was delineated relatively well by the algorithm. However, it can be seen that the segmentation had the tendency to leak away from

the nodal disease at the mediastinum, to include either the surrounding mediastinal soft tissue, or even at times the vessels. This was seen to occur both in cases without contrast administration, (figure 3.24f), as well as in cases with contrast-enhancement (figure 3.24e), depending on the level of contrast within the affected vessel.

Segmentation incongruent to the reference contours was also noted at the the chest wall (figure 3.24d_2), where musculature was included in the segmented regions on occasions.



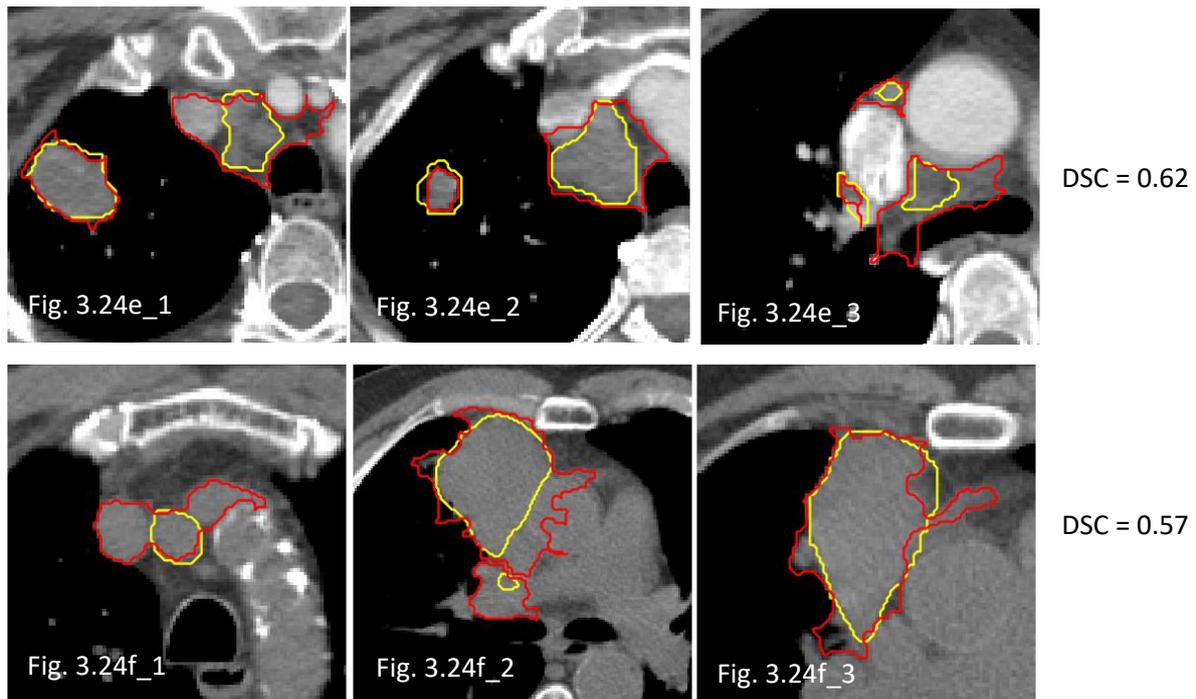
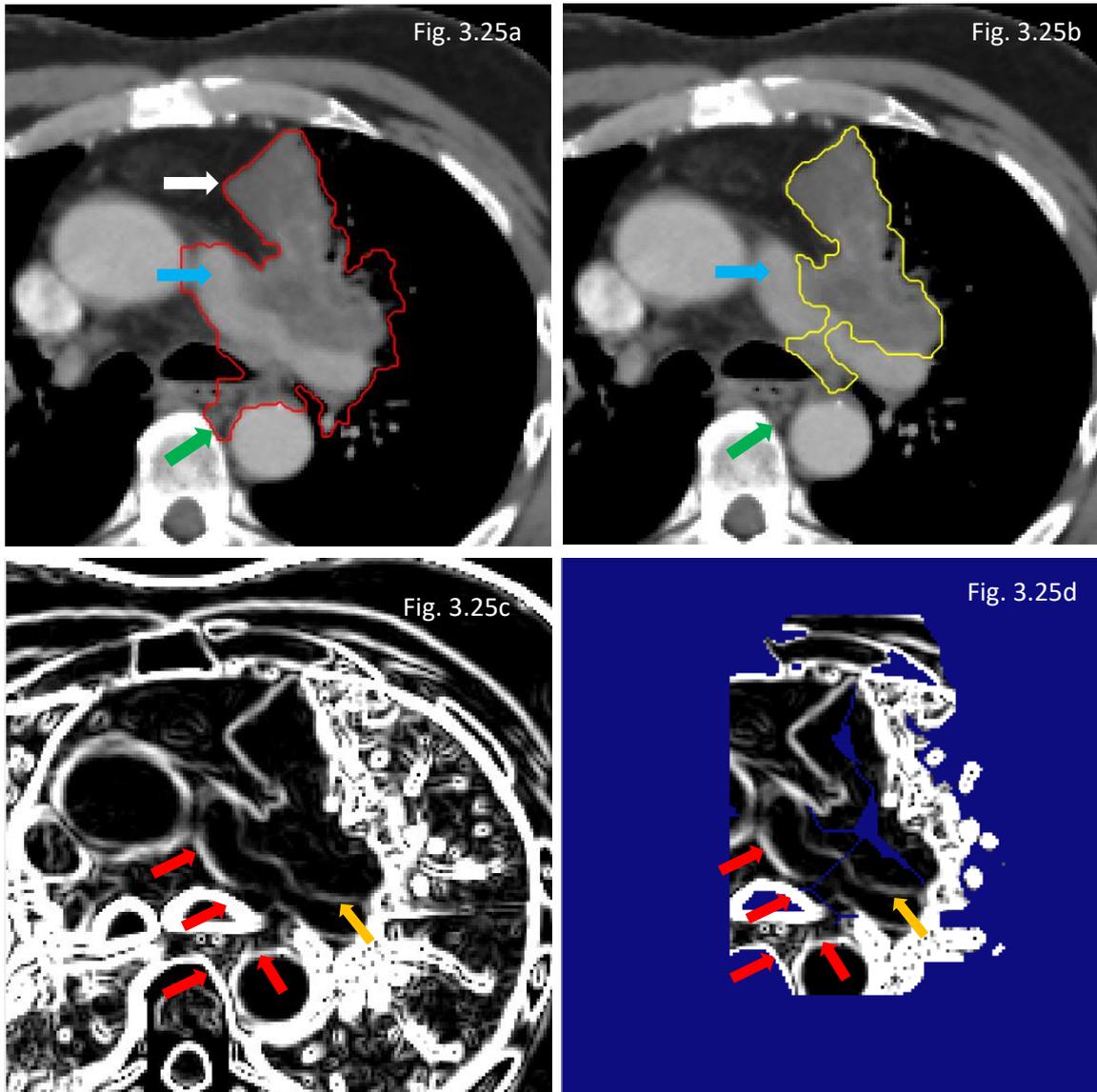


Figure 3.24. Watershed segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case)

This behavior can be explained on the closer inspection of the gradient maps and masks. In figure 3.25a – b, it can be seen that despite good adherence of the delineation to the tumour boundary at the anteromedial aspect of the tumour (white arrow), the whole of the left pulmonary artery was included in the segmented region (blue arrows in figures 3.25a and 3.25b). Although the small region of tumour between the artery and the left main bronchus was accurately included in the segmentation, there was further spillage towards the vertebral body (green arrows in figures 3.25a and 3.25b). This occurred despite the well-defined boundary between the vessel and the tumour that can be seen on the CT image. On the gradient image, this edge is also clearly apparent (orange arrow in figure 3.25c).

At the medial aspect of the tumour, there were a lot of competing gradients that affected the positioning of the watershed edge. The gradient magnitude around the boundary of the medial edge of the left pulmonary artery, left main bronchus, vertebral body and descending aorta (red arrows) were consistently higher than the magnitude at the medial tumour front. This resulted in the leakage of segmentation to the edge of highest gradient magnitude, instead of conforming to the boundary between the tumour and the left pulmonary artery, in spite its location being well within the search zone for the segmentation (figure 3.25d, with the superimposed masks in blue).



Figures 3.25a – d. Example of overestimation of the tumour region by the watershed approach. a) Watershed segmentation with inclusion of the left pulmonary artery (blue arrow), and extension towards the vertebral body posteriorly (green arrow). b) Gold standard contour. c) Gradient magnitude (Sobel operator). d) Superimposition of exclusion mask (blue regions) on gradient magnitude. Presence of edge in gradient image between the left pulmonary artery and the tumour depicted by orange arrow, and red arrows indicate locations of competing gradient magnitude.

From these evaluations, watershed segmentation was seen to produce segmentation that was congruent to the reference contours, at locations where the highest gradient magnitude was located at the tumour boundary within the search region. This occurred less commonly at the mediastinum, where competing gradient magnitude from the surrounding vessels and airways could affect the segmentation, resulting in leakage of the delineation into the surrounding tissues.

3.12.1.5 Exclusion of normal tissue structures

Leading on from the observations above, it was postulated that the performance of the segmentation could be improved through the application of further exclusion masks at the

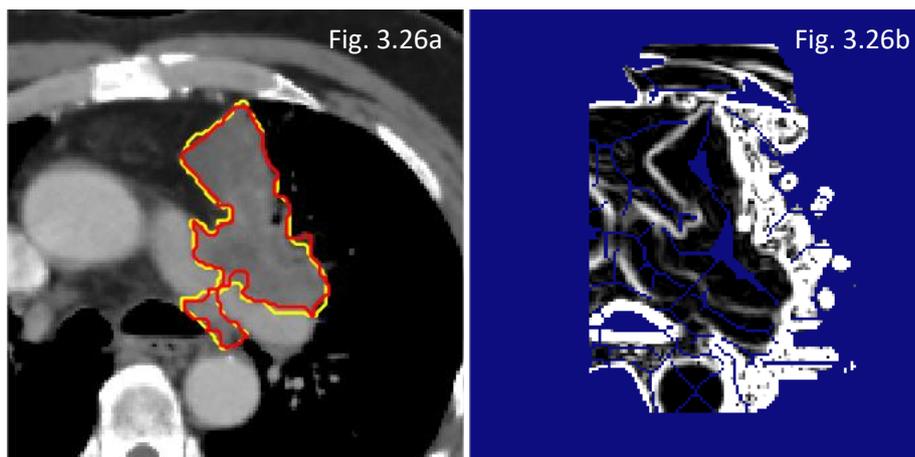
regions of competing gradient magnitude. Thus, a study on the impact of excluding normal tissue structures on the segmentation was performed, with the results shown in table 3.4.

There was a statistically significant improvement of the mean precision by more than 0.1 in the presence of further exclusion structures, with a corresponding improvement of the mean DSC from 0.76 ± 0.09 to 0.83 ± 0.07 . Although a statistically significant reduction in mean recall was obtained, the absolute difference was small at 0.02 and the recall score was still high at 0.94 ± 0.04 when the additional exclusion limits were applied.

	Exclusion structures absent	Exclusion structures present	<i>p</i> -value
Dice Similarity Coefficient	0.76 ± 0.09	0.83 ± 0.07	0.009
Recall	0.96 ± 0.04	0.94 ± 0.04	0.006
Precision	0.65 ± 0.12	0.76 ± 0.10	0.005
Time (seconds)	27.0 ± 7.2	63.6 ± 15.7	-

Table 3.4. Comparison of performance of watershed segmentation with and without further exclusion structures (Mann-Whitney test).

Figure 3.26 displays the same axial slice of the case in figure 3.25, where the watershed segmentation produced delineation similar to the reference contours.



Figures 3.26a – b. Same case and axial slice as figure 3.25, with corresponding exclusion masks (blue regions) overlaying gradient magnitude. Red outline – watershed segmentation; yellow outline – reference contours.

Although the incorporation of further exclusion structures improved the segmentation ability, the need to delineate more normal tissue structures for all cases undermined the time- and labour-saving benefits of the automatic process. Thus, other methods to exclude normal structures was explored.

3.12.1.6 Atlas-based generation of exclusion structures

Atlas-based segmentation of the normal tissues was explored as a means of generating the exclusion structures in an automatic fashion. An example of the results from a case where the segmentation was applied is shown in figure 3.27. In figures 3.27a, c, and e, it can be seen that quality of the bone, lung, trachea and spinal cord outlines were within acceptable clinical practice. For the bone outlines, there was some leakage of the outlines into the surrounding chest wall. There was also extension of the trachea contours into the main bronchi. However, other contours such as the vessels, mediastinum, oesophagus, heart and chest wall were less congruent with their respective structures (figures 3.27b, d and e). Similar inconsistencies were observed in cases with contrast-enhancement (figures 3.27g, h, and i). In fact, there were gross errors where regions of the tumour were incorrectly segmented as the vessel or mediastinum (figures 3.27b, d and h).

In addition to deformable registration, further pre- and post-processing was performed on the structures that were in the provided atlases within the programme (personal communication). Within the thorax, this applied to the lungs, trachea, bone and spinal cord. However, for the other atlas structures that were imported into the system, the segmentation results were based purely on the registration between the new image and the atlas image, without further post-processing.

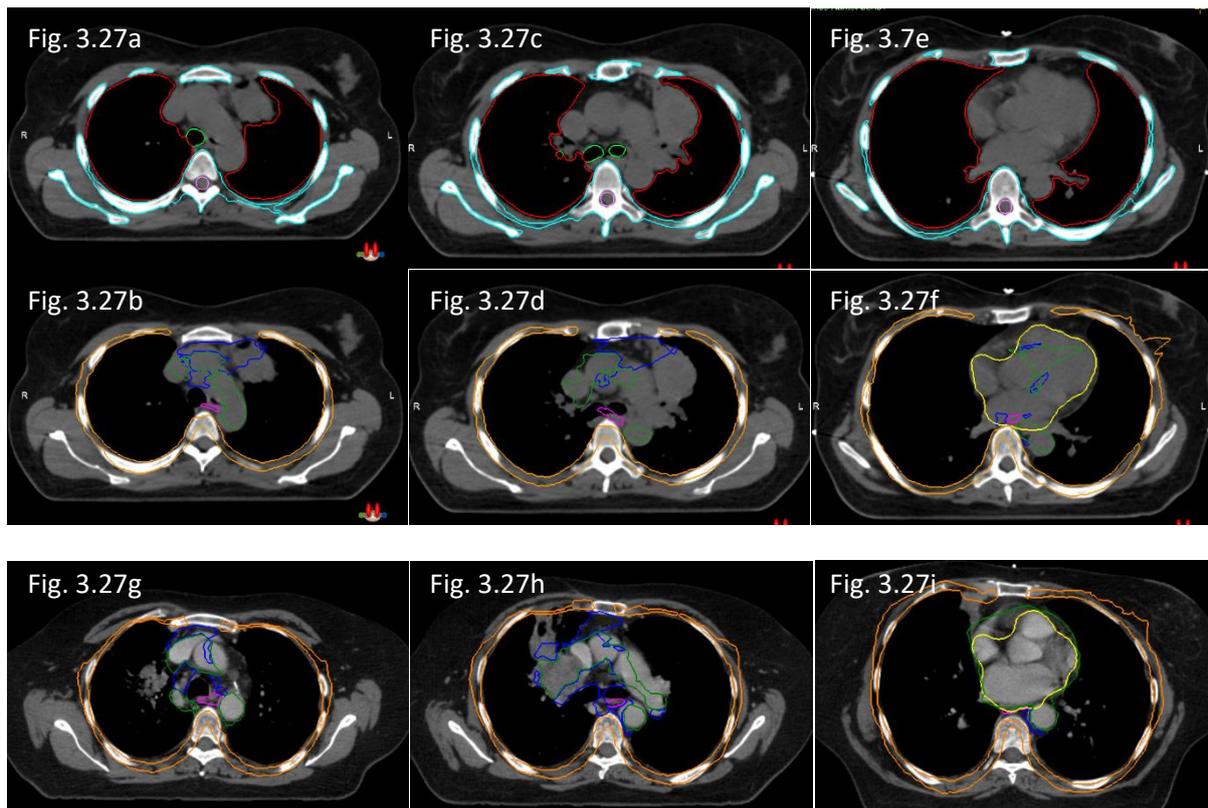


Figure 3.27. Atlas-based normal tissue segmentation. Cyan – bone; Red – lung; Light green – trachea; Pink – spinal cord; Orange – chestwall; Dark green – vessels; Dark blue – mediastinal soft tissue; yellow – heart. (a – f from case 1; g – i from case 2)

Although only a rough segmentation of the normal tissues was required to facilitate the watershed approach, because of the gross errors seen with the atlas-based segmentation, it was concluded that this workflow would not complement the watershed technique. Moreover, it

may actually worsen the outcome by incorrectly labelling the tumour as part of normal tissue. This avenue was therefore not pursued.

3.12.1.7 Semi-automated approach

The semi-automated approach was therefore explored as another means for introducing further exclusion limits, the performance of which is presented here. The line plot in figure 3.28 that shows the DSC across the three attempts at the semi-automatic implementation of the algorithm indicate that there is little variation in the DSC across the three attempts.

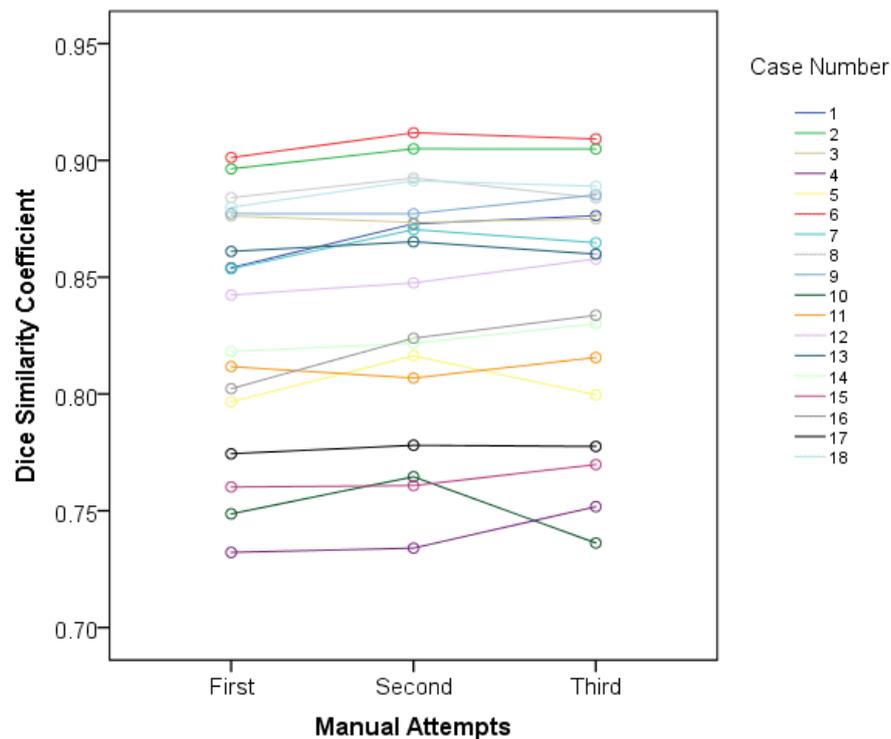


Figure 3.28. Dice similarity coefficients for three runs of semi-automated watershed segmentation.

	First run	Second run	Third run	Aggregate across three runs
Dice Similarity Coefficient	0.83 ± 0.05	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.05
Recall	0.95 ± 0.04	0.94 ± 0.04	0.94 ± 0.04	0.94 ± 0.04
Precision	0.76 ± 0.08	0.77 ± 0.08	0.77 ± 0.08	0.77 ± 0.08

Table 3.5. Individual and aggregate performance for semi-automated watershed segmentation.

The mean individual and aggregate scores are shown in table 3.5. On statistical testing using the Friedman test for repeatability, there was no statistical difference for recall ($\chi^2(2) = 0.78$, $p = 0.68$). However, DSC and precision were statistically significant ($p = 0.006$ and $p = 0.001$ respectively). In the post-hoc analysis (Wilcoxon signed-rank test with a Bonferroni correction (significance level set at $p < 0.017$)), there was no statistically significant difference between the second and third runs for both DSC and precision ($p = 0.56$ and $p = 0.50$ respectively). However, DSC and precision were found to be statistically significant between the first and second ($p = 0.002$ and $p = 0.001$ respectively), as well as the first and third runs ($p = 0.004$ and

$p = 0.002$ respectively). Despite the statistical difference seen, the absolute differences between the first and the other two runs for DSC and precision were very small, suggesting repeatability of the semi-automatic workflow. Nonetheless, the mean of the 3 runs were computed as the final results to dissipate bias.

The scores obtained for the fully automatic approach versus the semi-automatic approach is shown in table 3.6. There was a statistically significant improvement in the DSC and precision scores with the semi-automatic approach (Mann-Whitney test, $p = 0.004$ and $p = 0.002$ respectively), with an increase in the DSC by almost 0.1, and an improvement in the precision by more than 0.1. Similar to the results for the exclusion structures, the semi-automatic approach was associated with a slight worsening of the mean recall score, which although is statistically significant, is small in absolute terms at 0.02.

	Automatic	Semi-automatic	p -value
Dice Similarity Coefficient	0.76 ± 0.09	0.84 ± 0.05	0.004
Recall	0.96 ± 0.04	0.94 ± 0.04	0.009
Precision	0.65 ± 0.12	0.77 ± 0.08	0.002
Time (seconds)	27.0 ± 7.2	215.7 ± 73.3	-

Table 3.6. Comparison of performance of automatic and semi-automated watershed segmentation.

Figure 3.29 displays the same axial slice of the case in figures 3.25 and 3.26, where the segmentation results were similar to the reference contours. Minimal differences were seen between the three semi-automatic runs, despite differences in the placements of the points as seen in figures 3.29d, e and f.

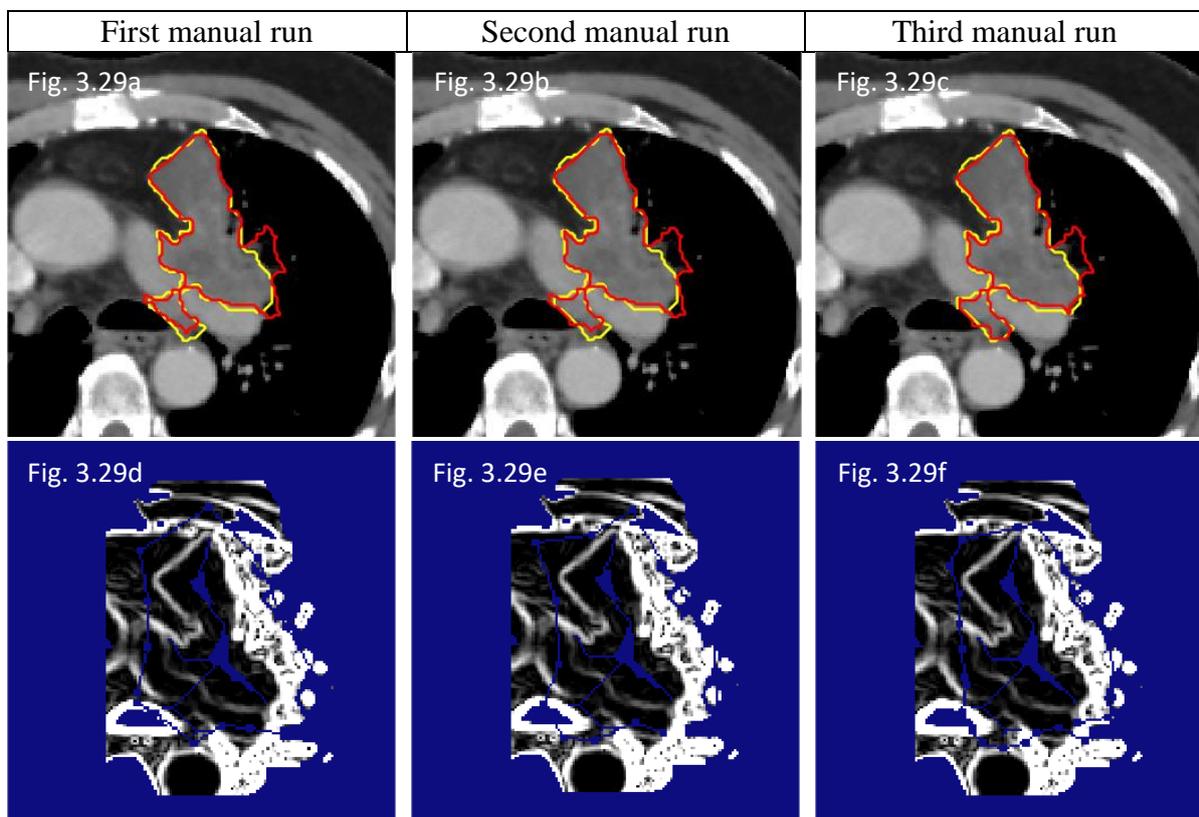


Figure 3.29. Same case and axial slice as figures 3.25 and 3.26, with corresponding exclusion masks (blue regions) and manually placed points (linearly connected) overlaying gradient magnitude. Red outline – watershed segmentation; yellow outline – reference contours.

These results show that the introduction of further manual input to narrow the search region for the segmentation improved the issue with segmentation leakage. This came at an added cost of increasing computational time, with a mean of 3.6 ± 1.2 mins. Although this is within clinically acceptable limits, further exploration of semi-automated approaches were not further continued in this work, as a fully-automated approach was preferred.

3.12.1.8 Validation data

Therefore, despite the improvement in performance of the watershed segmentation with manual input, to allow fair comparison of the segmentation technique with the other automated approaches, the fully automatic watershed segmentation was applied to the validation data. Similar results were observed between the runs (figure 3.30 and table 3.7). In keeping the results seen with the 18 subsample cases, good recall was observed, at the expense of precision. The estimated DSC for the watershed approach was 0.72 ± 0.10 , with a recall of 0.94 ± 0.05 and precision of 0.62 ± 0.12 across all three validation folds.

	Recall	Precision	DSC
Validation Run 1	0.94 ± 0.04	0.64 ± 0.13	0.73 ± 0.09
Validation Run 2	0.94 ± 0.04	0.64 ± 0.09	0.74 ± 0.07
Validation Run 3	0.94 ± 0.05	0.57 ± 0.13	0.68 ± 0.12
Aggregate across three runs	0.94 ± 0.04	0.62 ± 0.12	0.72 ± 0.10

Table 3.7. Mean performance of watershed segmentation on the validation datasets.

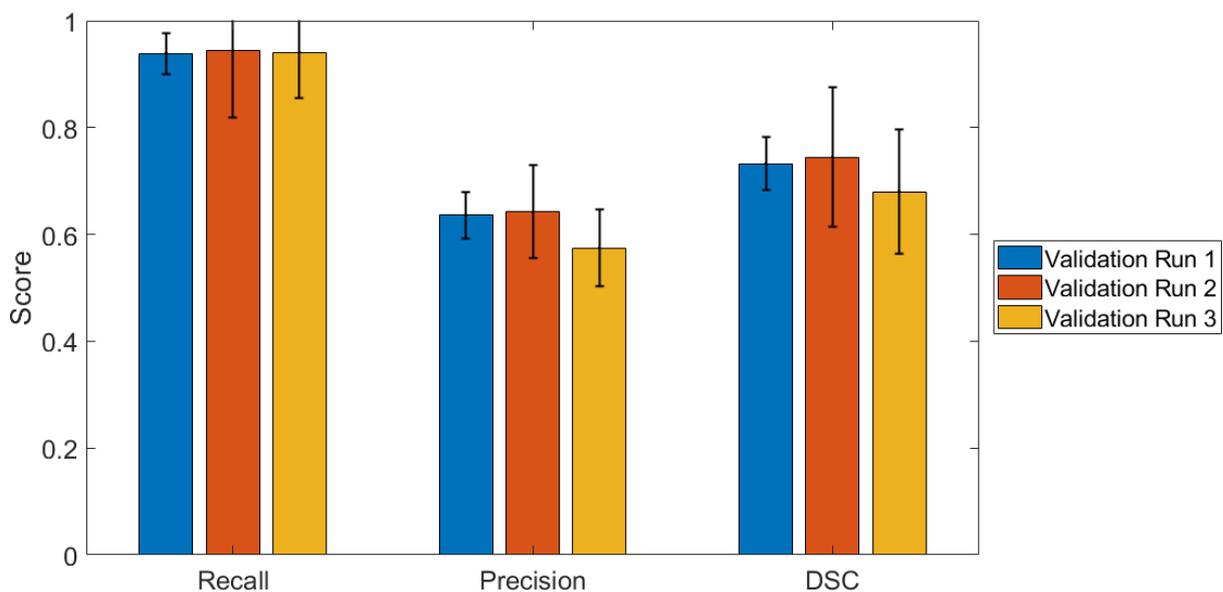


Figure 3.30. Mean performance of watershed segmentation on each fold of the validation datasets (error bars represent standard deviation).

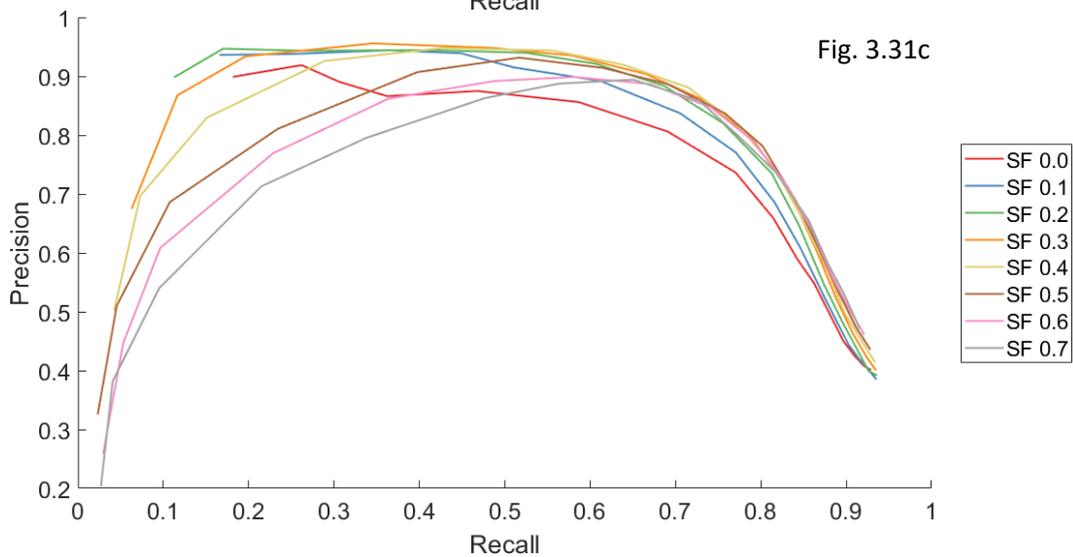
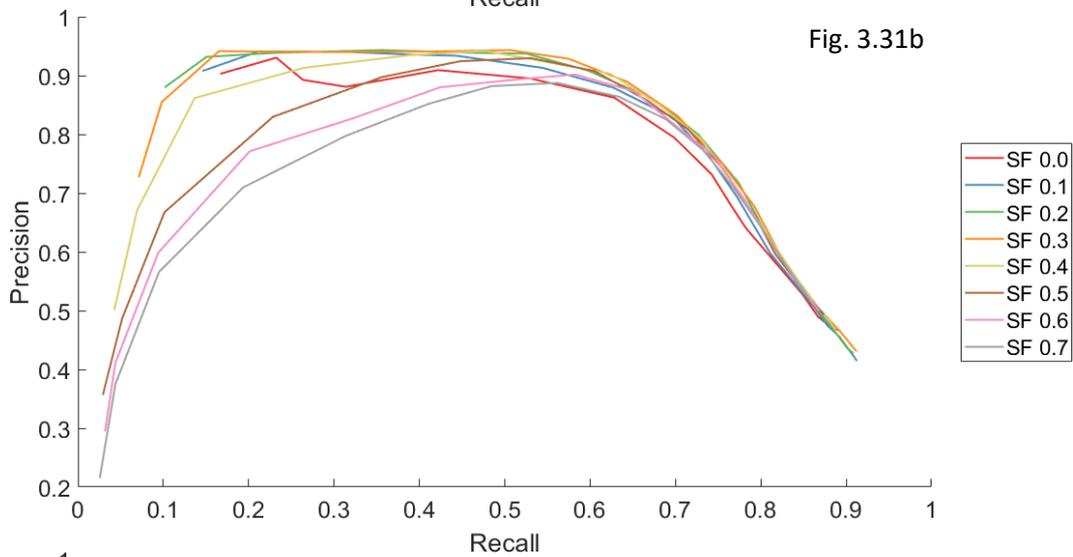
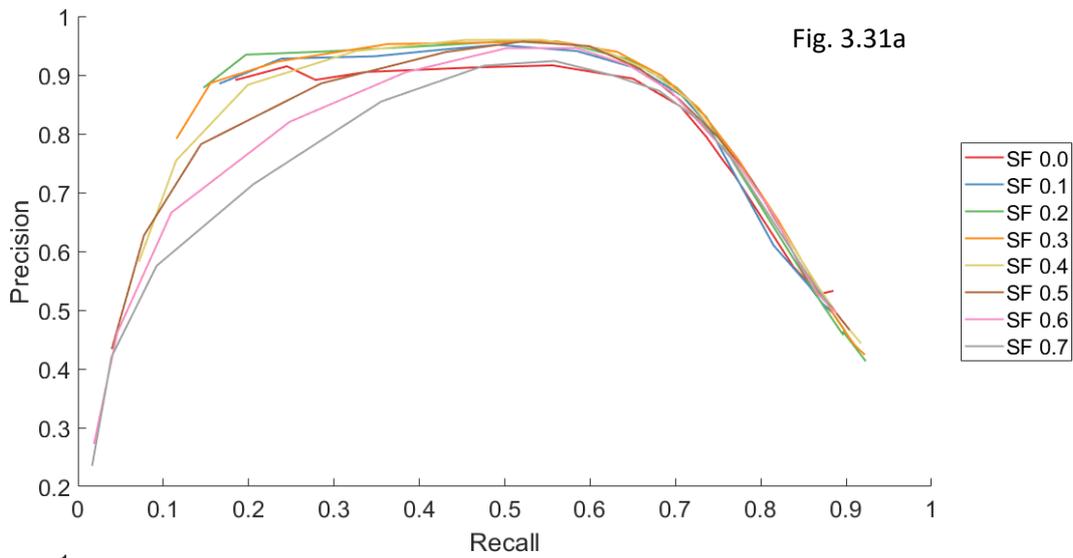
3.12.2 Chan-Vese active contour

3.12.2.1 Evaluation of mask initialisation and parameter settings on subsample cases

The parameter sweep and selection is shown for each of the three different types of initialisations for the 18 subsample cases, followed by an assessment of their performance based on their respective optimised parameters.

The precision recall plots (figure 3.31) for all three methods of mask initiation show a similar trend whereby an optimal balance was achieved between precision and recall for a small range of contraction bias, where values at the extremes result in poor precision, and/or poor recall. The large variation in precision and recall over the evaluated range of contraction bias suggests that the active contour propagation is highly sensitive to this parameter, whereby a small change in the parameter can lead to a potential large difference in the segmentation.

It was also interesting to note that near the optimal balance of precision and recall, both these measures appear to be less affected by the different smoothing factor within the evaluated range, denoted by the clustering of the plots at this part of the graph. This effect appears to be more apparent for the two eroded initiation masks than for the non-eroded circle, where smoothing factor values of greater than 0.5 performed less well.

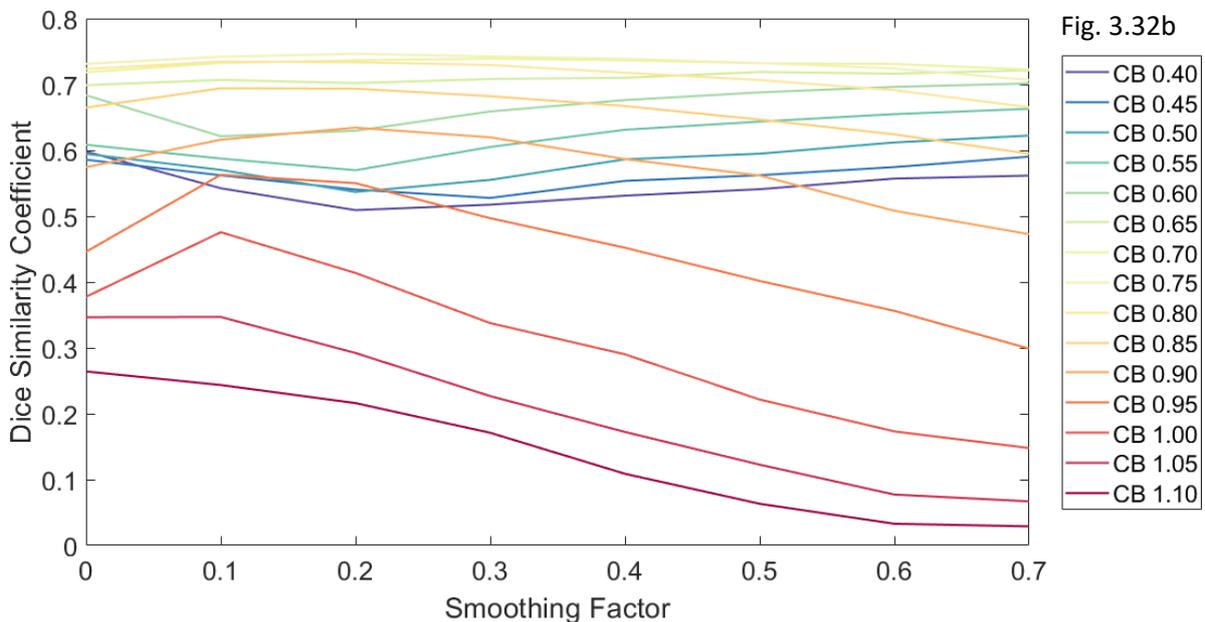
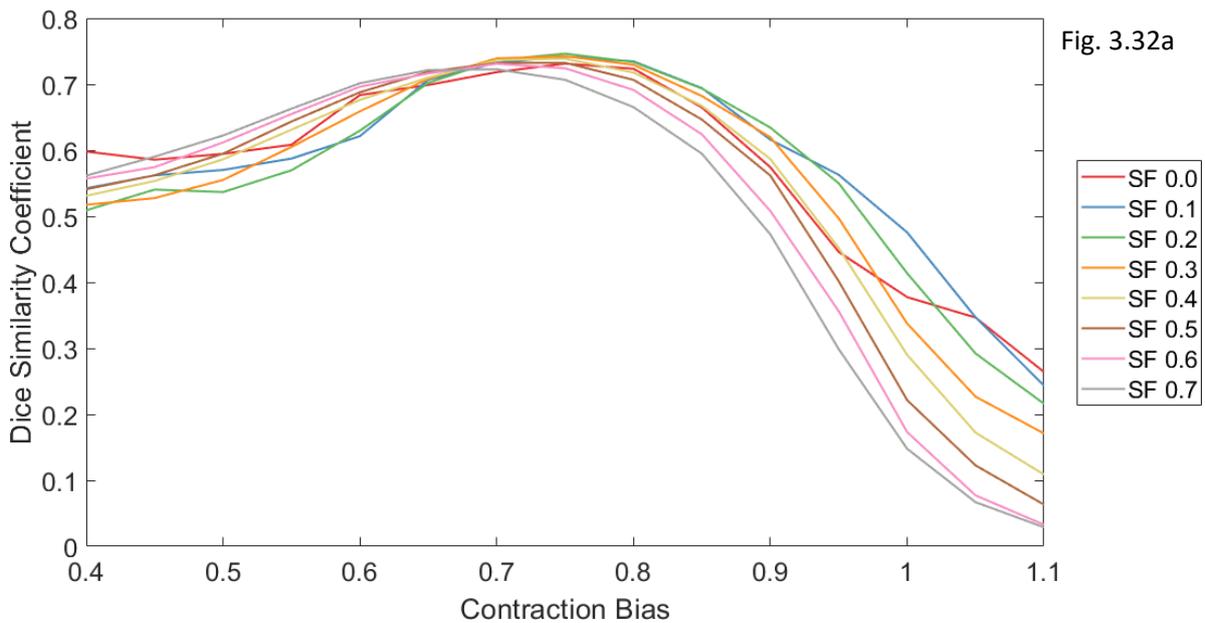


Figures 3.31a – c. Mean precision vs Mean recall plots for 18 training cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Convex polygon with 4-pixel erosion, b) Circle with 4-pixel erosion, c) Circle with no erosion (Chan-Vese active contour segmentation).

Following the results of the precision and recall plots, the DSC plots for the different initialisation masks were assessed to determine the optimal parameter settings, where the effect of alteration of the contraction bias and smoothing factors can be better visualised.

3.12.2.1.1 Initial mask: Convex polygon with 4-pixel erosion

The DSC plot for the convex polygon initialisation shows peak DSC scores of more than 0.7, at contraction bias between 0.7 and 0.8. Within this range, the smoothing factor had little influence on the DSC score.



Figures 3.32a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (Convex polygon initial mask with 4-pixel erosion; Chan-Vese active contour algorithm)

The highest achieved DSC for this initialisation was 0.75 ± 0.11 , with an optimal contraction bias at 0.75 and smoothing factor of 0.2.

3.12.2.1.2 Initial mask: Circle with 4-pixel erosion

Similar to the above, the highest DSC scores were achieved at a contraction bias around 0.7, where there was little influence from variation of the smoothing factor. Although the DSC scores were similar to the convex hull plots towards the extremes values of the contraction bias, lower peak DSC was obtained around the optimal contraction bias.

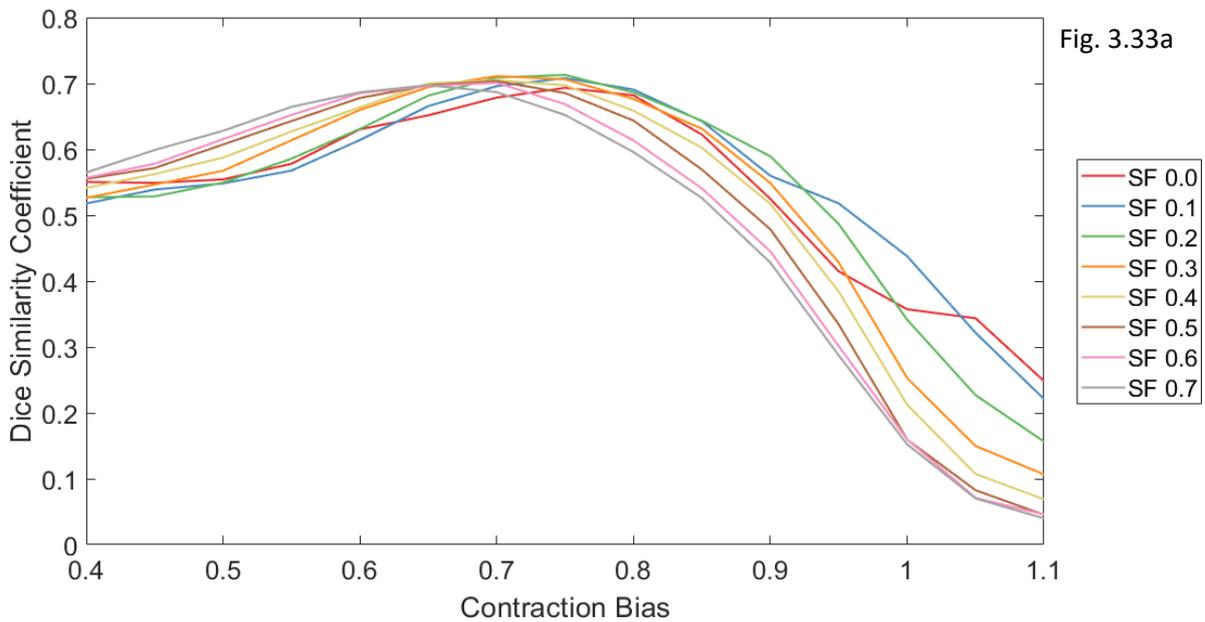


Fig. 3.33a

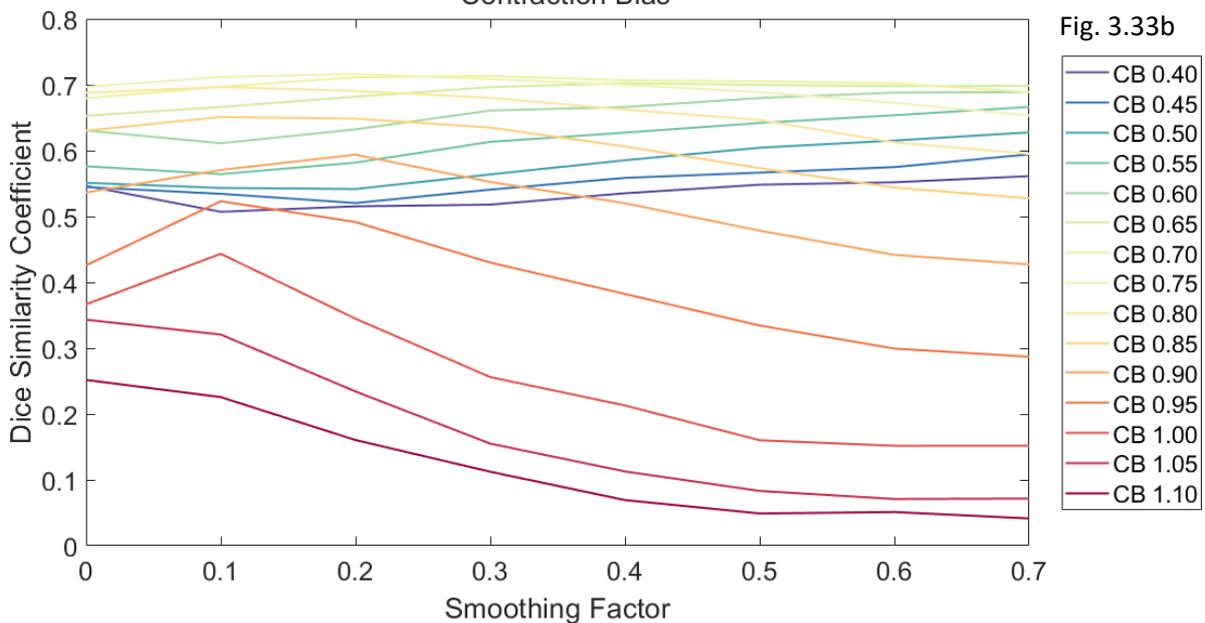


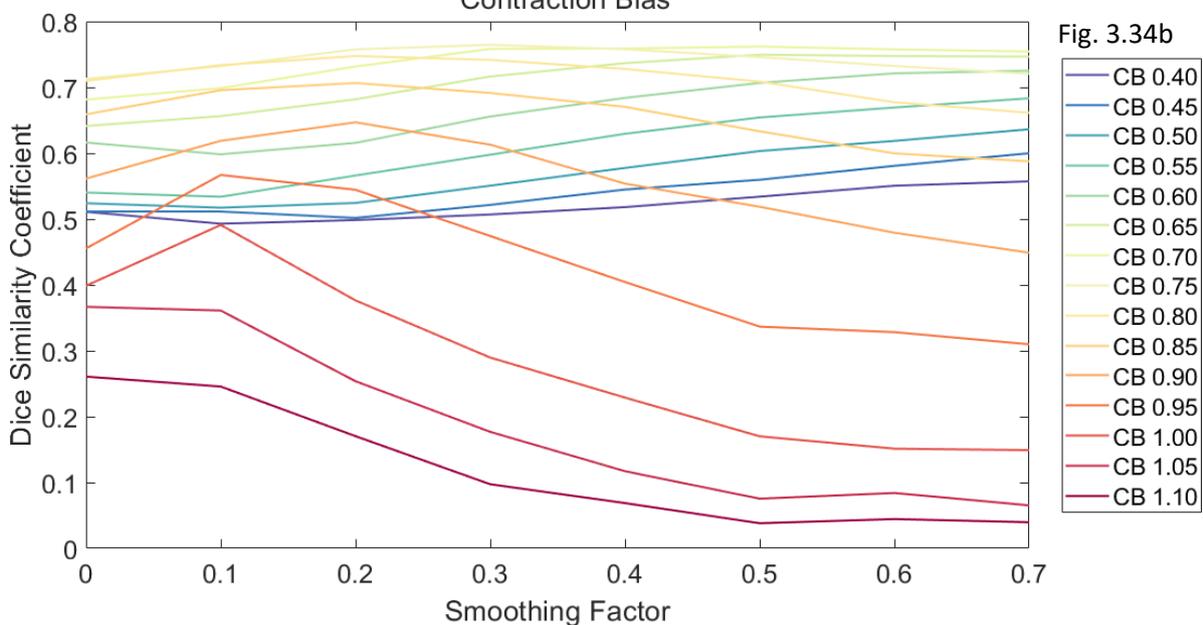
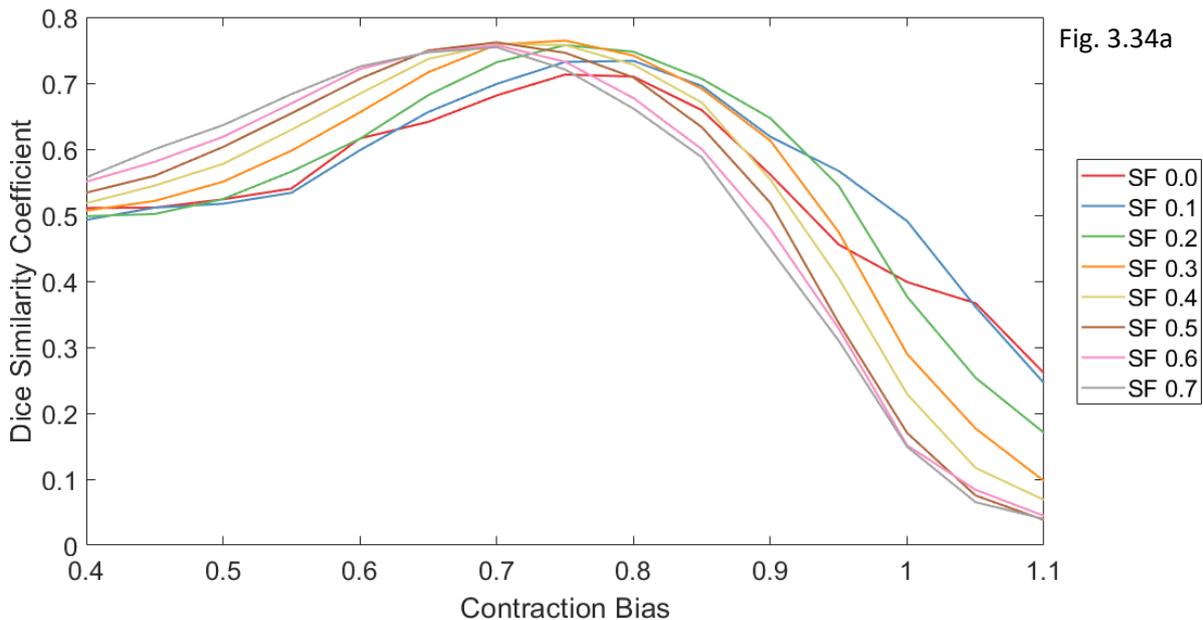
Fig. 3.33b

Figures 3.33a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (Circle initial mask with 4-pixel erosion; Chan-Vese active contour algorithm)

The highest achieved DSC for this initialisation was 0.72 ± 0.10 , with an optimal contraction bias at 0.75 and smoothing factor of 0.2.

3.12.2.1.3 Initial mask: Circle with no erosion

Again, for each of the smoothing factor parameters, the segmentation was largely affected by changes to the contraction bias, as can be seen by the wide range of DSC across the range of contraction bias in figure 3.34a. Values of contraction bias lower than 0.6 and higher than 0.85 resulted in poorer DSC across all the smoothing factor parameters, while DSC greater than 0.7 was achieved at contraction bias between 0.7 and 0.8. Between these values, there was little impact of variation of the smoothing factor on DSC as seen in figure 3.34b.



Figures 3.34a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 training cases (circle initial mask with no erosion; Chan-Vese active contour algorithm)

The highest achieved DSC for this initialisation was 0.76 ± 0.11 , with an optimal contraction bias at 0.75 and smoothing factor of 0.3.

A summary of the quantitative performance for the different initial masks is shown in table 3.8. With the use of parameters selected on the basis of the best DSC, higher precision than the recall was observed for all three initialisations.

	Optimal contraction bias	Optimal smoothing factor	Recall	Precision	DSC
Convex hull (4-pixel erosion)	0.75	0.2	0.70 ± 0.14	0.88 ± 0.08	0.75 ± 0.11
Circle (4-pixel erosion)	0.75	0.2	0.67 ± 0.15	0.86 ± 0.09	0.72 ± 0.10
Circle (No erosion)	0.75	0.3	0.74 ± 0.14	0.86 ± 0.08	0.76 ± 0.11

Table 3.8. Optimal parameter settings for Chan-Vese active contour segmentation with results of performance.

3.12.2.2 Qualitative assessment of segmentation performance

In some images, the different initial masks produced similar segmentation (figure 3.35). This occurred mainly in situations where the boundary of the masks was in close proximity to the tumour edge.

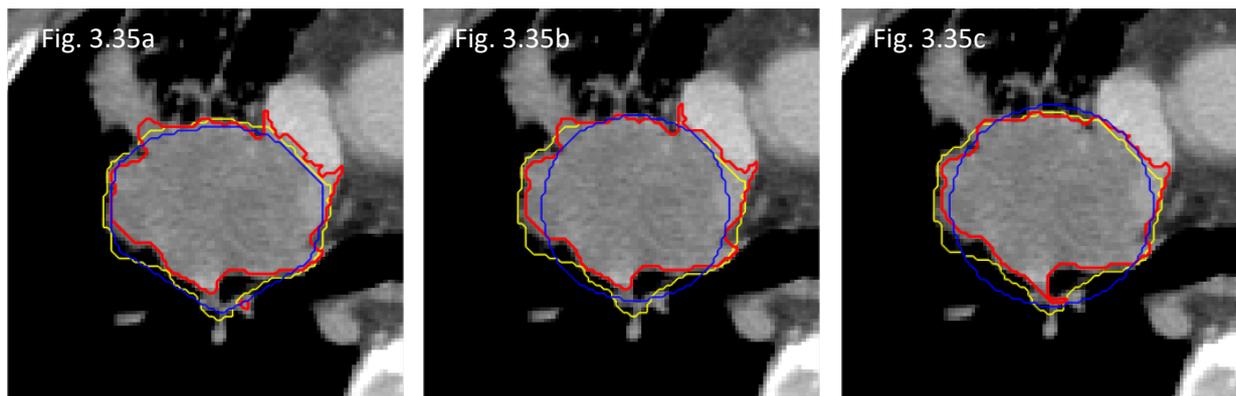


Figure 3.35. Example showing similar Chan-Vese active contour segmentation results using different initialisations a) Convex hull with 4-pixel erosion; b) Circle with 4-pixel erosion; c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.

However, differences in the behaviour of the Chan-Vese segmentation were observed between the different initialisations in a number of other cases.

Although the performance of the segmentation using the convex hull masks was comparable to the non-eroded circle quantitatively, two main issues were observed as illustrated in figures 3.36. In cases where there was little concavity of the submitted outlines, there was minimal change between the initial and the final contours using the convex hull mask. Consequently, this produced a final segmentation that was overfitted to the shape of the submitted contours. As shown in figures 3.37a and 3.37d, this effect could be in part due to parameter tuning. With the use of a different set of parameters, the contour evolved to a different shape to that of the initiating convex hull.

Similarities in the behaviour of the eroded and non-eroded circle masks were observed (figures 3.36a–c, suffix `_2` and `_3`). As the initial boundary for these masks was at a greater distance

away from the tumour edge, at the selected parameter settings, there was a tendency to draw the contours away from the tumour boundary into surrounding tissue. The contours were attracted to surroundings with density close to the tumour (e.g. muscle, non-contrast enhanced vessels, oesophagus), and to lesser extent regions with greater difference in density e.g. lung parenchyma, even without the use of a lower threshold.

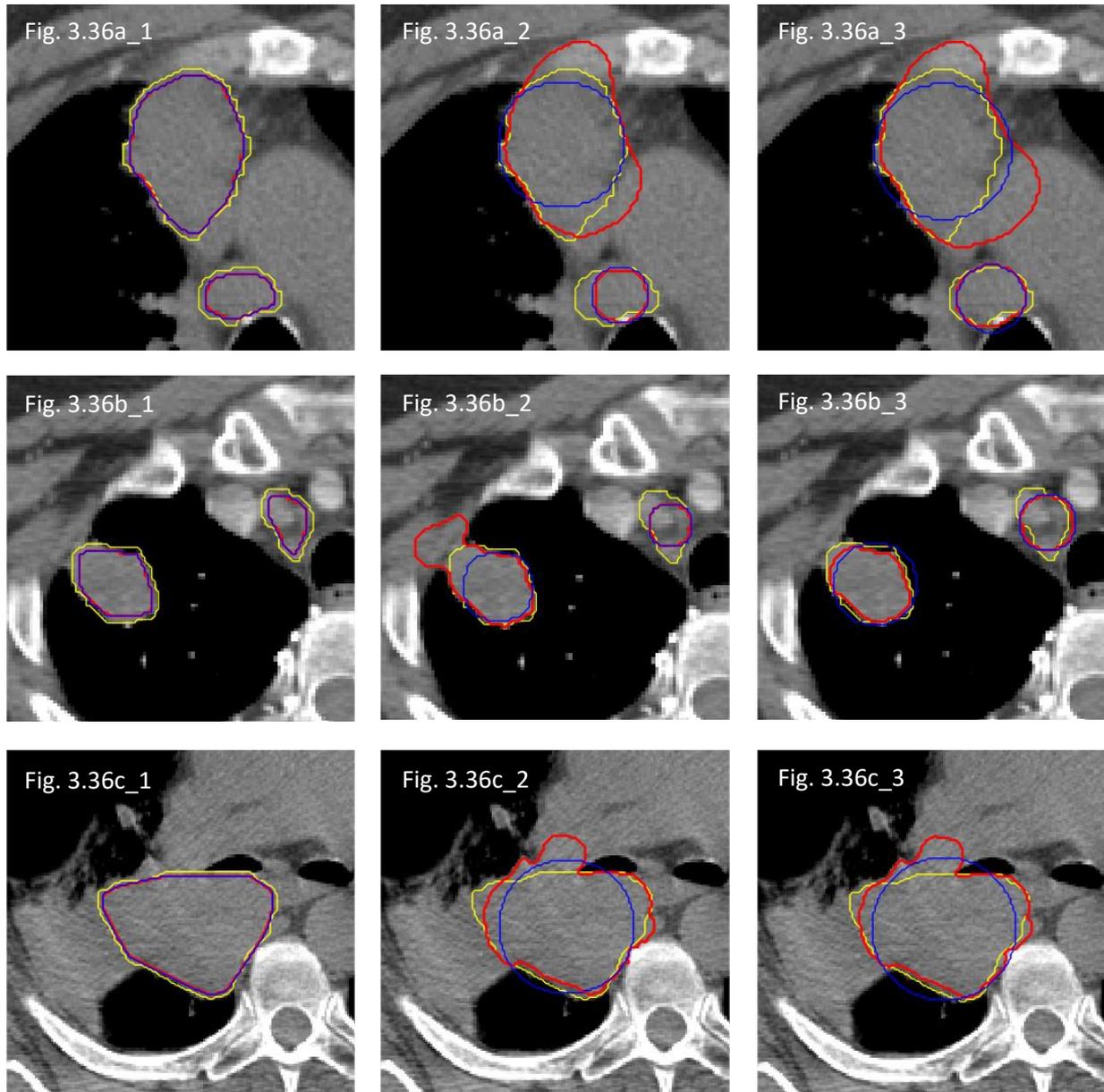


Figure 3.36. Examples (cases a to c) showing variation in Chan-Vese active contour segmentation results using different initialisations; suffix _1) Convex hull with 4-pixel erosion; suffix _2) Circle with 4-pixel erosion; suffix _3) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.

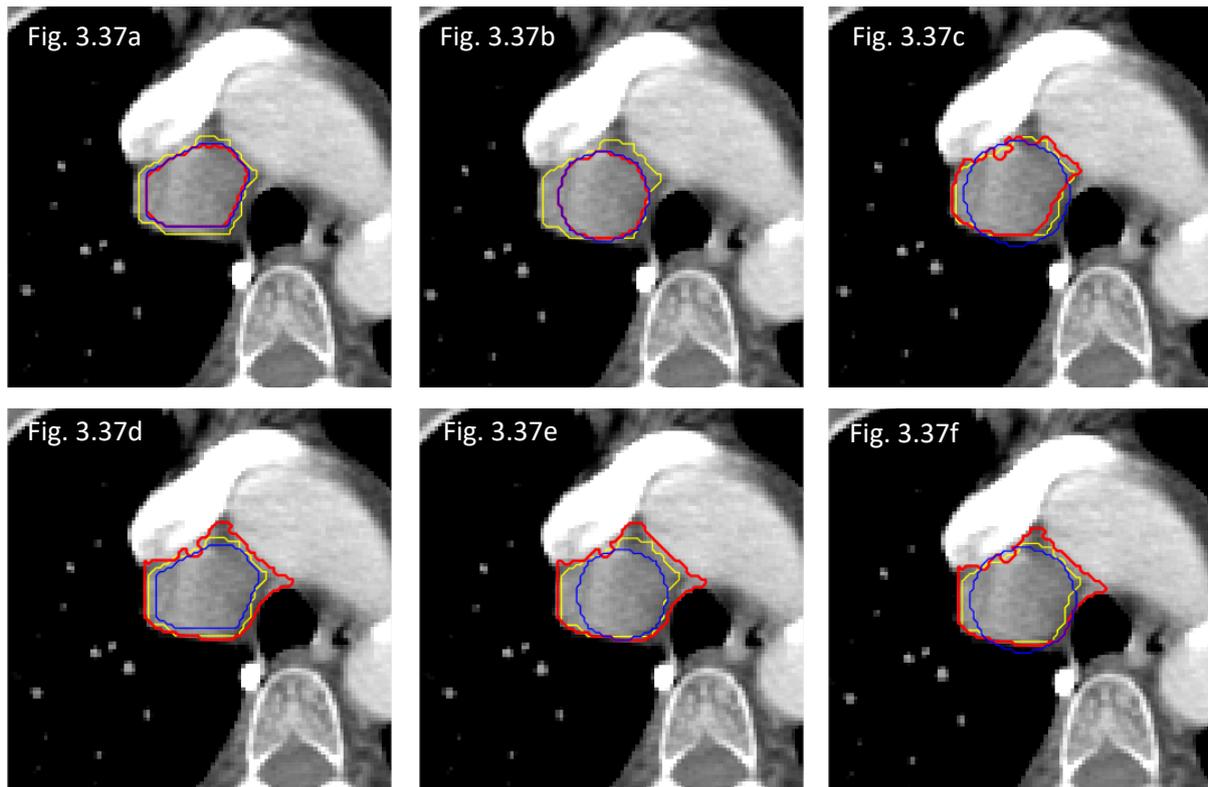


Figure 3.37. Comparison of Chan-Vese active contour segmentation in relation to different initial masks for the same axial slice, where the optimal parameter settings were used for figures a to c, whilst contraction bias of 0.6 and smoothing factor of 0.7 were used for figures d to f. a,b) Convex hull with 4-pixel erosion; b,e) Circle with 4-pixel erosion; a,d) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.

As seen in the use of the convex hull, parameter selection also affected the segmentation results. By choosing different parameter settings (e.g. contraction bias 0.6, smoothing factor 0.7), very similar segmentation results could be achieved for the same slices (figures 3.37 d - f) across the different initialisations.

Another behaviour of the Chan-Vese active contour is shown in figure 3.38, where the segmentation split into multiple contours, despite the use of a single initialisation mask. This contour evolution appeared to be due to its propensity to exclude regions of low density (lung parenchyma) resulting in it evolving ‘inwards’ and dividing. In this particular slice, in spite of the splitting capability, coverage of the tumour was still inadequate for the circle masks as the tumour boundary was at a distance from the initialisation edge.

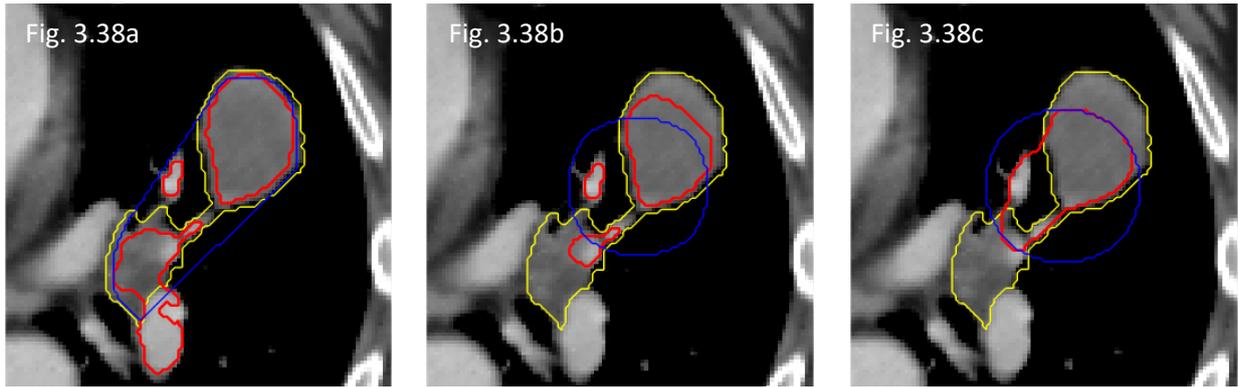


Figure 3.38. Example showing splitting of generated contour with Chan-Vese segmentation a) Convex hull with 4-pixel erosion; b) Circle with 4-pixel erosion; c) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.

On balance, it was deemed that it would be most appropriate to use the non-eroded circle mask to initialise the active contour segmentation. Even though this could result in less accurate segmentation for cases where the tumour boundary is not well approximated with a circle, there was less risk of overfitting.

3.12.2.3 Training using cross validation folds for parameter selection

Having established the workflow for the parameter selection and the use of the non-eroded circle as the initiation mask, the process was applied to the training data in their respective folds to determine the optimum parameter settings. Figure 3.39 shows the similar precision and recall curves obtained for all three training runs. The optimum settings for the three runs are summarised in table 3.9, where the highest mean DSC was obtained at the same settings across all three runs.

	Optimal contraction bias	Optimal smoothing factor	Recall	Precision	DSC
Training Run 1	0.75	0.3	0.71 ± 0.10	0.85 ± 0.06	0.75 ± 0.07
Training Run 2	0.75	0.3	0.66 ± 0.12	0.83 ± 0.08	0.71 ± 0.09
Training Run 3	0.75	0.3	0.69 ± 0.13	0.84 ± 0.07	0.73 ± 0.10

Table 3.9. Optimal parameter settings for Chan-Vese segmentation with results of performance.

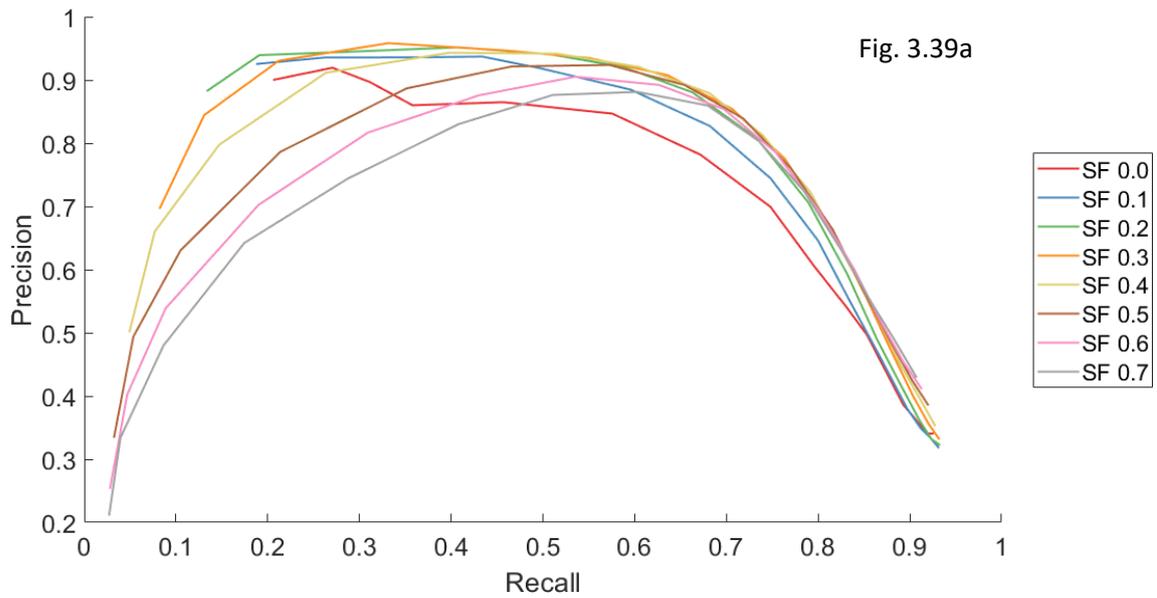


Fig. 3.39a

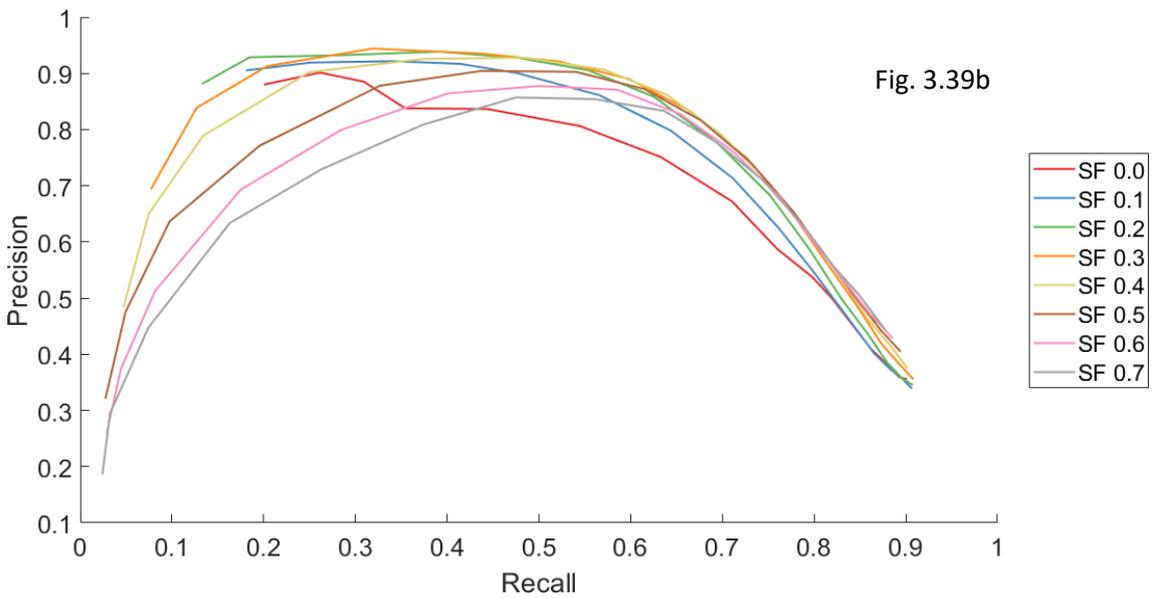


Fig. 3.39b

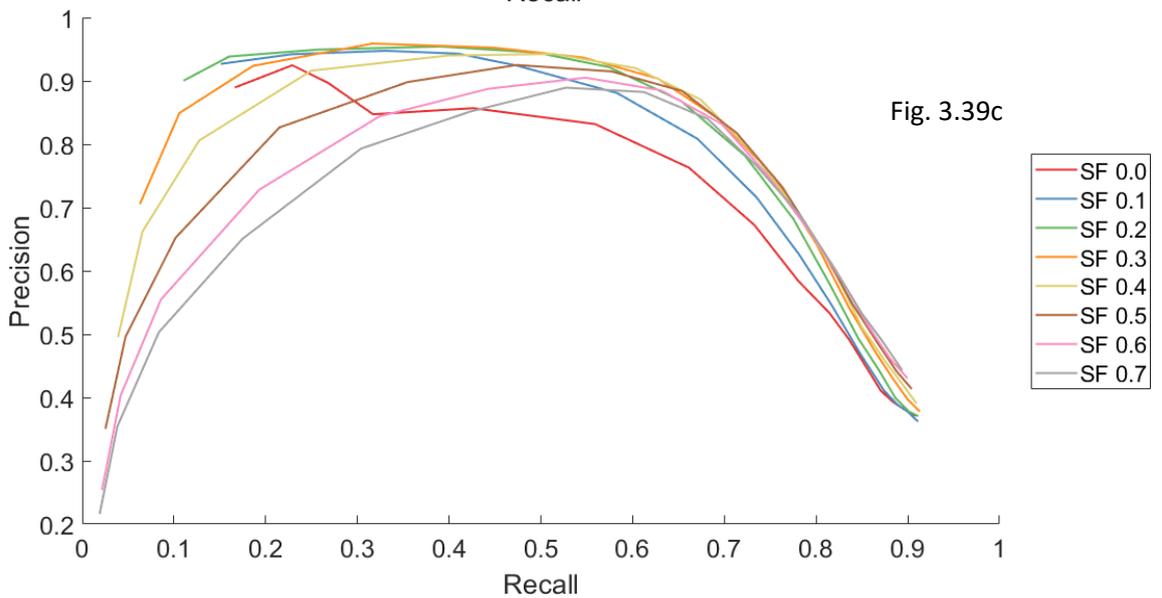


Fig. 3.39c

Figure 3.39. Mean precision-recall curves for Chan-Vese segmentation as function of contraction bias, for varying smoothing factors. a) Training Run 1; b) Training Run 2; c) Training Run 3.

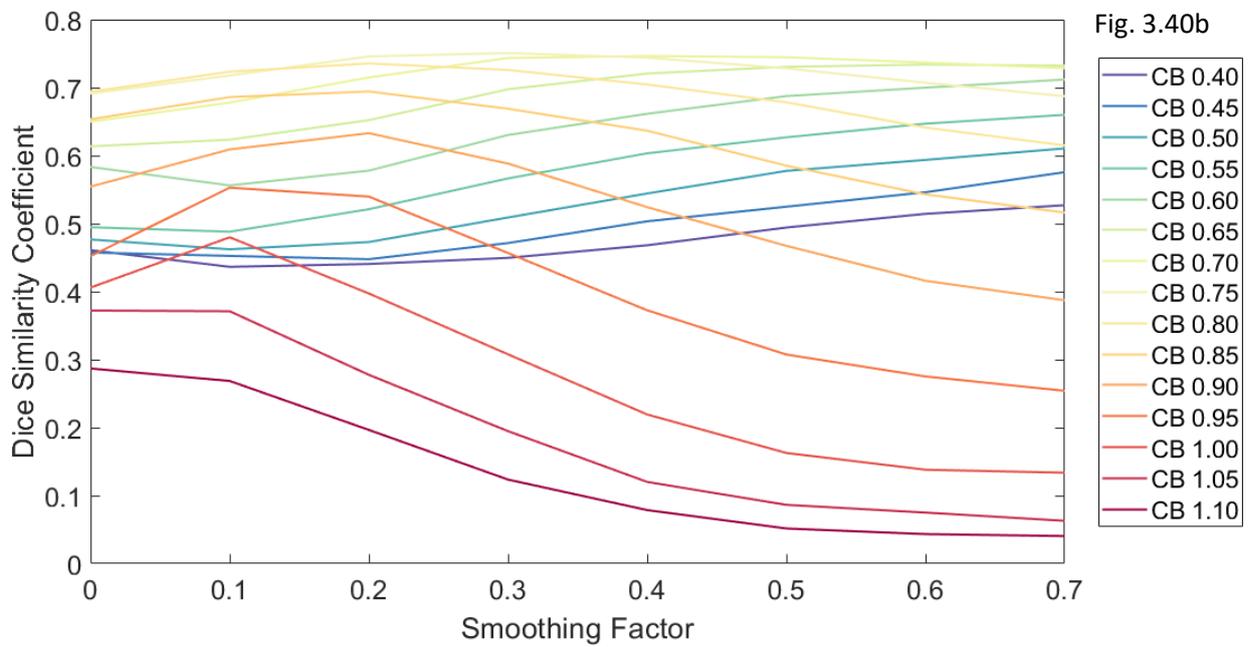
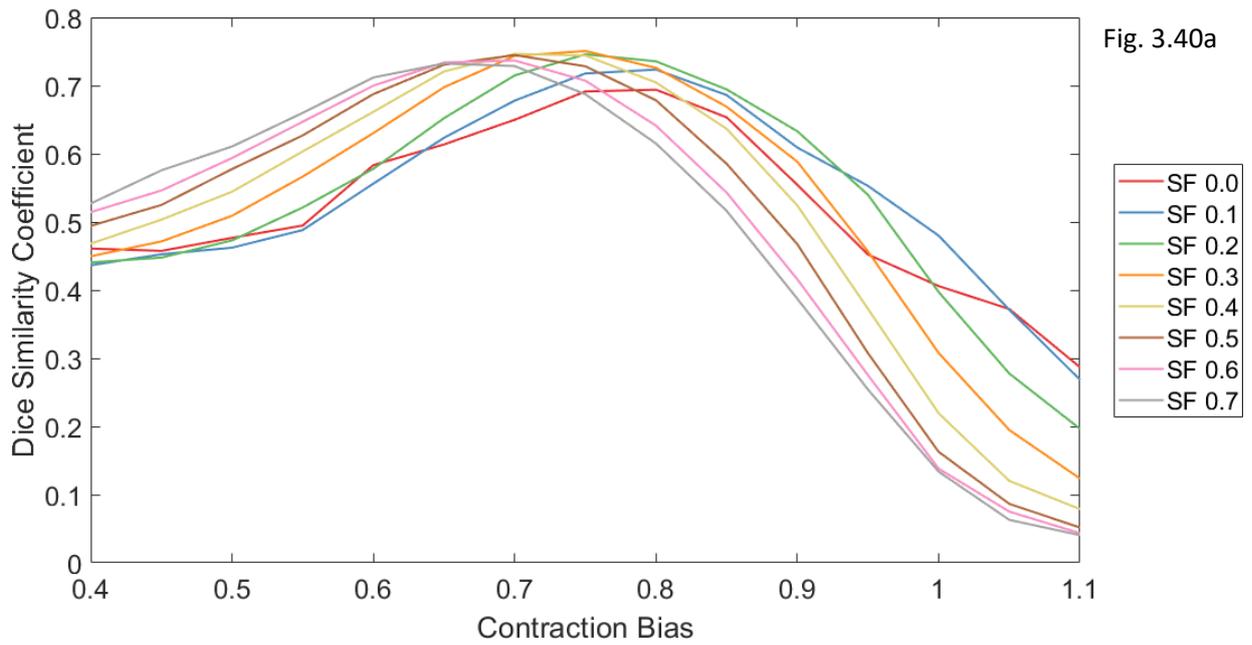


Figure 3.40. Mean Dice similarity coefficient for Chan-Vese segmentation of training run 1 displaying impact of different a) contraction bias b) smoothing factor.

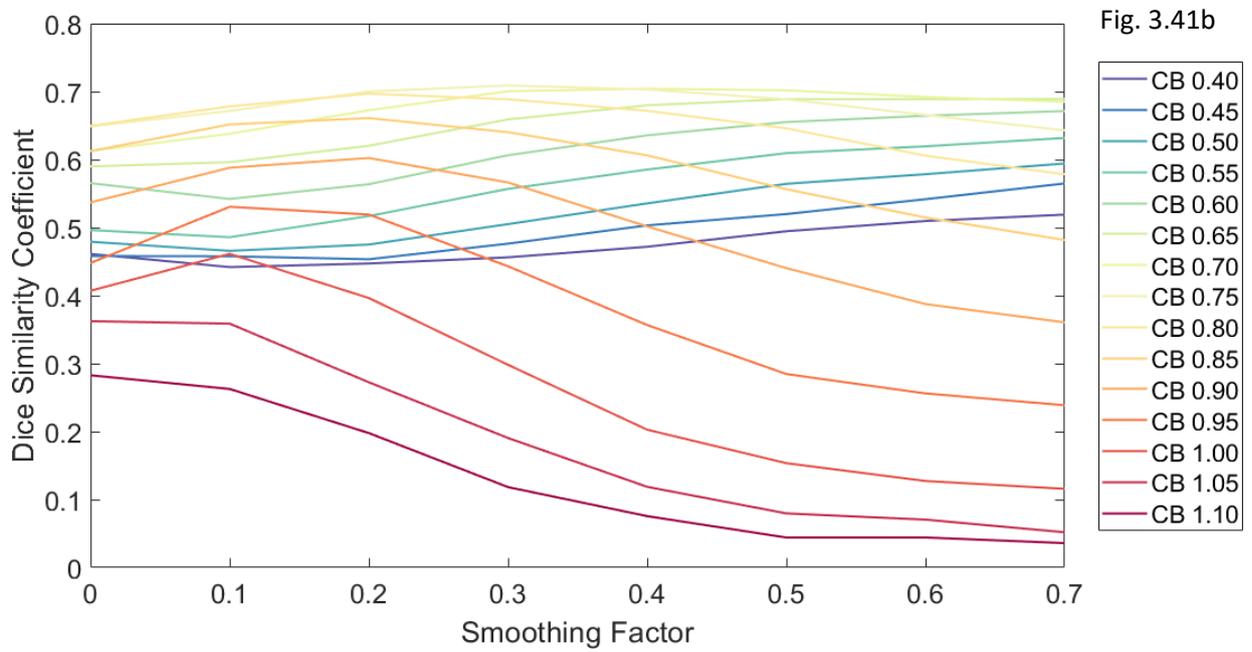
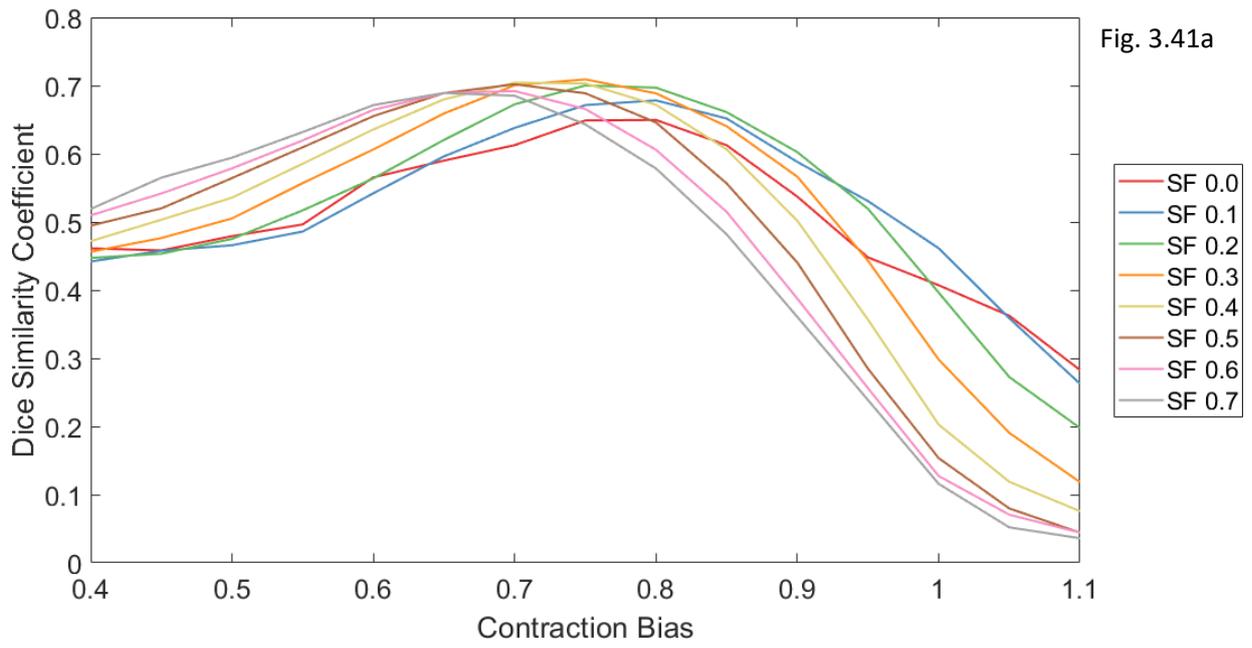


Figure 3.41. Mean Dice similarity coefficient for Chan-Vese segmentation of training run 2 displaying impact of different a) contraction bias b) smoothing factor.

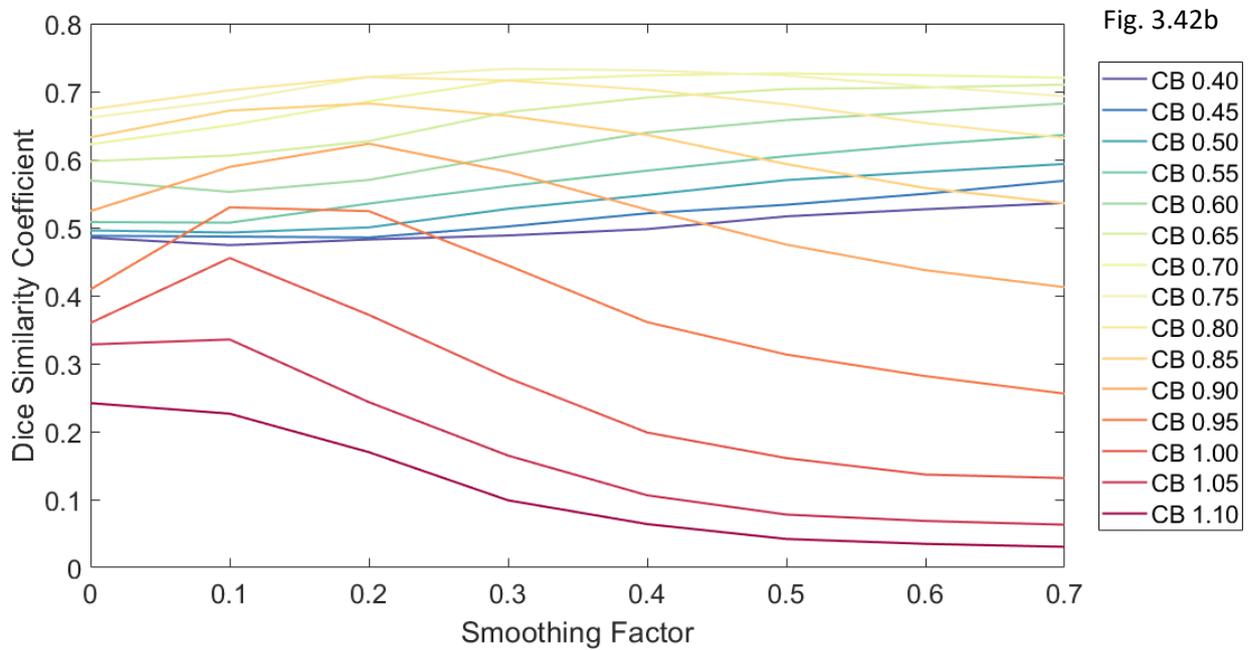
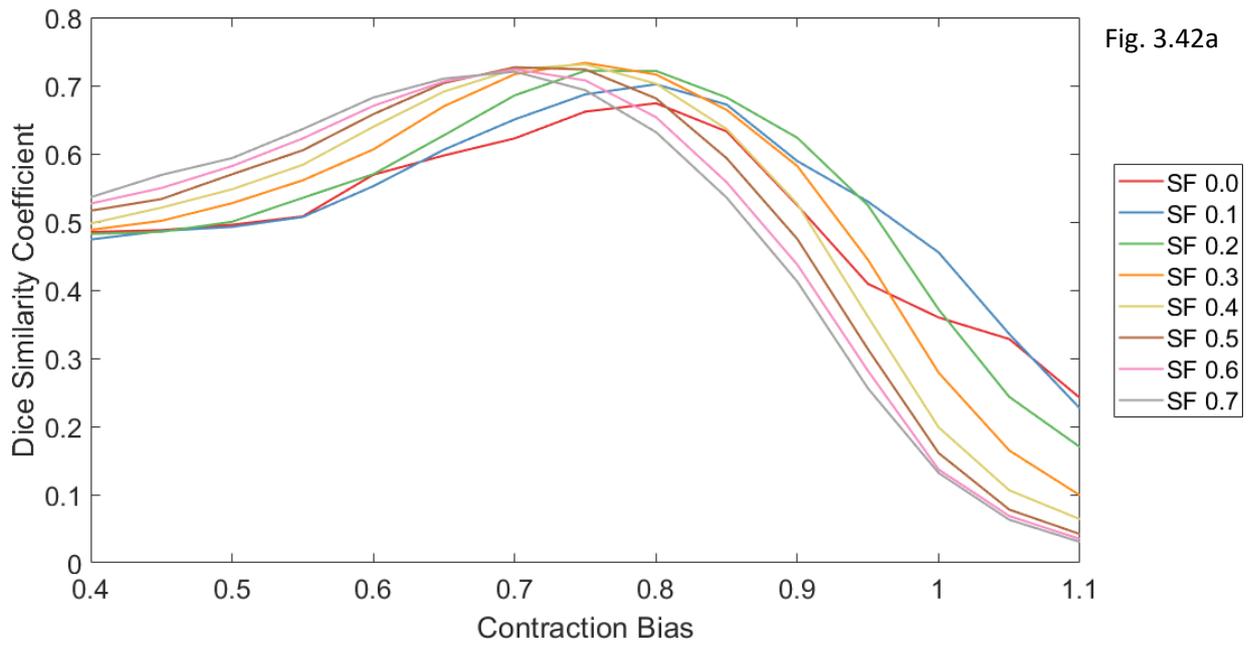


Figure 3.42. Mean Dice similarity coefficient for Chan-Vese segmentation of training run 3 displaying impact of different a) contraction bias b) smoothing factor.

3.12.2.4 Qualitative assessment of segmentation performance

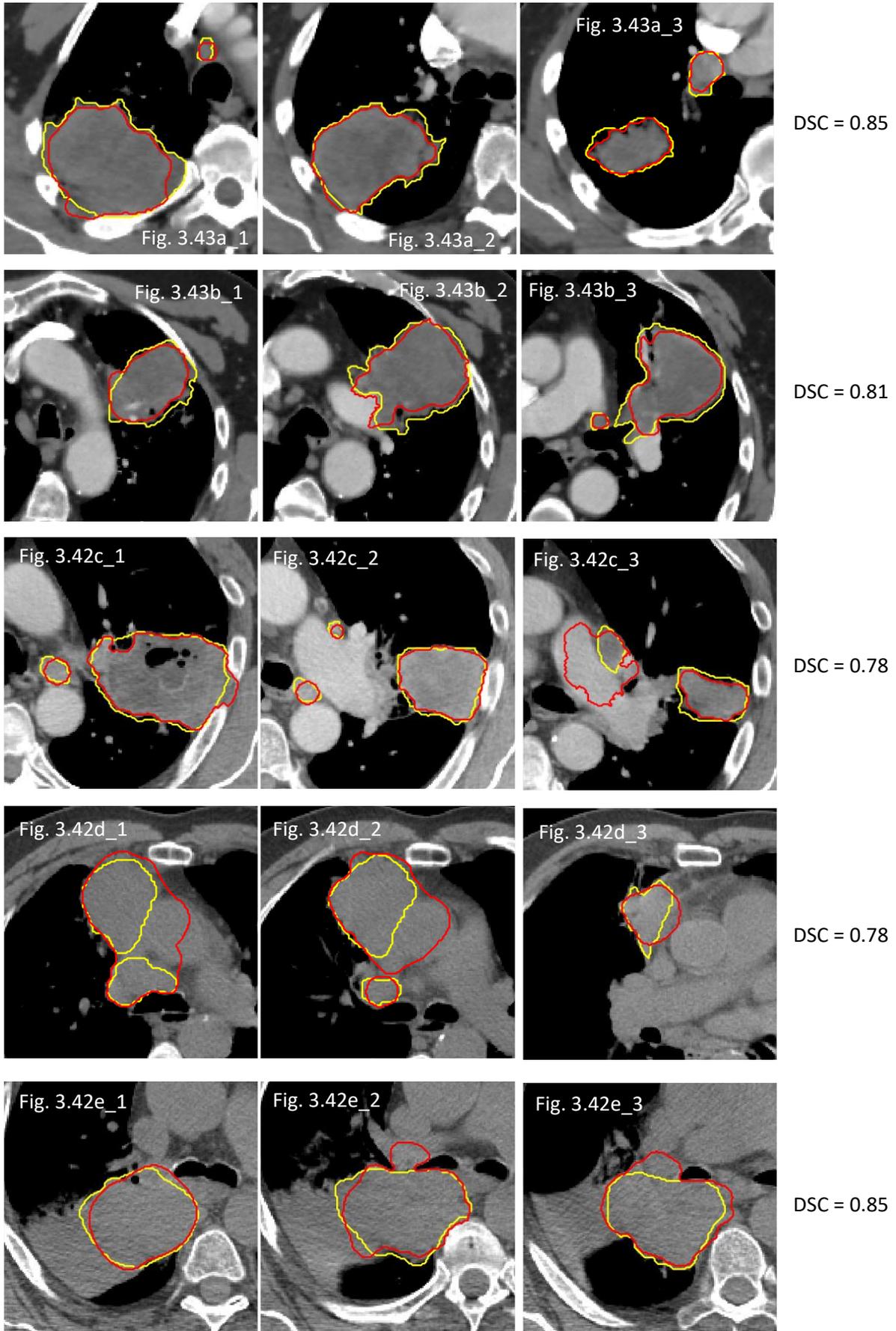
Smother contours were obtained with the Chan-Vese active contour segmentation as compared to those observed using the watershed approach. The Chan-Vese method performed well at GTV boundaries which were distinct, which included the tumour edge adjacent to the lung parenchyma, contrast-enhanced vessels and the ribs. In a number of cases, good separation of the tumour from the chest wall was also observed, although this may be dependent on the locality of the initial mask.

Where tumour was adjacent to lung parenchyma, it was noticed that the segmentation curve was generally placed on pixels of higher intensity values as compared to the reference contours, resulting in underestimation of the tumour at this front, the effect of which is more apparent when viewed on lung window levels. The effect of this was exacerbated in the presence of GGOs, which was seen to be contoured poorly. This was especially apparent in figure 3.43f where a large region of GGO was surrounding GTV, reflecting in a poor DSC score.

Its ability at including regions of cavitation was also location dependent. Cavities near the tumour boundary were generally excluded from the segmented region, whereas they tended to be included if situated towards the centre of the tumour (figures 3.43b–c).

Mixed performance was observed at the mediastinum. There were instances where good distinction was made between tumour and mediastinal fat, although for the small-volume mediastinal nodal disease, there was minimal change to the contours from the initialisation circles (figure 3.43c). The segmentation was limited in the presence of indistinct region boundaries. For example, leakage into non-contrast enhanced vessels was observed (figures 3.43d–e). The performance at adjacent lung collapse was also mixed, where good performance was seen in cases with good approximation of the true tumour edge by the initial mask.

There were also a minority of instances where the behaviour of the segmentation was seen to be erratic (figure 3.43c_3).



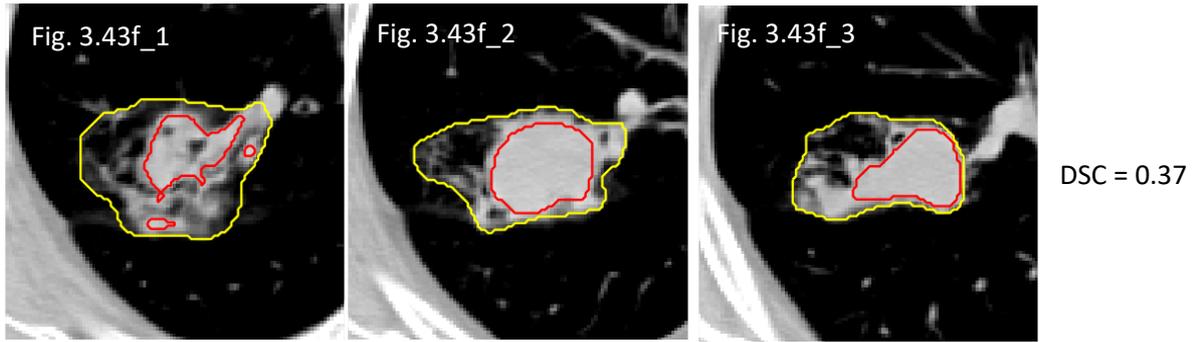


Figure 3.43. Chan-Vese active contour segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case)

3.12.2.5 Validation data

The application of the parameter settings from the training runs on the validation data is shown in figure 3.44 and table 3.10. Higher precision than recall was achieved for all three folds, with an estimated DSC of 0.73 ± 0.09 , recall of 0.69 ± 0.12 and precision of 0.84 ± 0.07 across the validation runs.

	Recall	Precision	DSC
Validation Run 1	0.65 ± 0.15	0.82 ± 0.08	0.69 ± 0.10
Validation Run 2	0.74 ± 0.10	0.86 ± 0.05	0.78 ± 0.07
Validation Run 3	0.68 ± 0.07	0.85 ± 0.08	0.76 ± 0.06
Aggregate across three runs	0.69 ± 0.12	0.84 ± 0.07	0.73 ± 0.09

Table 3.10. Performance of Chan-Vese active contour segmentation using contraction bias of 0.75 and smoothing factor of 0.3 on the validation datasets.

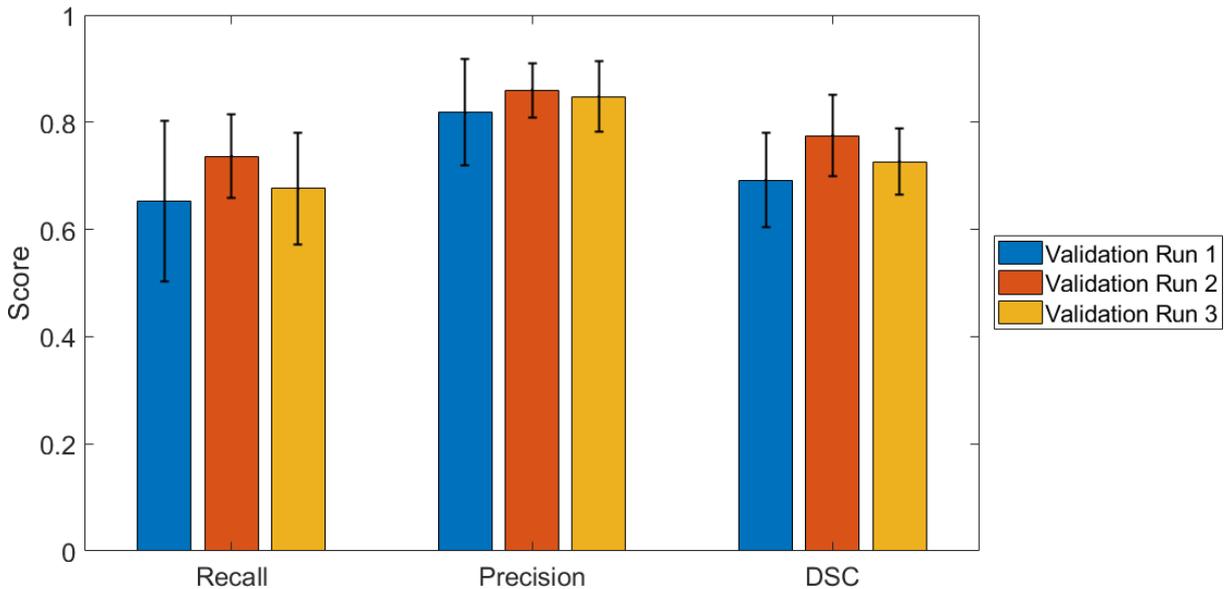


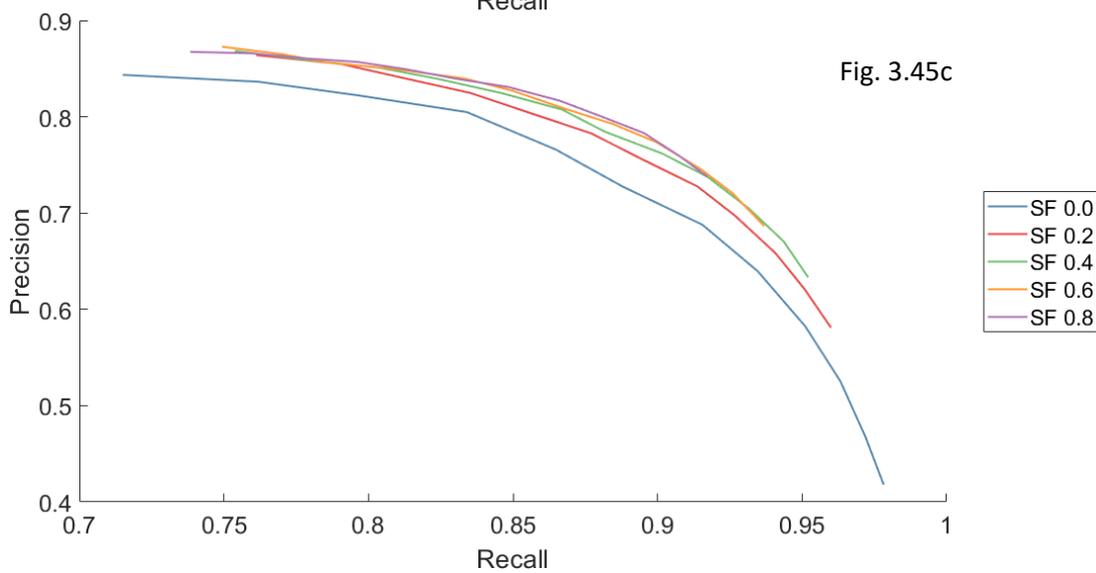
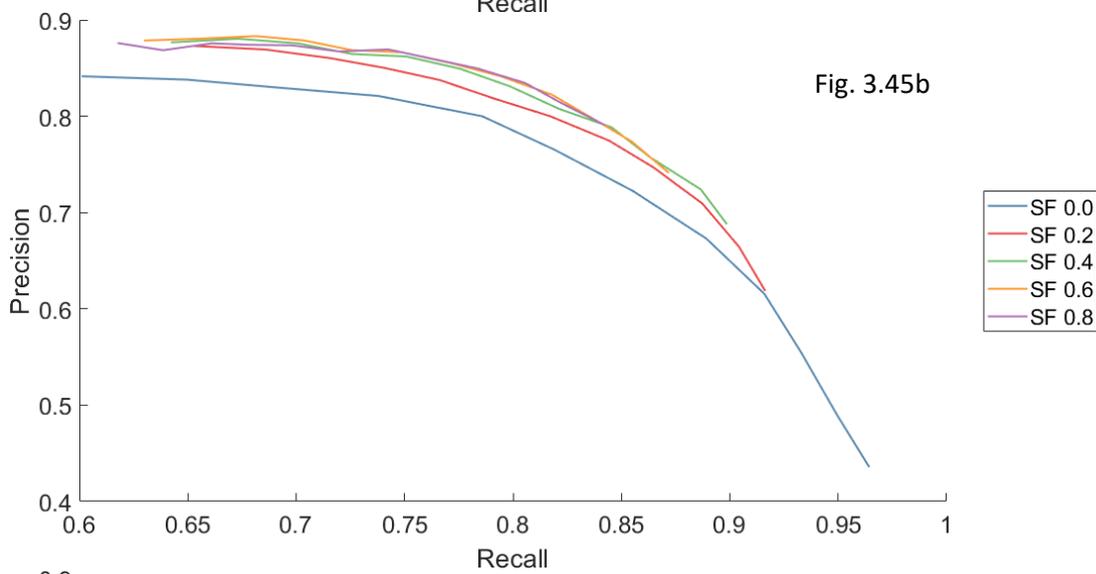
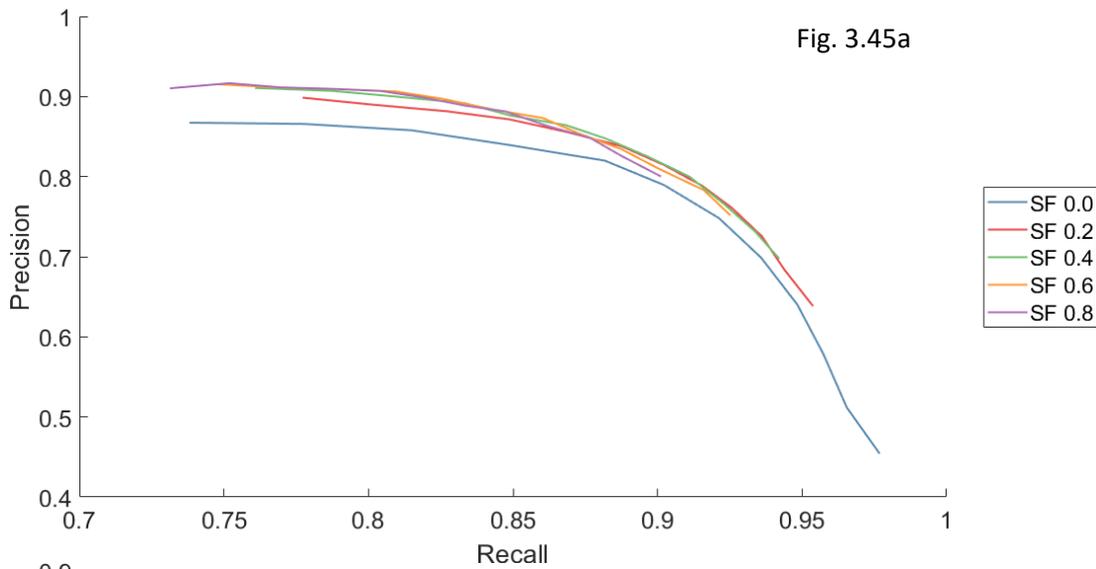
Figure 3.44. Mean performance of Chan-Vese active contour segmentation on each fold of the validation datasets (error bars represent standard deviation).

3.12.3 Edge-based active contour

3.12.3.1 Evaluation of mask initialisation and parameter settings on subsample cases

The precision recall curves for the three different mask initialisation is shown in figure 3.45. Across the different initialisation masks, a low smoothing factor of 0 was associated with poorer precision and recall. Additionally, recall appeared to be affected by the smoothing factor to a greater extent than Chan-Vese active contour segmentation. High smoothing factor was associated with lower values of recall. Better precision and poorer recall scores were achieved with larger contraction bias.

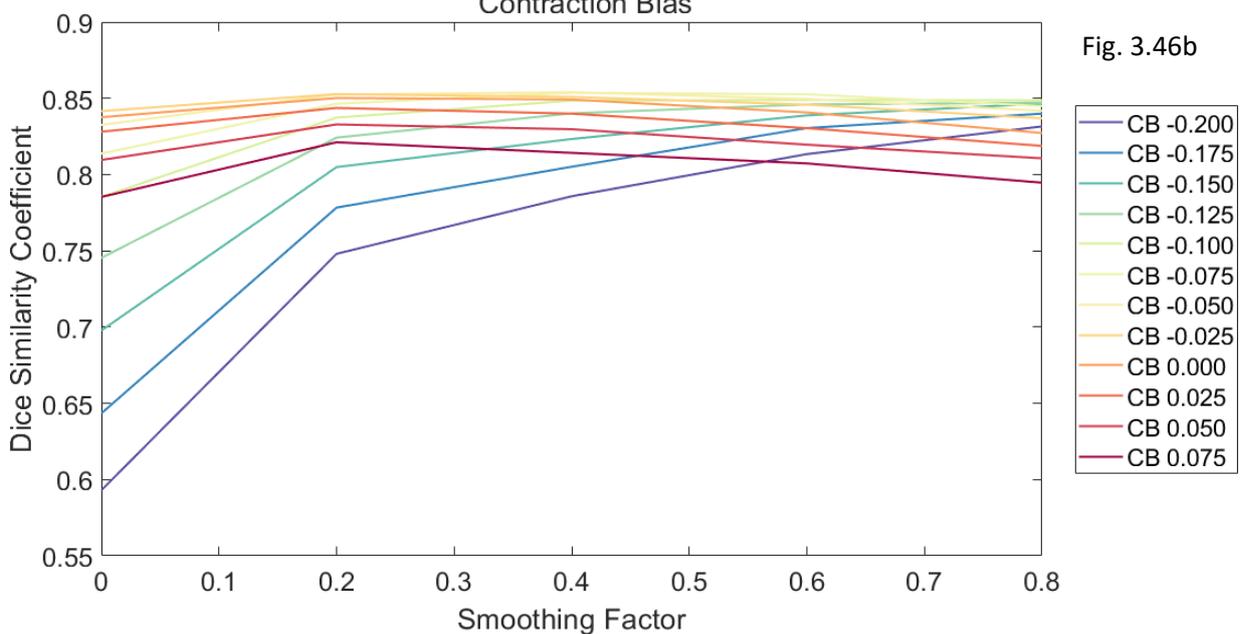
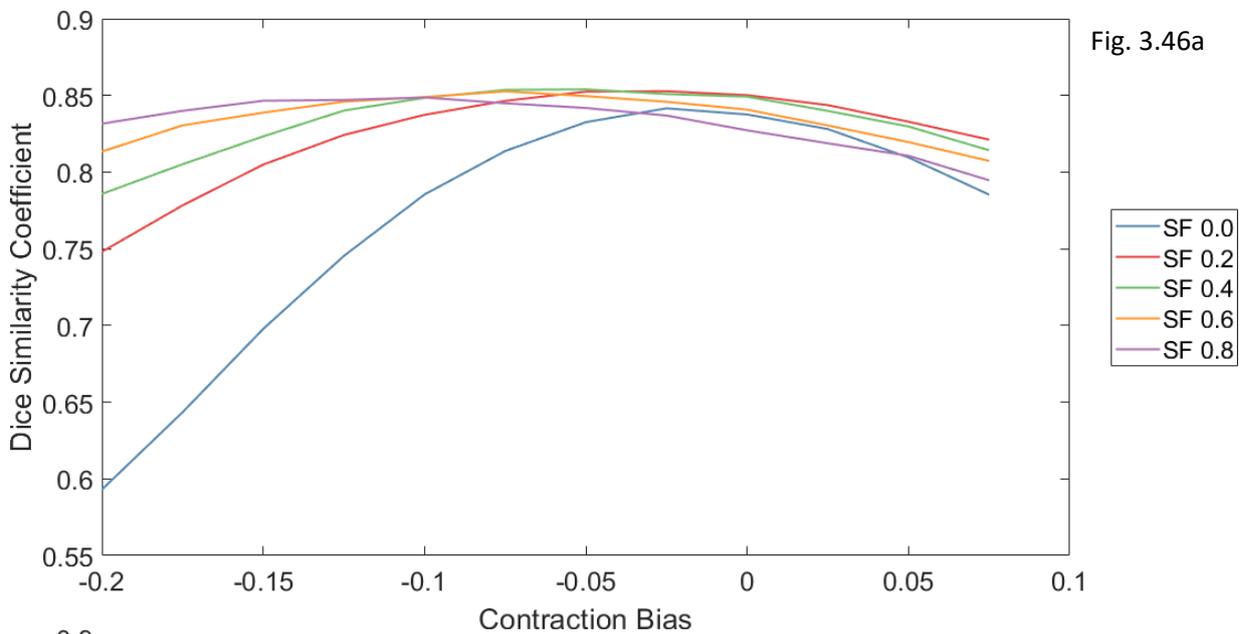
Across the parameter settings, the use of the convex hull masks was associated with higher precision compared to the initialisation with the circle masks. Lower rates of recall were achieved for the eroded circle mask compared to the two other initialisations.



Figures 3.45a – c. Mean precision versus recall plots for 18 subsample cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Convex polygon with 4-pixel erosion, b) Circle with 4-pixel erosion, c) Circle with no erosion (Edge-based active contour segmentation).

3.12.3.1.1 Initial mask: Convex polygon with 4-pixel erosion

Evaluation of the DSC plots for the convex polygon initialisation shows fairly good DSC scores of up to 0.85. This was dependent on both contraction bias as well as the smoothing factor, where higher scores were seen between a range of -0.1 to 0, and 0.2 to 0.6 respectively.



Figures 3.46a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Convex polygon initial mask with 4-pixel erosion; Edge-based active contour algorithm).

The peak DSC for this initialisation was 0.85 ± 0.09 , with an optimal contraction bias at -0.05 and smoothing factor of 0.4.

3.12.3.1.2 Initial mask: Circle with 4-pixel erosion

As compared to the initialisation with the convex hull, lower DSC scores were achieved with the use of the eroded circle mask. The best DSC scores were observed for contraction bias between -0.15 and -0.05, which was also dependent on the smoothing factor (between 0.2 to 0.8).

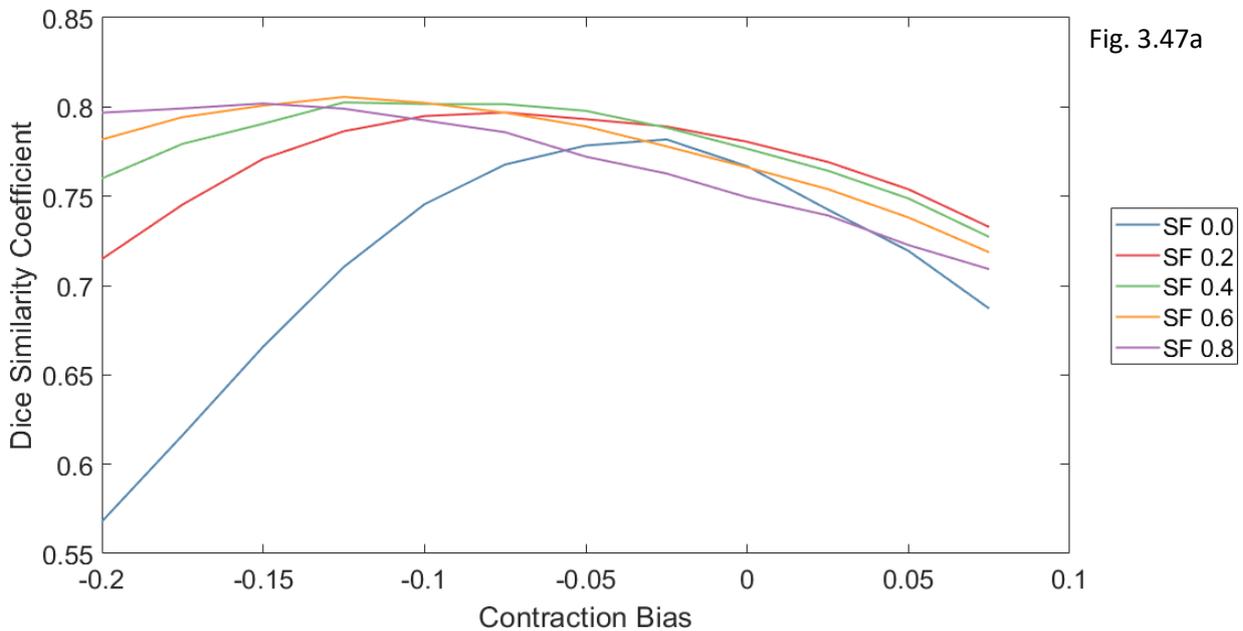


Fig. 3.47a

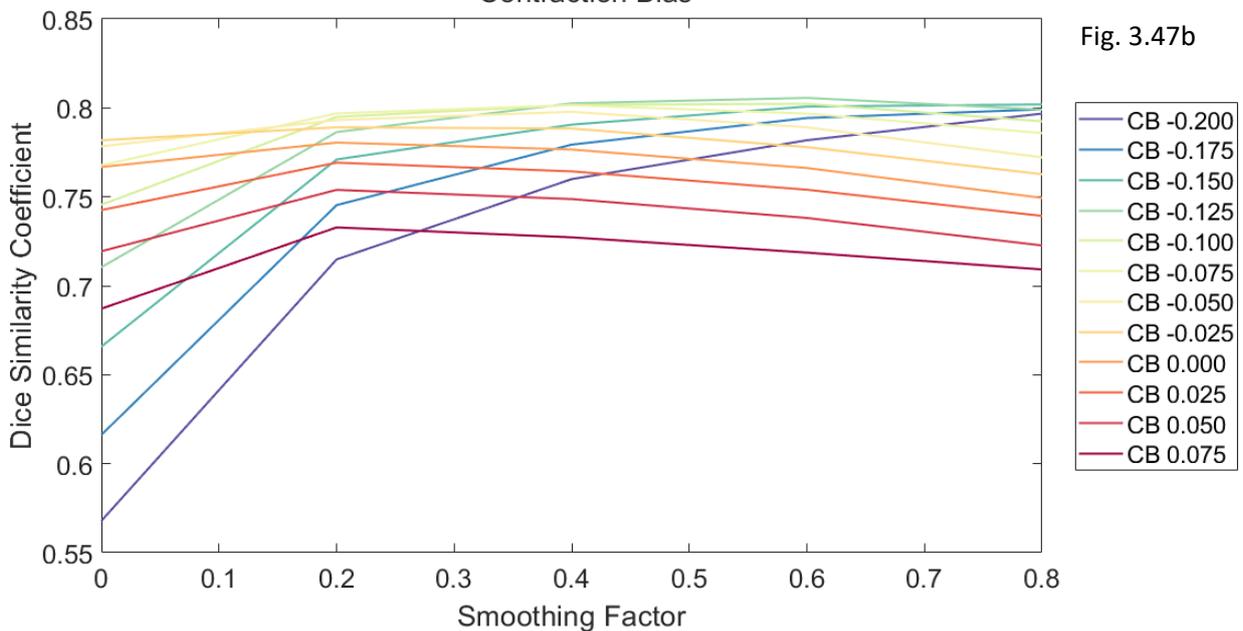


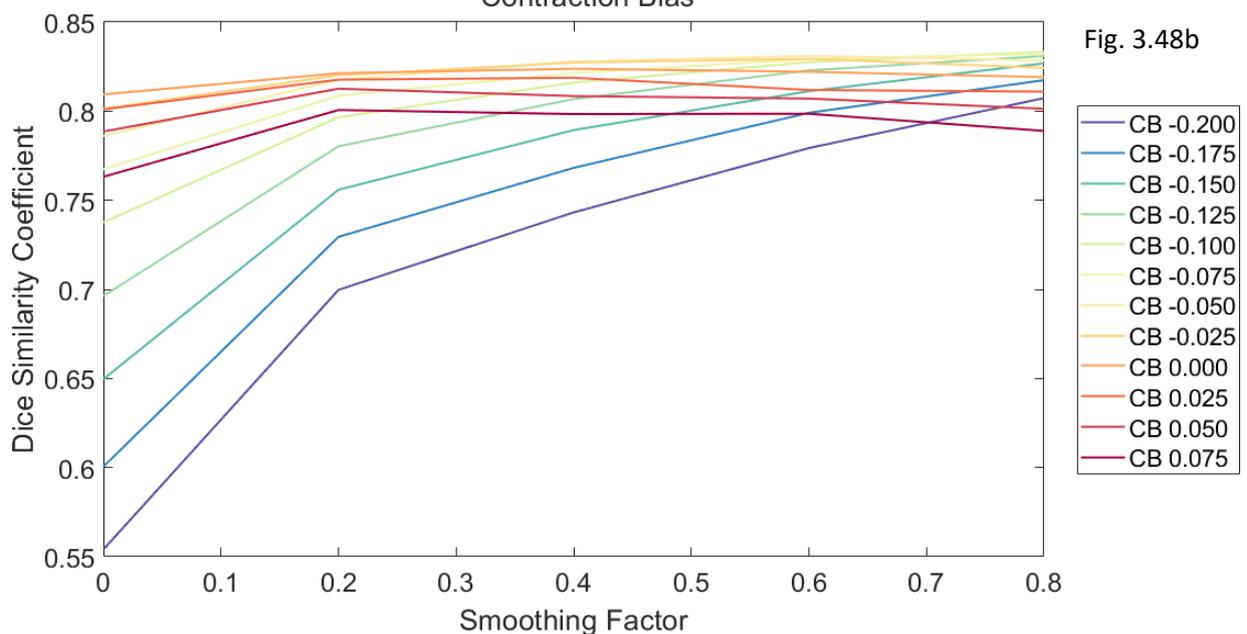
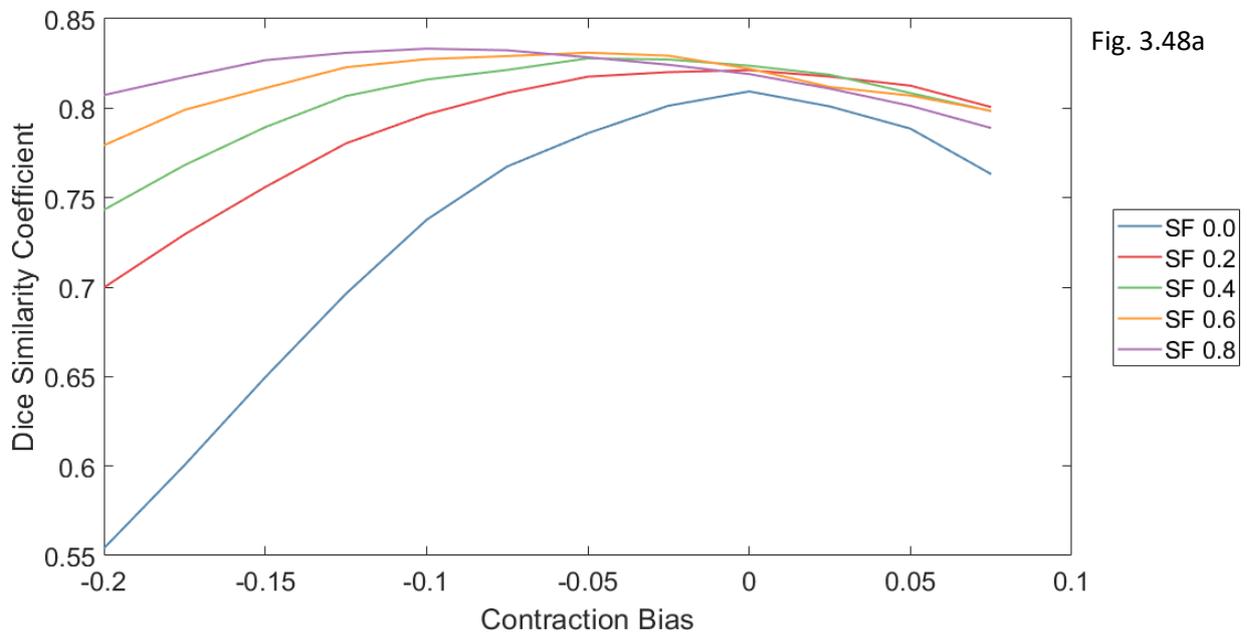
Fig. 3.47b

Figures 3.47a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Circle initial mask with 4-pixel erosion; Edge-based active contour algorithm).

The highest achieved DSC for this initialisation was 0.81 ± 0.07 , with an optimal contraction bias at -0.125 and smoothing factor of 0.6.

3.12.3.1.3 Initial mask: Circle with no erosion

In the use of a non-eroded initial circle mask, the best performance was observed at a contraction bias between -0.15 and 0. Similarly, this was also dependent on the smoothing factor, where higher DSC was observed between 0.4 and 0.8.



Figures 3.48a – b. Effect of contraction bias and smoothing factor on the mean Dice similarity coefficient for 18 subsample cases (Circle initial mask with no erosion; Edge-based active contour algorithm).

The highest achieved DSC for this initialisation was at 0.83 ± 0.09 , with an optimal contraction bias at -0.125 and smoothing factor of 0.8.

The quantitative performance for the different initial masks at their respective optimal settings is summarised in table 3.11. Similar recall and precision scores were achieved using the eroded convex hull and eroded circle masks, whilst the circle initial mask had higher mean recall compared to precision. All three methods of initialisation had a mean DSC of greater than 0.8.

	Optimal contraction bias	Optimal smoothing factor	Recall	Precision	DSC
Convex hull (4-pixel erosion)	-0.05	0.4	0.87 ± 0.09	0.86 ± 0.07	0.85 ± 0.09
Circle (4-pixel erosion)	-0.125	0.6	0.82 ± 0.09	0.82 ± 0.07	0.81 ± 0.07
Circle (No erosion)	-0.125	0.8	0.88 ± 0.06	0.80 ± 0.05	0.83 ± 0.09

Table 3.11. Optimal parameter settings for Edge-based active contour segmentation with results of performance.

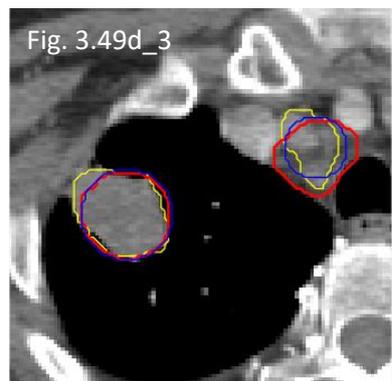
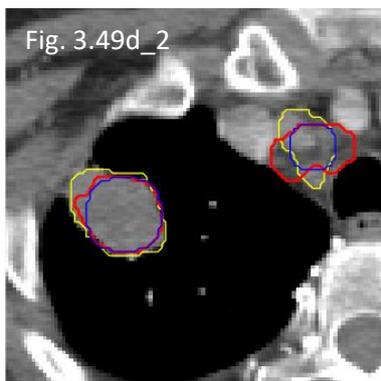
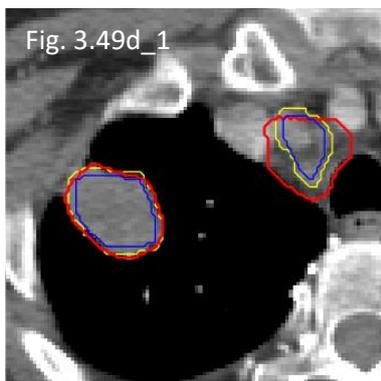
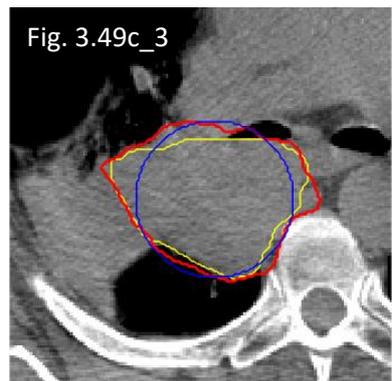
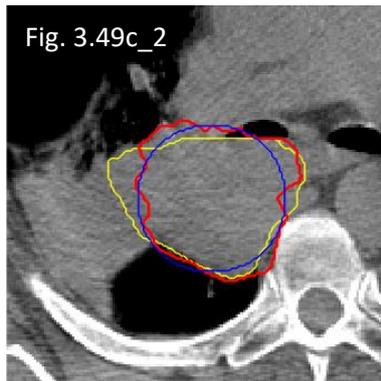
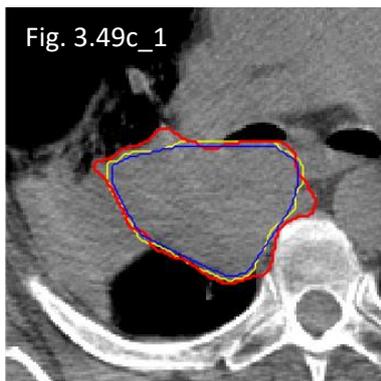
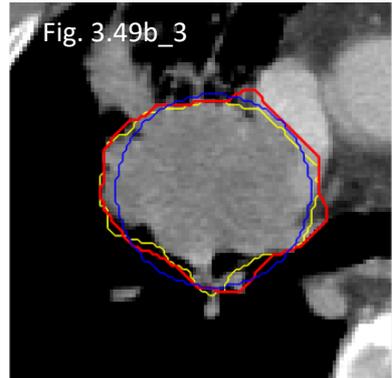
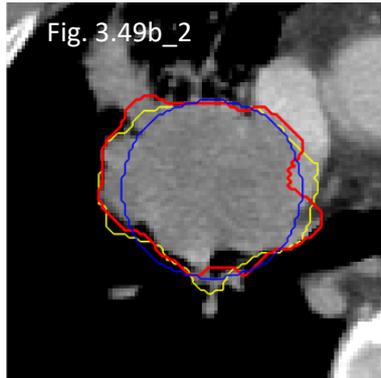
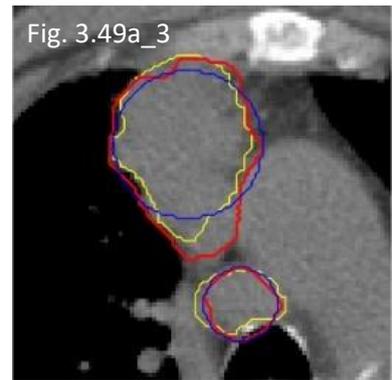
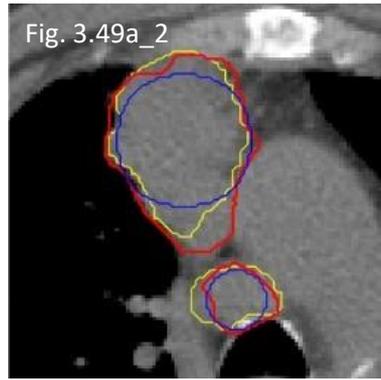
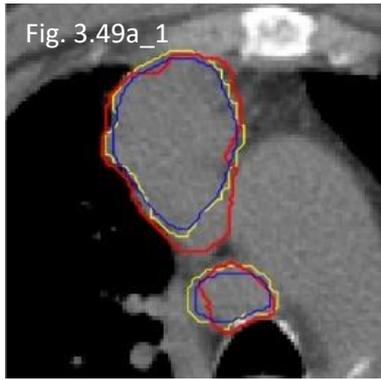
3.12.3.2 Qualitative assessment of segmentation performance

In spite of the good quantitative performance, assessment of the cases qualitatively revealed some differences in the segmentation behaviour between the different forms of initialisation.

The issue of overfitting with the use of the convex hull polygon was case dependent. For tumours with less irregularity and concavity, the initialising masks were closely adherent to the shape of the tumour (figure 3.49a_1) as compared to the circle masks where this was less of an issue (figures 3.49a_2 and 3.49a_3). The segmentation performance was observed to be case dependent as well. As shown in figures 3.49b_1 – 3, the final segmentation contours appeared to be minimally affected by the initialisation masks, with similar segmentation obtained using the different masks. Conversely for most other cases, in addition to the underlying edge information in the image, the final locations of the contours were affected by the position of the initial mask.

Another example illustrating this effect is shown in figures 3.49e_1 – 3. On this slice, all three initialisations were dissimilar to the submitted contour, which was dumbbell-shaped with a link between the tumour in the lung and the hilar nodal disease. Although the segmentation included both tumour regions when the convex hull mask was applied, regions of the lung parenchyma between the two regions of disease were also included, where the segmentation boundary was in close adherence to that of the initial mask. This effect was also observed when the two circle masks were used. As compared to the Chan-Vese approach, there was less tendency for the contour to split and as a result, lung parenchyma with regions of low intensity was not excluded.

The segmentation using the eroded circle mask did not encompass either the primary or nodal disease in full, as the initialisation was at a distance from the true tumour edge. This effect was also observed through the poor coverage of the nodal disease when the non-eroded circle was applied. Conversely, the non-eroded circle mask produced a segmentation with good coverage of the primary disease, as the initialisation edge was closer to the tumour boundary than that of the circle eroded mask. This demonstrates the tendency of the contour in locating edges close to the position of the initialisation. Additionally, the segmentation of the primary disease was very similar to that produced using the convex hull, suggesting that in addition to the initialisation position, edge information also contributed to the final location of the segmentation.



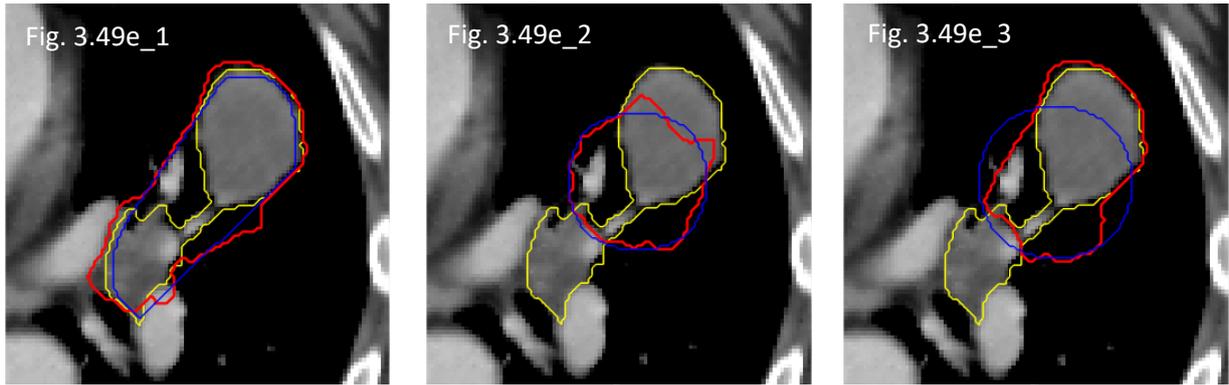


Figure 3.49. Examples (cases a to e) showing variation in edge-based active contour segmentation results using different initialisations; suffix _1) Convex hull with 4-pixel erosion; suffix _2) Circle with 4-pixel erosion; suffix _3) Circle with area equivalent to the submitted contour. Red outline – Active contour segmentation, Yellow outline – submitted contours, Blue outline – Initial mask.

The impact of parameter tuning on this algorithm was different to that observed with the Chan-Vese approach. Although the edge-based method was still affected by parameter tuning, there were less occurrences where the final contours were of close resemblance to the initial masks, even for small-volume nodal disease and initialisation through the eroded convex hull.

For most cases, better segmentation conformity was obtained when the initialisation edge was closer to the tumour boundary. Although the best performance was observed when initialisation was performed with the convex hull, it was felt that this initialisation would not be a fair comparison to the other segmentation techniques, due to the potential issue of overfitting. As the non-eroded circle mask produced acceptable results in the majority of situations, this was used in the rest of the study.

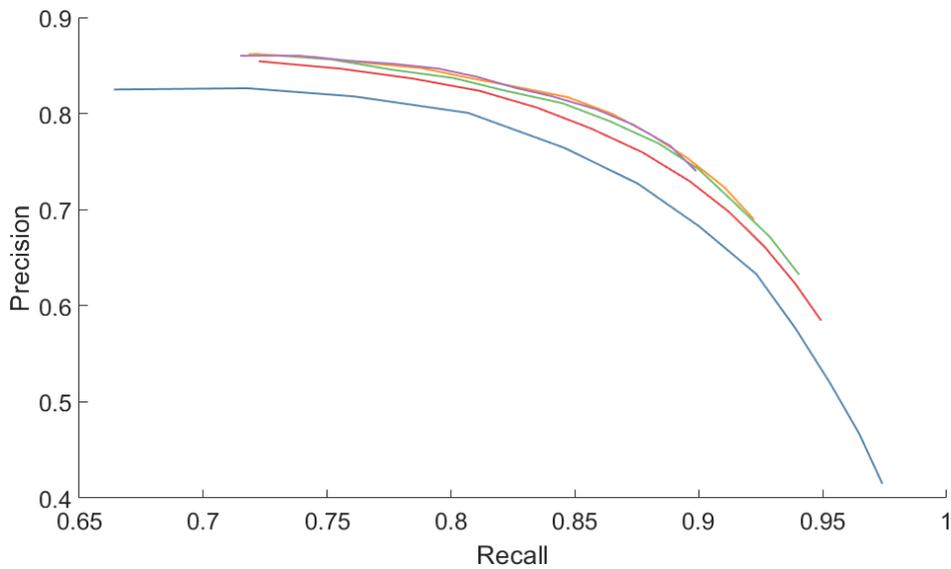
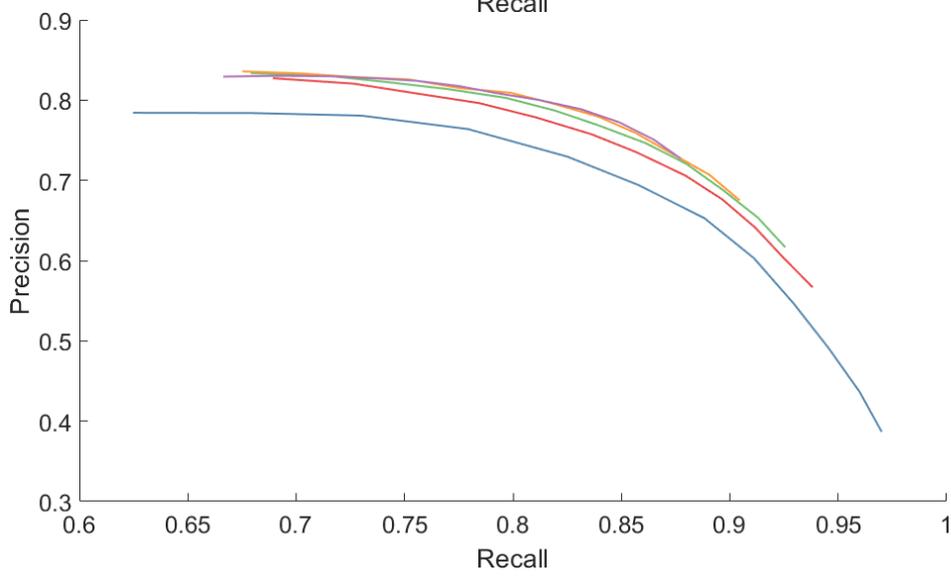
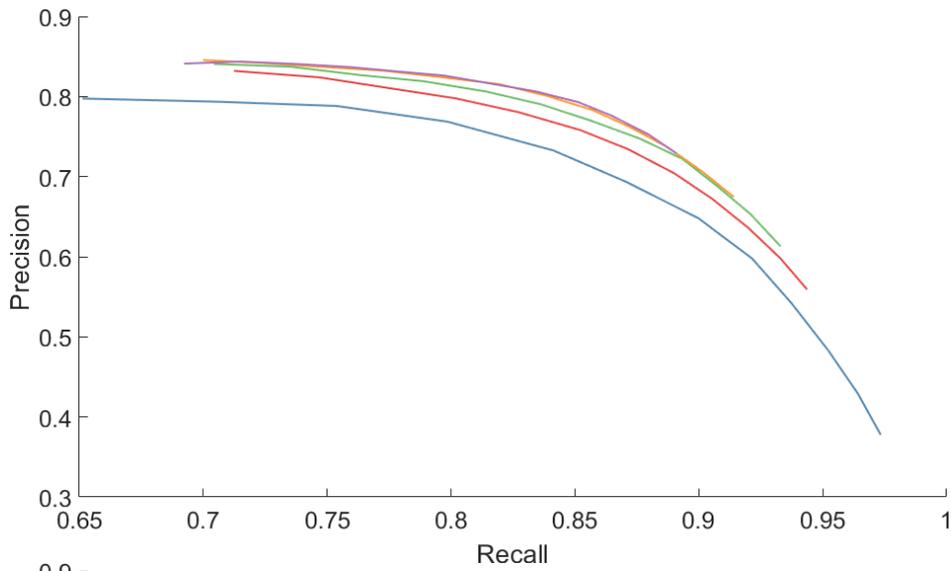
3.12.3.3 Training using cross validation folds for parameter selection

Assessment of the contraction bias for training runs 1 and 2 revealed that the highest DSC was achieved between a range of -0.15 and -0.1 with a smoothing factor of 0.8. However, the selection of the contraction bias in this range was very sensitive to the smoothing factor, with a large reduction in the DSC if the smoothing factor were decreased. Rather than to select the parameter based on the highest achieved DSC, it was decided that the optimum parameter settings should be chosen based not only on the DSC, but also on how sensitive the settings were to changes, such that the selected parameters would be best fit for most cases.

With this taken into account, it was decided that the optimal contraction bias was between -0.1 to -0.05, and a smoothing factor of 0.6, as shown in table 3.12, where similar performance was seen for all three training folds. From this, a contraction bias of -0.075 with a smoothing factor of 0.6 were chosen.

	Optimal contraction bias	Optimal smoothing factor	Recall	Precision	DSC
Training Run 1	-0.100	0.6	0.857 ± 0.067	0.783 ± 0.054	0.809 ± 0.056
	-0.075		0.839 ± 0.074	0.800 ± 0.052	0.810 ± 0.058
	-0.050		0.820 ± 0.076	0.815 ± 0.051	0.809 ± 0.060
Training Run 2	-0.100	0.6	0.840 ± 0.065	0.779 ± 0.054	0.799 ± 0.052
	-0.075		0.819 ± 0.073	0.795 ± 0.052	0.798 ± 0.055
	-0.050		0.799 ± 0.074	0.809 ± 0.049	0.795 ± 0.057
Training Run 3	-0.100	0.6	0.865 ± 0.051	0.800 ± 0.039	0.824 ± 0.039
	-0.075		0.847 ± 0.058	0.817 ± 0.037	0.825 ± 0.044
	-0.050		0.827 ± 0.062	0.828 ± 0.037	0.820 ± 0.049

Table 3.12. Optimal parameter settings for edge-based active contour segmentation with results of performance.



Figures 3.50a – c. Mean precision vs recall plots for training cases displaying the impact of variation of the contraction bias for each of the overlapping plots of different smoothing factors. a) Training run 1, b) Training run 2, c) Training run 3 (Edge-based active contour segmentation).

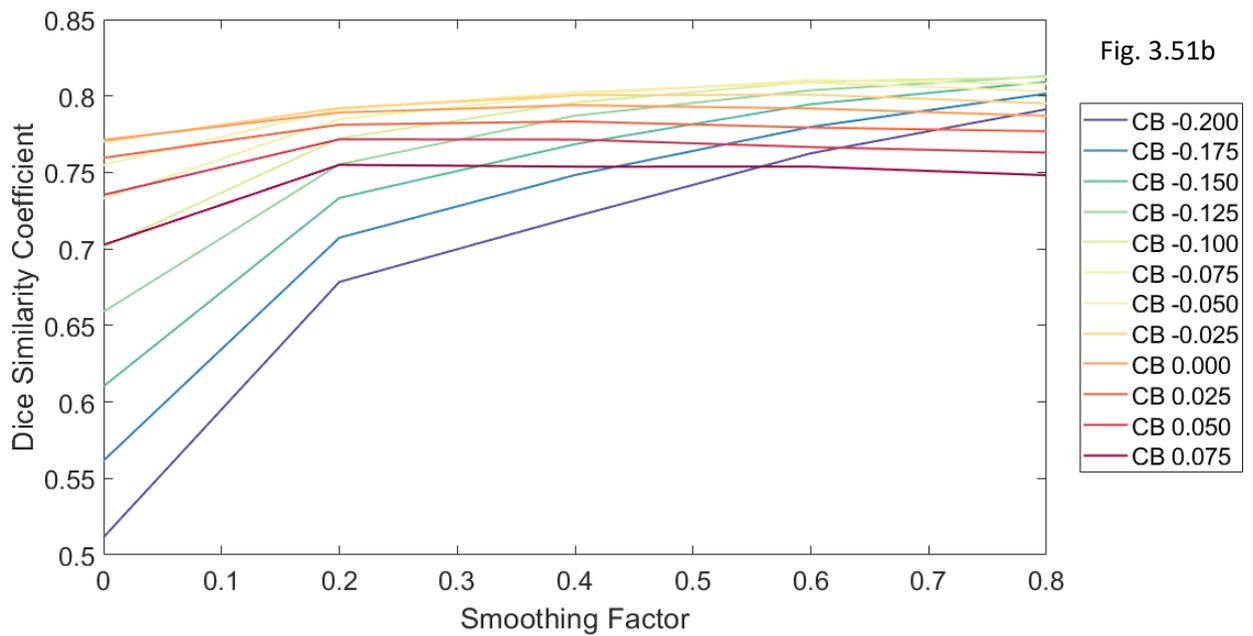
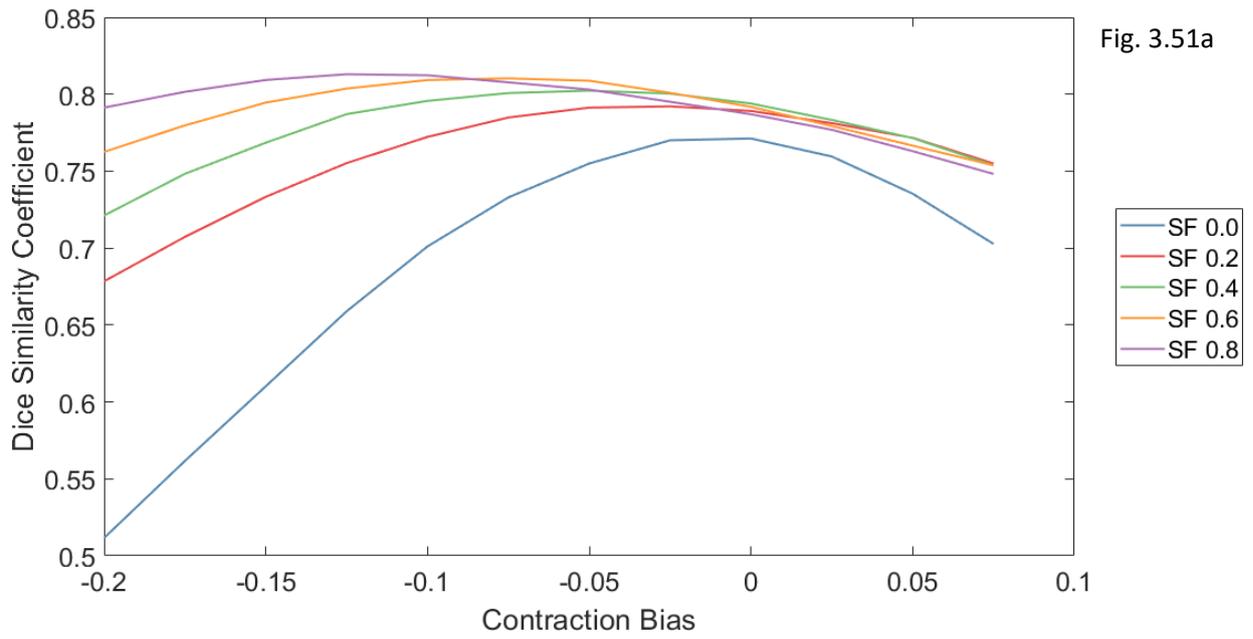


Figure 3.51. Mean Dice similarity coefficient for edge-based active contour segmentation of training run 1 displaying impact of different a) contraction bias b) smoothing factor.

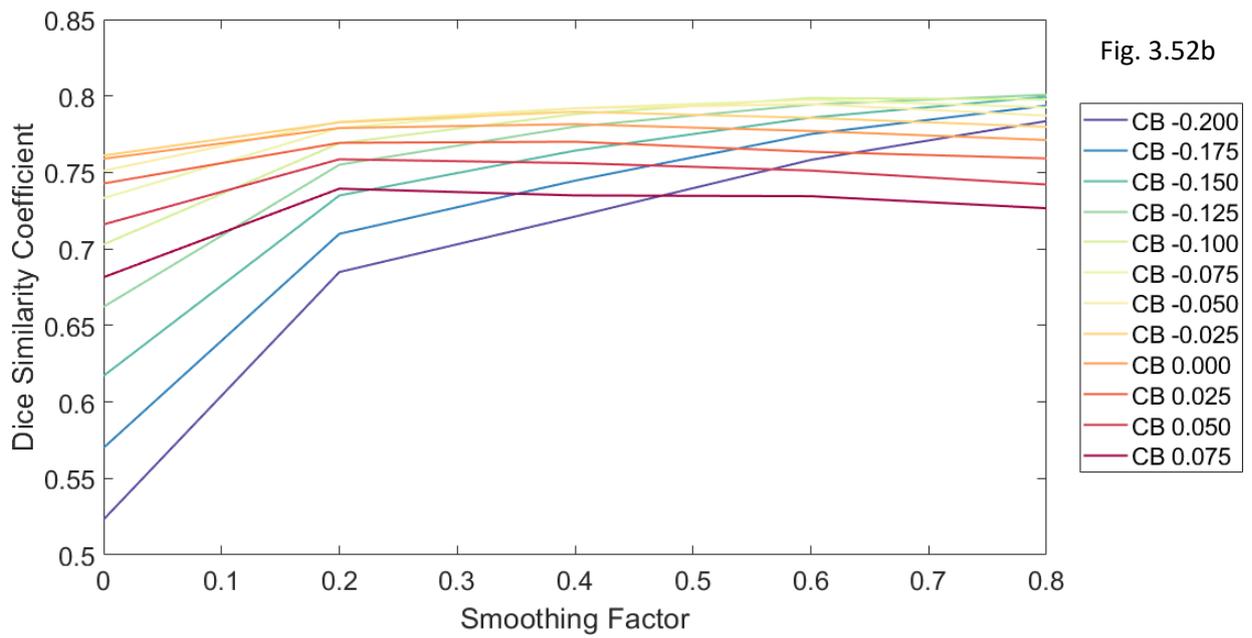
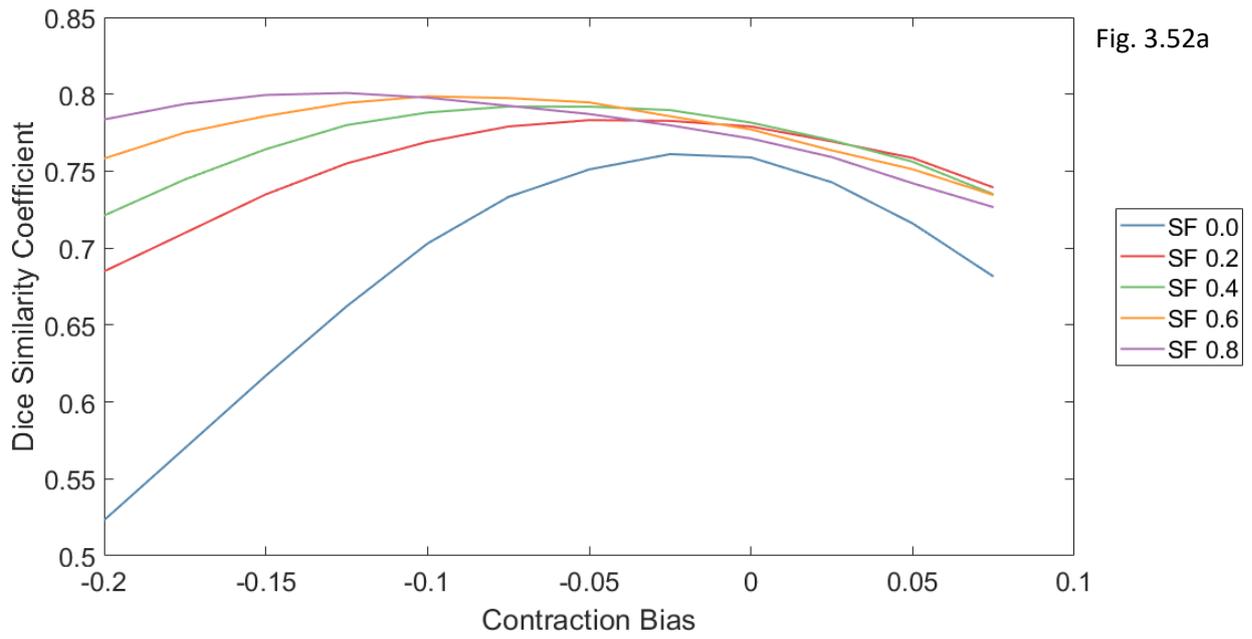


Figure 3.52. Mean Dice similarity coefficient for edge-based active contour segmentation of training run 2 displaying impact of different a) contraction bias b) smoothing factor.

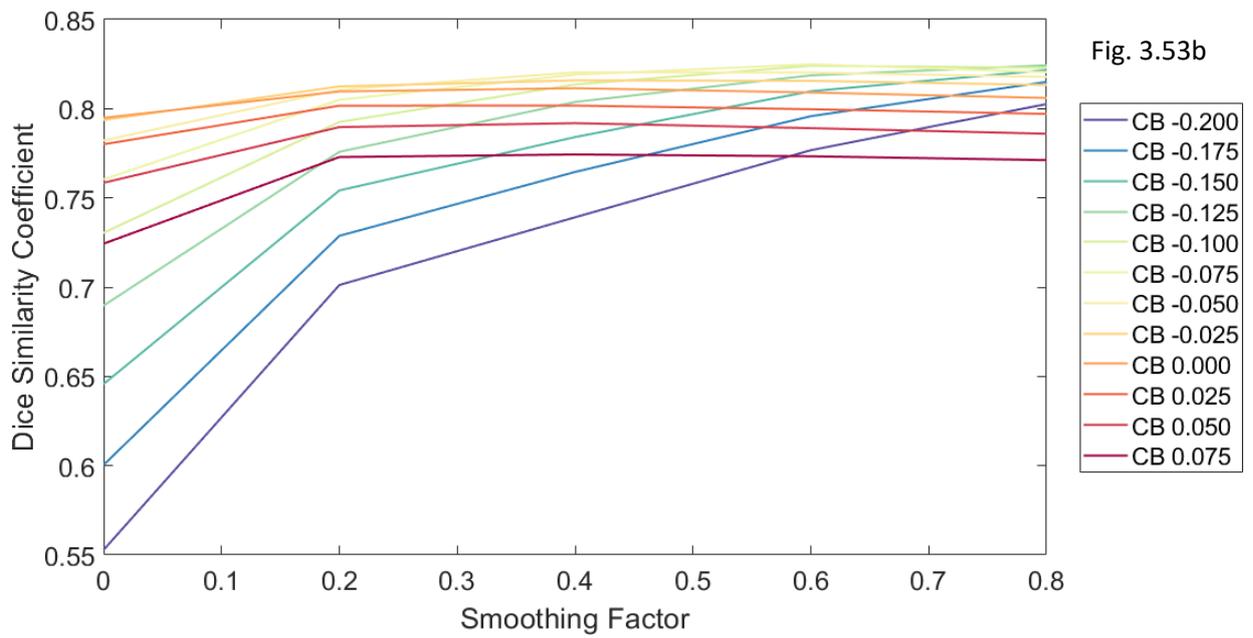
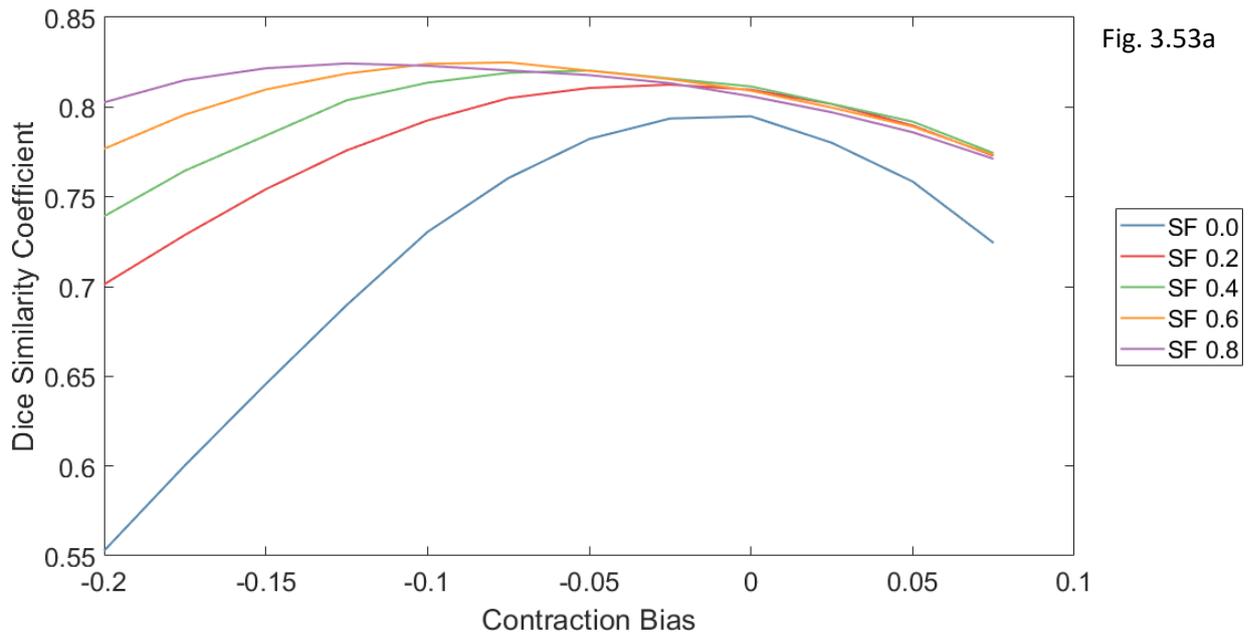


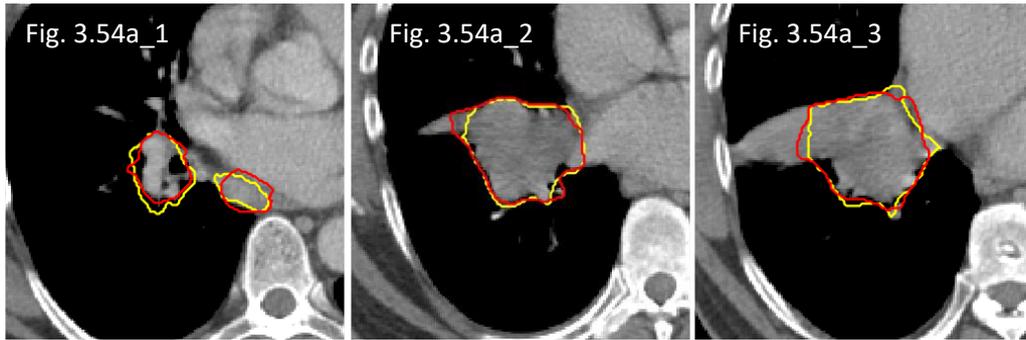
Figure 3.53. Mean Dice similarity coefficient for edge-based active contour segmentation of training run 3 displaying impact of different a) contraction bias b) smoothing factor.

3.12.3.4 Qualitative assessment of segmentation performance

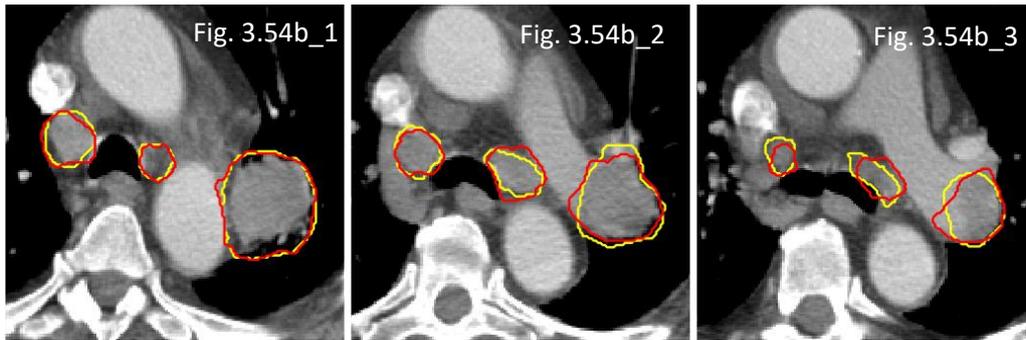
Like the Chan-Vese approach, the edge-based active contour produced segmentation that were generally smooth. Clinically acceptable segmentation achieved in a number of cases, examples of which are shown in figure 3.54, where distinction between tumour and contrast-enhanced vessels, ribs, chest wall musculature and lung parenchyma was seen. It was also able to perform the separation of adjacent lung collapse, as well as non-contrast enhanced vessels and mediastinal fat. Nevertheless, some segmentation had some regions that did not appear to correspond to the tumour edge information. This is apparent in figure 3.54c_3 where the segmentation included a portion of the main bronchi, as well as non-contrast enhanced superior vena cava. Figure 3.54d_1 shows another example of the contour being placed at a distance from the edge boundary (green arrow), and in 3.54f_3 there was leakage of the contour into lung parenchyma (orange arrow).

With this approach, regions of GGOs were seen to be appropriately included in the delineation in most parts (figure 3.54f). Additionally, cavitations were also handled well, at locations in the centre and at the periphery of the tumour (figure 3.54e).

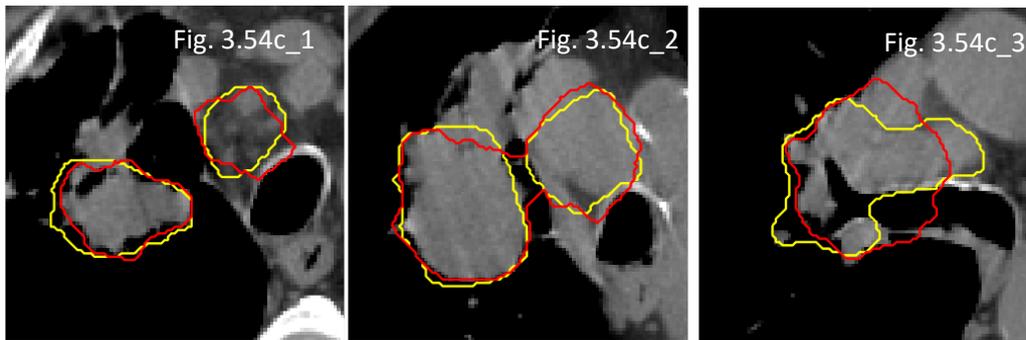
Unlike the Chan-Vese approach, there was less underestimation of the tumour boundary at the junction of the lung parenchyma.



DSC = 0.88



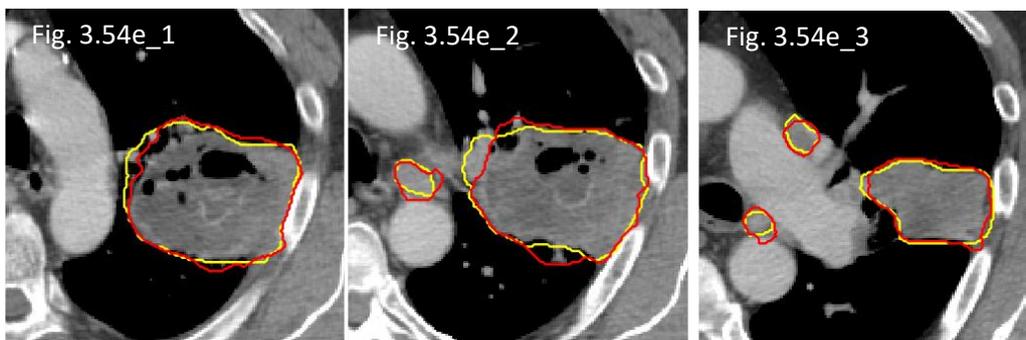
DSC = 0.83



DSC = 0.82



DSC = 0.78



DSC = 0.86

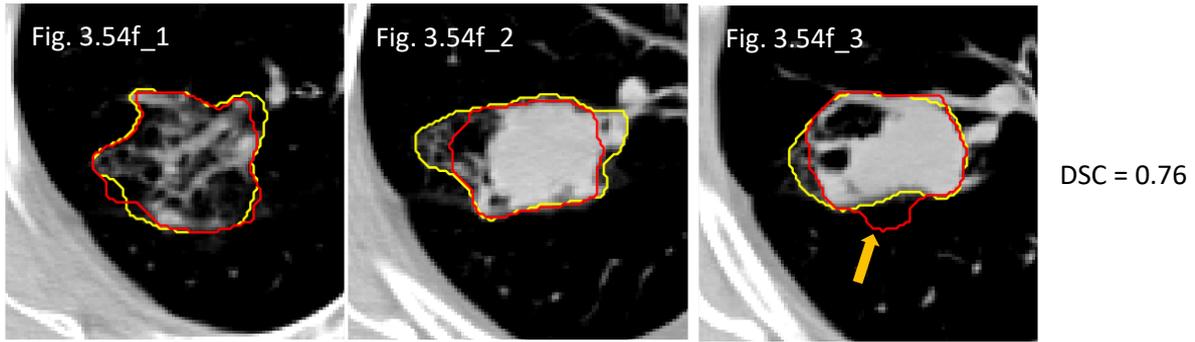


Figure 3.54. Edge-based active contour segmentation results (red outlines) for six representative training cases (a – f) versus reference contours (yellow outlines), with corresponding DSC for each case. Green and orange arrows indicate regions of poor conformity. (suffix _1 to 3 represent different axial slices for each case)

3.12.3.5 Validation data

The validation folds were processed using a contraction bias of -0.075 and smoothing factor of 0.6 (figure 3.55 and table 3.13). Although higher mean recall than precision was achieved for all three folds, the difference is smaller than the previous two segmentation techniques. The estimated DSC was 0.81 ± 0.05 across the three validation runs, with associated mean recall and precision of 0.84 ± 0.07 and 0.80 ± 0.05 respectively.

	Recall	Precision	DSC
Validation Run 1	0.83 ± 0.06	0.81 ± 0.04	0.81 ± 0.04
Validation Run 2	0.87 ± 0.05	0.82 ± 0.03	0.84 ± 0.04
Validation Run 3	0.81 ± 0.08	0.78 ± 0.06	0.78 ± 0.06
Aggregate across three runs	0.84 ± 0.07	0.80 ± 0.05	0.81 ± 0.05

Table 3.13. Performance of edge-based active contour segmentation with contraction bias of -0.075 and smoothing factor of 0.6 on the validation datasets.

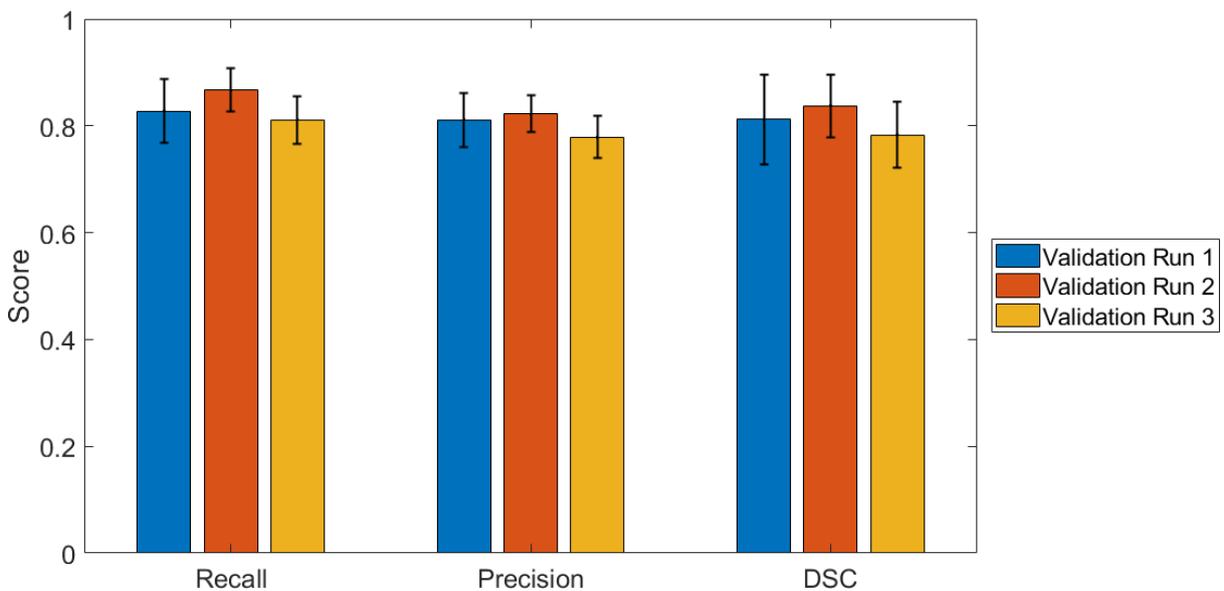


Figure 3.55. Mean performance of edge-based active contour segmentation on each fold of the validation datasets (error bars represent standard deviation).

3.12.4 Graph-cut segmentation

3.12.4.1 Superpixel generation on subsample cases

3.12.4.1.1 Performance of superpixel generation on subsample cases

The plots of the mean boundary recall and undersegmentation error in relation to the number of desired superpixels for the subsample 18 cases is shown in figures 3.56a-b. The scores for boundary recall increased with increasing number of desired superpixel regions for all SLIC compactness, where the maximum boundary recall of 1 was attained at $k = 13000$ for the non-adaptive SLIC algorithms. Greater number of superpixels was also associated with lower undersegmentation error.

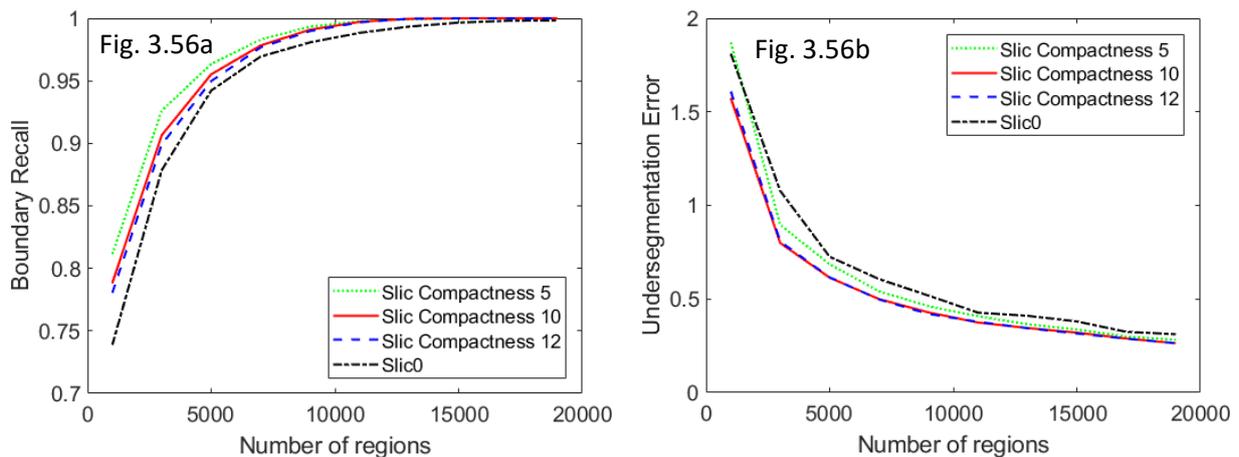


Figure 3.56. Impact of desired number of superpixel regions on a) mean boundary recall and b) mean undersegmentation error for different SLIC compactness.

The improvement in overlap of the superpixel boundary and the tumour edge with more superpixel regions can be visualised in figure 3.57, for all four SLIC algorithms. Whilst a low number of regions produced superpixels that did not conform well to the object boundary, large number of regions created more superpixels than necessary which increased the redundancy of the workflow. For this subgroup of cases, this occurred when the number of regions exceeded 10000.

On the other hand, for an image slice with two GTVs of different sizes and locations (figure 3.58), the minimum number of regions required to produce acceptable superpixels was seen to be affected by the size of the lesion. Although there was redundancy in the superpixels generated for the primary disease, the boundary of the nodal disease had a better fit with more superpixels.

Desired number of superpixels (k)			
$k = 3000$	$k = 7000$	$k = 11000$	$k = 15000$

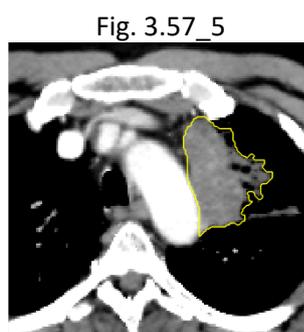
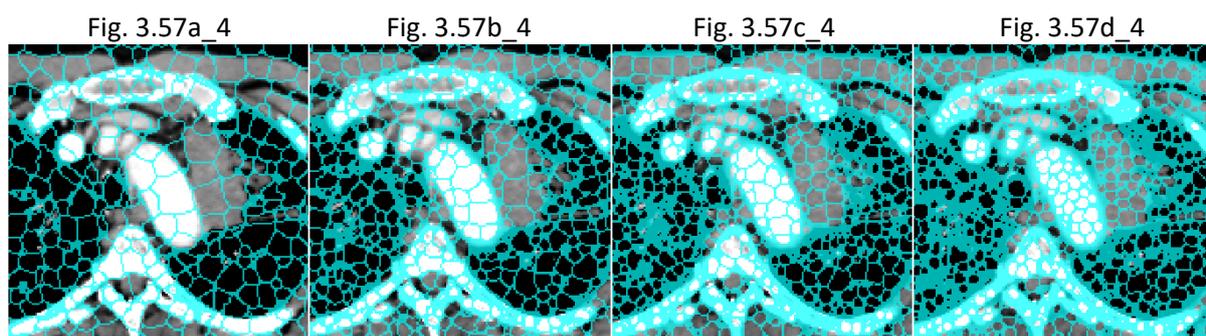
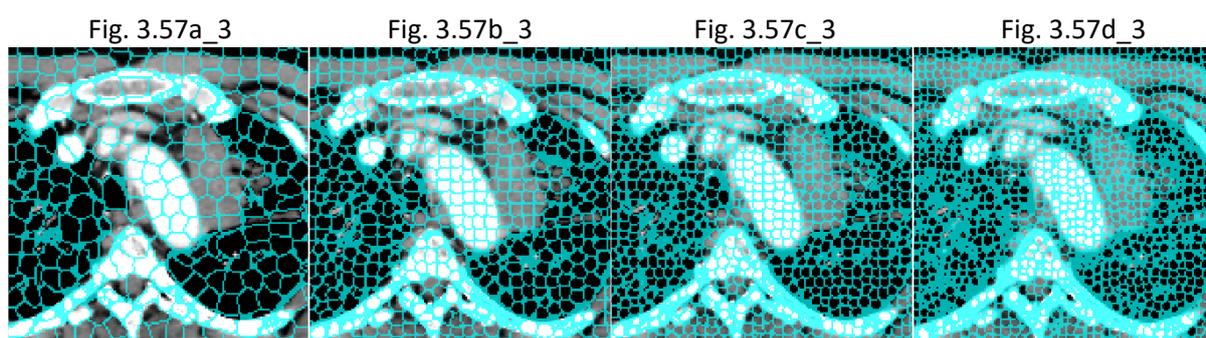
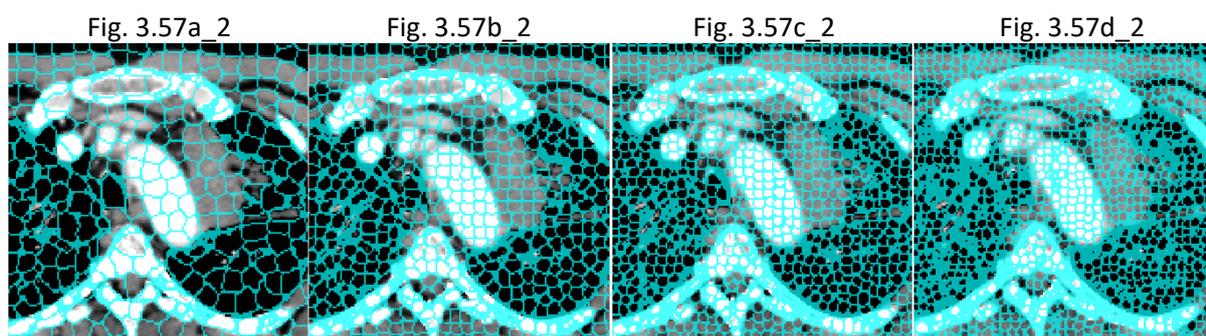
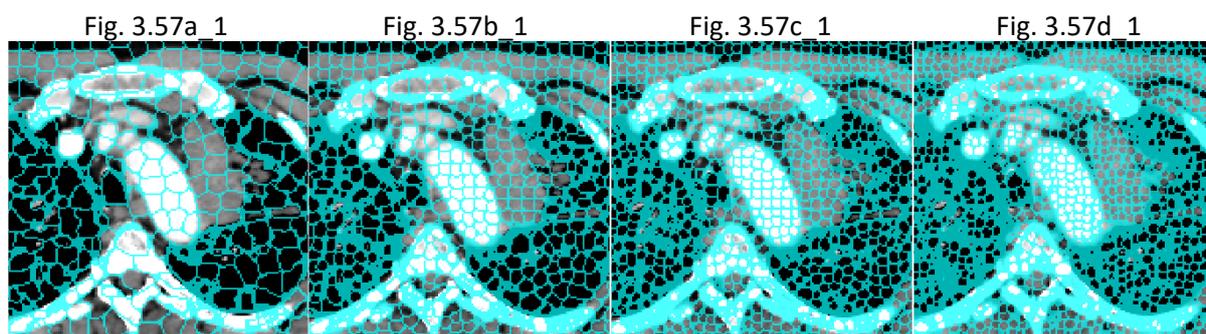


Figure 3.57a – d. Superpixels generated for one image slice showing impact of variation of number of regions a) 3000; b) 7000; c) 11000; d) 15000). Suffix_1) SLIC compactness 5; suffix_2) SLIC compactness 10; suffix_3) SLIC compactness 12; suffix_4) SLIC 0; Fig. 3.57_5) Reference outlines in yellow.

Desired number of superpixels (k)			
$k = 3000$	$k = 7000$	$k = 11000$	$k = 15000$

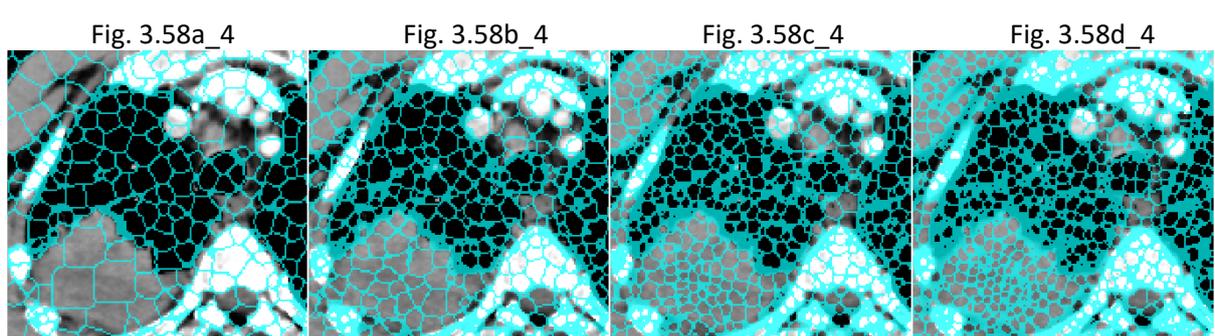
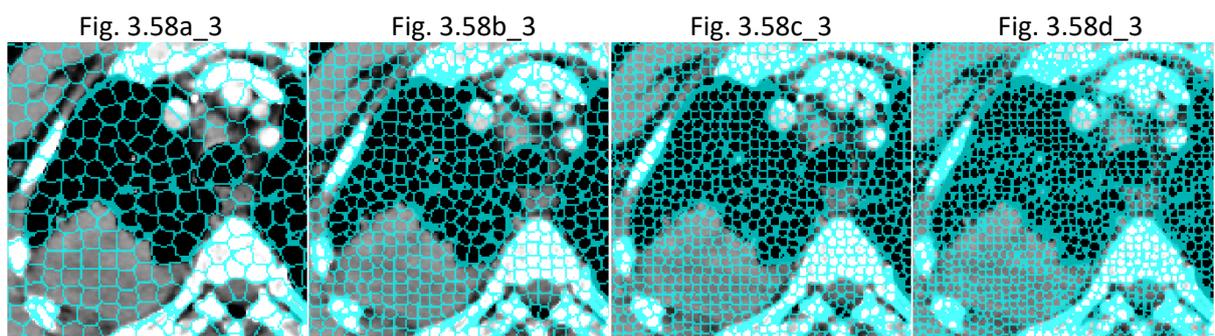
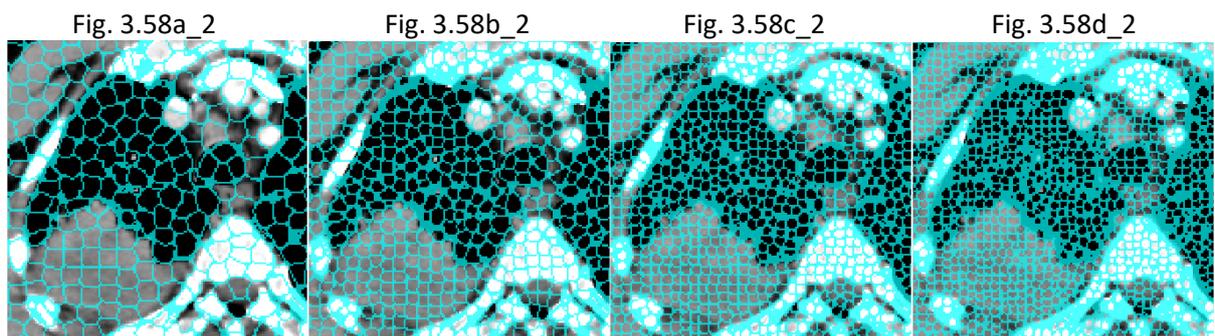
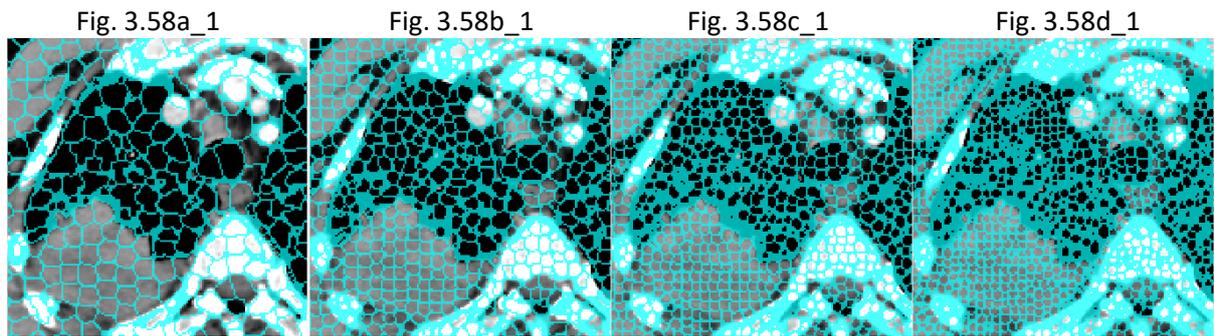


Figure 3.58a – d. Superpixels generated for one image slice with two tumour regions showing impact of variation of number of regions a) 3000; b) 7000; c) 11000; d) 15000). Suffix _1) SLIC compactness 5; suffix _2) SLIC compactness 10; suffix _3) SLIC compactness 12; suffix _4) SLIC 0; Fig. 3.58_5) Reference outlines in yellow.

The histogram in figure 3.59 shows that there was a wide spread of tumour sizes in this cohort, with a high proportion of disease at small pixel sizes of less than 150 pixels. This was taken into consideration in the decision on setting the minimum limit of the desired number of superpixels, to achieve a balance with the redundancy observed for larger disease. In a 512 x 512 image comprising of 262144 pixels, average superpixel sizes of 26, 37 and 52 pixels would be generated if 10000, 7000 and 5000 regions were selected respectively.

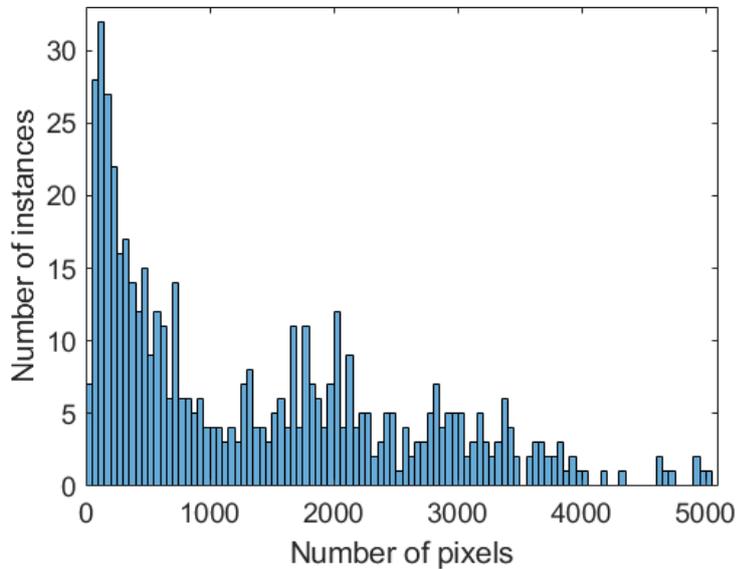


Figure 3.59. Size (number of pixels) of each tumour ROI in the axial plane across the 18 subsample cases.

In addition to the number of regions, the performance of the superpixel generation was also affected by the SLIC algorithm. The adaptive SLIC was found to create highly irregular regions, which was associated with poorer performance as compared to the other SLIC algorithms. SLIC with compactness of 5 was associated with more irregular regions as compared to SLIC 10 and 12, where their performance was similar.

3.12.4.1.2 Application of lazysnapping segmentation

The effect of the different SLIC algorithms and number of regions was assessed through the application of the lazysnapping segmentation with edge weight scaling factor of 500. Higher DSC, precision and recall scores were achieved with greater number of superpixel regions (figure 3.60). Relatively good recall scores with a mean greater than 0.86 was achieved when more than 5000 number of regions were used, although lower precision scores were obtained. All the SLIC algorithms achieved similar DSC scores. The three non-adaptive SLIC algorithms yielded recall and precision scores that were largely similar, while the adaptive algorithm produced higher precision and lower recall scores compared to the others.

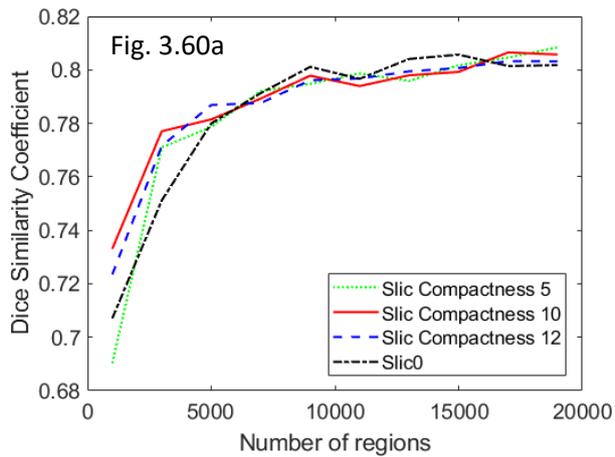
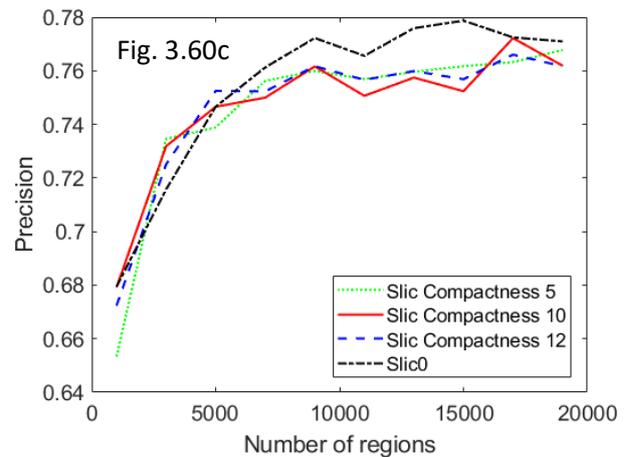
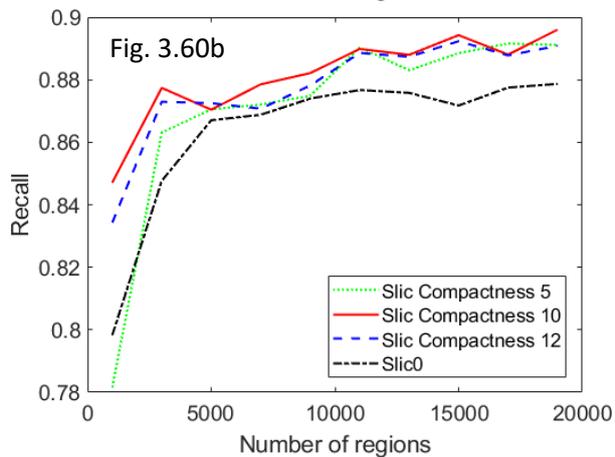


Figure 3.60. Performance of lazysnapping segmentation using superpixels in relation to the number of superpixel regions and different SLIC algorithms, showing results of a) Mean Dice similarity coefficient; b) Mean recall; c) Mean precision.



3.12.4.1.3 Qualitative assessment of lazysnapping segmentation

Inspection of the individual cases revealed differences in segmentation results with variation of the number of superpixel regions as well as the type of SLIC algorithm. Representative case slices for SLIC compactness of 5 and 10 are shown in figure 3.61. Between 5000 and 17000 number of superpixels, the segmentation performance was variable when reviewed qualitatively on a case-by-case basis, which did not exhibit an obvious trend with number of superpixels.

Nonetheless, the graph-cut segmentation was observed to work best at discriminating between regions with greater contrast, such as the bone (Case D), contrast-enhanced vessels (Case C) as well as the lung parenchyma (Cases A and B). Variation in performance was observed in regions with less distinct contrast at borders of the non-contrast enhanced vessels (Cases D, E and F), mediastinal tissue (Case F), and musculature in the chest wall (Case F).

This workflow also appeared to cope well with segmentation of GGOs as shown in cases B and C. A small region of GGO in case A was also included in the segmented region using SLIC compactness of 5.

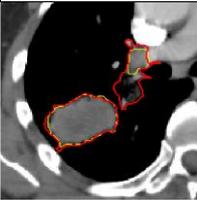
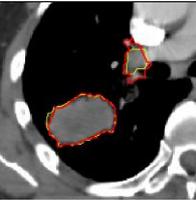
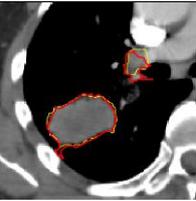
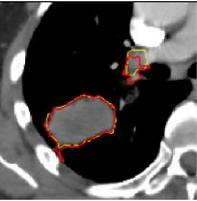
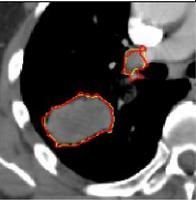
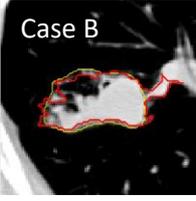
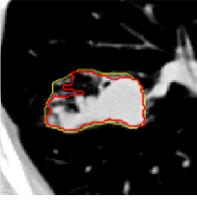
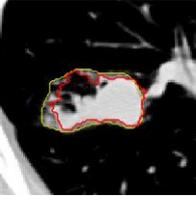
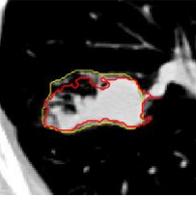
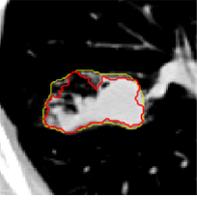
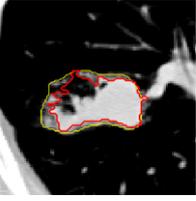
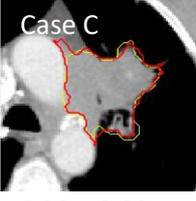
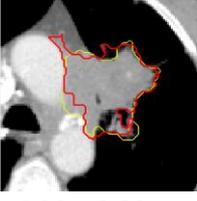
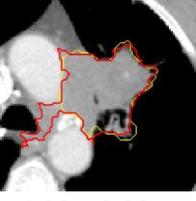
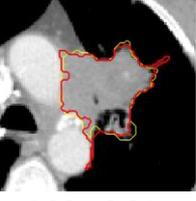
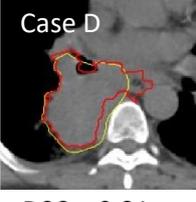
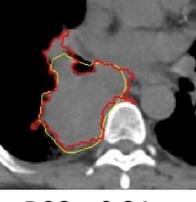
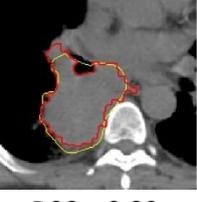
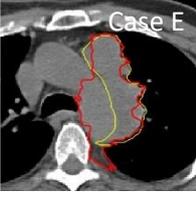
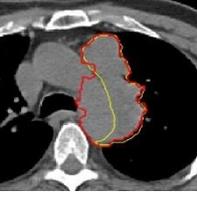
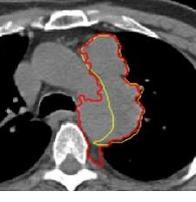
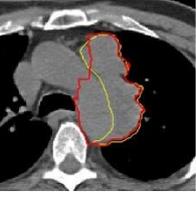
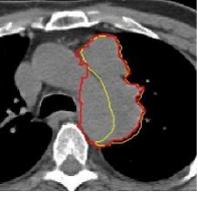
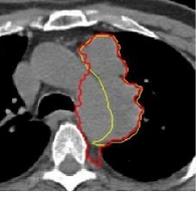
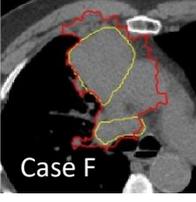
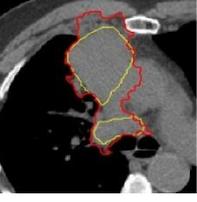
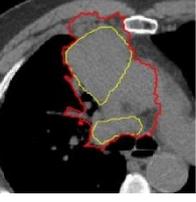
SLIC compactness = 5			SLIC compactness = 10		
$k = 5000$	$k = 11000$	$k = 17000$	$k = 5000$	$k = 11000$	$k = 17000$
Case A 					
DSC = 0.86	DSC = 0.89	DSC = 0.88	DSC = 0.87	DSC = 0.88	DSC = 0.89
Case B 					
DSC = 0.84	DSC = 0.84	DSC = 0.83	DSC = 0.83	DSC = 0.83	DSC = 0.82
Case C 					
DSC = 0.82	DSC = 0.82	DSC = 0.82	DSC = 0.83	DSC = 0.83	DSC = 0.84
Case D 					
DSC = 0.81	DSC = 0.82	DSC = 0.84	DSC = 0.81	DSC = 0.80	DSC = 0.82
Case E 					
DSC = 0.75	DSC = 0.78	DSC = 0.78	DSC = 0.73	DSC = 0.77	DSC = 0.79
Case F 					
DSC = 0.66	DSC = 0.71	DSC = 0.69	DSC = 0.66	DSC = 0.66	DSC = 0.67

Figure 3.61. Representative images slices of individual cases (A to F) for SLIC compactness 0 and 5, using number of superpixels $k = 5000, 11000$ and 17000 , with the associated DSC scores for each case.

3.12.4.1.4 Computational time

Despite the increase in the computation time for superpixel generation with increasing number of regions, the mean computational time for generation of the superpixels alone was very short, in the order of a small number of seconds for each case. However, the lazysnapping segmentation took significantly longer, where a linear increase in computation time (minutes) was seen with increasing number of superpixel regions. The mean difference in time taken to process each case was more than 10 minutes between the use of 5000 and 19000 number of superpixels.

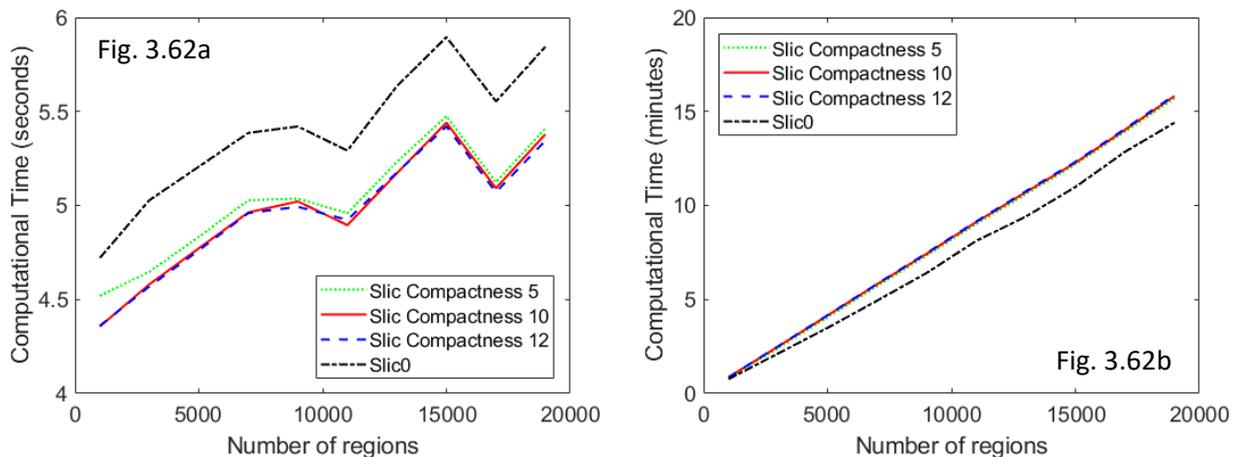


Figure 3.62. Mean computational processing time for individual cases a) superpixel generation alone b) superpixel generation and application of lazysnapping segmentation.

3.12.4.1.5 Summary from superpixel generation from subsample of 18 cases and superpixel parameter selection

A greater number of superpixel regions was associated with better boundary recall and undersegmentation scores, which when applied with graph-cut segmentation, was found to be associated with better mean DSC, recall and precision. The different SLIC algorithms also affected the segmentation results, albeit to a lesser extent as compared to the number of regions. When assessed qualitatively, there was variable segmentation performance when the different superpixel parameters were applied.

The superpixel generation took only a short number of seconds for the range of number of regions assessed, but there was a significant difference in the mean processing time for each case in the subsequent graph-cut segmentation, ranging from under two minutes for 1000 number of regions, to almost 17 minutes for 19000 number of regions.

Although a processing time of between 15 to 20 minutes would be acceptable on an individual case basis, this would translate to taking several days to process larger number of cases in the training cohort, which would be further compounded in the evaluation of the edge weight scale factor of the lazysnapping algorithm.

Thus, it was decided to extend this evaluation to the cross-validation folds of the training dataset to see if similar results for the superpixel parameters would be obtained. If so, suitable superpixel parameters would be chosen with the balance between the segmentation performance and the processing time. These values for the superpixel generation would then be fixed and used to evaluate the edge weight scale factor of the graph-cut algorithm.

3.12.4.2 Training using cross validation folds for superpixel and lazysnapping processes

3.12.4.2.1 Superpixel generation

Similar trends were yielded for the superpixel boundary recall and undersegmentation error in the training folds as previously observed. Although better performance was seen with increasing the desired number of superpixel regions, little further improvement in the performance was seen beyond 10000 regions.

For all three folds, it was observed that the adaptive SLIC algorithm did not perform as well as the other SLIC algorithms, with lower boundary recall and undersegmentation scores across the range of superpixel numbers. Compactness factor of 10 and 12 resulted in similar performance, whereas compactness of 5 had higher boundary recall but greater undersegmentation error at less than 10000 number of superpixel regions.

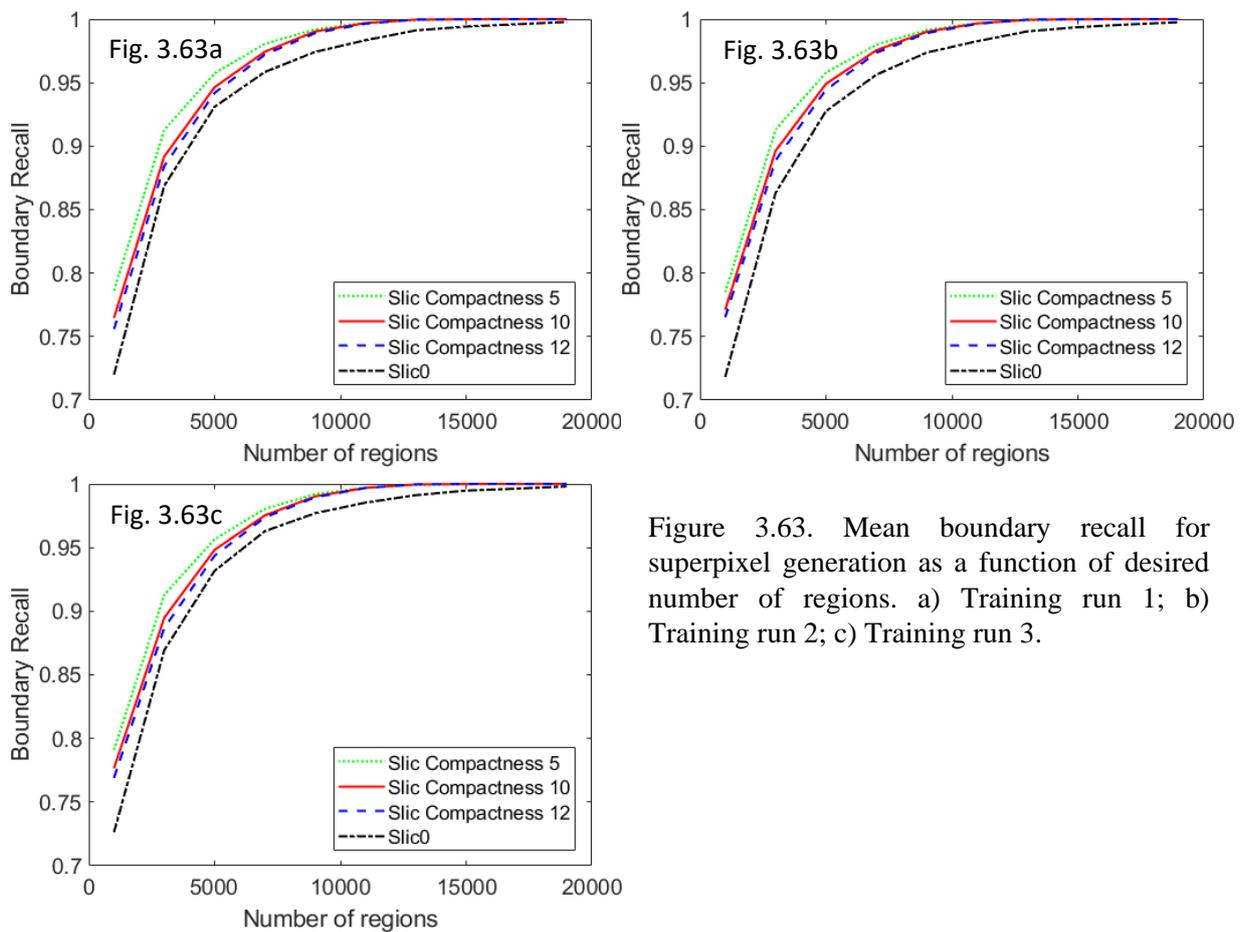


Figure 3.63. Mean boundary recall for superpixel generation as a function of desired number of regions. a) Training run 1; b) Training run 2; c) Training run 3.

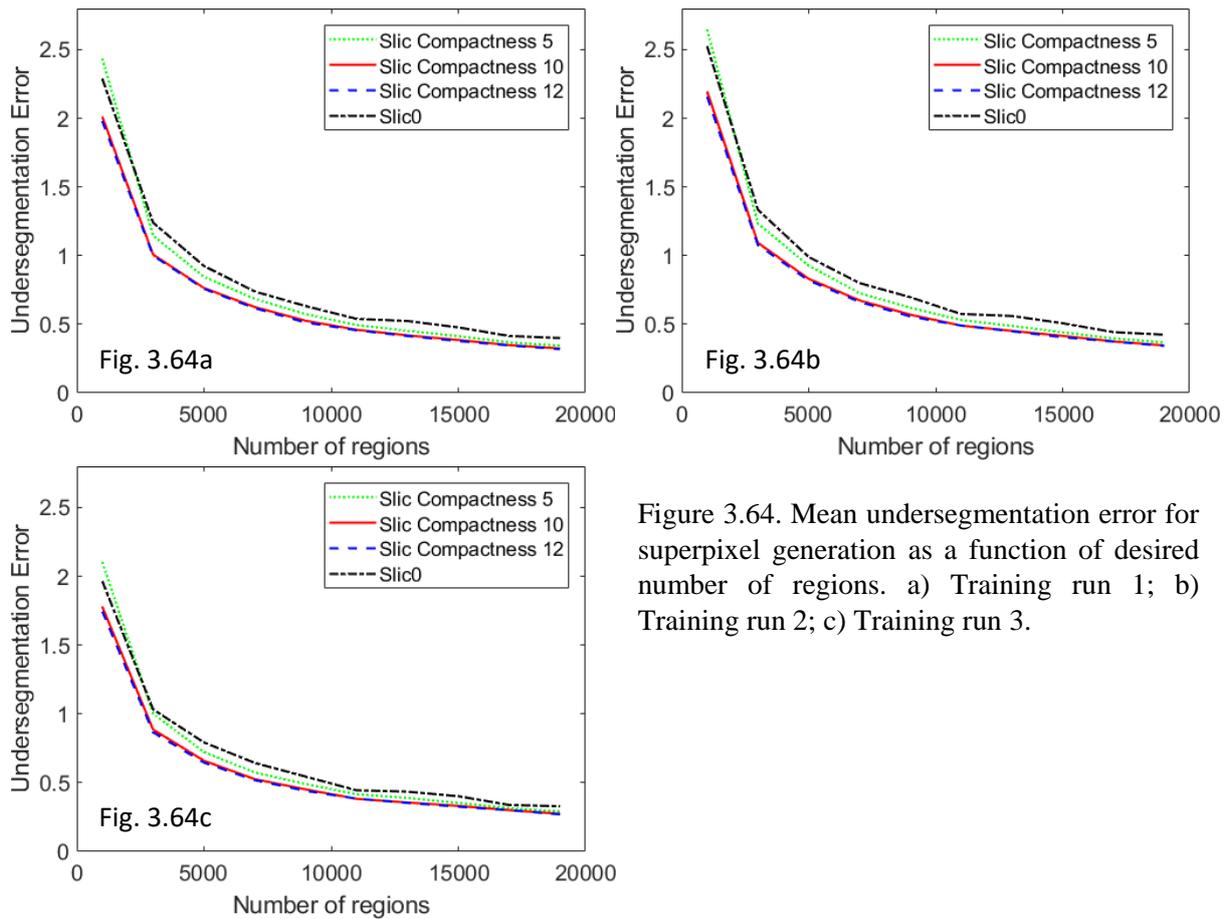


Figure 3.64. Mean undersegmentation error for superpixel generation as a function of desired number of regions. a) Training run 1; b) Training run 2; c) Training run 3.

3.12.4.2.2 Application of lazysnapping segmentation

Overall, all three training runs also displayed similar trends for mean DSC, recall and precision when the lazysnapping segmentation was applied, using an edge weight scale factor of 500. As the number of superpixel regions increased, DSC, recall and precision scores increased sharply until $k = 5000$, beyond which the improvement in the scores was more gradual. SLIC with compactness of 10 and 12 were associated with higher mean recall scores between k of 5000 and 10000 compared to SLIC0. Additionally, despite the higher mean boundary recall for the superpixel generation seen with SLIC with a compactness of 5 (figure 3.63), the mean recall scores were lower as compared to a compactness of 10 and 12 when the lazysnapping segmentation was applied (figure 3.66). The precision scores were similar between the different compactness factors.

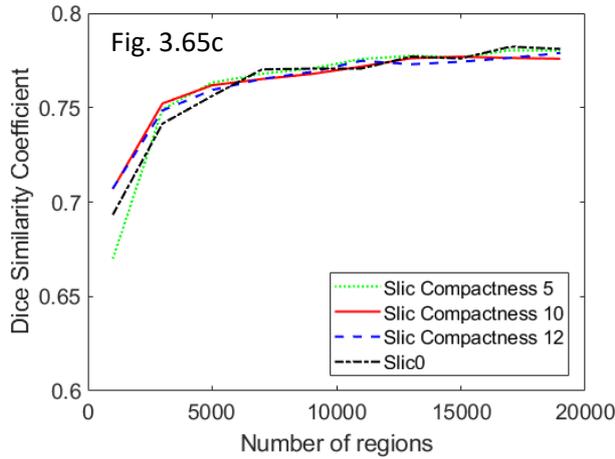
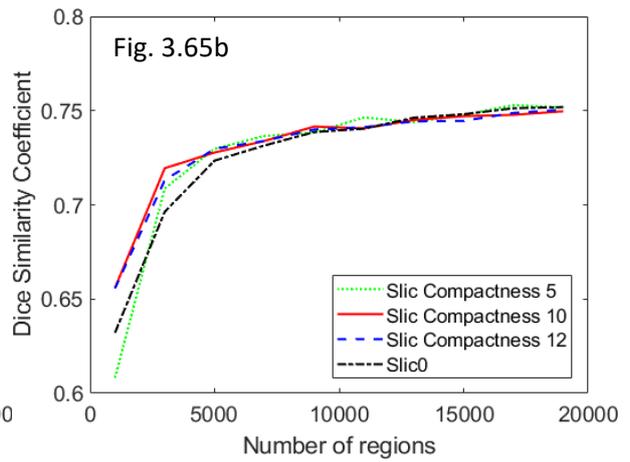
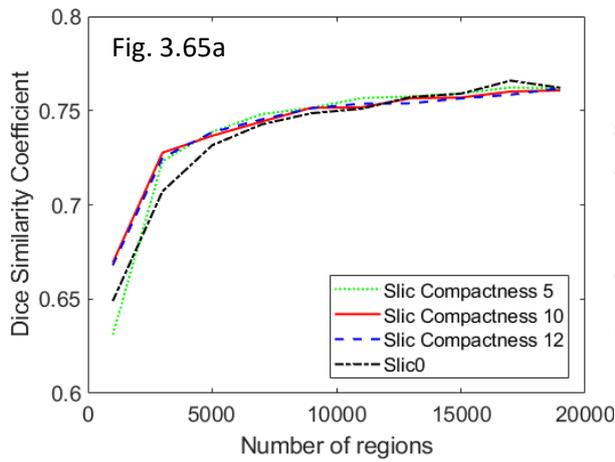


Figure 3.65. Mean Dice similarity coefficient in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.

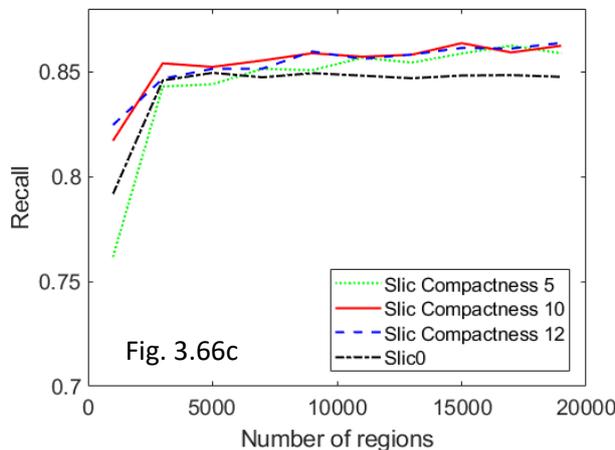
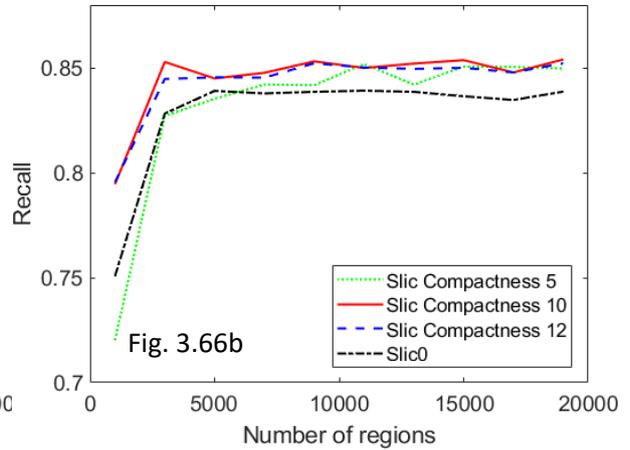
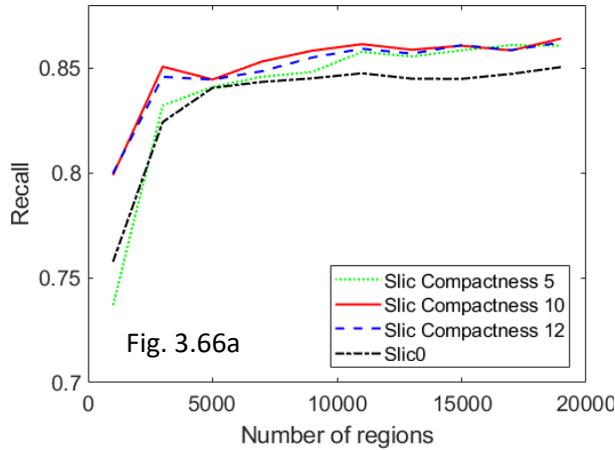


Figure 3.66. Mean recall scores in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.

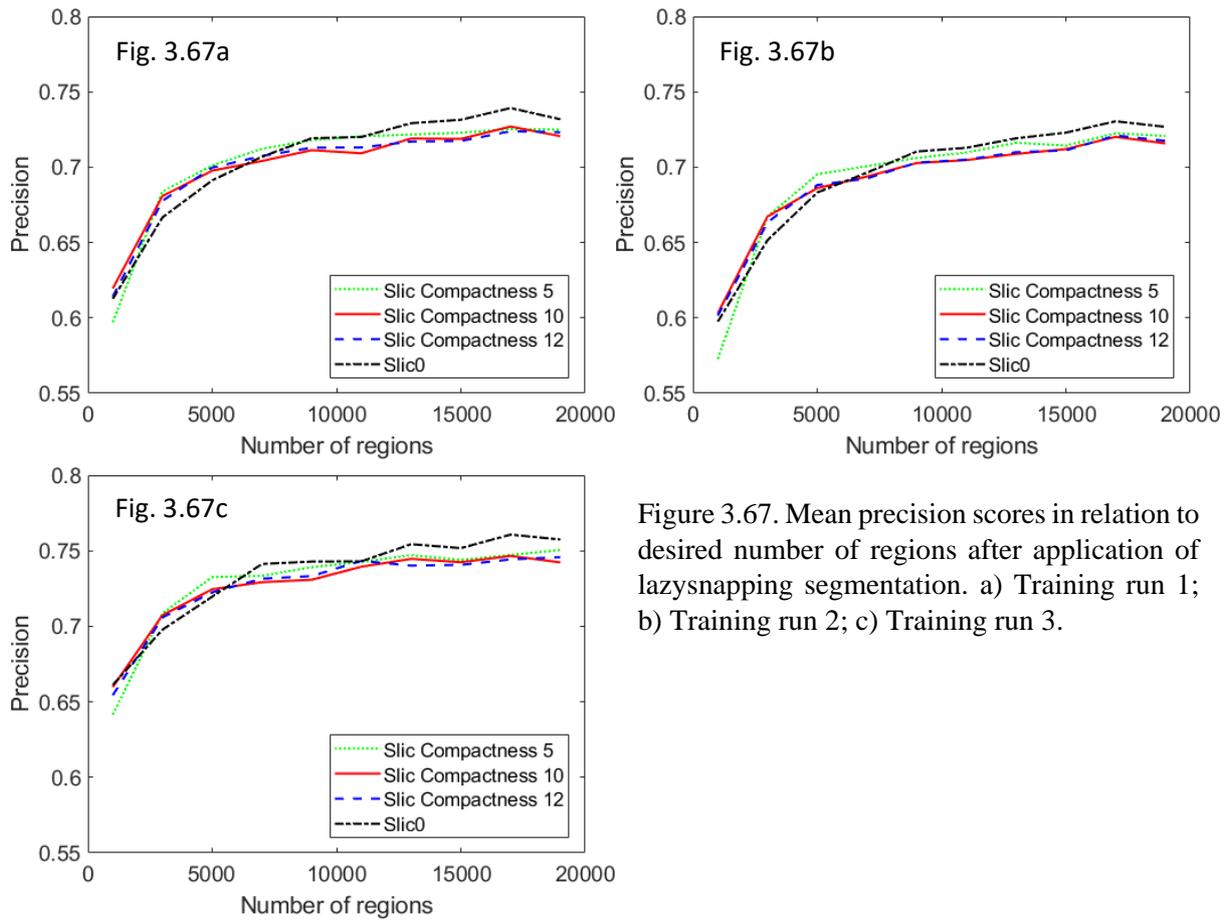


Figure 3.67. Mean precision scores in relation to desired number of regions after application of lazysnapping segmentation. a) Training run 1; b) Training run 2; c) Training run 3.

3.12.4.2.3 Computational time

Similar results for the computational time was obtained for the training folds as seen in the 18 subsample cases. Again, the mean computational time for the generation of superpixels for individual cases was in the order of short seconds across the number of regions. However, this was significantly greater in the application of the lazysnapping segmentation, which ranged from under 1 to 14 minutes.

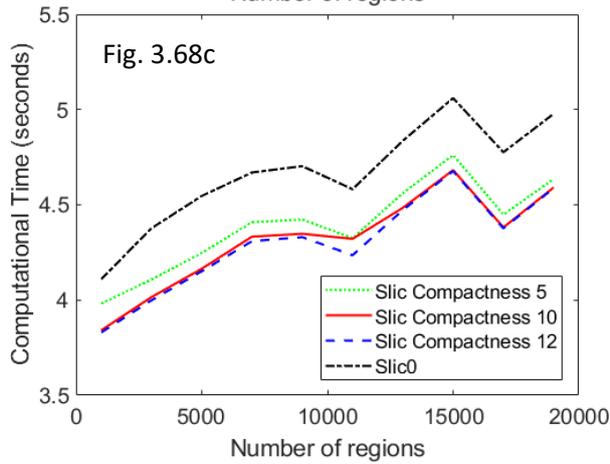
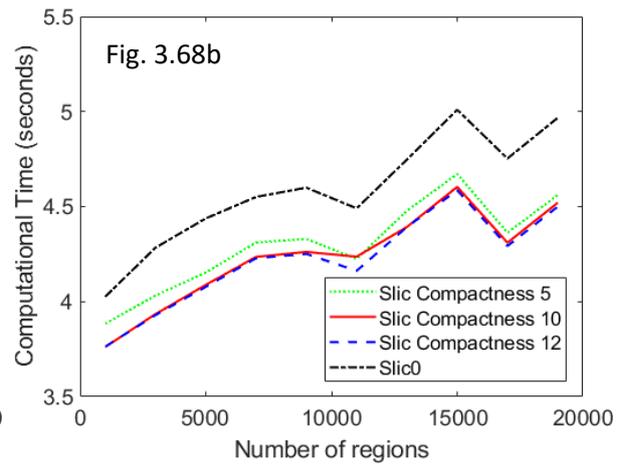
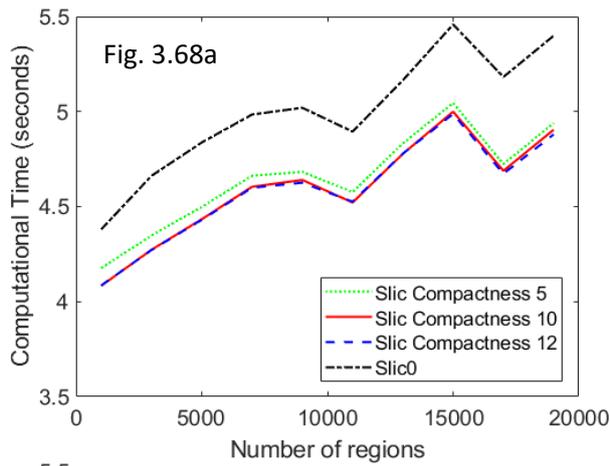


Figure 3.68. Mean computational time of individual cases for superpixel generation. a) Training run 1; b) Training run 2; c) Training run 3.

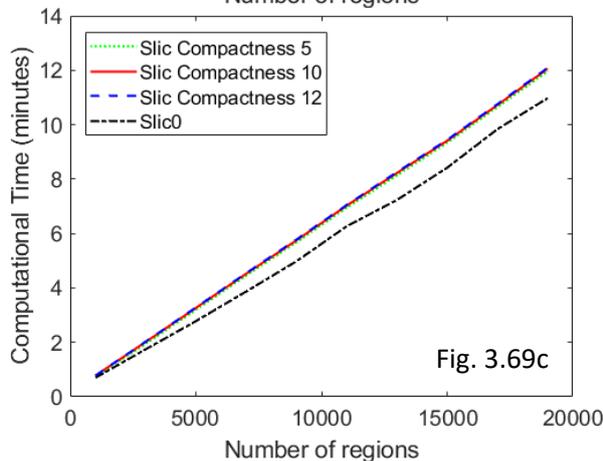
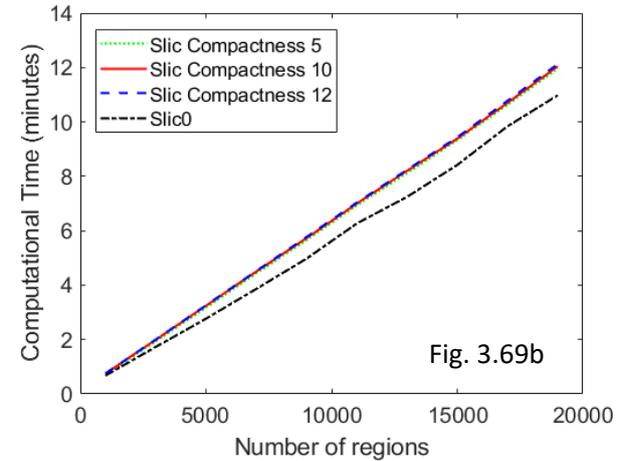
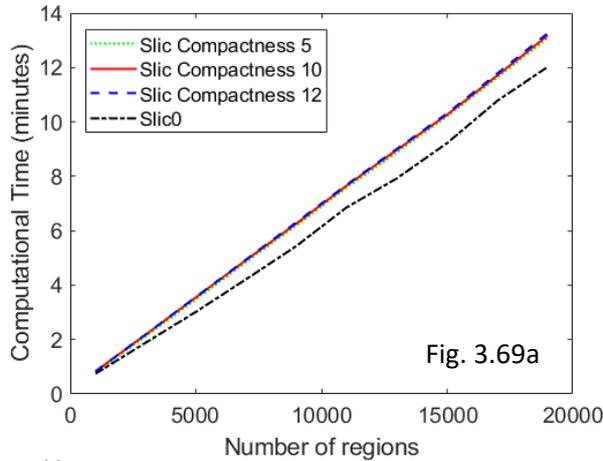


Figure 3.69. Mean computational time of individual cases for lazysnapping segmentation in relation to number of superpixel regions. a) Training run 1; b) Training run 2; c) Training run 3.

From these analyses, it was decided that 7000 number of regions would be used for further evaluation of the lazysnapping edge weight scale parameter. This allowed the balance between acceptable segmentation performance with a practical processing time. The SLIC algorithm with a compactness of 10 was also selected in preference to the others, based on the segmentation performance and the amount of superpixel irregularity observed. Coincidentally, this was the default setting in MATLAB.

3.12.4.3 Evaluation of edge weight scaling factor for lazysnapping algorithm

The impact of the edge weight scaling factor on the segmentation for each training fold is shown in figure 3.70. Increase in the edge weight scaling factor was associated with higher rates of recall but resulted in lower rates of precision. The best DSC was achieved at the lower edge weights, and the performance for scaling factors between 10 and 60 is displayed in table 3.14 to assist with parameter selection.

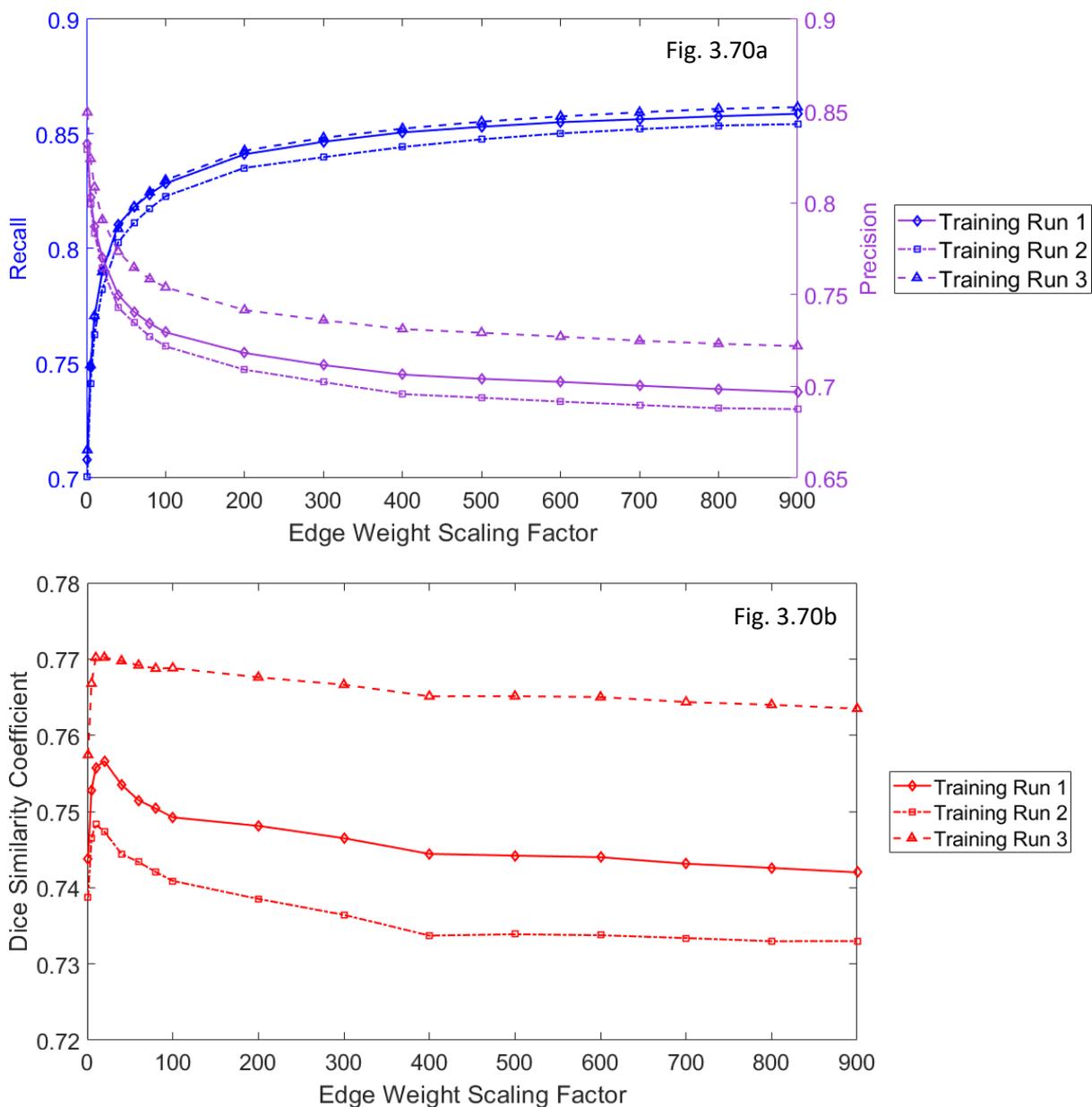


Figure 3.70. Performance with variation of edge weight scaling factors for the training runs. (Superpixel generation: $k = 7000$, SLIC $m = 10$) a) Mean recall and precision; b) Mean Dice similarity coefficient.

	Edge weight scaling factor	Dice Similarity Coefficient	Recall	Precision
Training run 1	10	0.76 ± 0.08	0.77 ± 0.08	0.79 ± 0.10
	20	0.76 ± 0.08	0.79 ± 0.08	0.77 ± 0.10
	40	0.75 ± 0.09	0.81 ± 0.07	0.75 ± 0.11
	60	0.75 ± 0.09	0.82 ± 0.07	0.74 ± 0.11
Training run 2	10	0.75 ± 0.07	0.76 ± 0.07	0.78 ± 0.10
	20	0.75 ± 0.07	0.78 ± 0.07	0.76 ± 0.10
	40	0.75 ± 0.08	0.80 ± 0.07	0.74 ± 0.10
	60	0.74 ± 0.08	0.81 ± 0.07	0.74 ± 0.10
Training run 3	10	0.77 ± 0.06	0.77 ± 0.07	0.81 ± 0.08
	20	0.77 ± 0.06	0.79 ± 0.07	0.79 ± 0.08
	40	0.77 ± 0.06	0.81 ± 0.07	0.77 ± 0.08
	60	0.77 ± 0.06	0.82 ± 0.06	0.77 ± 0.08

Table 3.14. Performance for the training runs with edge weight scaling factor between the range of 10 and 60. (Superpixel generation: $k = 7000$, SLIC $m = 10$)

Representative axial slices of the lazysnapping segmentation across a range of edge weight scale factors are shown in figure 3.71. Although lower edge weighting produced smaller segmented regions as compared to higher weighting factors (case C), it was observed that this scaling factor did not result in a great change in segmentation between a range of 10 to 900 for a number of cases when assessed visually. Low edge weight parameter was associated with underestimation at the tumour edge (cases C, D, F), although this effect was more pronounced at factors less than 10. In some cases, it was also observed that a low parameter was helpful in excluding vessels (cases D and E). Higher values tended to leak into surrounding tissues, such as the mediastinum and the chest wall (case F).

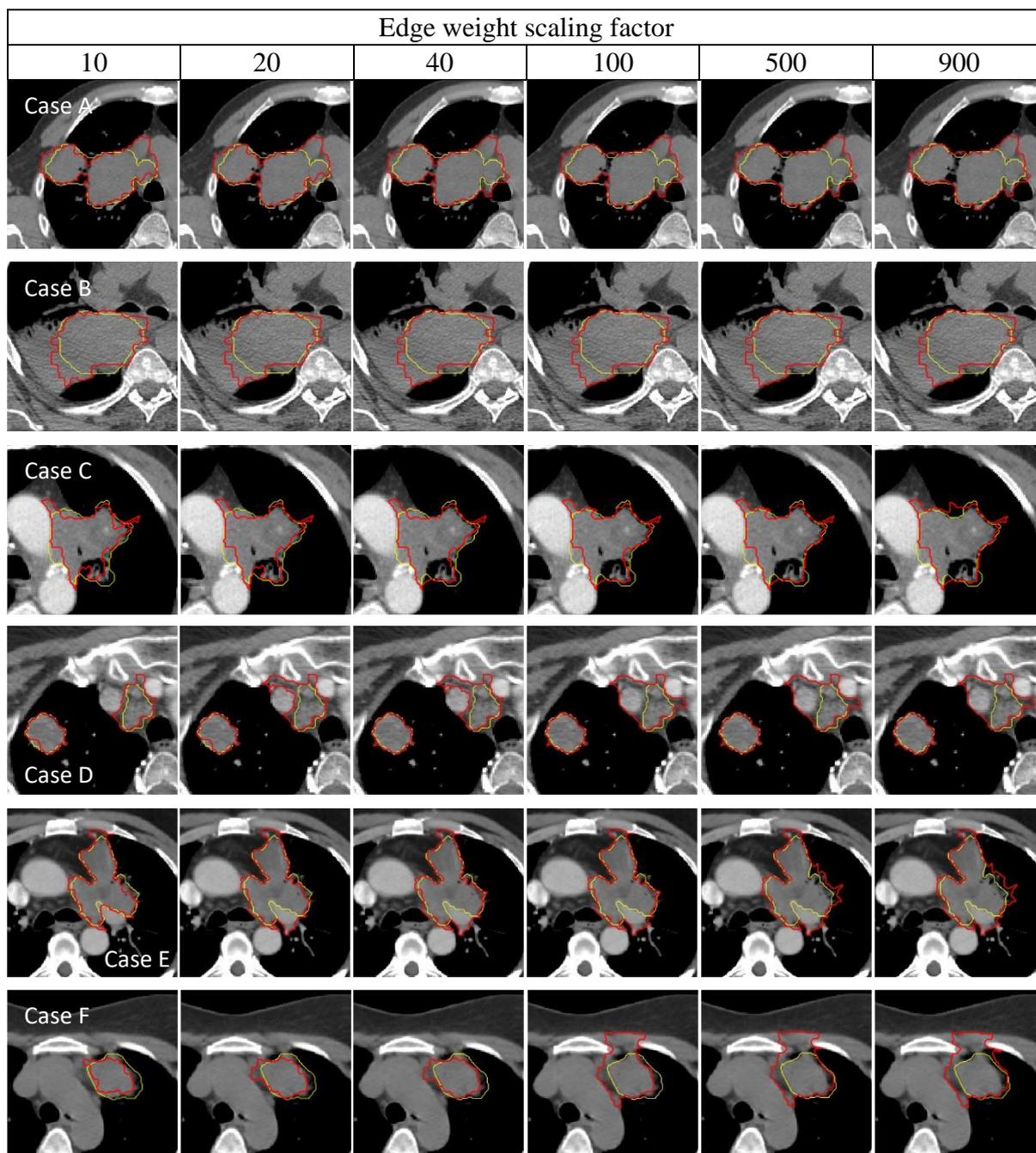


Figure 3.71. Representative images slices of individual cases (A to F) for edge weight scale factors of 10 to 900. (Superpixel generation: $k = 7000$, SLIC $m = 10$)

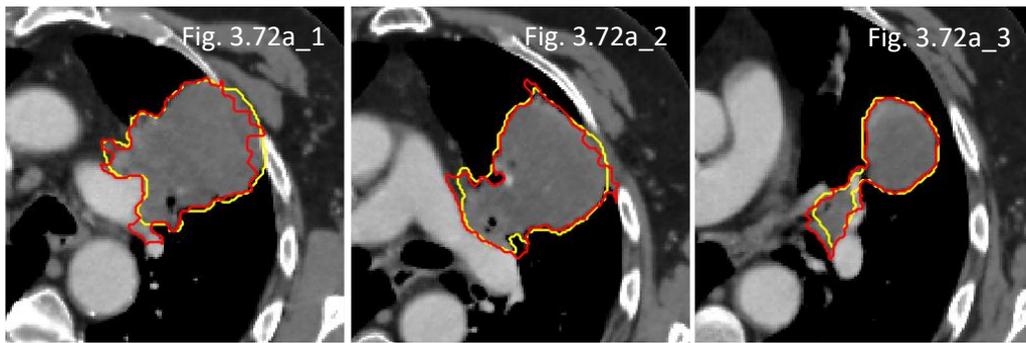
The highest mean DSC was seen for edge weight scale factors of 20 and 10 for runs 1 and 2 respectively, with tied mean DSC at factors 10 and 20 for run 3. However, a low scale factor would result in greater underestimation of the tumour with poorer recall rates. It was therefore decided that a scaling factor of 40 would be more appropriate for all three runs. This corresponded to higher recall scores and better tumour coverage, which was felt to be balanced with the associated precision scores. This also avoided the selection of a parameter close to the steep gradient fall off in the DSC plot, to help with applicability of the scale factor to new cases.

3.12.4.4 Qualitative assessment of segmentation performance

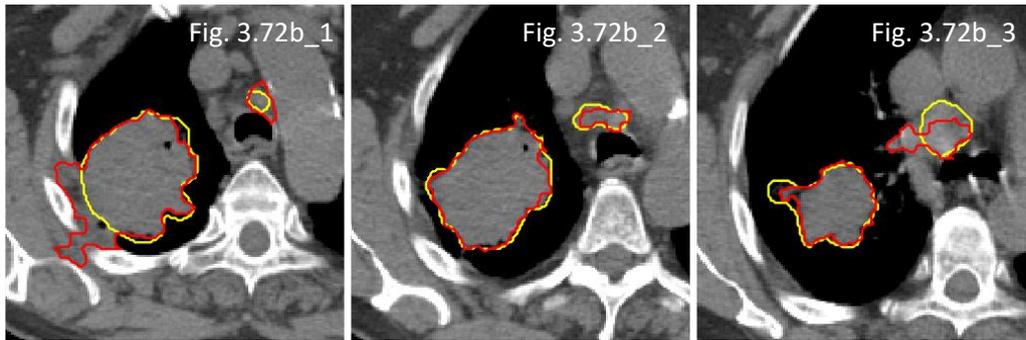
Like the watershed approach, the graph-cut segmentation was able to conform to irregularly shaped objects (figure 3.72a_3). In fact, most of the resultant segmentation had rather scalloped and jagged edges, unlike the smooth contours generated by the active contour approaches.

Good segmentation was observed at the region of the tumour bordering lung parenchyma and contrast-enhanced vessels. On most occasions, this technique had the capability to discriminate tumour from the chest wall (figures 3.72a_1 and 3.72a_2), though there were times where it did not work as well (figure 3.72b_1) with leakage of the segmentation into chest wall musculature. At the mediastinum, satisfactory delineation of nodal disease was seen (figures 3.72b_1 and 3.72b_2), although this was not consistent across all slices (figure 3.72b_3). In non-contrast enhanced scans, variable performance of the segmentation was obtained, with a mix of good and poorer results across different slices in the separation of the tumour from the vessels (figure 3.72e). Its ability to partition tumour from collapse was also inconsistent with some dependence on the background marker (figures 3.72c and 3.72d).

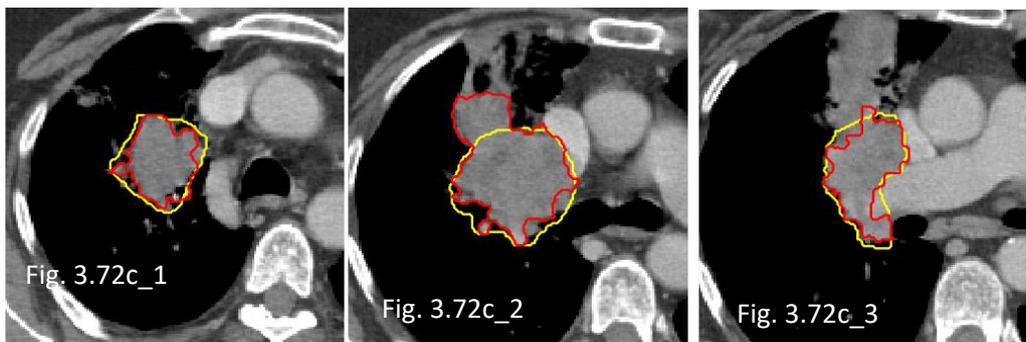
The graph-cut approach was able to perform in the presence of intra-tumoural cavities as well as regions of atelectasis on most occasions (figures 3.72d and 3.72f). However, there were instances where the segmented boundary was pulled towards the cavity edge rather than the tumour border (figure 3.72d_2 and 3.72f_3), especially for cavities located near the tumour periphery.



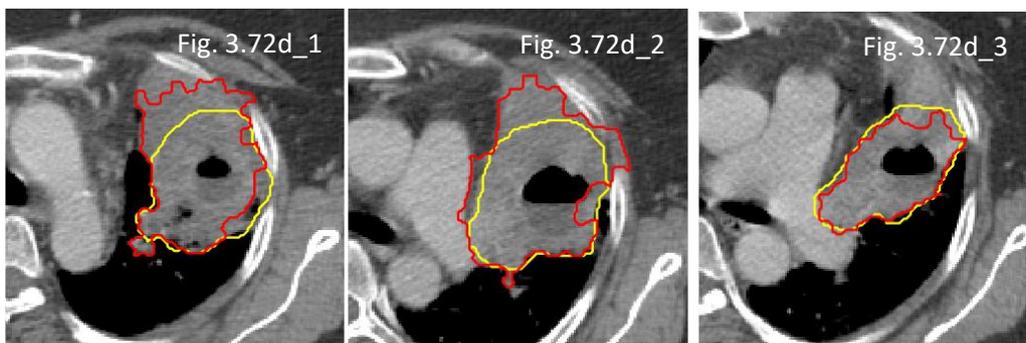
DSC = 0.89



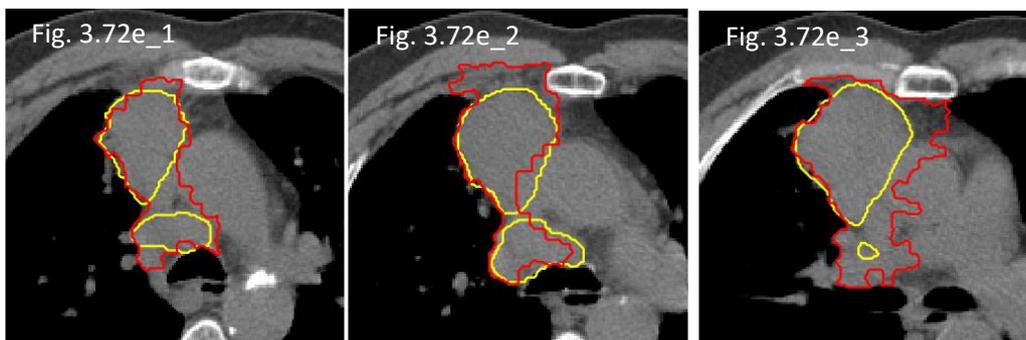
DSC = 0.84



DSC = 0.83



DSC = 0.73



DSC = 0.70

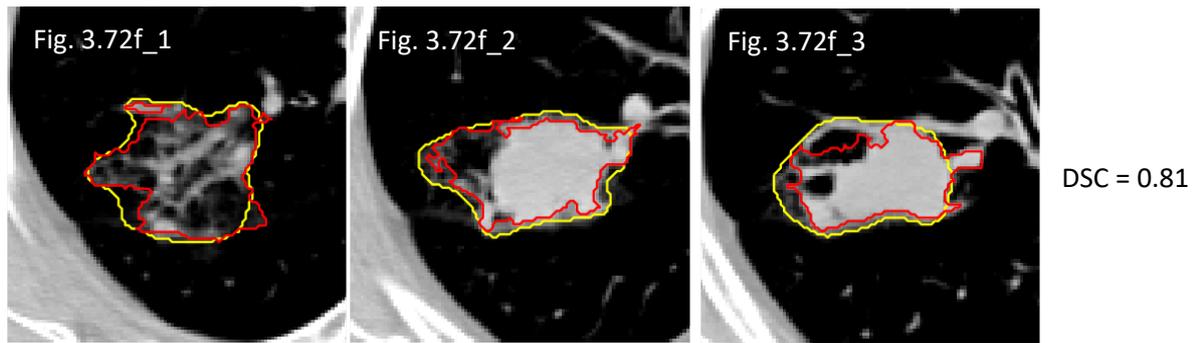


Figure 3.72. Graph-cut segmentation results (red outlines) for six representative training cases (a to f) versus reference contours (yellow outlines), with corresponding DSC for each case. (suffix _1 to 3 represent different axial slices for each case)

3.12.4.5 Validation data

A summary of the performance in the three validation folds using SLIC with compactness of 10, 7000 number of regions and graph-cut with an edge weight scale factor of 40 is shown in figure 3.73 and table 3.15, where high mean boundary recall and low mean undersegmentation errors were obtained for the superpixel generation. In terms of the estimated performance for the graph-cut segmentation, a DSC of 0.76 ± 0.08 , recall of 0.81 ± 0.07 and precision of 0.76 ± 0.10 were achieved across the three validation folds, where the computational time of individual cases were at 4.6 ± 1.5 mins.

	Validation run 1	Validation run 2	Validation run 3	Aggregate across three runs
Assessment of Superpixel				
Boundary Recall	0.98 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.98 ± 0.01
Undersegmentation Error	0.57 ± 0.24	0.48 ± 0.22	0.78 ± 0.37	0.61 ± 0.31
Assessment of Lazysnapping				
Dice Similarity Coefficient	0.76 ± 0.05	0.78 ± 0.07	0.73 ± 0.10	0.76 ± 0.08
Recall	0.80 ± 0.06	0.82 ± 0.07	0.80 ± 0.08	0.81 ± 0.07
Precision	0.77 ± 0.08	0.78 ± 0.08	0.72 ± 0.12	0.76 ± 0.10
Assessment of computational time (per case)				
Time for superpixel (seconds)	4.0 ± 1.1	4.7 ± 1.1	4.5 ± 1.6	4.4 ± 1.3
Time for lazysnapping (minutes)	4.1 ± 1.3	4.9 ± 1.3	4.9 ± 1.9	4.6 ± 1.5

Table 3.15. Performance of lazysnapping segmentation (edge weight parameter = 40) on superpixels (number of regions = 7000, SLIC compactness = 10) on the validation datasets.

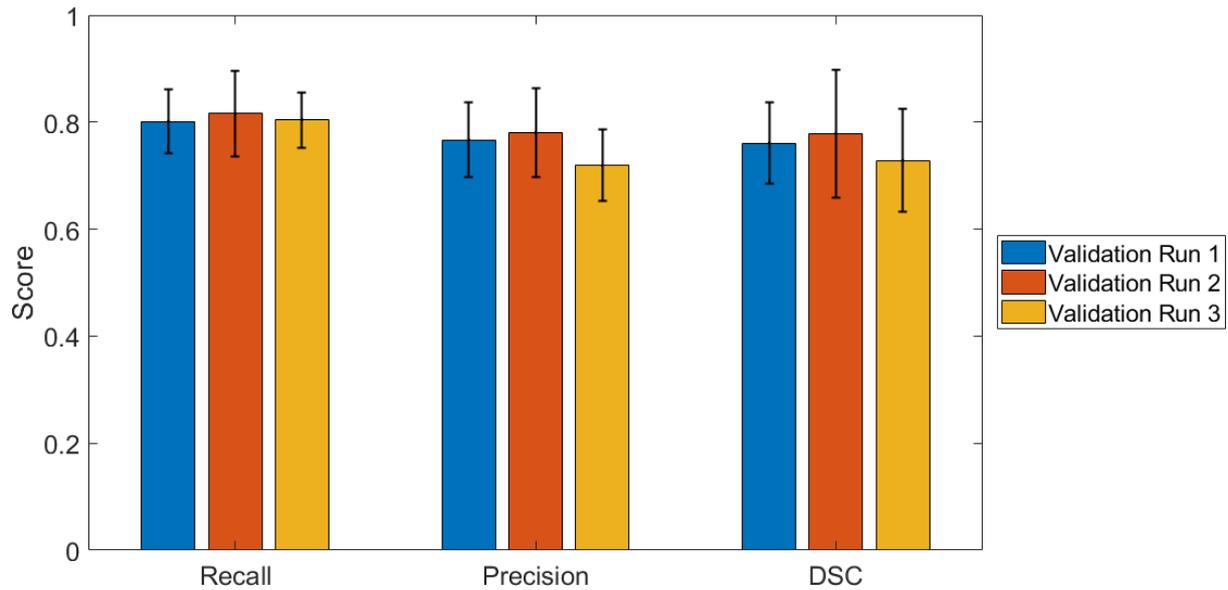


Figure 3.73. Mean performance of lazysnapping segmentation (edge weight parameter = 40) on superpixels (number of regions = 7000, SLIC compactness = 10) on each fold of the validation datasets. (error bars represent standard deviation)

Discussion

In this chapter, the development of fully automatic workflows for the initiation and segmentation of the lung tumours is presented, which includes the description of parameter selection for each of the cross-validation runs of the training dataset. An estimated performance from each of the validation datasets in the cross-validation runs is also reported. The evaluation the performance of each of the different technique is further elaborated in chapter 4, where the techniques are applied on an independent test dataset.

Chapter 4

Specific Aim C: Evaluation of different approaches to tumour segmentation

Introduction

This chapter describes the overall performance of the different segmentation approaches in the independent test data of 16 cases. The division of data, development of workflows and parameter tuning on the testing datasets is reported in chapter 3.

4.1 Summary of tasks

The same processes in chapter 3 were applied to the independent sample to assess the performance of the four different segmentation techniques.

Task C.1 Comparison of segmentation techniques on independent test dataset

Results

4.2 Task C.1 Comparison of overall performance between different segmentation techniques on independent test dataset

4.2.1 Overall performance

The overall performance for the different segmentation methods in the independent test is depicted in figure 4.1. Edge-based active contour achieved the highest mean DSC score of 0.80 ± 0.06 , followed by the graph-cut segmentation at 0.76 ± 0.06 , watershed at 0.72 ± 0.08 and Chan-Vese active contour at 0.71 ± 0.07 .

With a mean GMI and DI of 0.17 ± 0.06 and 0.20 ± 0.05 respectively, the edge-based active contour segmentation had a fairly balanced performance in terms of tumour coverage and avoidance of surrounding tissue. The graph-cut segmentation had similarly balanced values, though the scores were slightly higher at 0.20 ± 0.08 and 0.23 ± 0.08 respectively. Although the watershed segmentation was associated with an excellent GMI of 0.07 ± 0.03 , this was offset by a mean DI at 0.38 ± 0.10 , indicating that despite being good at encompassing the tumour region, the segmentation was not as precise compared to the other methods. Conversely, the Chan-Vese active contour was associated with higher mean GMI at 0.30 ± 0.10 versus the mean DI at 0.22 ± 0.10 .

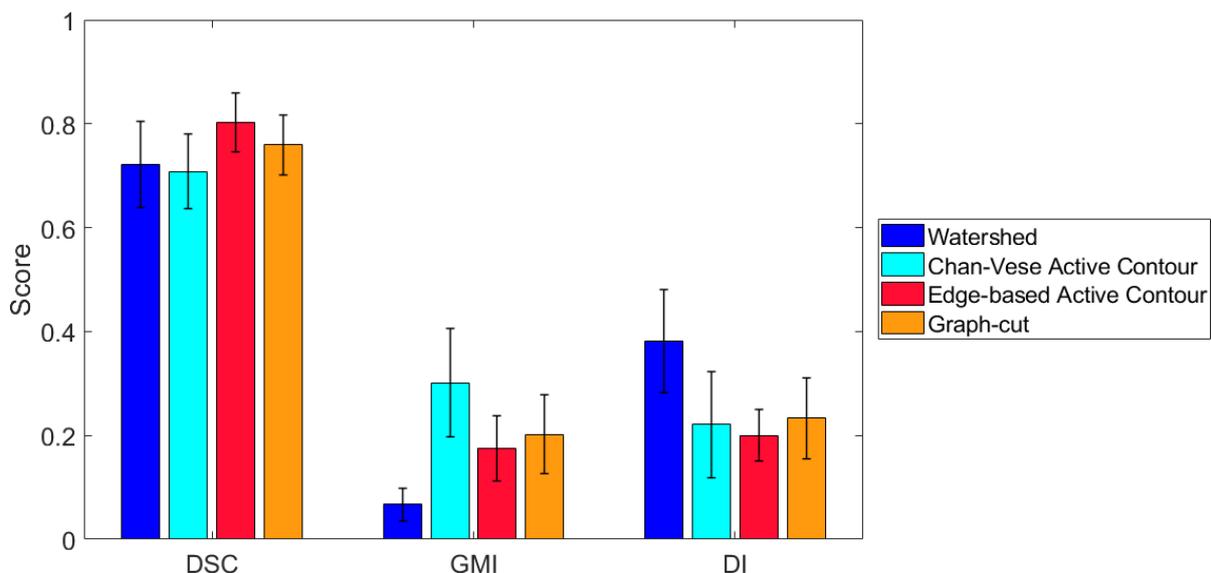


Figure 4.1. Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for independent test dataset.

4.2.2 Individual case performance

4.2.2.1 Quantitative analysis

Assessment of the conformity indices on an individual case basis revealed that the highest DSC scores were achieved using the edge-based active contour approach in 12/16 (75%) of the cases, although it was seen with the lowest GMI and DI in only 1/16 (6%) and 6/16 (38%) of the cases. Both graph-cut and Chan-Vese active contour achieved the highest DSC in 2/16 (13%) of the cases each. Based on the DSC, this indicated that the edge-based active contour method could achieve the most acceptable segmentation for the majority of the cases, although not for all.

Chan-Vese active contour had the highest GMI in 13/16 (81%) of the cases, indicating that it achieved poorer tumour coverage compared to the other approaches for these cases. On the other hand, the watershed approach had the lowest GMI in 15/16 (94%) of the cases. However, it had the highest DI in 14/16 (88%) of the cases, reflecting that although tumour coverage was good, it suffered from segmentation leakage.

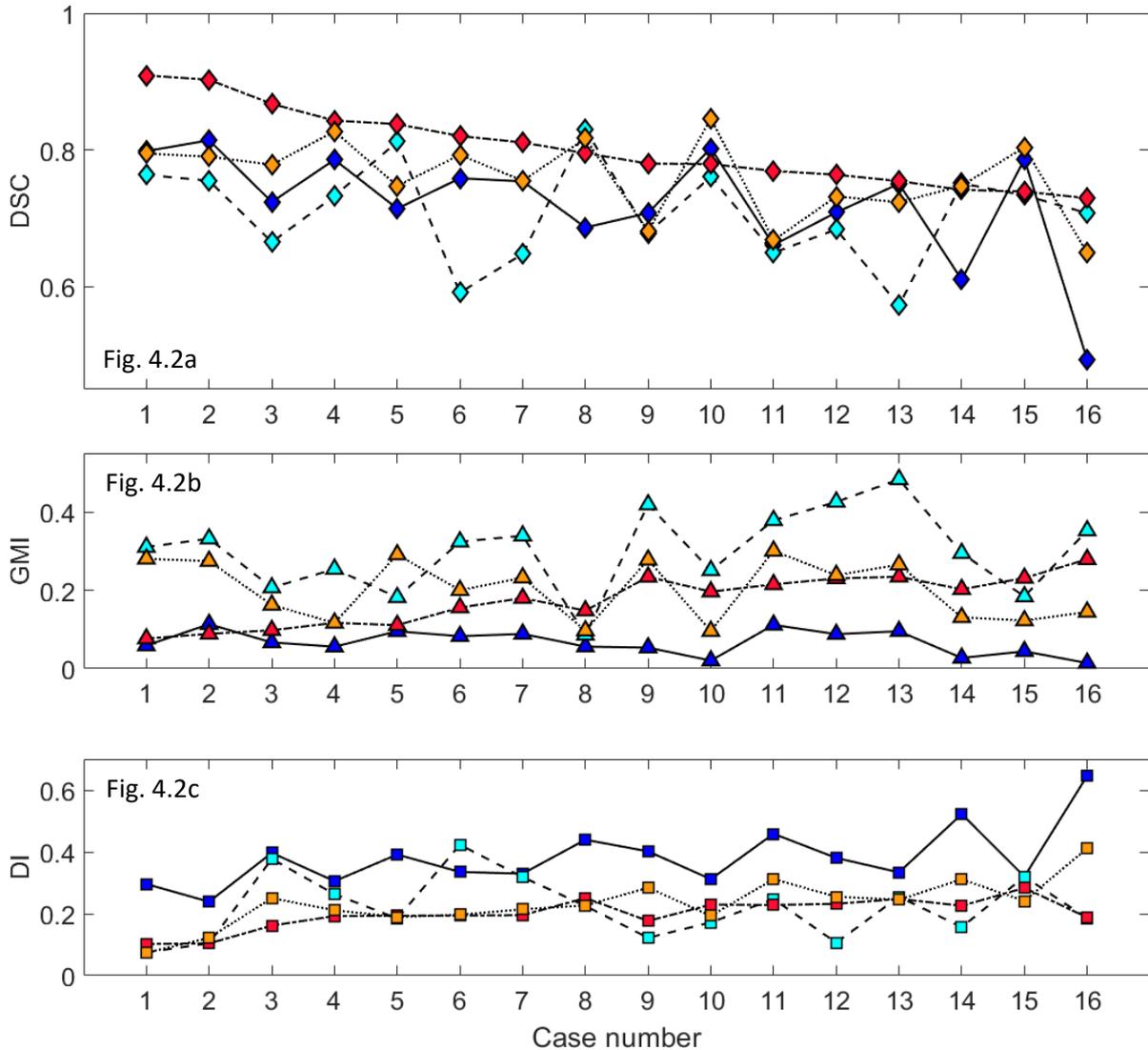
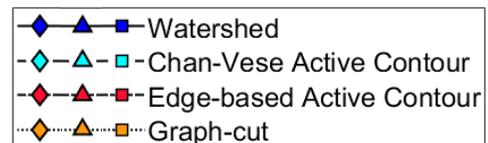


Figure 4.2. Conformity indices for segmentation of individual cases in comparison to reference contours. a) Dice similarity coefficient; b) Geographical miss index; c) Discordance index.



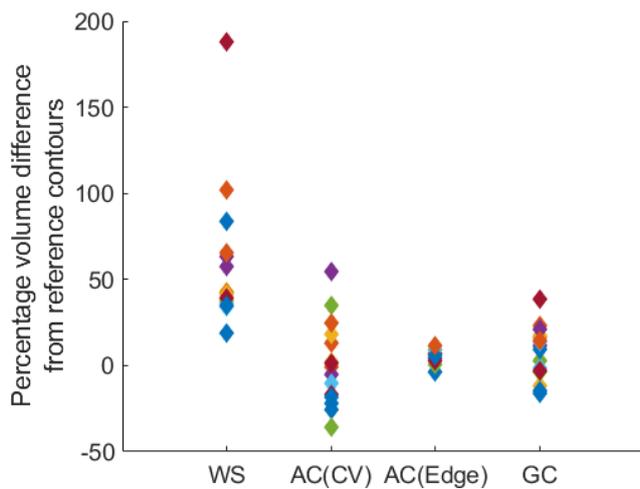


Figure 4.3. Percentage volume difference in relation to the reference contours for individual cases (mean and standard deviation shown in table). Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.

	Absolute Difference (cm ³)	Percentage Difference
WS	59.4 ± 51.8	58.2 ± 40.4
AC (CV)	6.8 ± 20.0	1.00 ± 23.8
AC (Edge)	9.8 ± 14.4	5.9 ± 3.9
GC	12.0 ± 24.3	7.4 ± 15.2

All the four segmentation techniques showed a tendency towards generation of volumes larger than the reference contours, as seen in figure 4.3. This effect was most pronounced for watershed segmentation, both in terms of percentage and absolute volume difference. The largest spread in volume difference was also seen with this technique, indicating that watershed produced volumes that were most dissimilar to the reference volumes, as compared to the other segmentation methods.

Although the Chan-Vese approach achieved a mean percentage volume difference close to 0, the spread of volume sizes was also relatively large judged by the standard deviation of 23%. Unlike the watershed approach, there were cases in which smaller volumes were produced by the Chan-Vese approach compared to the reference contours.

The mean percentage volume difference for the edge-based active contour approach was at 5.9%, which was associated with the smallest spread at 3.9%, suggesting that this approach resulted in volumes that were most similar to the reference volumes in terms of size.

The graph-cut technique was associated with a mean percentage volume difference of 7.4 ± 15.2%, where larger contours in comparison to the reference delineation were generated in the majority of the cases.

4.2.2.2 Qualitative analysis of segmentation performance

Representative cases are shown in figure 4.4 displaying the segmentation contours obtained with the different techniques are overlaid with the reference delineations. In general, the behaviours seen in the training phase were observed in this testing cohort.

Tumour coverage using the edge-based active contour was good in cases 2, 5, 1 and 3. Regions of GGOs and cavities were included appropriately in the segmentation for case 2, despite being located at the tumour periphery. The tumour was seen also to be appropriately covered with the Chan-Vese and graph-cut approaches in case 3, where the disease was juxtaposed to the non-enhanced pulmonary artery. However, the Chan-Vese approach failed to encompass most of the GGOs in case 2, while the graph-cut method managed to include this appropriately, barring a couple of slices. Like the training cases, underestimation of the tumour at the border of the lung parenchyma was seen for the Chan-Vese approach. This also occurred on some slices with the graph-cut approach at the mediastinal border (case 5).

In cases 16, 9 and 11 where mediastinal nodal disease was present, the tumour was also seen to be appropriately encompassed by the edge-based active contour in most parts. In contrast, the Chan-Vese algorithm generated contours approximated as a circle for small volume mediastinal disease. However, where the primary disease had a more elongated shape, the edge-based active contour approach failed to encompass the furthest tumour regions, which is most apparent in case 8. In such cases, the graph-cut method produced contours which were more congruent to the true tumour edge, due to its ability in generating contours which are more irregular.

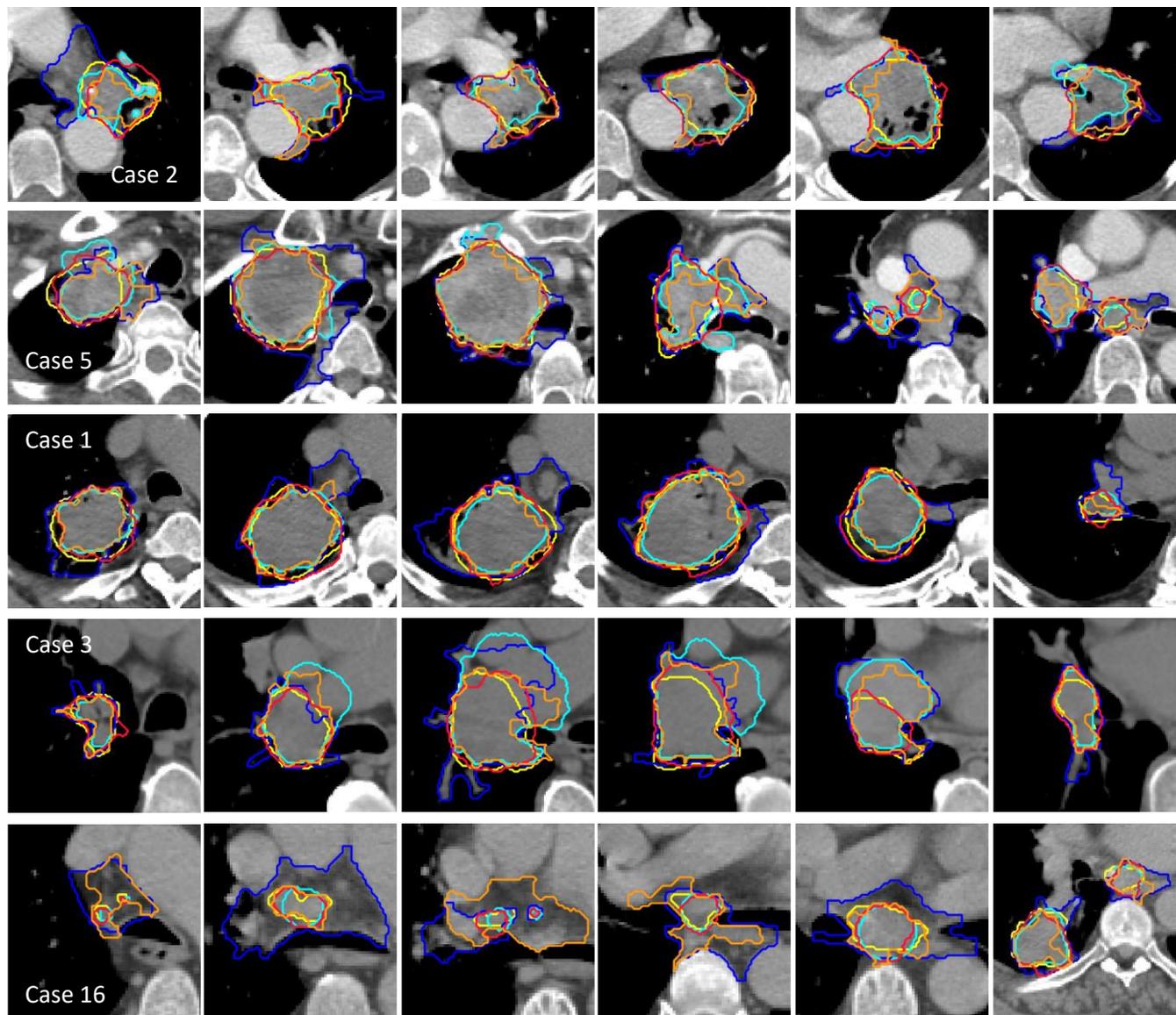
In concordance with the very low GMI scores, the watershed approach encompassed almost all tumour regions for all these cases, even in the presence of GGOs. However, this was at the expense of spillage of the contours into surrounding tissues, which was evident in most cases, albeit to different extents.

All four processes demonstrated the ability to differentiate regions of high contrast (e.g. ribs and contrast-enhanced vessels) from the tumour, which would have been in part due to the application of the exclusion masks. In case 15, there were some small errors at the ribs for with the edge-based active contour, though this was by far the minority of the observations. For adjacent vessels with low levels of contrast (cases 8, 11 and 15), leakage of the contours was seen with the Chan-Vese approach on some slices, which was not an issue seen with the other methods. Nevertheless, all four techniques had difficulty in distinguishing tumour from vessels where contrast was not administered (cases 3 and 10), though the edge-based active contour had the best performance in this setting. It was also seen to be best at approximating to the tumour in the presence of surrounding collapse in case 10.

The region of atelectasis in case 1 was appropriately excluded from the segmentation with the graph-cut and two active contour approaches but was included in the results with watershed technique. Leakage of the segmentation into the mediastinal fat was most apparent for the watershed approach followed by the graph-cut method (cases 16, 11, 8 and 5).

On some slices, the segmentation boundary with the edge-based active contour did not conform tightly to boundaries defining sharply contrasting regions, e.g. between lung parenchyma/airways and tumour (cases 10, 15 and 8). This is likely to be a consequence of the contour initialisation being placed at a distance from the tumour edge. On the other hand, there was better conformity at these boundaries with the Chan-Vese active contour approach, where the contour was seen to divide and included multiple areas of high intensity (first image in case 2). The watershed approach also produced segmentation congruent to the boundaries with sharp contrast, but it had the propensity to include vessels within the lung parenchyma (case 3), creating irregular outlines.

At the chest wall (cases 5 and 15), variable performance was observed. Leakage into the musculature was most evident with watershed, while the other techniques had mixed behaviour between the slices.



Segmentation	DSC	GMI	DI
Watershed	0.81	0.11	0.24
AC (CV)	0.75	0.33	0.10
AC (Edge)	0.90	0.09	0.10
Graph-cut	0.79	0.27	0.12
Watershed	0.71	0.10	0.39
AC (CV)	0.81	0.18	0.18
AC (Edge)	0.84	0.11	0.19
Graph-cut	0.75	0.29	0.19
Watershed	0.80	0.06	0.30
AC (CV)	0.76	0.31	0.08
AC (Edge)	0.91	0.08	0.10
Graph-cut	0.79	0.28	0.07
Watershed	0.72	0.07	0.40
AC (CV)	0.66	0.21	0.37
AC (Edge)	0.87	0.10	0.16
Graph-cut	0.78	0.16	0.25
Watershed	0.49	0.01	0.65
AC (CV)	0.71	0.35	0.19
AC (Edge)	0.73	0.28	0.19
Graph-cut	0.65	0.14	0.41

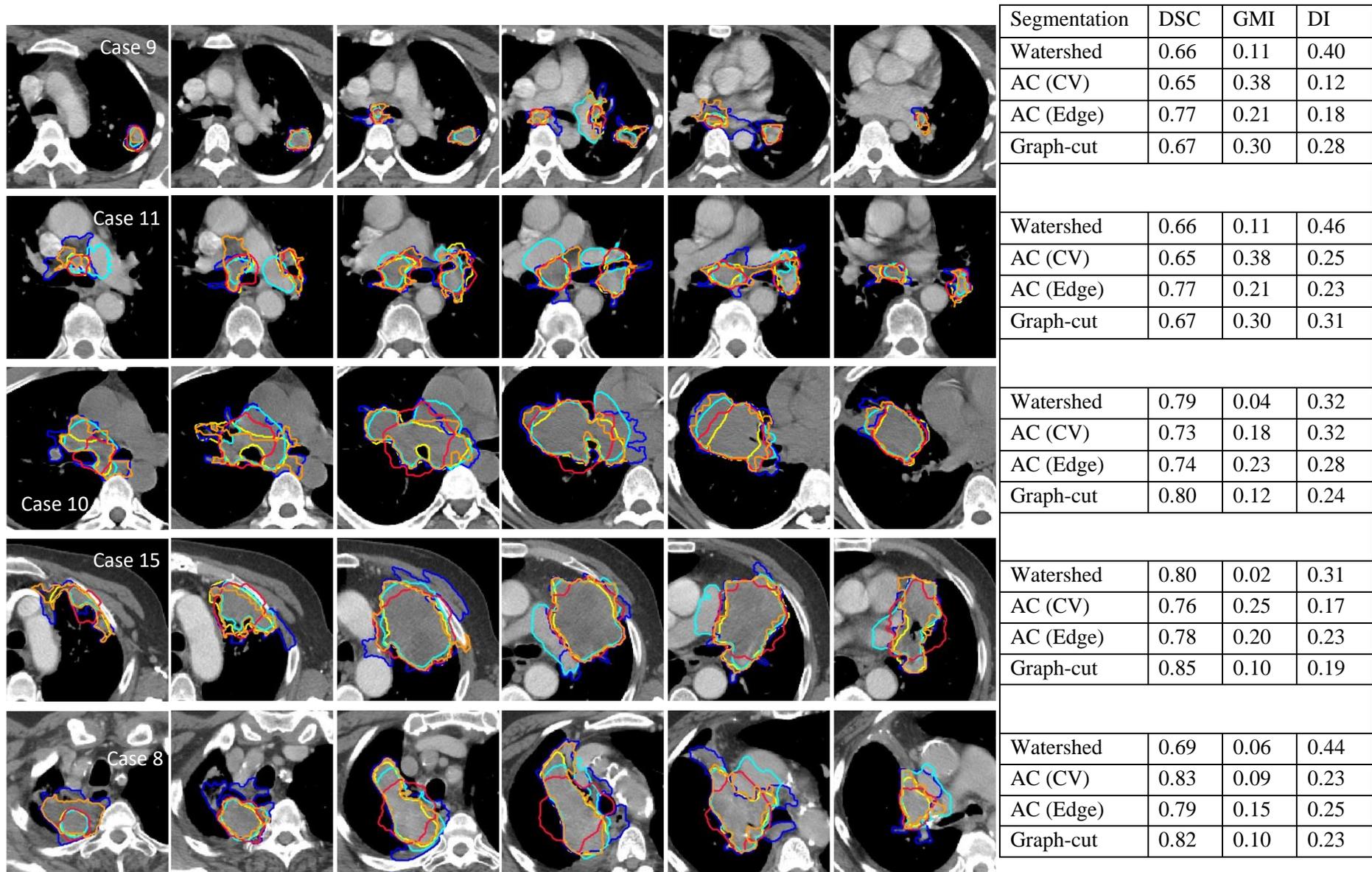


Figure 4.4. Segmentation results for representative individual cases. Blue – Watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow – reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour)

4.2.3 Computational time

There was a large difference in the mean computational time per case for the different segmentation methods. The watershed approach was the fastest where most cases were processed in under half a minute, followed by Chan-Vese active contour in under a minute. Most cases were processed within 2 to 4.5 minutes with edge-based active contour, and the graph-cut segmentation took the longest time between 8 and 20 minutes.

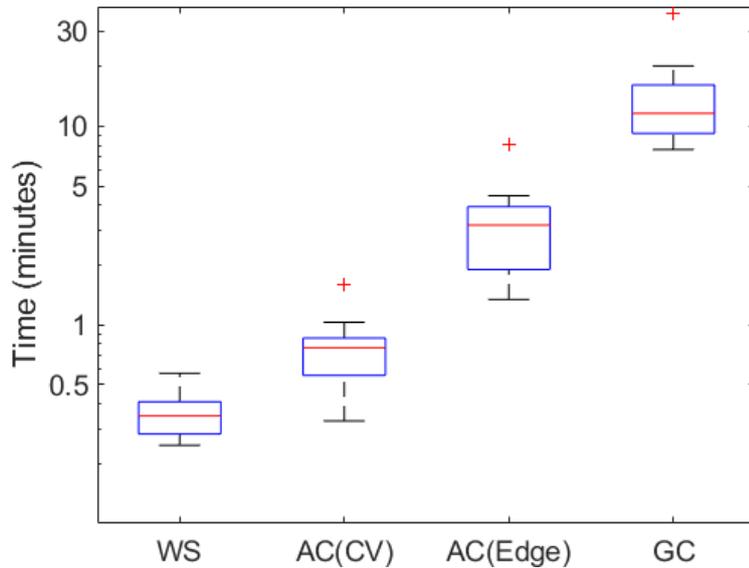


Figure 4.5. Time (minutes in logarithmic scale) for processing individual cases, with the mean and standard deviation shown in the table. (Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC).

WS	21 ± 5 seconds
AC (CV)	45 ± 18 seconds
AC (Edge)	3.2 ± 1.7 minutes
GC	13.9 ± 7.3 minutes

4.3 Summary of performance for different segmentation approaches

A summary of the observed performance and general behaviours of the different techniques specific to these workflows is displayed in table 4.1.

Technique	Parameter tuning	Use of prior	Speed	Quantitative analysis	General remarks	GGO	Cavitation	Elongated primary	Lung edge	Hilum/ Mediastinum	Chest wall	Exclusion of atelectasis	Nodal disease
Marker-controlled watershed	Not required	Internal and external limits	Fast (short seconds)	DI >> GMI (Higher recall, lower precision)	Dependent on competing gradients; Irregular contours	Acceptable	Acceptable	Acceptable	Acceptable	Poor in the face of competing gradients resulting in leakage into mediastinal fat and non-contrast enhanced vessels	Mixed performance dependent on bounding region	Mixed performance dependent on bounding region	Poor in the face of competing gradients (e.g. contrast-enhanced vessels, airways)
Chan-Vese active contour	Contraction bias; Smoothing factor	External limit; Size approximation for initialisation	Fast (long seconds)	GMI > DI (Lower recall, higher precision)	Dependent on location of initialising contour; Smooth contours; Ability to split and merge	Poor	Mixed performance if present at tumour boundary	Poor – sensitive to initialisation	Under-estimation of tumour	Poor with leakage into non-contrast enhanced vessels and vessel with low levels of contrast	Mixed performance, dependent on localising boundary	Generally acceptable, dependent on localising boundary	Dependent on parameter tuning
Edge-based active contour	Contraction bias; Smoothing factor	External limit; Size approximation for initialisation	Short minutes	DI ≥ GMI (Higher recall, lower precision)	Dependent on location of initialising contour; Smooth contours	Acceptable	Acceptable	Poor – sensitive to initialisation	Generally acceptable but occasional leakage into lung	Mixed performance (leakage into non-contrast enhanced vessels and airways dependent on localising boundary)	Mixed performance, dependent on localising boundary	Generally acceptable, dependent on localising boundary	Mostly acceptable
Graph-cut	Superpixel: number of regions, (non-adaptive) SLIC compactness; Lazysnapping : Edge weight scale factor	Foreground and background markers	Long minutes	DI ≥ GMI (Higher recall, lower precision)	Irregular contours	Acceptable	Mixed performance if present at tumour boundary	Acceptable	Acceptable	Mixed performance (leakage into mediastinal fat and non-contrast enhanced vessels)	Mixed performance	Mixed performance	Poor with leakage into mediastinal fat and non-contrast enhanced vessels

Table 4.1. Summary of the performance and general behaviours of the different techniques specific to the established workflows developed in this project.

Discussion

The performance of automatic tumour definition in a heterogeneous case mix of locally advanced NSCLC by the four different segmentation techniques was explored in this work. In both the development and independent testing phases, the edge-based active contour approach produced delineations most similar to the reference contours when judged by the mean DSC scores, followed by the graph cut approach.

There are several novel contributions in this work. Firstly, fully automatic processes were developed for the initiation and segmentation of the lung tumours based on the submitted contours, circumventing the need for any further manual input, in the context of generating reference volumes for quality assurance purposes. Additionally, the evaluation was conducted on heterogeneous clinical data from different centres with different scan acquisition parameters, as opposed to many of the reported segmentation studies where single data sources have been used. This study also explored the performance of the segmentation techniques in advanced lung cancer and encompassed disease with a wide spectrum of differing radiological appearances, reflecting the disease seen in the clinical context and thereby providing a better estimate of the performance of the segmentation in this setting. This also allowed an assessment of the behaviours of the different techniques at different tissue interfaces.

Incorporation of *a priori* knowledge with localising markers

The use of the submitted contours as the initial localising markers helped with all four segmentation processes on several folds. Introduction of *a priori* knowledge to assist with localising the segmentation to the region around the GTV facilitated the exclusion of regions which would be otherwise more challenging to segment, such as regions of adjacent lung collapse. This is based on the confidence that the submitted contours are in close proximity to the “true” tumour boundary, and thus can be used to locate and initialise the segmentation process. With the input of this information, it also enabled the segmentation processes to be fully-automated, whilst generating acceptable contours in a number of cases, albeit with varying successes with the different techniques. This was performed through the automatic generation of initialising masks for the active contour approaches and markers for watershed and graph-cut methods. In the case of watershed technique, oversegmentation was resolved through this process.

One concern with the use of the submitted contours in the segmentation process is the risk of overfitting of the resultant segmentation to the initiation locations, which would limit the aim of generating the contours, that is to act as the surrogate reference to which the clinician submissions would be assessed against. This was taken into account in the design of the different segmentation processes. Despite using the submitted contours to constrain the search region for the watershed and graph cut approaches, as the final segmentation results of the algorithms were based principally on the underlying image properties, the risk of overfitting to the shape of the submitted contours was low. On the other hand, segmentation using the active contour approaches is sensitive to the initialisation curve, which can influence the location and shape of the delineation results. To allay this impact, the initialisation was approximated with a circle of the same size as the submitted contours for both active contour methods thereby decreasing the reliance on the submissions.

Comparison of segmentation techniques

The different patterns of behaviour for the four segmentation techniques was made apparent through the use of a heterogeneous case mix of tumours involving different locations in the chest. Although some common trends were shared, each approach had their specific strengths and weaknesses, resulting in different methods outperforming the others dependent on which surrounding tissue was present at the particular tumour front. This mirrors the reports in the literature on observed differences in segmentation quality between software products applying different segmentation methods. For example in a clinical study comparing the segmentation accuracy of pulmonary metastases across six CAD systems, there were significant differences in the performance of the software packages, where between 71% and 86% of nodules were segmented adequately (409). Even with the input of manual correction in four of the packages, although there was an improvement to 98% for one of the packages, the performance in another was at 76%, suggesting variation in segmentation quality between different software with different segmentation techniques.

The results have shown that the watershed approach is a fast and quick technique. It worked well at tumour regions with surrounding lung parenchyma, but it suffered from leakage in the presence of gradient changes between normal tissues that were of higher magnitude than the tumour and its surroundings. This occurred at the mediastinum in the presence of vessels with high contrast, as well as at the airways with low density, and was also observed in some cases to cause segmentation errors at the chest wall in the presence of ribs. As a result of this, although good recall was seen with the watershed approach, this was at the expense of poor precision.

On the other hand, the graph cut approach was most expensive computationally, which was sped up through the application on pre-generated superpixels in order for the technique to be processed within acceptable time frames. This added to the complexity of the workflow and required additional steps to calibrate to the optimum parameters for superpixel generation, in addition to the parameters for the graph-cut algorithm. Nonetheless, acceptable segmentation quality was observed for a substantial proportion of cases, with a balance between its performance at accuracy and precision. Despite this, the behaviour of the graph-cut approach was inconsistent at some regions in the axial plane with poor concordance to contrast information, resulting in poor tumour coverage in these regions.

Although both the graph-cut and watershed applications did well in conforming to irregularly shaped tumours, they were susceptible to producing delineations that were more irregular than those typically constructed by human observers. Comparatively, the parabolic surfaces obtained through the active contour approaches better reflect the smooth geometry seen with manual delineations. As parameter tuning is paramount, one of the strengths of this work is in the systematic assessment and selection of the optimal parameters that were used, which is not typically reported in the literature. However, the need for parameter tuning to best fit to the whole dataset would contribute to some of the observed segmentation errors because of the difficulty in selecting generalisable parameters that fit all cases well. This was considered when performing the manual selection of the parameters, which was based primarily on the mean DSC scores. Instead of selecting the value associated with the best observed mean DSC, the variation of DSC, recall and precision with the parameters was taken into account. Moreover, although parameter-based methods enable control over the segmentation, the need for optimisation introduces a source of uncertainty to the process.

Comparison of quantitative reports of (semi-)automatic approaches in the literature

The direct comparison of these results to other studies is difficult due to the dissimilarities in the segmentation processes, variation in the cases analysed, as well as the differences in the assessment matrices used. Nonetheless, the quantitative results based on mean DSC seem to correlate with the observed performance of some of the (semi-)automated segmentation approaches reported in the literature.

For studies evaluating active contour approaches, Yip et al used an geodesic active contour algorithm in 3D Slicer that is also based on a level set formulation, where a median DSC of 0.6 was achieved for lung nodule segmentation (242). Way et al reported an overlap index of just under 0.6 (at 50% agreement for ground truth) based on delineation of lung nodules, using an initial k-means clustering approach followed by a 3D active contour refinement, where both 2D and 3D gradient information were incorporated as part of the external energy computation (241). Even taking into account the smaller tumour sizes in these studies that can yield lower scores, the mean DSC observed here for the edge-based active contour approach is comparatively higher at 0.80. Additionally, the achieved DSC scores for the edge-based and Chan-Vese approaches were similar to the results obtained by Yu et al in their nodal segmentation study (282). They did however, achieve higher DSC scores when an additional edge constraint was used along with the region-based snake.

Comparatively higher mean DSC scores were achieved by Shen et al (249) and Lermé et al (247) in their use of the graph cut approach, although both of these studies were semi-automatic and required user strokes for initialisation, which can explain the better performance seen. These studies were also performed with 4D and 3D techniques respectively, where the additional contextual information in the added planes can provide better results.

In the work by Tan et al, a semi-automatic approach with combined marker-controlled watershed segmentation followed by gradient-based active contour achieved a mean overlap index of 0.69 for clinical lung cancer cases (231). Yan et al obtained a mean overlap index of 0.83 in their semi-automatic watershed technique for segmenting nodal disease (287), although the mean score dipped to 0.73 when an automatic watershed approach was applied to sequential scans (288). Considering that DSC is usually a more optimistic estimate than overlap index, the performance of the automatic watershed approach was poorer here. However, the marker placement in the work by Yan et al was considerably smaller than what was used here, with external and internal margins at 7- and 3-pixels respectively (288).

Comparison with quantitative reports of manual delineation in the literature

These results appear to fall within the variability associated with manual delineation of advanced lung tumours when assessed quantitatively, with the mean DSC scores being at least as good as the reported DSC in studies assessing manual contours against a gold standard representative contour. The mean Kappa index, equivalent to the DSC score, was reported to be 0.72 in a study performed by Dewas et al in the delineation of locally advanced lung tumour (410). Louie et al reported that the mean inter-observer scores equivalent to DSC was at 0.512 ± 0.050 for delineation of the lung primary disease (411), where the assessment was performed in a pair-wise fashion. There was poorer conformity for nodal disease, where the equivalent mean DSC was at 0.29 ± 0.09 , although there may be some error in this estimation due to the smaller size of the nodal disease. Even so, these scores are lower compared to the results here,

despite their evaluation of disease at earlier clinical stages (Ib, II and IIIa) than this study. This suggests that automated approaches for advanced disease fall within the variability associated with manual contours, and may be used in the clinical context.

Qualitative performance

However, despite achieving comparable mean DSC scores with the different approaches, unsatisfactory delineation was observed qualitatively in a substantial proportion of cases. Because of the variety of the juxtaposed normal tissue surrounding the tumour, both good and poor segmentation could be present on a single axial slice, which is difficult to quantify and better described qualitatively. Having said this, the assessment of the DMI and DI scores in addition to DSC gave a better indication of the performance of the techniques.

Mixed performance was observed in many cases, with errors in segmentation that was dependent on the methods used. A common limitation to the techniques explored is the difficulty in segmentation at the mediastinum. As the watershed approach located regions of high gradient, mediastinal fat was often included in the segmentation. This was also a problem with graph-cut which displayed poor precision for nodal delineation. The Chan-Vese approach tended to balloon into regions of similar density, in particular vessels with no or little contrast. Although edge-based active contour exhibited good performance for the nodal disease, some errors were seen for the primary disease at the mediastinum where the localising boundary did not approximate well to the tumour edge.

It was also observed that there was some inconsistency with the segmented volumes with poor adherence on some slices but good conformity on others. Extension of the analysis from 2D to 3D with the addition of information from the Z-plane may improve this. Additionally, it would also allow an estimation of the cranial-caudal extent of the tumour, which is not provided with a 2D approach.

Reproducibility

One other limitation to this study is the lack of reproducibility assessment. In order to test the robustness of the workflows, it is important to evaluate for potential differences in results if different priors were used. This is crucial for application of the workflow in the wider context.

Need to improve segmentation techniques

In addition to conducting reproducibility studies, there is a need to improve the quality of the obtained segmentations. This is vital as the intended use of the generated contours is to act as a surrogate reference against manual contours for quality assessment.

The main difficulty in segmenting the lesions in advanced lung cancer is in its separation from adjacent tissues of similar density. It is a challenge to develop a robust and efficient means of handling the range of tumour appearance, size and location from the heterogeneity of cases. Many segmentation algorithms suffer from the presence of subtle lesion boundaries, limiting their ability to detect and define the lesion edge. In the work by Yip et al, minor to substantial manual adjustments were required for 87% of the semi-automated generated contours, with poor performance of the algorithm for lung nodules which were non-/part-solid, and for lesions with poorly defined boundaries (242).

To overcome this, methods based on shape approximation would be suitable for small ellipsoid nodules but less applicable for larger irregularly shaped lesions (264, 297). This issue was seen in here with the active contour approaches.

Others have used morphological methods to separate the tumour from adjacent high density structures, such as the work by Kostis et al (213) where smaller lesions were investigated, and Kuhnigk et al (224) where the approach was designed to fit larger lesions. Such methods capitalise on the unique geometry of tumours compared to background organs and typically involve a series of different steps to distinguish the surrounding organs from the tumour. For example, the chest wall can be approximated by a convex hull of a pre-segmented lung region (213, 214, 224). However, these approaches are unlikely to be generalisable to tumours at a multitude of locations in the chest, as the geometric properties of the tumour and their surrounding tissue is different at the peripheral chest wall, hilum and mediastinum, notwithstanding presence of nodal disease. It would also not be appropriate to base the segmentation on geometric properties in the case of disease infiltration into the chest wall.

Concurrent segmentation of the primary lesion and nodal disease is a challenge. To this end, Moltz et al designed separate interactive segmentation tools for pulmonary nodule and lymph node segmentation, where a watershed-based approach was added to the module in order to improve the segmentation leakage (214). This would be difficult to implement in the setting of advanced lung cancer, as the primary disease is often indistinguishable from the nodal disease at the hilum, which is typically contoured as a single volume.

It was seen from the initial experiments that the watershed approach can be improved with the definition of more exclusion structures. However, it had been difficult to find a robust automatic process to perform this. Thus, additional manual input into the developed algorithm was evaluated as a means to improve the performance of the watershed segmentation, where the mean precision was significantly increased compared to the fully-automatic approach. Many other groups that have achieved successes with semi-automated approach. For example, Velazquez et al (217) used the GrowCut semi-automatic algorithm in 3D Slicer on 20 cases of NSCLC with user seed initialisation and interaction. They achieved overlap fraction (overlapping region divided by the smallest region) of greater than 0.9 in the comparison of the computed-aided segmentations against the union of multiple manual delineations, and also found strong correlation (Spearman's correlation = 0.89) of the semi-automated segmentation to pathology specimens. The implementation of additional manual interaction in the watershed algorithm is relatively uncomplicated, and takes the form of external limits. Further manual internal limits were not required, as the automated means of internal marker placement was sufficient to generate good results. To extend this approach to graph-cut segmentation, there would be a need to define both foreground and background markers, improve both accuracy and precision.

Many other groups have used combination techniques to improve the segmentation performance. Based on the techniques explored in this study, avenues of further investigation include the application of the watershed and/or the graph-cut techniques prior to refining the segmentation with the active contour algorithms, which has the advantage of improving the smoothing of the final delineation. However, it is unclear as to whether such a combination approach would result in a more accurate and precise delineation based on the observed performance in this study. As the active contour approaches perform better with initialisation

close to the true tumour edge, the segmentation from watershed and graph-cut approaches may be too imprecise for this. The potential improvements obtained with combination techniques would also need to be evaluated against the offset of increased computational time.

Conclusions

Workflows for the automatic segmentation of advanced lung cancer was established through the application of a given prior using four image segmentation techniques. The edge-based active contour segmentation gave the best performance overall, with both tumour coverage and avoidance of normal tissue taken into consideration. Moreover, the case processing time of a short number of minutes was within acceptable limits. Despite producing acceptable segmentation results for a number of cases, there remained a significant proportion where unacceptable segmentation was produced. Errors in accuracy of irregularly shaped lesions and segmentation leakages at the mediastinum were observed. Other different approaches need to be explored to improve the quality of the segmentation.

Chapter 5

Specific Aim C: Assessment of performance of segmentation techniques in isolated peripheral tumours

Introduction

The performance of the segmentation techniques was explored in chapters 3 and 4 for disease located at different positions within the thorax. In the presence of multiple adjacent tissue types, variation in segmentation performance for the different techniques was observed. Following on from this work, the performance of the segmentation techniques for isolated peripheral lung tumours is assessed in this chapter where the tumour is surrounded by lung parenchyma in its entirety. The rationale for this is to evaluate how the automatic segmentation performance differ in the setting of less advanced disease, which most of the reports on automatic lung lesion segmentation are based on. Here, the established automated segmentation workflows are applied to isolated lung tumours. The hypothesis is that different segmentation performance between the different techniques would be observed in the presence of juxtaposition of a single tissue type, in contrast to the performance seen in the preceding two chapters.

In the absence of a different study cohort comprising solely of peripheral lung tumours, cases were extracted from the IDEAL trial and evaluated.

5.1 Summary of tasks

Task C.1 Comparison of segmentation techniques on isolated peripheral primary disease

- A) Evaluation on training dataset
- B) Evaluation on testing dataset

Task C.2 Comparison of segmentation techniques on isolated peripheral primary disease with dataset from different trial source

Task C.3 Comparison of segmentation techniques on advanced lung tumours versus isolated peripheral primary disease

Methods

5.2 Gold-standard reference ROIs

Cases with peripheral primary tumours surrounded by lung parenchyma were identified from the cohort. Duplicate GTV structures were created to isolate the primary tumour from the nodal disease, which were used for purposes of this work.

5.3 Segmentation workflow and assessment

These GTVs were processed using the same workflows that were established for the evaluation of the entire cohort. For the active contour and graph-cut approaches, the same optimal parameter settings were used.

The contour assessment was performed using the same quantitative measures (DSC, GMI and DI) and the quality of the segmentation was assessed visually.

5.4 Datasets

Using the same division of data as described in chapter 3 section 3.7, cases from the training dataset were identified and processed separately to the cases in the independent testing data. As the number of cases was small, data from the ISTART trial was also screened for suitability for use in this study. The ISTART trial is a multi-centre UK phase I/II trial of isotoxic accelerated radiotherapy in the treatment of patients with NSCLC, where 3D and 4D CT based delineation were permitted (412). As the ITV is delineated in 4D CT outlining rather than the GTV, cases using 4D CT for planning were excluded from this analysis.

5.5 Computational time

All the processes in this work were performed using an Intel Core i5-3317U CPU @ 1.70GHz, 4GB RAM on a Windows 10 64-bit environment.

Results

The performance of the different segmentation techniques is presented for the training dataset, followed by the independent test set. Subsequently, results from the different trial source is shown. The segmentation results for advanced and isolated peripheral lung tumours are summarised.

5.6 Task D.1 Comparison of segmentation techniques on isolated peripheral primary disease

5.6.1 Performance of training dataset

5.6.1.1 Segmentation performance

Ten cases in the total training dataset were identified where peripheral primaries were present. The mean volume of disease was considerably smaller than the whole dataset at $13.03 \pm 15.45 \text{ cm}^3$.

Marker-controlled watershed segmentation was associated with the highest DSC at 0.84 ± 0.04 , a good GMI of 0.08 ± 0.04 , and a DI of 0.21 ± 0.09 . Although the DI was the highest amongst the various approaches, as compared to the trends observed with the whole dataset in chapter 3, this score had improved by about half. Edge-based active contour and graph-cut had similar performance in the quantitative assessment, with better DI as compared to the watershed technique, but worse GMI. Overall, these approaches achieved a mean DSC of 0.79 ± 0.07 and 0.76 ± 0.04 respectively. The Chan-Vese active contour was seen to achieve the lowest mean DSC of 0.56 ± 0.13 , which was associated with a poor GMI 0.57 ± 0.14 , despite attaining a DI of 0.01 ± 0.01 .

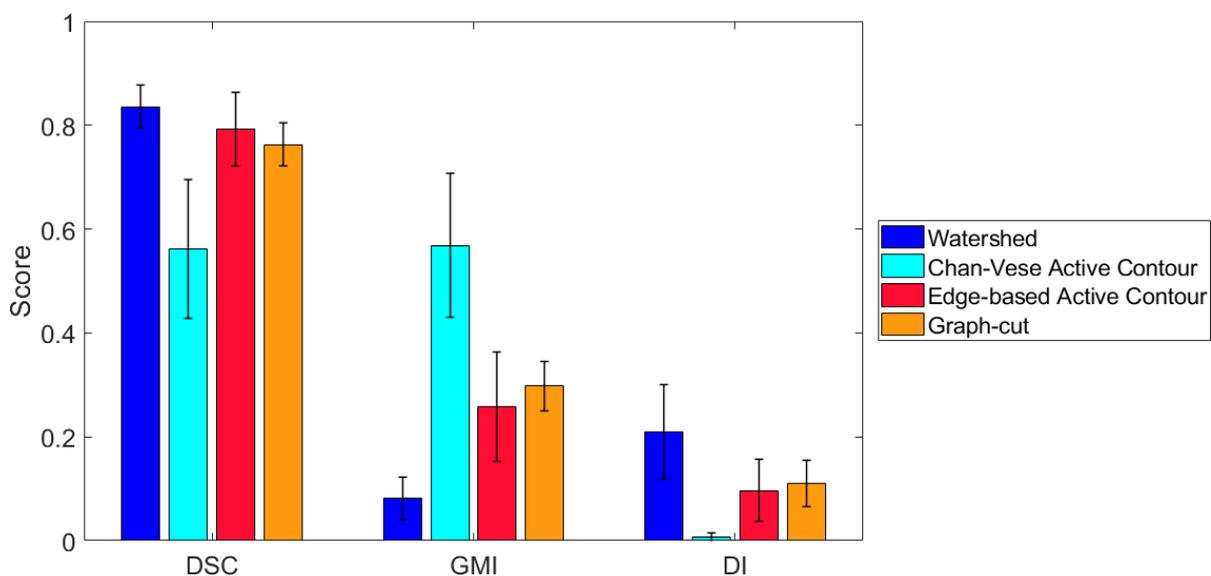


Figure 5.1. Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for peripheral lung primary disease within training set.

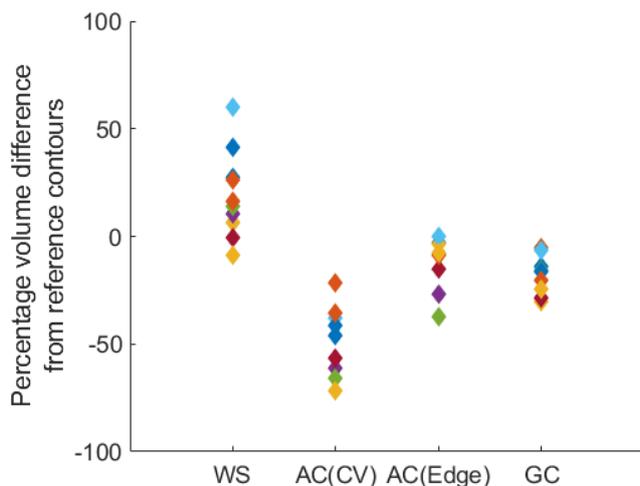
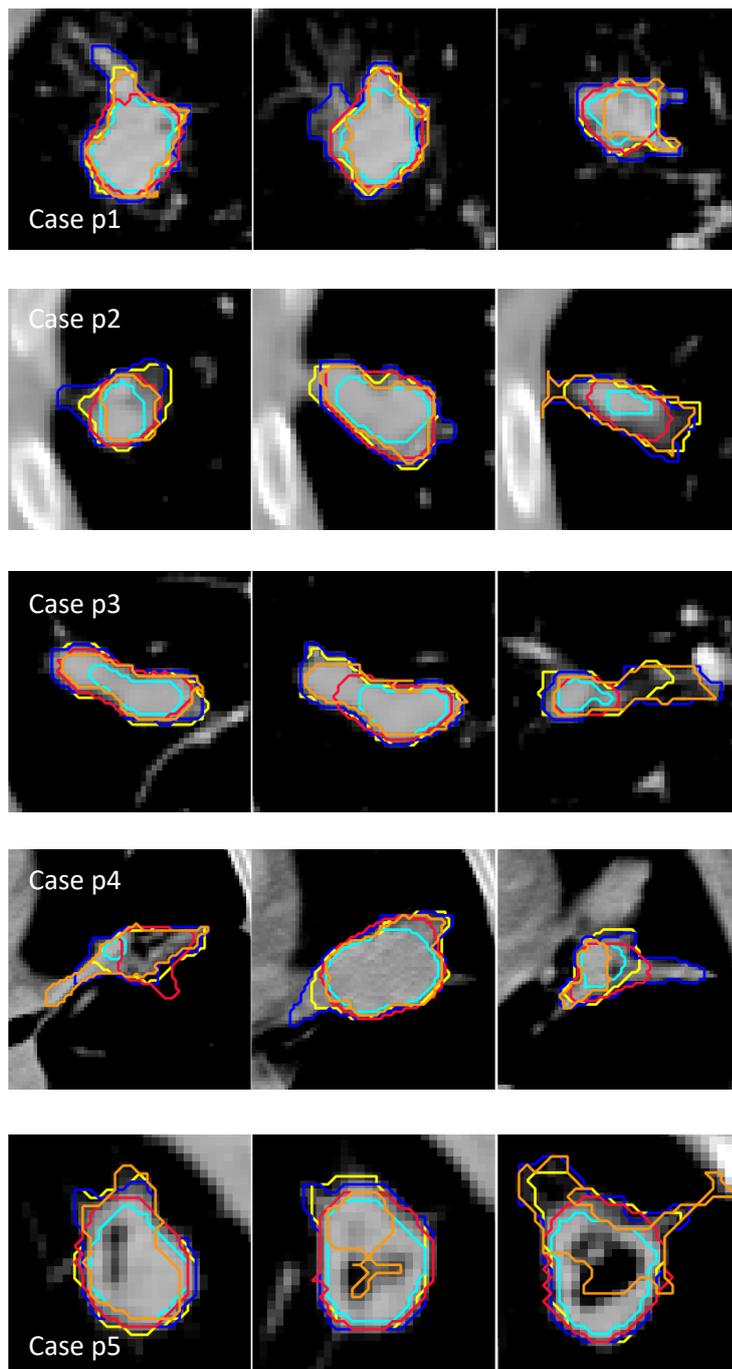


Figure 5.2. Percentage volume difference of peripheral primary disease in relation to the reference contours for individual cases (mean and standard deviation shown in table). Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.

	Absolute Difference (cm ³)	Percentage Difference
WS	1.28 ± 2.71	19.4 ± 20.3
AC (CV)	-7.24 ± 11.55	-49.8 ± 15.9
AC (Edge)	-1.12 ± 1.33	-11.0 ± 12.2
GC	-2.63 ± 4.00	-18.6 ± 8.7

Consistent with attaining the highest DI scores, watershed segmentation was seen to produce larger volumes as compared to the reference contours (figure 5.2), whereas the other techniques created contours smaller than the reference in general. In particular, the Chan-Vese algorithm was seen to generate volumes that were on average half the size of the reference volumes.

Visual inspection of the segmentation corroborated with the quantitative results. The smaller contours generated by the Chan-Vese approach was evident in a number of cases as seen in figure 5.3, where the delineation did not seem to extend to the edge of the tumour when viewed on lung windowing levels. In case p3, the contours failed to extend to the anterolateral extent of the tumour, as well as regions of atelectasis in cases p3 and p4. This was in keeping with the behaviour seen in chapter 3. Edge-based active contour generally produced clinically acceptable contours, though it was seen still to suffer from poor segmentation for disease that was not well approximated by a circle (second slice in case p3). The graph-cut approach was able to exclude regions of atelectasis appropriately and was able to produce contours that fit the irregular edge of the tumour. However, like previously, it was seen not to work as well in the presence of tumour cavitation (case p5). Overall, the marker-controlled watershed segmentation seemed to work the best in terms of tumour coverage and in its ability to include regions of atelectasis and cavity, despite its tendency to create larger contours. It also tended to adhere to regions of large gradient contrast. For example, the lateral edge of the contour was extended to the vessel edge in the third slice of case p3, and the segmentation was extended to include the band of atelectasis in the third slice of case p4.



Segmentation	DSC	GMI	DI
Watershed	0.78	0.04	0.34
AC (CV)	0.68	0.47	0.02
AC (Edge)	0.85	0.20	0.09
Graph-cut	0.78	0.28	0.14
Watershed	0.86	0.10	0.17
AC (CV)	0.42	0.70	0.00
AC (Edge)	0.69	0.43	0.02
Graph-cut	0.70	0.24	0.10
Watershed	0.84	0.07	0.21
AC (CV)	0.51	0.65	0.01
AC (Edge)	0.71	0.43	0.03
Graph-cut	0.80	0.24	0.13
Watershed	0.79	0.08	0.30
AC (CV)	0.55	0.58	0.02
AC (Edge)	0.78	0.28	0.11
Graph-cut	0.75	0.27	0.89
Watershed	0.90	0.11	0.09
AC (CV)	0.54	0.60	0.00
AC (Edge)	0.86	0.21	0.06
Graph-cut	0.73	0.36	0.09

Figure 5.3. Segmentation results for training dataset (peripheral primary disease; cases p1 – 5). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour)

5.6.1.2 Computational time

As this analysis involved smaller volumes, the computation time for processing whole cases was shorter than previously seen. Watershed segmentation produced the fastest results in under 15 seconds, whereas most of the tumours were processed by graph-cut segmentation in less than 5 minutes. Cases were segmented by the two active contour approaches in less than one minute.

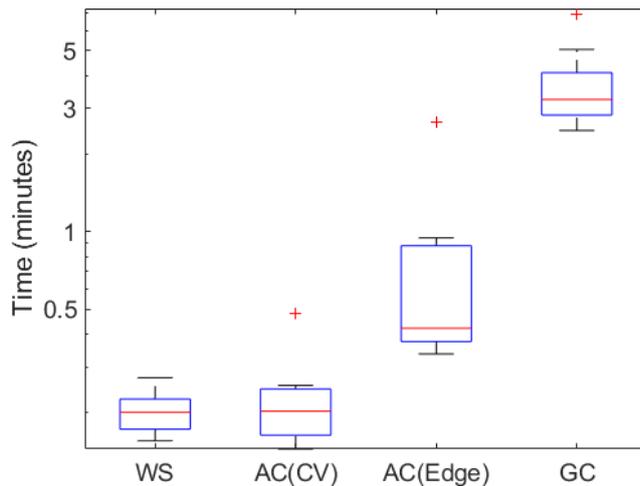
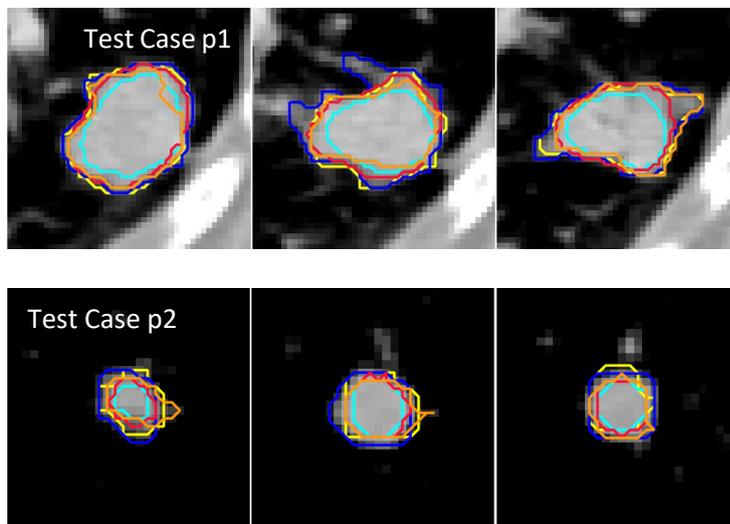


Figure 5.4. Time (minutes in logarithmic scale) for processing individual cases (peripheral primary disease only), with the mean and standard deviation shown in the table. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.

WS	12 ± 2 seconds
AC (CV)	14 ± 6 seconds
AC (Edge)	45 ± 42 seconds
GC	3.8 ± 1.4 minutes

5.6.2 Performance of independent test dataset

There were only two cases in the testing dataset that had isolated peripheral primary disease, and the segmentation results from these cases are shown in figure 5.5. Similar to the observations in the training dataset, the watershed approach had the best DSC amongst the various segmentation techniques. Again, the tendency for segmentation leakage was seen (second slice in test case p1) but overall there was greatest consistency with the reference contours using this approach. Moreover, the lowest GMI scores were associated with this approach.



Segmentation	DSC	GMI	DI
Watershed	0.85	0.06	0.19
AC (CV)	0.68	0.48	0.00
AC (Edge)	0.84	0.24	0.05
Graph-cut	0.76	0.29	0.11
Watershed	0.85	0.08	0.20
AC (CV)	0.66	0.50	0.01
AC (Edge)	0.78	0.35	0.02
Graph-cut	0.82	0.25	0.10

Figure 5.5. Segmentation results for testing dataset (peripheral primary disease; test cases p1 – 2). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour)

The reference volumes of test case p1 and p2 were 8.85cm³ and 0.64cm³ respectively. The watershed approach produced contours larger than the segmentation volume, though this was seen to be small in absolute terms (table 5.1). This technique also produced volumes most

similar to the reference judged by the percentage volume difference and was also associated with short computational times.

	Absolute volume difference from reference contours (cm ³)		Percentage volume difference from reference contours		Processing time	
	Test case p1	Test case p2	Test case p1	Test case p2	Test case p1	Test case p2
Watershed	1.40	0.09	15.8	14.7	13 sec	9 sec
AC (CV)	-3.95	-0.31	-44.6	-48.0	11 sec	7 sec
AC (Edge)	-1.26	-0.20	-14.3	-31.6	35 sec	9 sec
Graph-cut	-1.73	-0.10	-19.5	-15.9	4.6 min	1.2 min

Table 5.1. Absolute and percentage volume difference of peripheral primary disease from reference contours for individual cases, with the respective computational processing time.

5.7 Task D.2 Comparison of segmentation techniques on isolated peripheral primary disease with dataset from different trial source

5.7.1 Clinical and scanning parameters of ISTART dataset

There were thirteen cases in the ISTART dataset with isolated peripheral primary disease. Three cases were excluded as the tumour delineation was based on 4D CT outlining. Of the remaining ten cases, two cases contained two primary lesions, resulting in a total of twelve lesions for the segmentation assessment with a mean volume of 8.29 ± 11.75 cm³. Contrast-enhancement was used in six out of the ten cases.

The scanning parameters for the ten cases are shown in table 5.2, where two out of the four centres were new centres that did not participate in the IDEAL trial. Variation in CT scanning practices were present across the centres.

Centre	No. of cases	CT Manufacturer	Slice Thickness (mm)	Pixel Spacing (mm)	Tube Voltage (kV)	Modulating or fixed current	Tube current (mA) Mean (range)
J	5	Siemens	3	0.875 to 0.977	120	Modulating	155.5 (90 – 230)
K	2	GE Medical Systems	2.5	0.977	120	Modulating	244.8 (185 – 305)
F	1	Philips	3	1.172	120	Modulating	117.1
	1	Unknown	3	1.172	120	Fixed	37
G	1	Philips	3	1.172	120	Modulating	139.6

Table 5.2. Scanning parameters for the 10 evaluated cases in the ISTART dataset.

5.7.2 Performance of ISTART dataset

5.7.2.1 Overall performance

Similar results were obtained for the ISTART dataset to the cohort from the IDEAL trial. Watershed segmentation was the best performing technique at a DSC of 0.81 ± 0.05 , GMI of 0.12 ± 0.07 and DI of 0.22 ± 0.09 . Edge-based active contour achieved the next best DSC at 0.75 ± 0.11 with a GMI of 0.34 ± 0.14 and DI of 0.06 ± 0.04 , followed closely by graph-cut with a DSC of 0.73 ± 0.08 , GMI of 0.31 ± 0.08 and DI of 0.14 ± 0.10 . The Chan-Vese approach scored the lowest DSC at 0.48 ± 0.14 , with a GMI of 0.66 ± 0.12 and DI of 0.00 ± 0.01 .

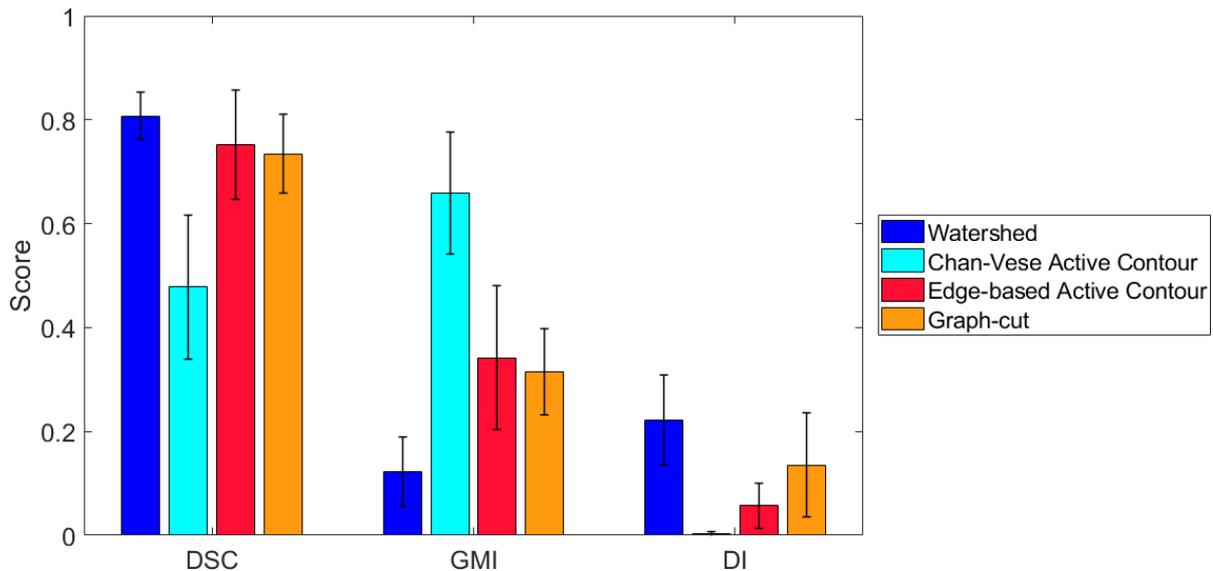


Figure 5.6. Performance of segmentation methods displaying mean DSC, GMI and DI scores (error bars represent standard deviation) for peripheral lung primary disease with ISTART dataset.

Although the watershed approach produced the largest volumes, they were most consistent with the reference standards compared to the other techniques. Chan-Vese segmentation produced the smallest contours that were less than half the size of the reference volumes for the majority of cases.

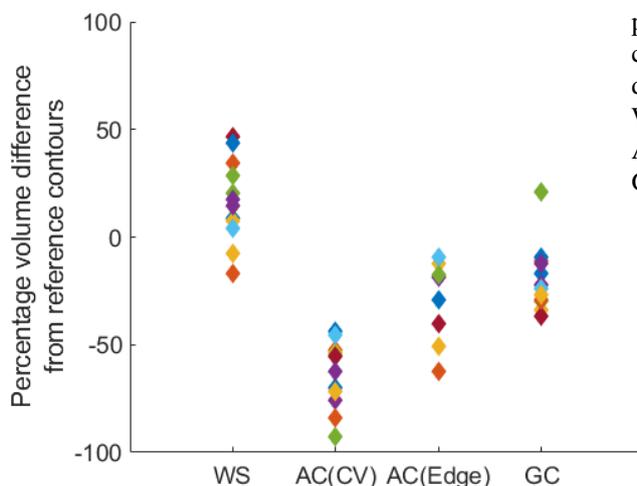


Figure 5.7. Percentage volume difference of peripheral primary disease in relation to the reference contours for individual cases (mean and standard deviation shown in table) of ISTART dataset. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.

	Absolute Difference (cm ³)	Percentage Difference
WS	0.99 ± 1.32	16.8 ± 19.4
AC (CV)	-4.22 ± 5.29	-63.5 ± 15.4
AC (Edge)	-1.21 ± 1.11	-27.2 ± 17.5
GC	-1.73 ± 2.76	-19.2 ± 15.6

5.7.2.2 Individual case performance

5.7.2.2.1 Quantitative analysis

The highest DSC score was achieved by the watershed approach in 8/12 (66%) of the cases. In three cases where the watershed approach achieved the second highest DSC score, the margin was very small. Additionally, across all the cases, the watershed technique consistently achieved the lowest GMI score. This was offset by it being associated with the highest DI in the majority of cases (11/12, 92%). In contrast, Chan-Vese active contour was seen to have the highest GMI scores across all the cases, resulting in the lowest DSC in all but one case.

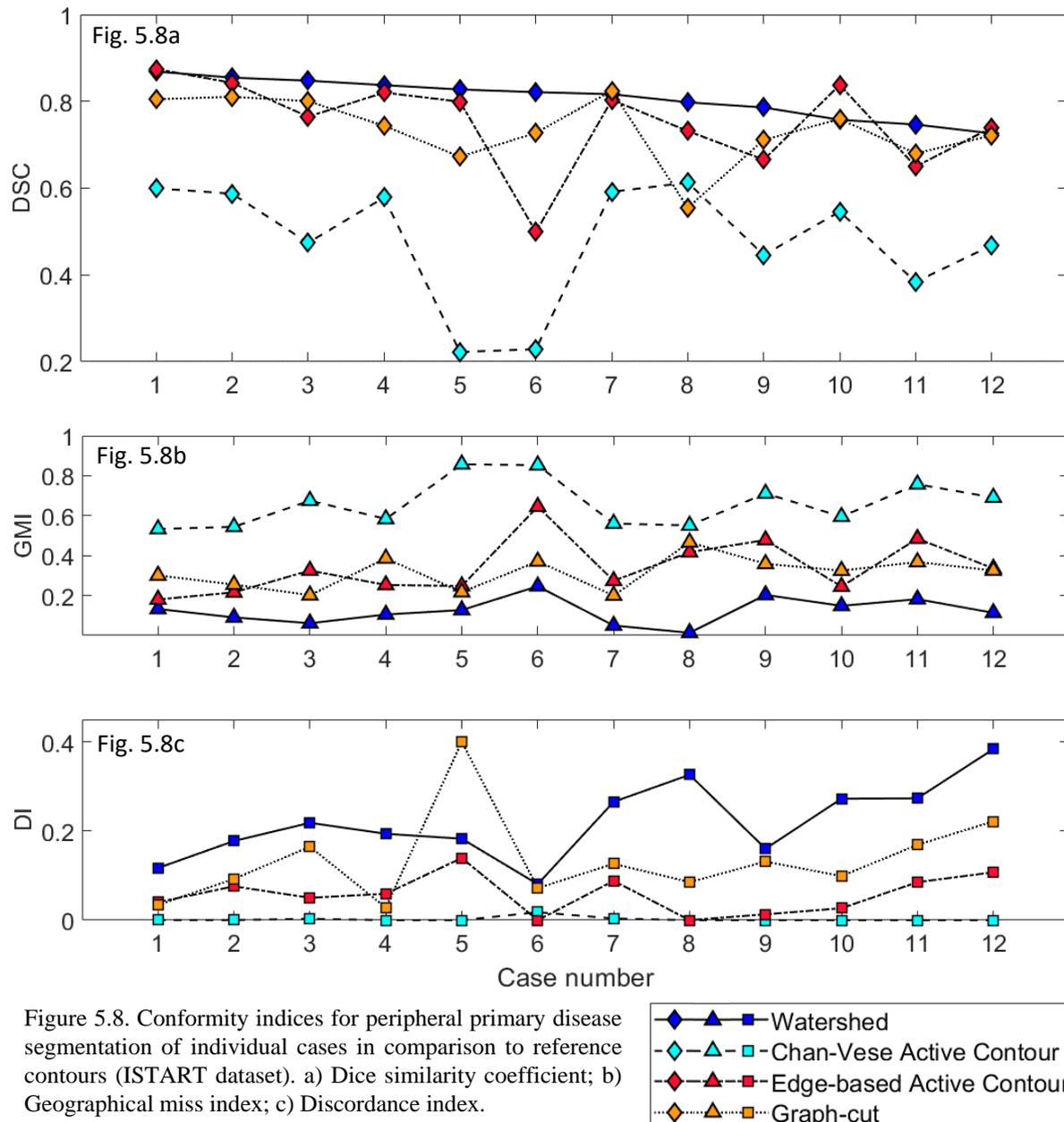


Figure 5.8. Conformity indices for peripheral primary disease segmentation of individual cases in comparison to reference contours (ISTART dataset). a) Dice similarity coefficient; b) Geographical miss index; c) Discordance index.

5.7.2.2.2 Qualitative analysis of segmentation performance

The best tumour coverage was seen with the watershed approach, with appropriate inclusion of patchy GGOs (ISTART Case p2, slice 3). Although the segmentation had the propensity to extend to chest wall in the case of juxtapleural lesions (ISTART Case p9 and p10, slice 1) and

include surrounding vasculature (ISTART Case p10, slice 3), clinically acceptable segmentation was observed in the majority of cases. Graph-cut performed well in terms of tumour coverage for the solid components but it failed at including patchy GGOs within the segmentation consistently (ISTART cases p4 and p10). The edge-based active contour approach fared better in this regard, although underestimation of the tumour was seen for lesions that did not approximate well to the localising boundary (ISTART Case p2, slice 3). and the Chan-Vese approach produced smaller volumes that appeared to be in concordance to contrast information seen in mediastinal window levels rather than in lung window levels.

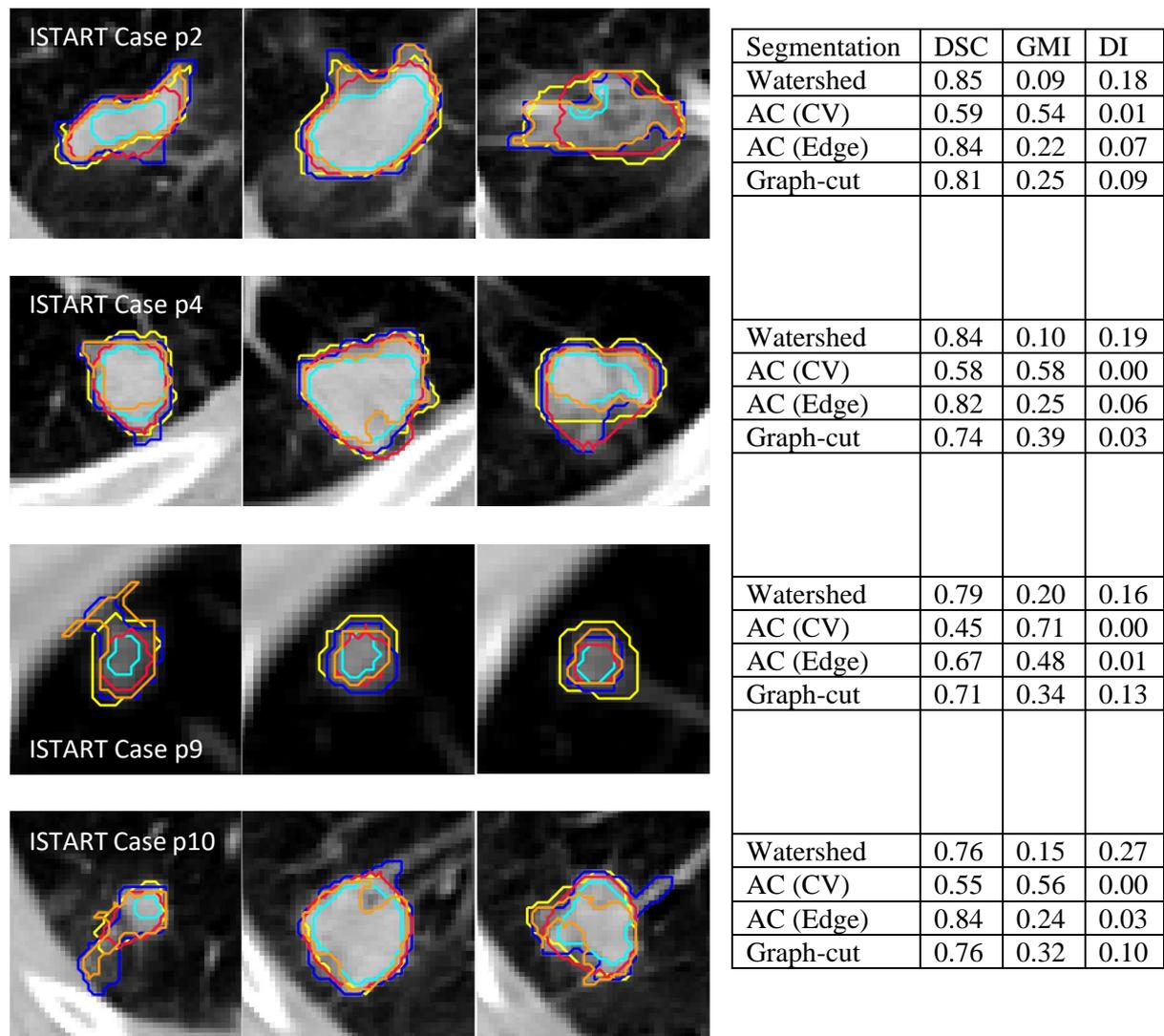


Figure 5.9. Segmentation results for ISTART dataset (peripheral primary disease; ISTART cases p1 – 10). Dark blue – watershed; Cyan – Chan-Vese active contour; Red – Edge-based active contour; Orange – Graph-cut; Yellow- reference contours. (Table: AC(CV) – Chan-Vese active contour; AC(Edge) – Edge-based active contour)

5.7.2.3 Computational time

The processing time for whole cases was similar for the watershed and Chan-Vese in the order of short seconds, followed by edge-based active contour application in under a minute for the majority of cases. Graph-cut segmentation took the longest time with a mean of 3.1 ± 2.2 minutes.

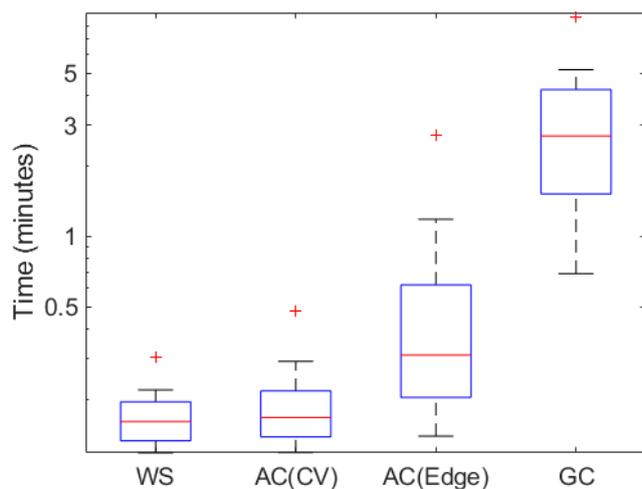


Figure 5.10. Time (minutes in logarithmic scale) for processing individual cases of ISTART dataset (peripheral primary disease only), with the mean and standard deviation shown in the table. Watershed – WS; Chan-Vese active contour – AC(CV); Edge-based active contour – AC(Edge); Graph-cut – GC.

WS	10 ± 3 seconds
AC (CV)	12 ± 6 seconds
AC (Edge)	36 ± 44 seconds
GC	3.1 ± 2.2 minutes

5.8 Task D.3 Comparison of segmentation of advanced lung tumour versus isolated peripheral lung tumours

A summary of the qualitative measures for the segmentation of advanced lung tumours from chapter 3 and the isolated peripheral primary disease from this chapter is produced in table 5.3. Marker-controlled watershed was the best overall performing technique for the segmentation of isolated lung tumours across all the evaluated measures, in contrast to the edge-based active contour approach that was identified as the most appropriate technique for advanced disease. Taking into consideration that a skew towards lower conformity indices would be obtained for smaller volumes, the performance for the graph-cut approach appears to be similar in both settings. Chan-Vese active contour, however, performed poorly for isolated peripheral lung tumours despite its computational speed.

	Volume (cm ³)	Marker-controlled watershed					Chan-Vese active contour					Edge-based active contour					Graph-cut				
		DSC	GMI	DI	% volume difference	Time	DSC	GMI	DI	% volume difference	Time	DSC	GMI	DI	% volume difference	Time	DSC	GMI	DI	% volume difference	Time
Training Dataset																					
Advanced disease (63 cases)	116.24 ± 87.35	0.72 ± 0.10	0.06 ± 0.04	0.38 ± 0.12	-	27.0 ± 7.2 sec	0.73 ± 0.09	0.31 ± 0.12	0.16 ± 0.07	-	-	0.81 ± 0.05	0.16 ± 0.07	0.20 ± 0.05	-	-	0.76 ± 0.08	0.19 ± 0.07	0.14 ± 0.10	-	-
Peripheral disease (10 cases)	13.03 ± 15.45	0.84 ± 0.04	0.08 ± 0.04	0.21 ± 0.09	19.4 ± 20.3	12 ± 2 sec	0.56 ± 0.13	0.57 ± 0.14	0.01 ± 0.01	-49.8 ± 15.9	14 ± 6 sec	0.79 ± 0.07	0.26 ± 0.10	0.30 ± 0.05	-11.0 ± 12.2	45 ± 42 sec	0.76 ± 0.04	0.10 ± 0.06	0.11 ± 0.04	-18.6 ± 8.7	3.8 ± 1.4 min
Independent Dataset(s)																					
Advanced disease (16 cases)	124.87 ± 139.26	0.72 ± 0.08	0.07 ± 0.03	0.38 ± 0.10	58.2 ± 40.4	21 ± 5 sec	0.71 ± 0.07	0.30 ± 0.10	0.22 ± 0.10	1.00 ± 23.8	45 ± 18 sec	0.80 ± 0.06	0.17 ± 0.06	0.20 ± 0.05	5.9 ± 3.9	3.2 ± 1.7 min	0.76 ± 0.06	0.20 ± 0.08	0.23 ± 0.08	7.4 ± 15.2	13.9 ± 7.3 min
Peripheral disease (2 cases)	8.85; 0.64	0.85; 0.85	0.06; 0.08	0.19; 0.20	15.8; 14.7	13 sec; 9 sec	0.68; 0.66	0.48; 0.50	0.00; 0.01	-44.6; -48.0	11 sec; 7 sec	0.84; 0.78	0.24; 0.35	0.05; 0.02	-14.3; -31.6	35 sec; 9 sec	0.76; 0.82	0.29; 0.25	0.11; 0.10	-19.5; -15.9	4.6 min; 1.2 min
Peripheral disease (Different trial source; 12 cases)	8.29 ± 11.75	0.81 ± 0.05	0.12 ± 0.07	0.22 ± 0.09	16.8 ± 19.4	10 ± 3 sec	0.48 ± 0.14	0.66 ± 0.12	0.00 ± 0.01	-63.5 ± 15.4	12 ± 6 sec	0.75 ± 0.11	0.34 ± 0.14	0.06 ± 0.04	-27.2 ± 17.5	36 ± 44 sec	0.73 ± 0.08	0.31 ± 0.08	0.14 ± 0.10	-19.2 ± 15.6	3.1 ± 2.2 min

Table 5.3. Summary of qualitative measures for segmentation of advanced and isolated peripheral lung tumours. (Computational time displayed for processes using an Intel Core i5-3317U CPU @ 1.70GHz, 4GB RAM on a Windows 10 64-bit environment.)

Discussion

These results support that the performance of the different techniques is variable depending on the setting in which they are used. Although the watershed approach was limited in its use at the mediastinum due to competing gradients from other organs, good concordance to clinically acceptable delineations were achieved between the lung parenchyma and tumour interface. The ability to include patchy GGOs in the segmentation is an important and valuable aspect of this approach, which is commonly seen at the periphery of lung tumours. On the other hand, Chan-Vese active contour modelled poorly to the true tumour edge. At the parameter settings that were used, it was consistently underestimating the tumour boundary and appeared to conform to the perceived tumour edge on mediastinal window levels rather than on lung window levels that is used clinically.

As marker-controlled watershed segmentation is fast and relatively inexpensive computationally compared to the other explored techniques, it is an attractive approach to adopt. With the small sizes typically associated with isolated peripheral lung tumours, even shorter processing times are required to complete the segmentation.

In the clinical context, the inter-observer delineation variability of isolated peripheral tumour volumes is small, as shown in the study by Persson et al where the mean equivalent Jaccard score for pair-wise comparisons was at 0.72 ± 0.09 (413). Taking into account the different conformity measure, the achieved DSC seen with watershed approach appears to be comparable. Studies designed to compare automatic and manual contours need to be performed to give a better estimate for this, in addition to reproducibility studies to validate the reliability of this approach. Methods to improve the precision at the pleural surface and trimming of vascular without undermining the tumour coverage should also be investigated, which are beyond the remit of this project.

Conclusions

The different techniques applied in different clinical settings produced variation in the observed segmentation performance. The developed workflows using marker-controlled watershed segmentation was the overall best performing approach in the setting of isolated peripheral lung tumour segmentation.

Chapter 6

Specific Aim E: Evaluation of texture features in classification of tumour and non-tumour regions

Introduction

Building on from the segmentation techniques described in the preceding two chapters, the potential for application of texture features in the segmentation task is explored in this and the next chapter.

In the thorax, texture cues have been used for a variety of different purposes. Texture features have been assessed in lung parenchyma to detect abnormal lung tissue (414) and disease such as emphysema (415-420), fibrosis (421-423) and pneumonitis (424, 425). A recent review has summarised the application of texture cues in the setting of lung cancers (426), which include the detection of aggressive tumours (183), prognostication (427, 428) and evaluation of treatment response (429). Texture descriptors form the basis of radiomics research, where associations with gene-expression profiles have been shown that it can be used as a predictive biomarker (430).

Texture descriptors have also been used to perform segmentation tasks as discussed in chapter 1 section 1.9.7. The segmentation of sub-regions within part-solid pulmonary nodules has been performed using gradient and 3D intensity texture features along with shape information (431). For lung tissue, Korfiatis et al developed a method for segmenting lung parenchyma affected by interstitial pneumonia through SVM classification of border regions based on wavelet features and intensity values following initial lung segmentation using k-means clustering (432). These studies support the notion that texture descriptors may be helpful in segmentation of tumours.

Here, the test for the hypothesis that texture descriptors can be helpful in the binary classification of GTV and its immediate surrounding region is described. Out of the different approaches to analyse texture, statistical, model and signal processing methods have been most commonly applied to medical imaging. Thus, these texture features are evaluated in this work and elaborated as follows:

A) Statistical: Gradient-derived features

These features describe the spatial variation of the intensity values, whereby high gradient values are assigned to regions of stark transitions of scale levels whilst subtle transitions are represented by low gradient values. The absolute gradient is sensitive to changes in the scale magnitude between neighbouring pixels and can then be assessed based on histogram-derived parameters.

B) Statistical: Co-occurrence matrix

The different combinations of intensities are evaluated for pairs of pixels in a given direction and number of pixels apart. This information is then tabulated into what is known as a co-occurrence matrix, which can then be evaluated through statistical measures such as the energy, contrast, entropy, homogeneity and correlation. For example, contrast refers to the magnitude of the difference on scale values within the objects of the image, whereas entropy characterises

the homogeneity of the pixel distribution within the ROI with respect to the particular direction or orientation, i.e. a measure of the disorder within the region.

C) Statistical: Run-length matrix

Information on the frequency in which the same intensity value recurs in a given direction is captured in the form of a matrix for each region. Computation of different run-lengths can be then constructed for different directions, which can in turn generate new texture parameters.

D) Signal processing: Wavelet features

The representation of information in the form of wavelets stems from the field of signal processing. The principles behind this is similar to how the Fourier transform works, where instead of processing the data directly on the original source, a transformation is applied to identify and evaluate the subunits of which the original data is made up from. A signal (or in this work the intensity values of an image) can be presented in terms of its frequency, and the application of such a transform would change this information to being represented in the time/spatial domain in the form of waves. As any signal can be represented by a summation of different sine and cosine waves within the time/spatial domain, analysis of any image in terms of pixel value frequencies can then be analysed implicitly in this form.

A limitation to the Fourier transform is in the difficulty with the evaluation of the waves at a snapshot in time/space, because of the uncertainty trade-off between frequency and time/space. Wavelet analysis overcomes this issue where the instantaneous frequency can be determined. Instead of decomposing the signal into an infinite wave, the wavelet transform deconstructs the information into wavelets which are limited by time/space.

In other words, wavelet analysis provides another means of evaluating the frequency content of the image, i.e. how fast the grey-level value of a 2D image varies. A high spatial frequency is assigned to a region with many variations of the grey-level values and a greater number of peaks and troughs. If the scale values vary slowly, being almost the same throughout the ROI, it is represented by a low spatial frequency. Each pixel is given a set of numbers (wavelet coefficient) that describe the frequency content at that point in the image on a particular set of scales (i.e. size of the region evaluated).

E) Model: Autoregressive model-based parameters

This describes the amount of regularity and repetition that is present in a region in terms of fineness and coarseness. If the texture is coarse, the autoregressive function will drop off slowly; otherwise, it will fall off rapidly. The principle behind the autoregressive model is that it assumes that pixel intensity values can be estimated from on the weighted sum of their immediate neighbouring pixels with the presence of other parameter such as noise. Thus, parameters that are derived from autoregressive modelling describe the relationship of neighbouring pixels, as well as the standard deviation of noise.

6.1 Summary of tasks

Firstly, an evaluation was performed to ascertain if these descriptors can be used to distinguish tumour from non-tumour regions. The performance of the classification of the regions based on multiple texture features was compared to use of the single most discriminatory texture parameter, as well as the mean as the sole feature.

Additionally, further comparisons were made to evaluate the applicability of the classification models on distinguishing between GTV and tissue at a distance away.

Task E.1 Classification of GTV versus adjacent tissue based on multiple texture features

Task E.2 Comparison of texture classification versus classification based on mean value and classification based on most discriminatory texture parameter

Task E.3 Impact of classification of GTV versus tissue at a distance away

Methods

The process for feature computation, selection and classification using a) Linear discriminant analysis (LDA) (section 6.8.1) and b) k-NN classification (section 6.9.1) is described.

6.2 Study design - Feature selection and classification using LDA in MaZda

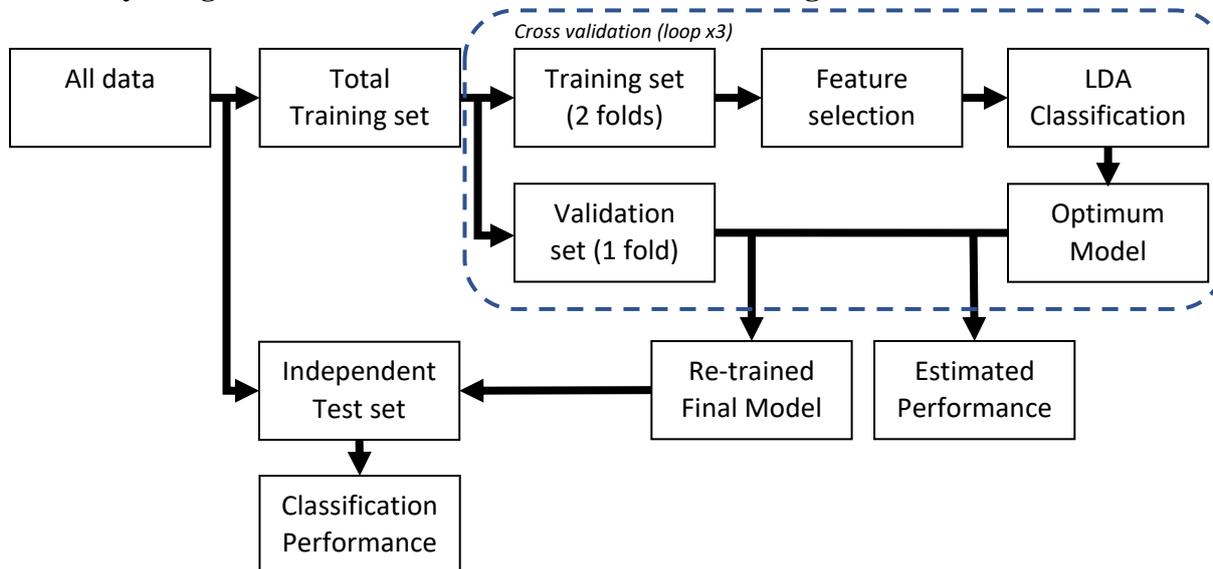


Figure 6.1. Design for LDA classification.

The design of the workflow and division of data is shown in figure 6.1. The same dataset of 79 cases was used for this analysis, using the same division of the dataset that was applied in the segmentation work with an independent test set (16 cases; 830 samples) separate to the training data (63 cases; 3164 samples), detailed in table 6.1. Additionally, the same 3-fold cross-validation partitioning was used, whereby feature selection and training of the LDA classifier was performed on the training sets.

	Cases	Slices	Number of samples/ROIs		Cases	Mean number of samples/ROIs
Independent test set	16	415	830			
Total training set	63	1582	3164	Validation set	21	1055
				Training set	42	2109

Table 6.1. Division of data and sample sizes for feature selection and LDA classification.

6.3 Processing of CT images

CT images were imported into MATLAB as per the workflow in the segmentation work. As CT pixel values were kept between -1000 and 1000 HU, there could be up to 2001 different intensity values within an image. For texture features where frequency of recurrence of particular intensity value is calculated (e.g. co-occurrence matrix), having too many different intensity values can undermine the texture representation. Thus, in order to generate texture features that would be representative of the region class, the intensity values of the images were discretized. This was performed through the re-sampling of the intensity values of each of the

pixel of the CT into equally-spaced bin widths of 25 HUs, resulting in up to 80 different values. Not only did this serve as a normalisation step, the process also helped with reduction of image noise. This approach is similar to the work performed by Aerts et al (430). The resultant images were then exported as bitmap (BMP) files for processing in MaZda.

6.4 Generation of ROIs – BMP files

The reference contours were imported into MATLAB using the same workflow as the segmentation analysis, which were used to generate ROIs for the texture evaluation. For each axial slice, three ROIs were defined. This consisted of the GTV, an annulus of 10-pixel width around the GTV, and another 10-pixel width annulus which was positioned at a distance of 10 pixels away from the GTV. To create the annulus structures, morphological dilation was performed using a square structuring element of 20-pixel width, followed by a removal of the overlapping inner region (i.e. GTV excluded for the annulus adjacent to the GTV; GTV and annulus adjacent to the GTV excluded for the annulus at 10-pixel distance away from the GTV). This ensured that the regions were non-overlapping.

For GTVs located near the edge of the lung, there was a possibility that the morphological expansions could dilate into regions exterior to the body contour, especially in thin subjects. To avoid this, the body contour was segmented using Otsu's thresholding (`imbinarize`), the complement of which was used to exclude regions exterior to the body contour.

The resultant masks were then exported as separate BMP files.

6.5 Generation of ROIs – ROI files

The analysis of the texture parameters for each region was a multi-step process using the software package MaZda v4.6 (Technical University of Lodz, Poland) (433, 434). Although BMP files could be imported directly into MaZda, in order for the texture parameters to be computed, a separate ROI file for each of the axial slice had to be created from the BMP files, which was then used for the texture calculations.

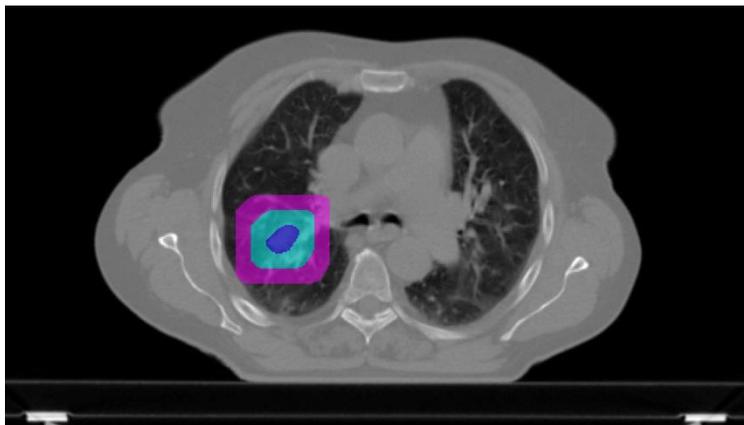


Figure 6.2. Example of an axial slice of a CT image normalised and discretised into equally spaced bin widths of 25 HUs as displayed in MaZda. The ROIs for analysis is represented by the coloured regions (Blue = GTV, Aqua = 10-pixel width annulus adjacent to GTV, Pink = 10-pixel width annulus at 10-pixel distance away from GTV).

In MaZda, up to 16 different ROIs within an axial slice could be defined (with each ROI associated with a particular bitmap). For this work, three different bitmaps were used for the three ROIs. In the preliminary phase of the project, this process was performed manually, where the BMP files of the ROIs were imported into MaZda separately, which were then used to select the ROIs in the relevant bitmap (figure 6.2). To improve the efficiency of this process, an independent automated GUI was developed by Jena R which was used to create the ROI files from the BMP masks.

The CT images were then loaded separately with the corresponding ROI file, and the texture parameters were calculated for each region and saved out in individual reports.

6.6 Generation of texture features

A total of six different classes of statistical parameters was calculated using MaZda (table 6.2).

Feature class	Abbreviation	Type of features	Number of features
First order histogram statistics	HIS	Mean Variance Skewness Kurtosis 1 st percentile 10 th percentile 50 th percentile 90 th percentile 99 th percentile	9
Co-occurrence matrix	COM	Angular second moment Contrast Correlation Sum of squares Inverse difference moment Sum average Sum variance Sum entropy Entropy Difference variance Difference entropy	11 (in four directions, and interpixel distance from one to five in each direction), i.e. total 220
Run-length matrix	RLM	Run length nonuniformity Grey level nonuniformity Long run emphasis Short run emphasis Fraction of image in runs	5 (in four directions); i.e. total 20
Absolute gradient	GRA	Mean Variance Skewness Kurtosis Percentage of pixels with non-zero gradient	5
Autoregressive model	ARM	Theta 1 Theta 2 Theta 3 Theta 4 Sigma	5
Wavelet transform	WAV	Sub-band LL Sub-band LH Sub-band HL Sub-band HH	4 (with 5 scales), i.e. total 20

Table 6.2. Classes and sub-categories of histogram and texture features used for this work.

i. First-order histogram statistics

There were nine first-order statistics that were computed, which are not technically texture descriptors by definition. However, they were included in the feature list, as they are classical descriptors that are commonly used (e.g. mean, variance), and can contribute to the segregation of the different regions.

ii. Co-occurrence matrix

Eleven types of statistics were computed for the co-occurrence matrix in four directions, at interpixel distance from one to five in each direction. This gave rise to 220 parameters.

iii. Run-length matrix

Five types of statistics describing the run length matrix 5 in four directions were calculated, resulting in 20 parameters.

iv. Absolute gradient

Five types of statistics describing the absolute gradient of the region were computed.

v. Autoregressive model

Four parameters (theta 1 to 4) describe the inter-pixel relationship, whilst sigma represents the standard deviation of noise.

vi. Wavelet features

At each scale (magnification), the energy of the Harr wavelet was computed from filtering with combinations of low-pass and high-pass filters. Up to five scales were calculated, giving rise to a potential of 20 features.

A total of 279 features were therefore calculated for each dataset. Calculation of the run-length matrix, co-occurrence matrix and the gradient features were performed at 7 bits/pixel in keeping with the re-binning of the data values described above. Wavelet transform was performed at the default setting of 8 bits/pixel.

The analysis of texture features revealed that the Harr wavelet could not be computed between scales 3 and 5 due to the small size of some of the ROIs (nodal structures). These parameters were therefore excluded from the rest of the analysis, leaving a total of 267 features for each dataset.

6.7 Feature reduction

Subsequently, feature reduction was performed to produce a smaller number of features to be used in the classification stage. This was essential as it would be difficult to predict which parameters would be most useful for texture classification when working with a large number of features. As some texture parameters would be highly correlated to another, these features would not carry any new information to help the classifier, and hence would not greatly contribute to the partitioning of the data. Moreover, having a large number of features would require very large number of data samples to provide statistically reliable discrimination, and can also increase further computation time. Thus, it was important to reduce the number yet at the same time select the texture parameters that would best distinguish one region from another,

to provide a more compact set of features that would be useful for texture discrimination and classification.

For this work, feature selection was performed using a combination of manual and statistical methods. Three statistical filter-based methods were used for feature selection, each of which chose ten different parameters. For each of these methods, scores were computed for the individual parameters which were then used to rank each feature, from which the top ten parameters were selected.

The Fisher score is a commonly used supervised feature selection method, where the Fisher criterion is computed as the ratio of the between-class to the within-class variance. Features with higher inter-class than intra-class scatter would have higher scores and ranks. Another method for feature selection is based on minimisation of both the classification error probability (POE) and the average correlation coefficient (ACC). This is based on calculating the percentage of misclassified samples for each of the features taking into account the misclassification rate for all the other features. The features with the lowest error rates are selected. The third method is the mutual information (MI) measure that models the dependence between two or more random variables. If the variables (texture feature and class category) are correlated, the MI measure is high.

The selection of features was based on their performance at distinguishing the GTV from the adjacent tissue (i.e. not on the ROIs at 10-pixel distance). When the three statistical methods were performed sequentially, it was observed that there was significant overlap of parameters between the different feature list. To avoid this, selection with all three statistical methods was performed concurrently. However, with this approach, the respective calculated scores in which the selection was based on was not made available, and thus the selection was accepted *de facto*.

Further evaluation and reduction of the feature list was performed manually after comparison of the feature lists between the training runs.

6.8 Classification – COST B11 programme

After feature selection, the computed statistics were imported into the COST B11 program v3.3 (Technical University of Lodz, Poland) (434), which is an accompanying module to MaZda that supports the quantitative analysis as well as classification of the computed features.

Data handling within the B11 module was constrained to a maximum of 30 features. Additionally, further reduction of the training data had to be performed as B11 only supported analysis of less than 2000 samples for each evaluation. Therefore, 999 reports were chosen at random from each of the training datasets, equating to 1998 data samples for each run. This corresponded to excluding 95 image slices for training run 1, 13 slices for training run 2, and 59 slices for training run 3. This step was not required for the validation datasets as there were fewer number of image slices for analysis.

To account for differences in scaling between the texture features, feature standardisation was performed, where each predictor data was centred and scaled according to the respective means and standard deviations.

6.8.1 LDA classification

Despite having performed feature selection to reduce the number of features for evaluation, the resultant feature list was still too large for visual representation of the data. Within B11, further dimensional reduction was available using LDA, where a linear transform matrix was computed from the data, such that the ratio of the determinants is maximized. Transformation of the original data by means of this matrix produced the most discriminatory feature (MDF), where the multi-dimensional raw data was distilled into a single dimension that made visualisation and analysis more manageable.

The linear separability coefficient was computed and used as an indicator of the usefulness of LDA for discrimination. This was defined as the largest eigenvalue of the ratio of the between-class scatter matrix and the total scatter matrix, where a value close to 1 indicated better linear separability of the data. The Fisher coefficient was also calculated (ratio of mean-squared between-class distance to the mean-squared within-class distance), where higher scores are associated with better discrimination of the categories.

Within B11, classification of the categories based on the MDF was performed using a 1-NN classifier, which was applied to the training runs.

6.9 Classification - MATLAB

The computed texture features were exported from MaZda into MATLAB via Excel. Principle component analysis (PCA) was applied to the training dataset to allow better visualisation of the separability based on a lower dimensionality of the multi-variate data.

With the increased functionality within MATLAB, the data was assessed to see if discriminant analysis would be a good classification model, by checking if the assumptions for discriminant analysis were met. In addition to LDA, quadratic discriminant analysis (QDA) was also assessed, where the separation of the classes was based on a quadratic rather than a linear surface. Models for both LDA and QDA were built using the training data. To assess the LDA model for its assumption of equal covariance matrices, the Bartlett test was applied. On the other hand, this was not required for QDA as the covariance matrices are computed separately for each class. However, both LDA and QDA assume that the data fit a Gaussian mixture model. This was tested separately for LDA and QDA by Q-Q plots and the Mardia kurtosis test.

6.9.1 k-NN classification

The k-NN algorithm was applied as a classification tool for the data. This is a non-parametric classification method, where no assumptions on the underlying data distribution were made. The algorithm works by deciding on the class for the data point in question based on the class of the surrounding neighbour(s). For example, in a 5-nearest neighbour classifier, the category of a data point would be set as the majority class of its 5 nearest data points. The number of neighbours (k) is the main parameter that has to be specified for this classification method.

Other hyperparameters of the algorithm could also be modified. For this work, the default hyperparameter settings in MATLAB were used. This included an equal distance weighting function (where all neighbours were evaluated with equal weighting, independent of the distance to the data point in question), as well as the use of the Euclidean distance in the calculation of the distance metric. Feature standardisation was also performed, where each

predictor data was centred and scaled according to the respective means and standard deviations.

Training of the classification models was performed to decide on the optimal number of neighbours for the final classifier, between a range of 1 to 20. As the sample size was no longer a limiting factor, all the samples in data was used. Additionally, in order to train the classification model, select the optimal neighbour size and to provide an estimate of the classification performance, a nested cross-validation approach was also adopted (see section 6.9.1.1). This change to the study design was performed to produce an unbiased estimate of the classification. After the models based on the optimised parameters were built, evaluation of their performance was carried out using the validation datasets derived from each of the cross-validation folds.

Re-training of the whole training dataset was performed to select the best performing classification parameter to build the final classification model. This was then applied to the testing dataset, to assess the performance of the classifier to unseen data.

6.9.1.1 Study design - classification using k-NN classifier in MATLAB

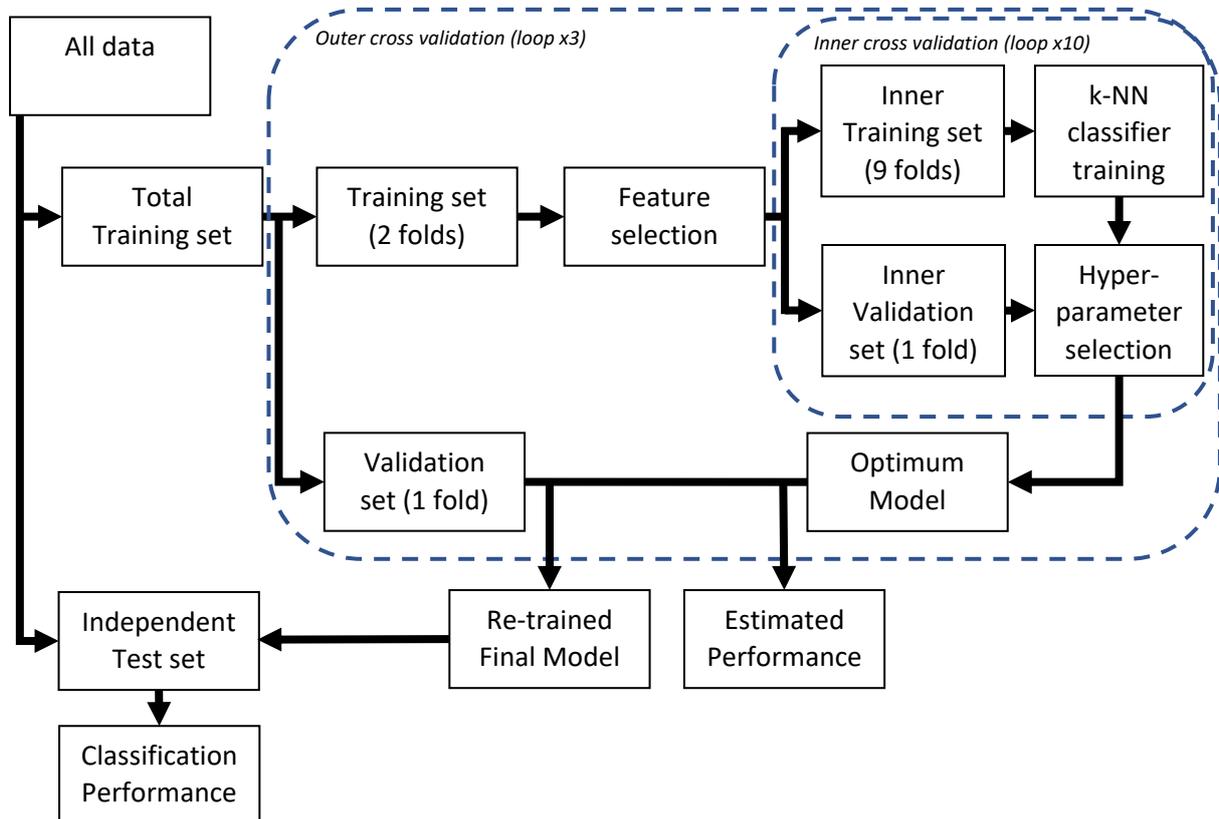


Figure 6.3. Design for k-NN classification.

In addition to applying the outer 3-fold cross-validation for feature selection, a further inner (nested) 10-fold cross-validation was applied to each of the outer training set as part of the model fitting procedure to decide on the best number of neighbours to use in the estimation of the classification performance (see figure 6.3). This was assessed based on the aggregate scores for the misclassification error and variance, taking into account model complexity. The final models were built after re-training on the total training data (3-fold cross validation), which

were applied to the independent test set. The same cross-validation folds were used to train the different classifiers to allow fair comparison between the different parameters. Details of sample sizes are given in table 6.3.

	Cases	Slices	Number of samples		Cases	Mean number of samples		Mean number of samples
Independent test set	16	415	830					
Total training set	63	1582	3164	Validation set	21	1055		
				Training set	42	2109	Nested Validation set	211
							Nested Training set	1898

Table 6.3. Division of data and sample sizes for k-NN classification.

6.10 Classifier assessment

To assess the performance of the classification models, the misclassification error was computed for all phases (training, validation and testing), defined as

$$Misclassification\ Error = \frac{FP + FN}{TP + TN + FP + FN} \quad [6.1]$$

where TP denotes true positives, TN denotes false negatives, FP denotes false positives and FN denotes false negatives. Tumour regions were considered as the positive class in the binary classification. The misclassification error, which is the inverse of the accuracy measurement, gives an overall assessment of how well the classification has performed, where a low score is associated with good classification. The misclassification error of the training data is also known as the resubstitution loss. Although this usually gives an optimistic estimate of the predictive error on new data, it is useful for estimating the bias and variance for parameter tuning when assessed with the errors with validation/testing cohorts. Model performance was estimated based on the misclassification error of the validation/testing sets.

Additionally, the sensitivity, defined as

$$Sensitivity = \frac{TP}{TP + FN} \quad [6.2]$$

and the specificity, defined as

$$Specificity = \frac{TN}{TN + FP} \quad [6.3]$$

were also calculated, where high values are associated with better classification. As the classification for this work involved binary classes (GTV versus either the adjacent tissue or the tissue at 10 pixels distance), the positive class for this work was defined as the GTV, whilst the negative class was either the adjacent tissue or the tissue at 10 pixels distance in the relevant experiments.

Receiver operating characteristic (ROC) curves were also generated to assess the classification models.

6.11 Experiments

6.11.1 GTV versus adjacent non-tumour tissue

The classification models were trained and assessed with the use of multiple texture features. Additionally, separate training and assessment of the classification using a) the mean value and b) the most discriminatory texture feature as the sole parameter were carried out with the same workflow as that for the multiple texture features.

After selection of the optimum parameter based on the inner cross-validation, this was used to estimate the performance of the classifier using the different feature sets using the validation sets of the outer cross-validation.

The whole training data was then used to build the final classification model and tested on the independent test set.

6.11.2 Classification of GTV versus tissue at a distance away

The models that were built based on discriminating GTV from the adjacent tissue was also applied to datasets with features derived from the GTV versus the tissue at 10-pixels distance. Further training of the models was not performed using this data. Instead, the predictive value of the classification was assessed by applying these samples to the classifier, from both the training and the independent testing datasets.

Classification models for the multiple texture features, mean value and the most discriminatory texture parameter were compared.

Results

6.12 Feature selection

The top ten feature parameters from the sequential application of the Fisher, POE + ACC and MI selection with the associated scores for each run is shown in table 6.4. Within each training run, there were a number of features that consistently ranked highly across the different statistical calculations (denoted by ^). These were the sum variances calculated across the different directions and pixel distances.

Training run	Fisher score		Classification error probability + Average correlation coefficients		Mutual Information	
	Feature	Score	Feature	Score	Feature	Score
Run 1	Perc.10%	16.16	S(4,4)SumVarnc^	0.253	S(3,3)SumVarnc^	0.874
	S(0,4)SumVarnc^	15.85	S(5,5)SumVarnc^	0.403	S(2,2)SumVarnc^	0.872
	S(3,3)SumVarnc^	15.84	S(5,5)SumEntrp	0.411	S(0,4)SumVarnc^	0.869
	S(0,3)SumVarnc^	15.77	S(0,1)Correlat	0.414	S(4,4)SumVarnc^	0.869
	S(2,2)SumVarnc^	15.74	S(0,5)SumVarnc^	0.427	S(0,5)SumVarnc^	0.868
	S(0,5)SumVarnc^	15.71	S(4,-4)SumVarnc	0.427	S(0,3)SumVarnc^	0.865
	S(4,4)SumVarnc^	15.53	S(3,-3)SumVarnc^	0.423	S(4,0)SumVarnc	0.861
	S(0,2)SumVarnc	15.40	135dr_GLvNonU	0.431	S(3,0)SumVarnc	0.860
	S(1,1)SumVarnc	15.03	S(0,3)SumAverg	0.446	S(3,-3)SumVarnc^	0.860
	S(0,4)SumOfSqs	14.99	S(4,4)SumEntrp	0.451	S(5,5)SumVarnc^	0.856
Run 2	Feature	Score	Feature	Score	Feature	Score
	S(0,3)SumVarnc^	12.11	S(3,0)SumVarnc^	0.397	S(0,4)SumVarnc^	0.781
	S(0,4)SumVarnc^	12.07	S(0,4)SumVarnc^	0.416	S(0,5)SumVarnc^	0.781
	S(0,2)SumVarnc	11.94	45dgr_RLNonUni	0.439	S(0,3)SumVarnc^	0.780
	S(0,5)SumVarnc^	11.88	45dgr_LngREmph	0.468	S(3,-3)SumVarnc^	0.776
	S(2,2)SumVarnc^	11.84	S(2,0)SumVarnc	0.487	S(4,-4)SumVarnc	0.775
	S(2,-2)SumVarnc^	11.80	S(4,0)SumVarnc^	0.502	S(5,-5)SumVarnc	0.774
	S(3,3)SumVarnc^	11.68	S(5,5)SumEntrp	0.505	S(4,0)SumVarnc^	0.774
	Perc.10%	11.63	S(0,5)SumVarnc^	0.512	S(2,-2)SumVarnc^	0.774
	S(3,-3)SumVarnc^	11.59	135dr_GLvNonU	0.516	S(2,2)SumVarnc^	0.774
S(1,1)SumVarnc	11.59	S(5,5)SumAverg	0.522	S(3,3)SumVarnc^	0.774	
Run 3	Feature	Score	Feature	Score	Feature	Score
	S(0,3)SumVarnc^	12.74	S(2,0)SumVarnc^	0.423	S(0,4)SumVarnc^	0.8
	S(0,4)SumVarnc^	12.73	S(0,4)SumVarnc^	0.457	S(0,5)SumVarnc^	0.799
	S(0,2)SumVarnc^	12.56	45dgr_LngREmph	0.473	S(3,0)SumVarnc^	0.798
	S(0,5)SumVarnc^	12.56	S(1,0)Correlat	0.475	S(4,0)SumVarnc	0.795
	S(2,2)SumVarnc^	12.54	S(0,5)SumVarnc^	0.483	S(0,3)SumVarnc^	0.795
	S(3,3)SumVarnc^	12.41	Vertl_GLvNonU	0.490	S(3,3)SumVarnc^	0.795
	S(1,1)SumVarnc	12.26	S(3,0)SumVarnc^	0.492	S(2,0)SumVarnc^	0.795
	S(0,3)SumOfSqs	12.24	S(2,2)SumVarnc^	0.493	S(2,2)SumVarnc^	0.795
	S(0,4)SumOfSqs	12.24	S(5,5)SumEntrp	0.498	S(2,-2)SumVarnc	0.794
S(0,5)SumOfSqs	12.21	S(5,5)SumAverg	0.509	S(0,2)SumVarnc^	0.792	

Table 6.4. Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied sequentially for each cross-validation training set, with associated scores. ^ denotes features common in at least one other feature selection method in each run.

Using the combination of three feature selection methods, the top thirty features from each cross-validation training set in their ranked order is shown in table 6.5. Within each training run, the majority of the parameters comprised of the different computations of the sum variance and the sum of squares, which was seen across all three training runs (24 out of 30 in runs 1 and 2, 25 out of 30 in run 3). Although these parameters were determined to be most discriminatory when assessed individually, there was concern that this set of features, if highly correlated, would not complement each other to yield the best classification, and hence impact on the classification results.

Rank	Training set 1	Training set 2	Training set 3
1	S(3,3)SumVarnc	S(0,4)SumVarnc	S(0,4)SumVarnc
2	S(2,2)SumVarnc	S(0,5)SumVarnc	S(0,5)SumVarnc
3	S(0,4)SumVarnc	S(0,3)SumVarnc	S(3,0)SumVarnc
4	S(4,4)SumVarnc	S(3,-3)SumVarnc	S(4,0)SumVarnc
5	S(0,5)SumVarnc	S(4,-4)SumVarnc	S(0,3)SumVarnc
6	S(0,3)SumVarnc	S(5,-5)SumVarnc	S(3,3)SumVarnc
7	S(4,0)SumVarnc	S(4,0)SumVarnc	S(2,0)SumVarnc
8	S(3,0)SumVarnc	S(2,-2)SumVarnc	S(2,2)SumVarnc
9	S(3,-3)SumVarnc	S(2,2)SumVarnc	S(2,-2)SumVarnc
10	S(5,5)SumVarnc	S(3,3)SumVarnc	S(0,2)SumVarnc
11	S(4,-4)SumVarnc	S(3,0)SumVarnc	S(1,0)SumVarnc
12	S(5,-5)SumVarnc	S(2,0)SumVarnc	S(4,4)SumVarnc
13	S(5,0)SumVarnc	45dgr_RLNonUni	S(1,-1)SumVarnc
14	S(5,5)SumEntrp	45dgr_LngREmph	Vertl_GLevNonU
15	S(0,1)Correlat	S(4,4)SumVarnc	45dgr_LngREmph
16	S(2,-2)SumVarnc	S(0,2)SumVarnc	S(1,0)Correlat
17	135dr_GLevNonU	S(5,0)SumVarnc	S(1,1)SumVarnc
18	S(0,3)SumAverg	S(5,5)SumEntrp	S(5,0)SumVarnc
19	S(0,2)SumVarnc	135dr_GLevNonU	S(5,5)SumEntrp
20	S(4,4)SumEntrp	S(5,5)SumAverg	S(5,5)SumAverg
21	Perc.10%	Perc.10%	S(0,3)SumOfSqs
22	S(1,1)SumVarnc	S(1,1)SumVarnc	S(0,4)SumOfSqs
23	S(0,4)SumOfSqs	S(1,-1)SumVarnc	S(0,5)SumOfSqs
24	S(0,5)SumOfSqs	S(0,1)SumVarnc	S(0,2)SumOfSqs
25	S(0,3)SumOfSqs	S(0,2)SumOfSqs	S(0,1)SumVarnc
26	S(2,0)SumVarnc	S(0,3)SumOfSqs	S(2,2)SumOfSqs
27	S(0,2)SumOfSqs	S(0,4)SumOfSqs	S(1,1)SumOfSqs
28	S(2,2)SumOfSqs	S(1,1)SumOfSqs	S(3,3)SumOfSqs
29	S(0,1)SumVarnc	S(0,1)SumOfSqs	S(0,1)SumOfSqs
30	S(1,-1)SumVarnc	S(1,0)SumVarnc	S(1,-1)SumOfSqs

Table 6.5. Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied in combination to yield the top thirty ranked texture parameters for each cross-validation training set.

For the two parameters sum variance and sum of squares, bivariate correlation analysis was therefore performed between each of the computed feature (4 different directions, and between 1 to 5 pixel-distances). The features were found to be highly correlated (Pearson's correlation between 0.983 to 1 for sum variance, and 0.991 to 1 for sum of squares). This suggested that there would be a lot of redundancy if the classification is based on this list, at the expense of other features which may better contribute to the partitioning.

It was therefore decided to exclude pixel distances 2, 3, and 5 of the co-occurrence matrix in the feature selection. Pixel distance 4 was kept as it was highly ranked in the lists above. Pixel distance 1 was also retained, as it is commonly computed in other texture analysis work. Moreover, inclusion of these distances permitted the evaluation of pixels closer together and at a distance, with potentially complementary information.

The filter-based feature selection was repeated with the exclusion of the co-occurrence matrix at pixel distances 2, 3, and 5, which is shown in table 6.6. Features common to all three training runs is denoted by *, and ** for parameters present in two runs. 25 out of the 30 features were present in all three runs, with 2 being common between two runs, which were retained in the feature selection. Although parameters for the sum average and correlation were present in at least two runs, they were associated with different directions and pixel distances. It was decided empirically to select S(4,0) sum average, and S(1,0) correlation, making a total of 29 features.

	Training run 1	Training run 2	Training run 3
1	S(0,4)SumVarnc*	S(0,4)SumVarnc*	S(0,4)SumVarnc*
2	S(4,4)SumVarnc*	S(4,-4)SumVarnc*	S(4,0)SumVarnc*
3	S(4,0)SumVarnc*	S(4,0)SumVarnc*	S(4,-4)SumVarnc*
4	S(4,-4)SumVarnc*	S(4,4)SumVarnc*	S(1,1)SumVarnc*
5	S(0,1)SumVarnc*	S(1,1)SumVarnc*	S(4,4)SumVarnc*
6	S(1,1)SumVarnc*	S(1,-1)SumVarnc*	S(1,-1)SumVarnc*
7	S(1,-1)SumVarnc*	S(0,1)SumVarnc*	S(1,1)SumOfSqs*
8	S(0,4)SumOfSqs*	S(1,0)SumVarnc*	S(1,0)SumVarnc*
9	S(1,0)SumVarnc*	S(0,4)SumOfSqs*	S(0,1)SumVarnc*
10	S(1,-1)SumOfSqs*	S(1,-1)SumOfSqs*	S(0,1)SumOfSqs*
11	S(4,4)SumOfSqs*	S(1,0)SumOfSqs*	S(1,0)SumOfSqs*
12	S(1,1)SumOfSqs*	135dr_GLevNonU**	S(1,-1)SumOfSqs*
13	S(0,1)Correlat	S(4,4)SumOfSqs*	S(4,0)SumOfSqs*
14	S(4,4)SumEntrp*	S(4,4)SumEntrp*	45dgr_LngREmph**
15	135dr_GLevNonU**	S(0,1)SumOfSqs*	Vertl_GLevNonU
16	S(0,1)SumOfSqs*	45dgr_LngREmph**	S(1,0)Correlat
17	S(4,0)SumOfSqs*	S(1,1)SumOfSqs*	S(4,4)SumEntrp*
18	S(1,0)SumOfSqs*	Variance*	Variance*
19	S(0,4)SumAverg	45dgr_RLNonUni	S(0,4)SumOfSqs*
20	Kurtosis	S(4,0)SumAverg	S(1,0)SumAverg
21	Perc.10%*	Perc.10%*	S(4,4)SumOfSqs*
22	S(4,-4)SumOfSqs*	S(4,0)SumOfSqs*	Perc.10%*
23	Variance*	S(4,-4)SumOfSqs*	S(4,-4)SumOfSqs*
24	S(4,0)SumEntrp*	S(4,0)SumEntrp*	S(4,0)SumEntrp*
25	S(0,4)SumEntrp*	S(0,4)SumEntrp*	S(4,-4)SumEntrp*
26	S(1,1)SumEntrp*	S(1,1)SumEntrp*	S(0,4)SumEntrp*
27	S(4,-4)SumEntrp*	S(1,-1)SumEntrp*	S(1,1)SumEntrp*
28	S(1,-1)SumEntrp*	S(4,-4)SumEntrp*	S(1,-1)SumEntrp*
29	S(4,4)Entropy*	S(4,4)Entropy*	S(4,-4)Entropy
30	S(0,1)SumEntrp	S(4,0)Entropy	S(4,4)Entropy*

Table 6.6. Filter-based feature selection (Fisher, Classification error probability + Average correlation coefficients and Mutual Information) applied in combination to yield the top thirty ranked texture parameters for each cross-validation training set (exclusion of co-occurrence matrix at pixel distances 2, 3 and 5). *denotes features present in all three runs; ** denotes features present in two runs.

For the 29 chosen features thus far, the Mann-Whitney U test was performed to assess for differences in the means between the ROIs. All the features exhibited statistically significant

differences (p -value = 0.000) except the parameter 135dr_GLevNonU, which was associated with a p -value of 0.093. As this is not statistically significant at the 5% level, this parameter was excluded.

The final list of 28 features is shown in table 6.7, which was used for classification.

Feature list	Feature Class
Variance	First-order statistics
Perc.10%	First-order statistics
S(1,0)Correlat	Co-occurrence matrix
S(1,0)SumOfSqs	Co-occurrence matrix
S(1,0)SumVarnc	Co-occurrence matrix
S(0,1)SumOfSqs	Co-occurrence matrix
S(0,1)SumVarnc	Co-occurrence matrix
S(1,1)SumEntrp	Co-occurrence matrix
S(1,1)SumOfSqs	Co-occurrence matrix
S(1,1)SumVarnc	Co-occurrence matrix
S(1,-1)SumEntrp	Co-occurrence matrix
S(1,-1)SumOfSqs	Co-occurrence matrix
S(1,-1)SumVarnc	Co-occurrence matrix
S(4,0)SumAverg	Co-occurrence matrix
S(4,0)SumEntrp	Co-occurrence matrix
S(4,0)SumOfSqs	Co-occurrence matrix
S(4,0)SumVarnc	Co-occurrence matrix
S(0,4)SumEntrp	Co-occurrence matrix
S(0,4)SumOfSqs	Co-occurrence matrix
S(0,4)SumVarnc	Co-occurrence matrix
S(4,4)Entropy	Co-occurrence matrix
S(4,4)SumOfSqs	Co-occurrence matrix
S(4,4)SumVarnc	Co-occurrence matrix
S(4,4)SumEntrp	Co-occurrence matrix
S(4,-4)SumEntrp	Co-occurrence matrix
S(4,-4)SumOfSqs	Co-occurrence matrix
S(4,-4)SumVarnc	Co-occurrence matrix
45dgr_LngREmph	Run-length

Table 6.7. Final list of 28 features used for classification analysis.

6.13 Task E.1 Classification of GTV versus adjacent non-tumour tissue with multiple texture features

6.13.1 LDA classification

Figure 6.4 shows the LDA classification in relation to the MDF for each of the training runs. The data points for the GTV and surrounding region seem to cluster towards the respective ends for the MDF scales, despite there being a small amount of overlap of the data points near the centre of the graphs. This suggests that the two groups may be separated relatively well based on the MDF.

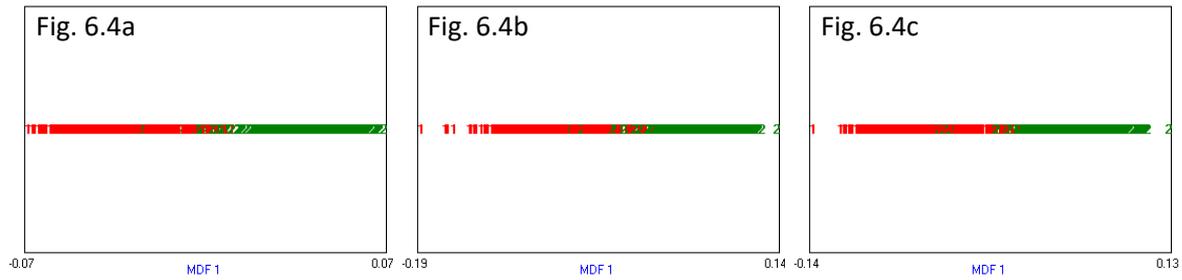


Figure 6.4. Results of LDA with respect to MDF, red “1” denotes GTV data points, green “2” denotes data points for region surrounding GTV. A) Training run 1; b) Training run 2; c) Training run 3.

The results of the classification for each of the training runs is presented in table 6.8. The linear separability coefficient was relatively high, indicating that the groups can be distinguished through LDA. High levels of accuracy, sensitivity and specificity were achieved in the identification of the tumour from the adjacent tissue in all three runs.

	Linear Separability	Fisher Coefficient	Accuracy	Sensitivity	Specificity
Training Run 1	0.91	38.7	98.00%	98.00%	98.00%
Training Run 2	0.87	27.3	97.35%	97.30%	97.40%
Training Run 3	0.87	26.8	97.35%	97.40%	97.30%

Table 6.8. Classification results for three training runs using LDA and 1-NN classifier.

Despite the promising results for LDA classification in the B11 module, it became apparent that this process would be difficult to apply to the validation runs, as well as the testing cohort using B11. Although the MDF could be calculated for the other runs, to allow for fair classification, calculation of the MDF for the new runs should be performed with the same transformation as applied to the training data. Additionally, it was not possible to assess if the assumptions in using LDA as a feature extraction tool were met within B11. As both of these functionalities were not present in the B11 module, further analysis of the validation and testing data was not performed here.

6.13.2 Principle component analysis

Figure 6.5 shows the spread of the training data after application of PCA, with respect to the two most discriminatory dimensions. Similar to the LDA plots, there was clustering of the data points for the two different regions, suggesting that separability between the groups could be achieved. Rather than applying the classifier on the untransformed texture parameter values, alternatively, the classification could be based on the most expressive features (MEFs) generated from the PCA. However, it was unclear as to how the transformation of the data into different dimensions would be affected by the different scales associated with computation of the different texture parameters. Although normalisation of the data can mitigate the effect of parameters with scales of higher magnitude, this results in a spread of the influence of these parameters across more principal components, requiring more principal components to represent the data. Interpretability of the results would also be more difficult with PCA. Additionally, misclassification could still arise, as judged from the overlap of the data points between the two groups from the visualisation of the data from the top three principle components, suggesting that optimum classification solution based on PCA was not linear. For these reasons, PCA was not used for classification.

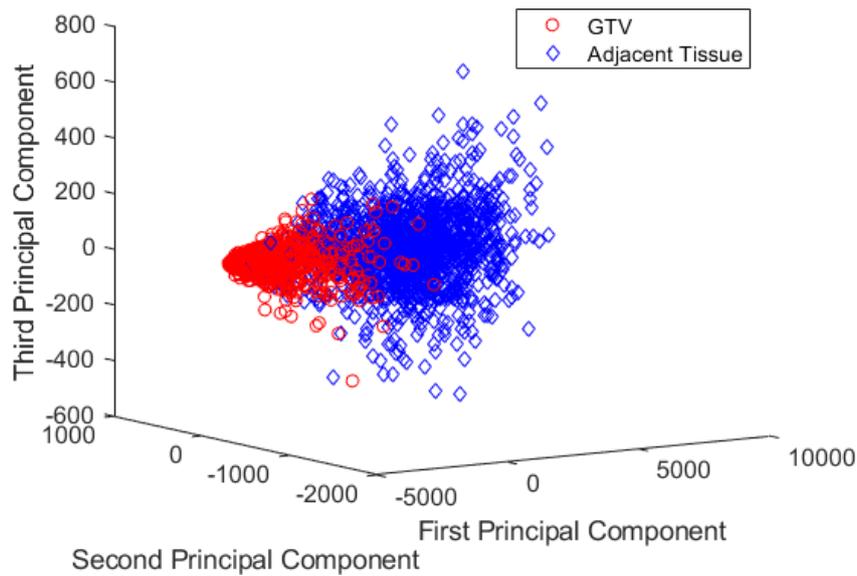


Figure 6.5. Visualisation of texture parameters separability between GTV and adjacent tissue, based on the first, second and third principal components.

6.13.3 Assumptions for discriminant analysis

The potential use of LDA for classification was further explored in MATLAB following the promising results from the B11 module.

However, for LDA, a p -value of 0 was achieved with the Bartlett test, rejecting the hypothesis of equal covariance matrices for the modelled data. As the criterion of a single Gaussian covariance was not met, the application of LDA was probably unsuitable.

The QQ plots for the LDA and QDA models are shown in figure 6.6, where it can be seen that the distributions for both discriminants did not fit the gaussian mixture model. This was confirmed through the Mardia kurtosis test where a p -value of 0 was obtained for both LDA and QDA, implying that the data was not consistent with a multivariate normal distribution for both approaches.

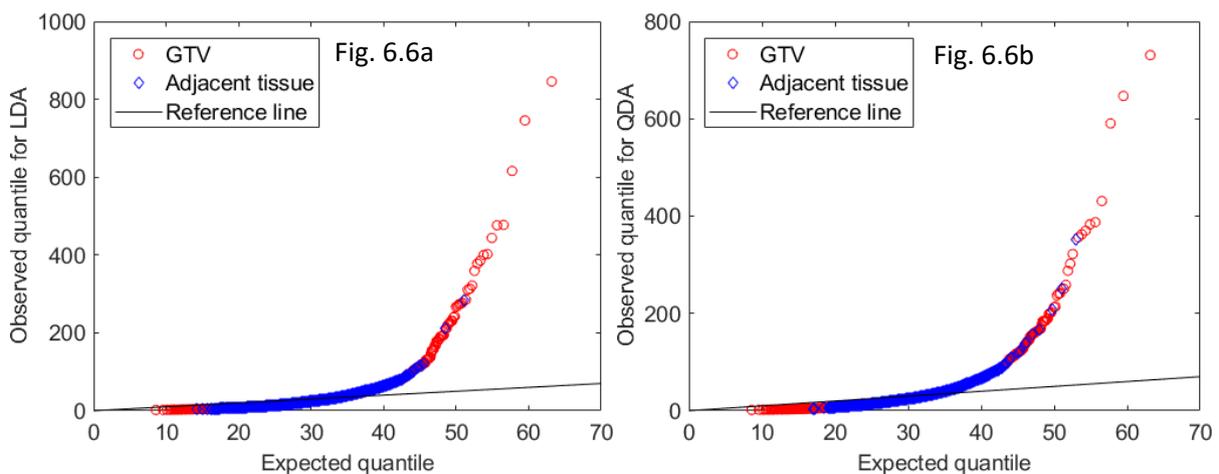


Figure 6.6. QQ plots for assessment of distribution for training data a) Linear discriminants b) Quadratic discriminants.

From these tests, it was decided that it would not be appropriate to pursue further classification with discriminant analysis.

6.13.4 k-nearest neighbours classification optimisation

Plots for all cross-validation folds that were used to select the optimal value of k are shown in appendix A.1.

Although $k = 1$ was associated with the best misclassification score in run 1, the use of a 1-NN classifier was associated with a high variance, and hence may not predict well for new testing data which the classifier has not seen. The next lowest mean classification error in training run 1 was at $k = 5$. Training runs 2 and 3 achieved the lowest mean classification error at $k = 5$. Based on this, as well as the acceptable variance seen in the above plots, $k = 5$ was selected as the optimum neighbour size for all three runs. The training error rates are shown in table 6.9.

Training Run	Misclassification error	Sensitivity	Specificity
1	$2.3 \pm 1.0\%$	$98.2 \pm 1.6\%$	$97.2 \pm 0.9\%$
2	$1.9 \pm 1.0\%$	$97.5 \pm 1.2\%$	$98.7 \pm 1.5\%$
3	$1.8 \pm 0.5\%$	$98.6 \pm 1.2\%$	$97.8 \pm 1.6\%$

Table 6.9. Error rates with 5-nearest neighbour classifier on nested validation data using multiple texture features.

6.14 Task E.2 Classification of GTV versus adjacent non-tumour tissue with single texture feature (most discriminatory feature and mean)

6.14.1 Sum Variance (0,4) as sole feature

The sum variance at 4-pixel distance and 45° direction (Sum variance (0,4)) was found to be the most discriminative texture feature in the training data (table 6.6). As described earlier, the values of sum variance (0,4) was found to be statistically different between GTV and the adjacent tissue from the Mann Whitney U test (p -value = 0.000), where the values for the GTV was lower than those for the adjacent tissue (figure 6.7).

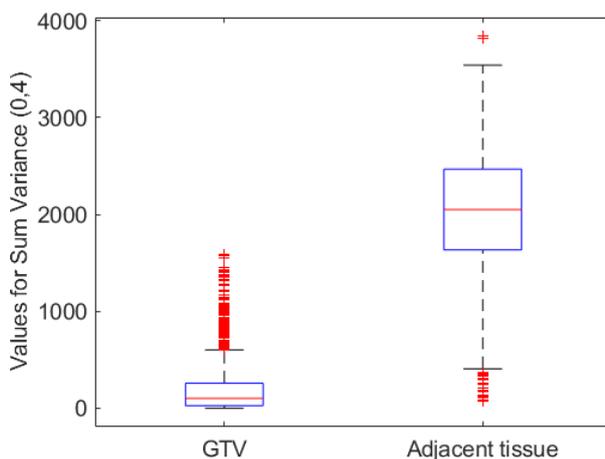


Figure 6.7. Boxplot comparing the values for sum variance of the GTV versus the adjacent tissues at 4-pixel distance and 45° direction.

6.14.1.1 k-nearest neighbours classification optimisation

Plots for all cross-validation folds that were used to select the optimal value of k with the use of sum variance (0,4) are shown in appendix A.2.

A summary of the neighbourhood size and best achieved misclassification error for each outer cross-validation run is shown in table 6.10.

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	19	$5.5 \pm 1.3\%$	$95.8 \pm 2.0\%$	$93.1 \pm 2.4\%$
2	17	$3.7 \pm 1.3\%$	$97.4 \pm 1.4\%$	$95.3 \pm 2.1\%$
3	19	$5.3 \pm 1.2\%$	$96.6 \pm 2.5\%$	$92.6 \pm 2.6\%$

Table 6.10. Error rates with optimum parameters on nested validation data using sum variance (4-pixel distance and 45° direction) for classification.

6.14.2 Mean as sole feature

The GTV was associated with higher mean values as compared to the adjacent tissue, as shown in the box plot in figure 6.8, which was statistically significant using the Mann Whitney U test at a p -value of 0.000. This indicated that the mean values may be useful in the partitioning of the two groups.

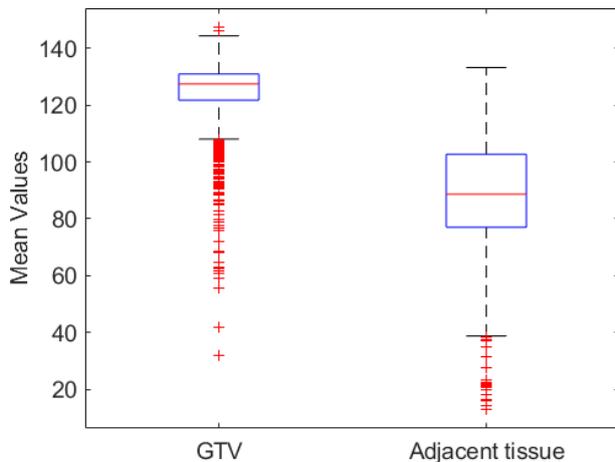


Figure 6.8. Boxplot comparing the mean values of the GTV versus the adjacent tissues.

6.14.2.1 k-nearest neighbours classification optimisation

Plots for all cross-validation folds that were used to select the optimal value of k with the use of mean as the sole feature are shown in appendix A.3.

A summary of the neighbourhood size and best achieved misclassification error for each outer cross-validation run is shown in table 6.11.

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	19	$8.9 \pm 2.2\%$	$90.5 \pm 3.5\%$	$91.5 \pm 2.6\%$
2	20	$10.3 \pm 2.3\%$	$88.0 \pm 3.9\%$	$91.5 \pm 2.0\%$
3	20	$10.5 \pm 1.6\%$	$87.9 \pm 3.6\%$	$90.9 \pm 2.8\%$

Table 6.11. Error rates with optimum parameters on nested validation data using mean values for classification.

6.14.3 Comparison of k-NN classification for multiple texture feature set versus sum variance and mean value as sole feature

6.14.3.1 Estimated classifier performance

The estimation of the optimised classifier performance for the different texture sets is shown in table 6.12, and the associated ROC in figure 6.9.

All three methods of classification showed good performance, though the best performance was seen in the use of the multiple texture feature set, with a very low error rate of $2.8 \pm 0.5\%$. This was followed by the use of sum variance (0,4) where the misclassification error was doubled, and lastly the use of the mean value, where an error rate of $10.5 \pm 1.5\%$ was seen.

The sensitivity and specificity scores were similar for the multiple features and sum variance (0,4), as compared to the use of the mean value. The results for misclassification error and sensitivity were statistically significant between the three groups analysed by the Friedman test (p -value = 0.05), and there was a trend towards statistical significance for specificity (p -value = 0.097).

Texture feature	Misclassification error	Sensitivity	Specificity	AUC
Multiple (28) texture features	$2.8 \pm 0.5\%$	$97.5 \pm 1.1\%$	$96.9 \pm 1.9\%$	0.988 ± 0.005
Sum Variance (0,4) as sole feature	$5.1 \pm 1.4\%$	$96.1 \pm 2.1\%$	$93.7 \pm 2.9\%$	0.984 ± 0.009
Mean value as sole feature	$10.5 \pm 1.5\%$	$88.6 \pm 5.5\%$	$90.5 \pm 4.0\%$	0.945 ± 0.018
p -value	0.050	0.050	0.097	0.050

Table 6.12. Estimate of classification performance with optimised classifiers parameters.

This was corroborated with the ROC curves as shown in figure 6.9, where good performance was seen with the use of the mean as the sole parameter (mean area under curve (AUC) = 0.945). Better performance was observed when the texture parameter sum variance (0,4) was used (mean AUC 0.984). The best classification results were seen with the use of the multiple texture feature set was used (mean AUC = 0.988).

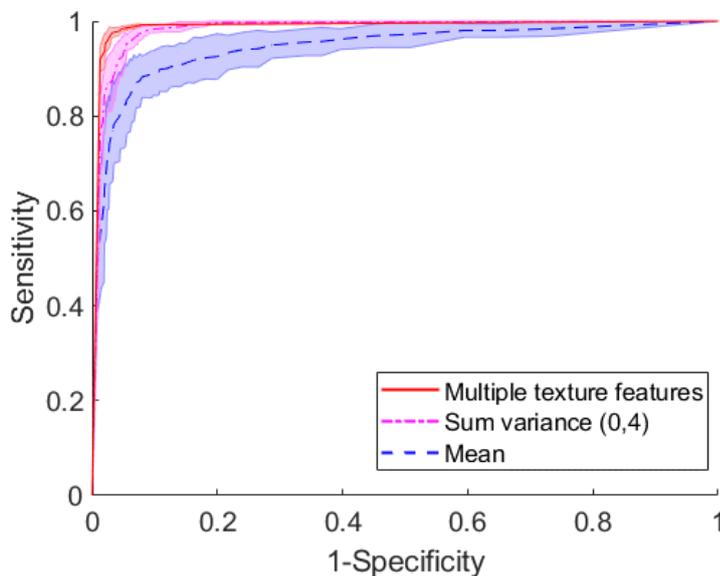


Figure 6.9. Mean ROC curves comparing performance of k-NN classification on training data using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value). Shaded regions represent confidence interval.

6.14.3.2 Re-optimisation of final classification models

The final models were re-trained using the outer 3-fold cross validation, with the performance curves shown in Appendix B.1.

The optimum error rates of the re-training data and trained final model are shown in table 6.13 and 6.14 respectively.

	Neighbour size	Misclassification error	Sensitivity	Specificity
Multiple (28) texture features	5	2.8 ± 0.5%	97.5 ± 1.1%	96.9 ± 1.9%
Sum Variance (0,4) as sole feature	17	5.0 ± 1.6%	96.1 ± 1.7%	94.0 ± 2.7%
Mean as sole feature	19	10.0 ± 1.7%	88.4 ± 5.3%	91.6 ± 2.2%

Table 6.13. Optimum error rates from re-training using optimum parameters.

	Misclassification error	Sensitivity	Specificity
Multiple (28) texture features	1.4%	98.8%	98.4%
Sum Variance (0,4) as sole feature	4.7%	96.7%	93.9%
Mean as sole feature	9.7%	88.4%	92.2%

Table 6.14. Error rates of trained data associated with final models using optimum parameters.

6.14.3.3 Independent test set

The results of the classification on the independent testing data for the multiple texture feature set and sole features (sum variance (0,4) and mean in table 6.15 with their corresponding ROC curves in figure 6.10. As compared to the training and validation data, the testing data was associated with a slight increase in the misclassification rates, indicating that despite the cross-validation, there was an element of overfitting in the training of the model. Nonetheless, the lowest rate of misclassification was seen with the use of multiple texture features, where the highest sensitivity and specificity were achieved. This was followed by the sum variance (0,4) as the sole feature, with the poorest performance observed when using the mean as the only parameter. This corroborated with the training data in that the classification of the GTV and the adjacent tissue was most accurate when multiple texture features were used.

	Misclassification error	Sensitivity	Specificity	AUC
Multiple (28) texture features	4.7 %	95.7 %	94.9 %	0.978
Sum Variance (0,4) as sole feature	6.8 %	94.0 %	92.5 %	0.973
Mean as sole feature	11.7 %	88.4 %	88.2%	0.949

Table 6.15. Performance of final k-nearest neighbour classifier on independent test set for multiple texture features, sum variance (0,4) as sole parameter, and mean value as sole feature.

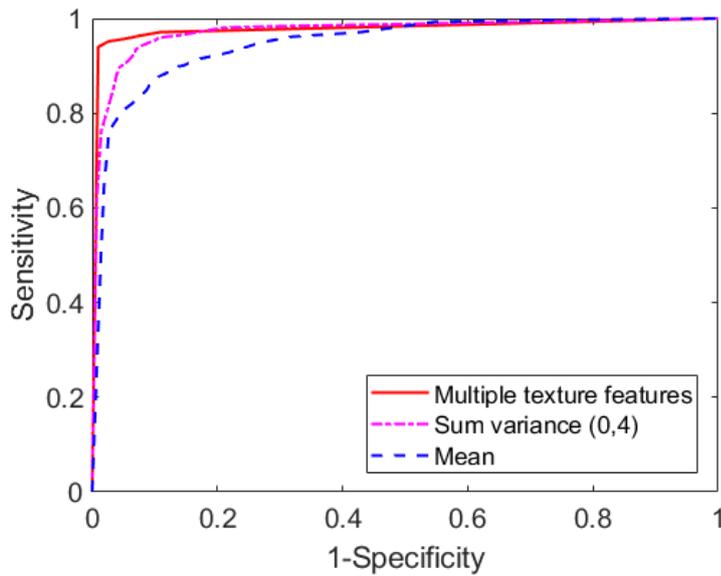


Figure 6.10. ROC curves comparing performance of k-NN classification on independent test set using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value).

6.15 Task E.3 Classification of GTV versus tissue at a distance away

The partitioning of GTV and tissue at a distance away was performed to assess the applicability of the model on discriminating tumour from other non-tumour tissue.

6.15.1 Application of training data

There was no difference in the sensitivity between classification of the GTV and adjacent tissue, versus GTV and the tissue at 10-pixel distance away for all three feature sets. This was expected, as the features extracted from the GTV were no different between them.

Similar misclassification error and specificity scores were seen for use of the mean value, which suggests that the classification model worked equally well when applied to tissue at a distance away. Lower misclassification error and higher specificity scores were obtained when the multiple texture feature and sum variance (0,4) as a single feature were applied, suggesting that the classification models were more effective at partitioning more distant tissue.

		Multiple (28) texture features	Sum Variance (0,4) as sole feature	Mean value as sole feature	<i>p</i> -value
Misclassification error	Adjacent tissue	2.8 ± 0.5%	5.1 ± 1.4%	10.5 ± 1.5%	0.050
	Tissue 10 pixels away	1.8 ± 0.2%	2.9 ± 0.8%	10.3 ± 1.8%	0.050
Sensitivity	Adjacent tissue	97.5 ± 1.3%	96.1 ± 2.1%	88.6 ± 5.5%	N.A.
	Tissue 10 pixels away	97.5 ± 1.3%	96.1 ± 2.1%	88.6 ± 5.5%	
Specificity	Adjacent tissue	96.8 ± 1.9%	93.7 ± 2.9%	90.5 ± 4.0%	0.097
	Tissue 10 pixels away	98.9 ± 1.0%	98.1 ± 0.6%	90.8 ± 2.7%	0.050
AUC	Adjacent tissue	0.988 ± 0.005	0.984 ± 0.009	0.945 ± 0.018	0.050
	Tissue 10 pixels away	0.995 ± 0.001	0.995 ± 0.002	0.947 ± 0.020	0.097

Table 6.16. Comparison of performance of optimised k-nearest neighbour classifier on classifying GTV versus adjacent tissue, and tissue at 10-pixel distance using training data.

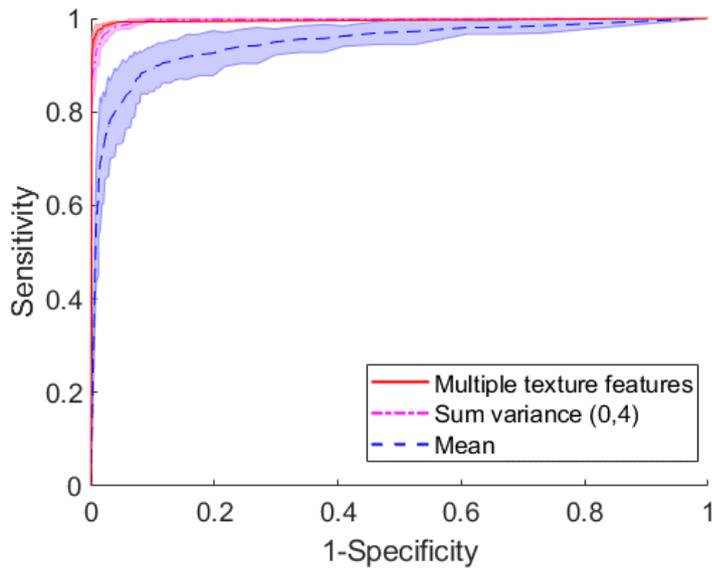


Figure 6.11. Mean ROC curves comparing performance of trained k-NN classifier on testing data using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value) in the classification of GTV versus tissue at 10-pixel distance away, where models were trained on features derived from GTV and adjacent tissue. Shaded regions represent confidence interval.

The classification performance of the GTV versus tissue at 10-pixel distance away using the re-trained final models is shown in table 6.17.

	Misclassification error	Sensitivity	Specificity
Multiple (28) texture features	0.9 %	98.8 %	99.4 %
Sum Variance (0,4) as sole feature	2.7 %	96.7 %	97.9 %
Mean as sole feature	9.8 %	88.4 %	91.9 %

Table 6.17. Errors in application of data for GTV versus tissue at 10-pixels on re-trained final models.

6.15.2 Independent test set

Application of the independent test set on the respective trained classification models corroborated with the results seen above, where the poorest performance was observed when the mean value was applied. Use of sum variance (0,4) as the sole texture parameter was as sensitive as the use of the multiple texture feature set, though it produced slightly inferior results for accuracy and specificity.

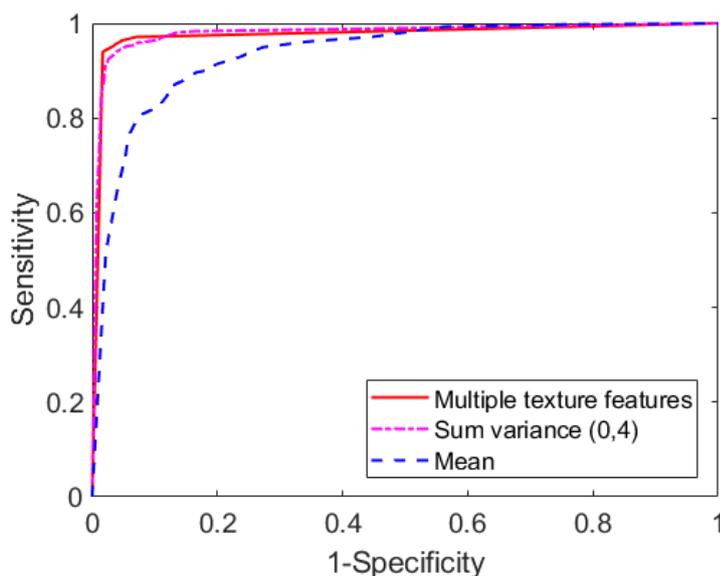


Figure 6.12. ROC curves comparing performance of trained k-NN classifier on independent test set using multiple texture features versus sole features (sum variance (4-pixel distance and 45° direction) and mean value) in the classification of GTV versus tissue at 10-pixel distance away, where models were trained on features derived from GTV and adjacent tissue.

	Misclassification error	Sensitivity	Specificity	AUC
Multiple (28) texture features	4.1 %	95.7 %	96.1 %	0.975
Sum Variance (0,4) as sole feature	4.9 %	96.1 %	94.0 %	0.979
Mean as sole feature	13.5 %	88.4 %	84.6 %	0.935

Table 6.18. Performance of k-nearest neighbour classifier on classifying GTV versus tissue at 10-pixel distance away, using multiple texture features, sum variance (0,4) as sole parameter, and mean value as sole feature on independent test set. Classification models were trained based on classification of GTV versus adjacent tissue for the respective feature sets.

Discussion

In the evaluation of the use of single features for classification, better classification was obtained when the best discriminating texture descriptor was used, in this case sum variance (0,4), as compared to the mean to distinguish GTV from the surrounding tissue. However, the use of multiple texture features outperformed the use of single parameters. This evaluation shows that texture features are better at discriminating tumour from adjacent tissue, where less errors were observed than the using the mean HU alone. Stability of the classification models was also demonstrated through their ability at distinguishing tumour from regions further away.

This suggests that image texture descriptors contain useful information about tumour regions beyond standard first-order statistical evaluation. In this study, classification was performed using large ROIs encompassing whole tumour regions and their surrounding tissue. This is a relatively easy classification task, which can explain the good performance observed here. However, in order to apply texture descriptors to segmentation, there is need to evaluate their performance in the partitioning of smaller regions, which is a more difficult task. Moreover, the effect of tumour boundary regions also needs to be assessed, i.e. how excluding part of the tumour and how inclusion of the surrounding tissues impact the classification.

It was interesting to note that the variance and entropy descriptors of the co-occurrence matrices ranked highly in the feature selection process. These features describe how varied and disordered the co-occurrence matrices are, which were found to be higher for the surrounding tissues compared to the tumour. This indicates that there was greater heterogeneity in the surrounding regions than the tumour, which is likely due to the multiple tissue types present in the non-tumour region.

In this work, a combination of three different filter selection methods were used for feature selection. Advantages of the filter feature selection techniques include the speed and ease of computation. Also, as the ranking and selection criteria of the filter methods are independent of the classification and training procedures, they are more generalizable and therefore more robust to overfitting compared to other methods. On the other hand, as each feature is considered independently with filter methods, redundant features may be selected which can impact the classification results. Although the performance with the multiple feature set was superior to the single highest ranked feature, the difference in performance was small, suggesting that there was a lot of redundancy in the multiple feature set used.

Thus, further evaluation on the selection of features should be performed to assess how a different or a smaller feature set would affect classification. Generally, a smaller feature set would be preferred over a larger set to decrease the risk of overfitting and computational expense. However, as the coefficients of the ranking criteria was not available through MaZda with the combination feature selection approach, further reduction in the feature list would have been arbitrary.

Slightly higher error rates were observed when the classification was performed on the independent test set. Although this trend is typically observed in model testing, an optimistic bias may have been introduced in the estimate of the classification performance from the feature selection process. The final methodological design comprised of an inner cross-validation to select the best hyperparameter, followed by an outer cross-validation to estimate the classification errors. This nested cross-validation approach was aimed at reducing overfitting errors and to allow for the error estimate to be unbiased (435). However, as feature

selection was not performed within the inner cross-validation folds, some bias may have been introduced into the cross-validated estimate of the classification performance (436).

Nonetheless, these results indicate that texture features are better at the binary discrimination of tumour versus non-tumour regions than single parameters. These findings support the subsequent evaluation of texture features in segregating tumour from non-tumour with smaller ROIs, towards the goal of texture segmentation.

Conclusions

Excellent discrimination of tumour from non-tumour regions was achieved using a multiple texture feature set, which showed better performance than the use of a single best discriminatory texture feature or the mean as a single feature. There is a need to assess the classification performance on smaller ROIs in order to develop the workflows for texture segmentation.

Chapter 7

Specific Aim F: Evaluation of tumour and non-tumour discrimination with variation of distance from tumour boundary and ROI sizes

Introduction

The results in the preceding chapter show some promise that texture features can be helpful in the discrimination of tumour from its surrounding tissue. As an extension from the previous work, it was decided that the next step would be to explore the ability of the classifier in the partitioning of tumour versus non-tumour regions based on smaller ROIs. This is an important step in the development of a texture segmentation process, where partitioning of tumour and non-tumour regions is carried out on smaller regions (clusters of pixels/voxels), as texture descriptors are dependent on the size of the evaluated ROIs (i.e. degree of magnification). In addition to assessing the discrimination of pure tumour and non-tumour regions, an evaluation at the tumour boundary was also undertaken in order to assess the classification performance where ROIs are composed of varying amounts of tumour and non-tumour tissue, heading towards the goal of tumour segmentation using texture features.

Ideally, a study design using an overlapping sliding window approach within a bounding region around the GTV for ROI selection would be performed. Each of the window would constitute an ROI, on which the textural features would be computed. Such a method would be most thorough, as it would involve the evaluation of all the pixels and their surroundings within the search region. However, large quantities of ROIs would be generated using this methodology. It would not be pragmatic to perform the computation of the textural features in MaZda using the established workflow as it involved preparing, importing and processing individual files for each ROI. Even if the ROI selection tool were used within MaZda, as it only allows for up to 16 different ROIs to be placed on a single slice, such an evaluation would not be feasible.

Thus, automation of the whole process using MATLAB was explored initially. These procedures were set up, where thirteen feature maps based on first-order statistics, gradient and Euclidean distance were created by tiling a filter of varying sizes (3- to 9-pixel squares) across the image for both training and testing data. Subsequently, using the training data, classification based on the texture maps was performed to assign each pixel according to their class membership. The test data was then applied to the trained classification model. Additionally, the pixels and their indices of the test data were also extracted, and from the results of the class membership, mapping of tumour and non-tumour regions were performed on a pixel-wise basis. With the ability to map the predicted pixels back into its spatial context using the extracted indices, the process of tumour segmentation was therefore performed on a pixel-wise basis. However, this workflow was limited by the textural computation facilities in MATLAB, which was not as extensive as compared to MaZda. Computation of individual textural maps using the sliding window approach was feasible in MaZda, which could then be imported into MATLAB. The disadvantage of this process is the computational and manual expense in generating and exporting the individual feature maps for each feature, for each image slice of each case. More importantly, the bit range of the exported texture map data from MaZda (bmp files) was not sufficient for the calculated features.

Instead of performing a binary classification, multi-class classification was also briefly explored using the methods above. This involved the discrimination of tumour from the eight

individual tissue types (see list in chapter 2 table 2.1). Although this was achievable using a small number of cases (three cases), the processing time for training the classification model increased significantly when it was extended to larger numbers, taking several days. Issues with imbalanced classes was also observed.

It was decided that the computation of the texture features should be performed in MaZda in order to generate the range of texture parameters. Because of this, rather than performing a pixel-wise evaluation through a sliding window approach, the analysis was undertaken using a subsample of ROIs selected randomly, and a binary classification system was applied. While MaZda was used for texture parameter generation, procedures for ROI generation, processing and classification were performed in MATLAB.

7.1 Summary of tasks

Task F.1 Assessment of classification of GTV versus non-tumour tissue, using ROIs (8- and 16-pixel squares) with no overlap at GTV boundary (i.e. pure tumour versus pure non-tumour tissue)

Task F.2 Assessment of classification of GTV versus non-tumour boundary, using ROIs (8- and 16-pixel squares) with overlap at GTV boundary (i.e. tumour versus non-tumour tissue, both with contamination of tissue from the other class)

Methods

7.2 Generation of ROIs – BMP files

The importing and processing of the CT image files and the reference contours were described in section 6.3, which were then used to generate ROI files. For each image slice, five ROIs of 8- and 16-pixel square sizes were defined within each of the respective regions. This was performed by using a random number generator to select the indices of the ROI centroids from a list of indices extracted from the respective regions, from which an expansion was applied to create the ROIs. For the ROIs consisting of pure tissue types, it was ensured that the GTV and non-tumour ROIs were placed completely within their own regions. This was performed through appropriate expansion and erosion of the GTV mask file with logical operators to limit the list of indices of the respective regions, followed by a further check for overlap to ensure that erroneous ROIs were not created. The GTV ROIs were further processed to exclude non-tumour regions, where GTV contours were smaller than the generated ROIs. The non-tumour region was limited to a 20-pixel rim surrounding the GTV. Like the earlier work, the body contour was extracted from the non-tumour region using Otsu's thresholding.

The boundary ROIs were created with a similar process, from the region defined through appropriate expansion and erosion of the GTV mask. A square 8-pixel structuring element was used in the expansion/erosion from reference contours for the 16-pixel square ROIs, whilst a 4-pixel structuring element was used for the 8-pixel ROIs. Logical operators were performed on the generated ROIs to determine the number of pixels which overlapped with the GTV and non-tumour region. Boundary ROIs within the GTV were defined at $> 50\%$ and $< 100\%$ overlap with GTV, whilst boundary ROIs within the non-tumour region were defined at $> 50\%$ and $< 100\%$ overlap with the non-tumour region. Image slices with smaller GTV than ROIs were excluded, as it would not be possible to generate boundary GTV according to the above definition.

An example of an image slice with the corresponding ROIs is shown in figure 7.1.

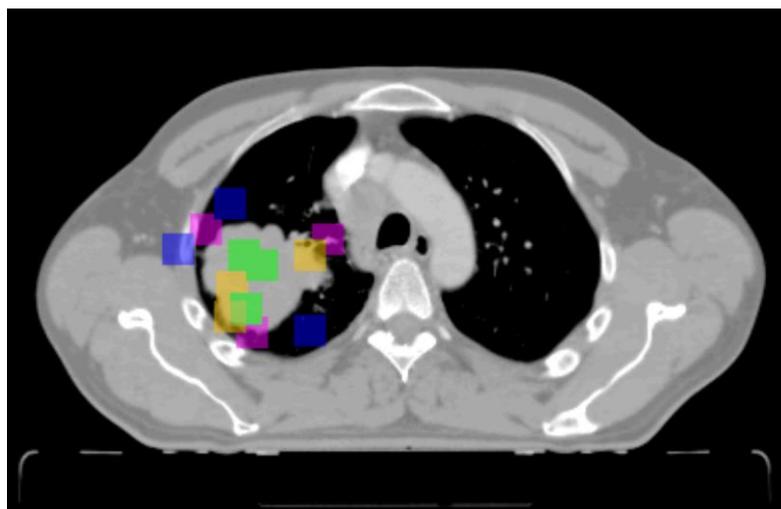


Figure 7.1. Example of a representative slice for ROI placement (16-pixel square size shown). For illustration, three ROI samples are displayed for each group, though five samples for each group was used in all the experiments. Green: Non-boundary tumour tissue; Blue: Non-boundary non-tumour tissue; Orange: Boundary tumour tissue; Pink: Boundary non-tumour tissue.

The resultant masks were then exported as separate BMP files, from which ROI files were created and processed as per the workflow described in chapter 6 section 6.5.

7.2 Generation of texture features

Wavelet parameters between scales 3 and 5 were not computed. Co-occurrence matrices at 2,3 and 5 distances were also not considered. Thus, for the 16-pixel square ROIs, a total of 135 texture features were computed (see chapter 6 section 6.6 for feature list).

For the 8-pixel square ROIs, 130 texture features were used. This comprised of the same list as that used for the 16-pixel square ROIs apart from the five autoregressive model parameters as these could not be computed.

7.3 Feature selection

The feature selection process was modified in this work based on the observed results in chapter 6. To avoid feature redundancy and to keep the feature list as small as possible, it was decided that a sequential feature selection method would be used. This belongs to the wrapper feature selection approach, where subsets of features are analysed iteratively to decide on whether the feature in question would be included or excluded from the list. In sequential feature selection, each feature is added with each iteration and kept, if the feature improves the performance of the model. k-NN classification was applied as the predictive model, using the same number of neighbours for both parameter tuning and testing (see workflow in section 7.5). In other words, the criterion for feature selection was based on the performance of the feature using the k-NN classifier, thereby optimising the performance of the classification. Unlike filter methods where the number of included features had to be specified upfront, this is not required for wrapper methods where the number of features is determined by its impact on classification.

Disadvantages of wrapper feature methods include their tendency to overfit, as well as being computationally expensive. To reduce this, a pre-processing step was incorporated using a filter-based method. The class-separability criterion was defined as AUC of the ROC curve, which is a non-parametric test i.e. no assumptions on the distribution of the underlying data were made. The top 70 features were then passed on to the sequential feature selection process (figure 7.2).

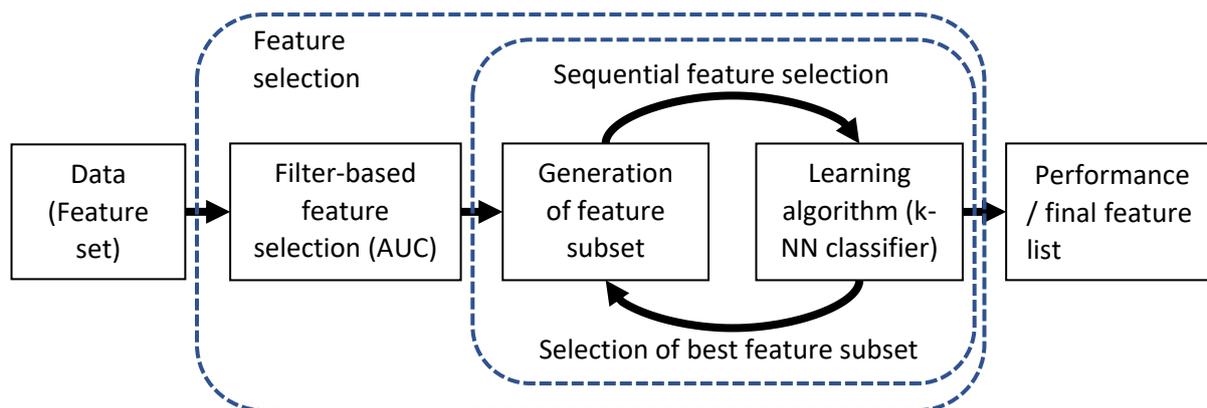


Figure 7.2. Diagrammatic representation of feature selection workflow.

7.4 k-NN classification and classifier assessment

The same process for parameter tuning and assessment of performance was carried out as previously described (chapter 6 section 6.9.1). Although all values of neighbourhood size from 1 to 20 were computed, only odd values were considered in the selection process to avoid issues with ties in classification class.

7.5 Study design – classification using k-NN classifier

The workflow for classification tuning and assessment is shown in figure 7.3.

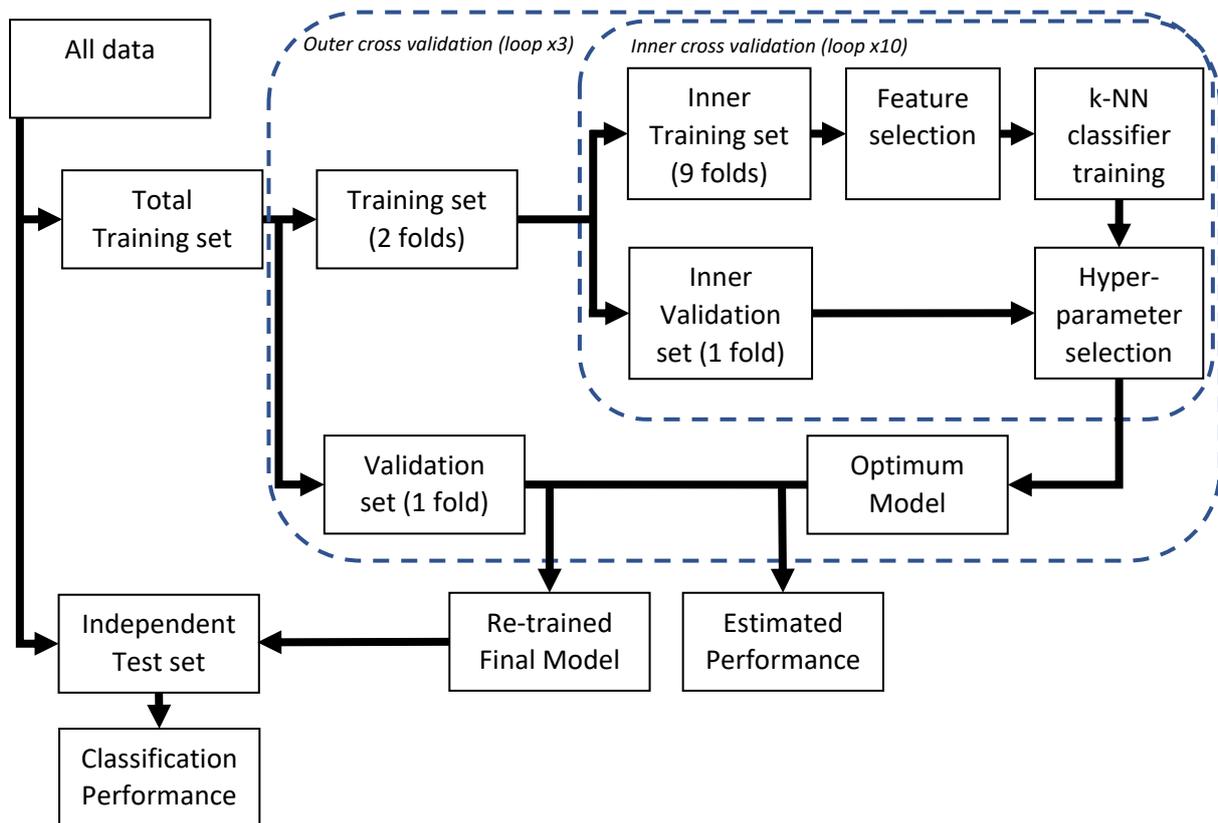


Figure 7.3. Workflow for k-NN classification.

The workflow was similar to that used in the preceding chapter, except the incorporation of the feature selection process in the inner cross-validation loop. This design would give a better estimate of the classification performance from the outer cross-validation runs, which was then corroborated with the independent testing set on the final model.

The same folds for splitting the data into the independent test set and training set (including both outer and inner cross-validations) were used to maintain consistency throughout all the experiments. A summary of the sample sizes for the various runs is shown in table 7.1.

	Cases	Slices	Samples		Cases	Mean number of samples		Mean number of samples
Independent test set	16	415	4150					
Total training set	63	1582	15820	Validation set	21	5273		
				Training set	42	10547	Nested Validation set	1055
							Nested Training set	9492

Table 7.1. Division of data and sample sizes for classification tuning and testing.

For the analysis of the boundary regions, slices with smaller GTV than ROIs were excluded, resulting in 15790 samples for the total training set for the 16-pixel analysis, and 15175 samples for the 8-pixel analysis. There were 4030 samples used for independent testing for both region sizes.

7.6 Experiments

7.6.1 Application of GTV versus non-tumour tissue on classification model built on whole tumour/non-tumour regions

In a preliminary experiment, the total training set for ROIs comprising of pure tumour and non-tumour regions (16-pixel size) was applied to the previously built final model based on whole tumour/non-tumour regions. Despite achieving a good sensitivity score of 98%, there was a high false positive rate and an associated specificity at 52%, resulting in a high misclassification error of 25%, which was significantly poorer than the previous results. Possible explanations for this include poor representation of the data for the model, issues with classifier fitting (including parameter tuning), poor feature selection, or that classification of smaller ROIs is indeed more difficult.

Thus, new classification models were tuned and built for these new datasets to assess the performance in the classification of these regions.

7.6.2 Classification of GTV versus non-tumour tissue for non-boundary ROIs

The classification models were built after optimising for the neighbourhood size using the reduced feature set, to discriminate tumour and non-tumour ROIs which were placed away from the tumour boundary. These regions consisted of either pure tumour or non-tumour tissue, with no overlapping pixels (referred as non-boundary ROIs in this report). 16- and 8-pixel square ROIs were evaluated.

7.6.3 Classification of GTV versus non-tumour tissue at tumour boundary

Additionally, the same optimisation and model building procedure was performed to classify tumour and non-tumour ROIs which were placed at the tumour boundary (referred as boundary ROIs in this report). All of these ROIs comprised of a mix of tumour and non-tumour tissue, where each ROI comprised of more than 50% (but less than 100%) overlap with their corresponding class. Both 16- and 8- pixel ROIs were also assessed.

Results

Classifier training, parameter tuning and feature selection performed within the nested cross-validation loops is reported for each of the experiments (non-boundary and boundary regions; 16- and 8-pixel ROIs), followed by the estimated classification performance and feature selection for the four experiments.

Parameter selection for the final model is subsequently reported, followed by the results of the independent test set including its performance and selected features.

7.7 Parameter tuning: Nested cross-validation

7.7.1 Task F.1 Assessment of classification of GTV versus non-tumour tissue at non-boundary region

7.7.1.1 ROI size: 16-pixel square

Similar error rates and trends were observed for all three training runs, with the nested validation error rates levelling off beyond k of 5 (figure 7.4). There was some element of overfitting to the nested training data, judging by the variance of the curves. This was larger at smaller neighbour sizes but was also present at higher values of k , though this was felt to be small. Nonetheless, with similar error rates, there was parameter stability at higher values of k . Within increasing neighbour size, all three runs showed higher sensitivity at the expense of specificity (figure 7.5).

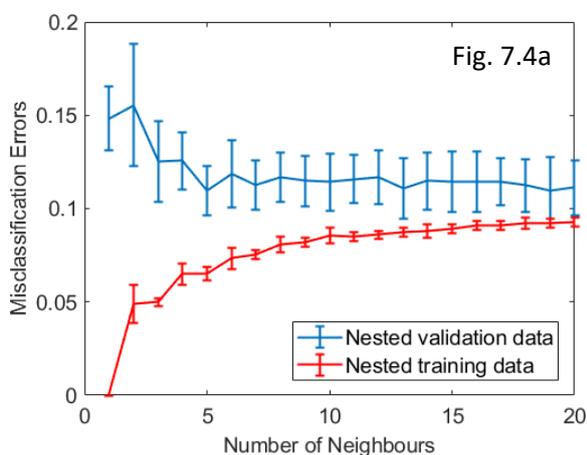
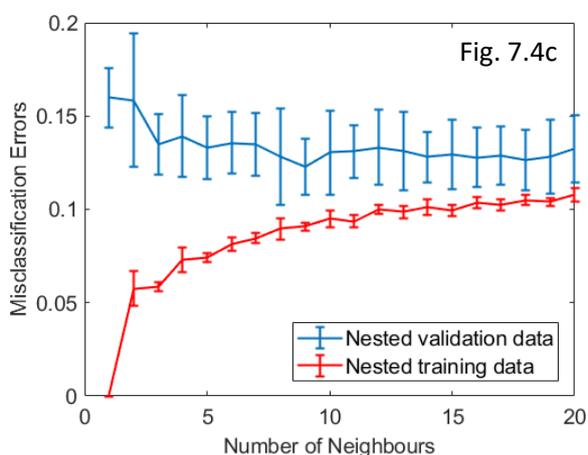
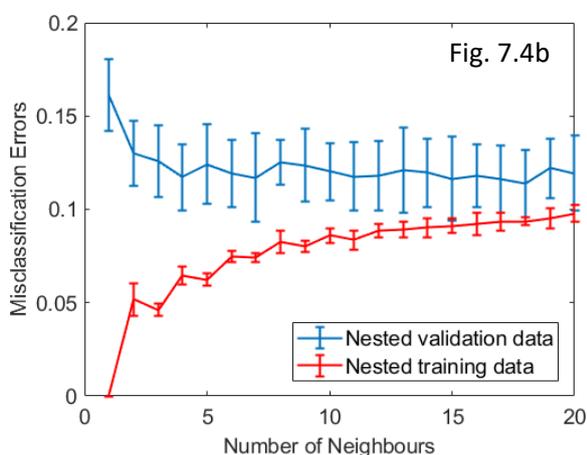


Figure 7.4. Mean misclassification errors for k -NN classification of non-boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



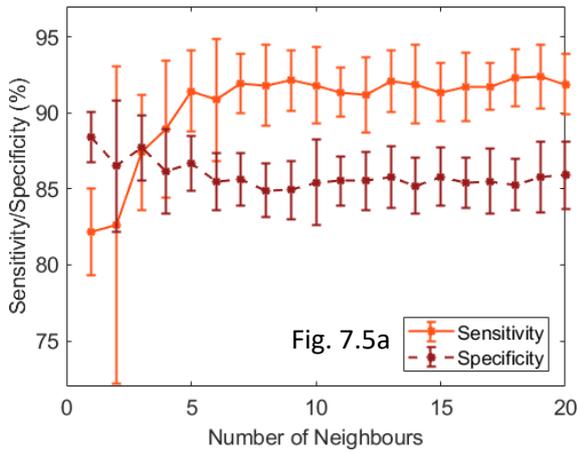


Figure 7.5. Mean sensitivity and specificity plots of nested validation data for k-NN classification of non-boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.

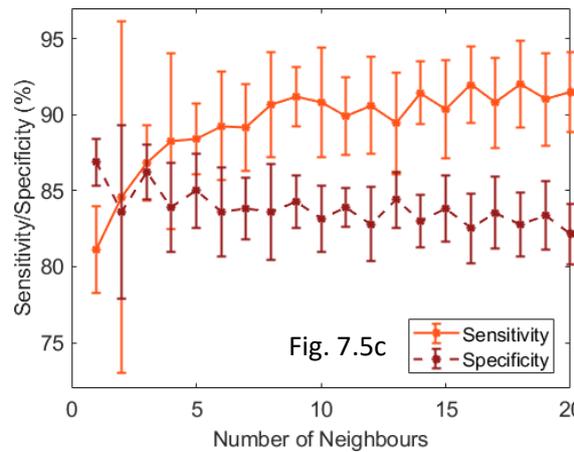
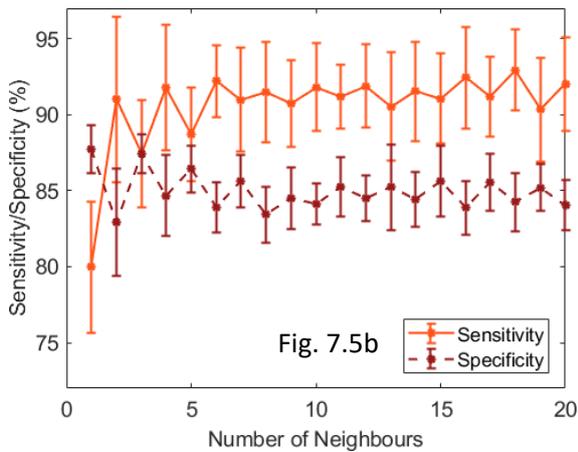


Figure 7.6 shows the number of selected features for the three runs. There appears to be some consistency in the number of features chosen between the three runs, with little variation with the neighbour size.

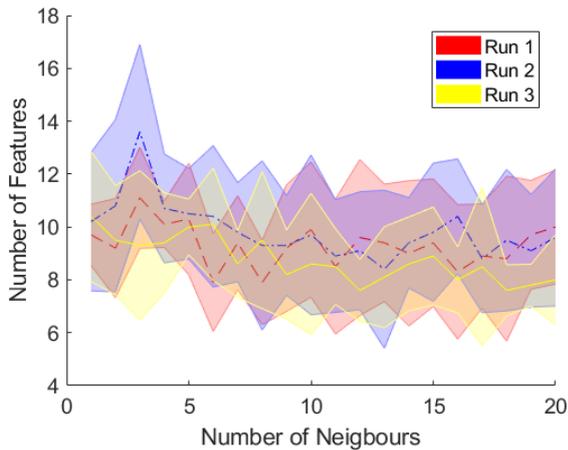


Figure 7.6. Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Non-boundary regions with ROI size of 16-pixel square). Shaded regions represent standard deviation.

The number of neighbours associated with the lowest nested validation misclassification error is shown in table 7.2. Despite the differences in the neighbour sizes across the three runs, the error rates were similar. These parameters were used to build the models based on the training data, in order to estimate the model performance in the validation set.

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	5	$11.0 \pm 1.3\%$	$91.4 \pm 2.7\%$	$86.7 \pm 1.8\%$
2	17	$11.6 \pm 1.8\%$	$91.2 \pm 2.6\%$	$85.6 \pm 1.9\%$
3	9	$12.3 \pm 1.5\%$	$91.2 \pm 1.9\%$	$84.3 \pm 1.7\%$

Table 7.2. Error rates with optimum parameters on nested validation data (Non-boundary regions with ROI size of 16-pixel square).

7.7.1.2 ROI size: 8-pixel square

Similar trends for bias and variance results were obtained for all three cross-validation runs in the tuning of the hyperparameters and features, with levelling of the curve beyond a neighbourhood size of 10 (figure 7.7). The minimum misclassification errors associated with the first and second cross-validation runs were at $k = 20$ ($14.5 \pm 1.7\%$) and $k = 14$ ($14.7 \pm 1.4\%$) respectively, though odd number of neighbours were selected for building the classifier instead. The sensitivity and specificity plots also levelled at k greater than 10 (figure 7.8).

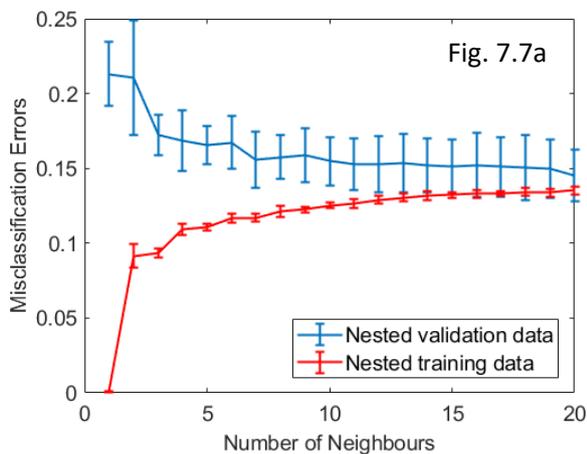
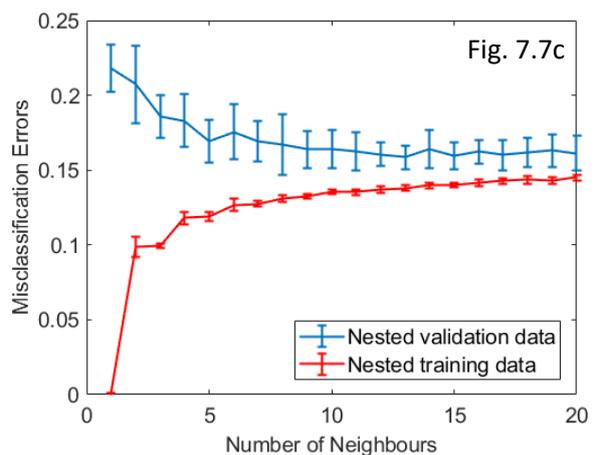
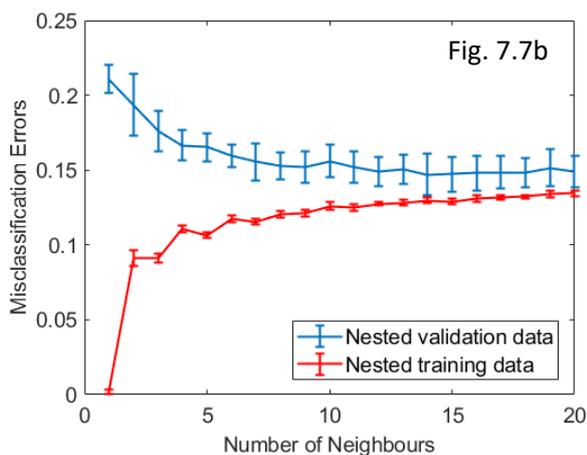


Figure 7.7. Mean misclassification errors for k-NN classification of non-boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



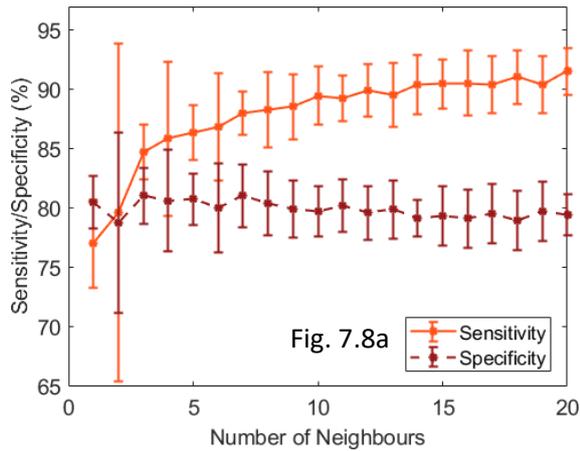
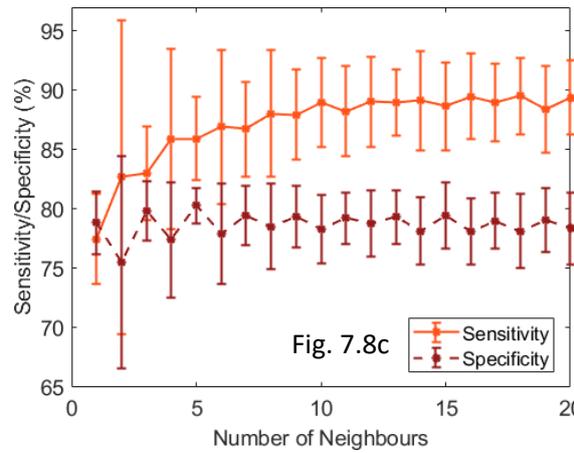
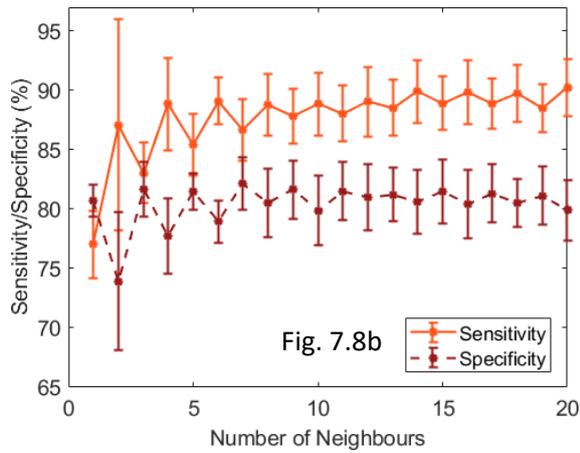


Figure 7.8. Mean sensitivity and specificity plots of nested validation data for k-NN classification of non-boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



Smaller number of features were selected compared to 16-pixel ROIs. Again, similar number of features were selected across the different neighbour sizes.

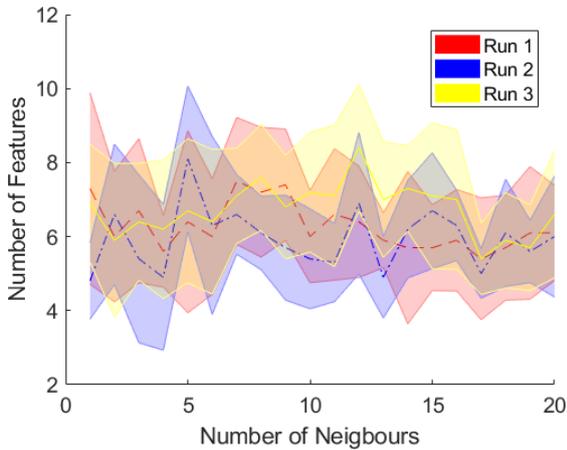


Figure 7.9. Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Non-boundary regions with ROI size of 8-pixel square). Shaded regions represent standard deviation.

The selected number of neighbours and their associated error rates is shown in table 7.3.

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	19	$15.09 \pm 2.0\%$	$90.4 \pm 2.4\%$	$79.8 \pm 2.5\%$
2	15	$14.7 \pm 1.2\%$	$88.9 \pm 2.3\%$	$81.5 \pm 2.7\%$
3	13	$12.3 \pm 1.5\%$	$91.2 \pm 1.9\%$	$84.3 \pm 1.7\%$

Table 7.3. Error rates with optimum parameters on nested validation data (Non-boundary regions with ROI size of 8-pixel square).

7.7.2 Task F.2 Assessment of classification of GTV versus non-tumour tissue at boundary region

7.7.2.1 ROI size: 16-pixel

Parameter stability was achieved at k greater than 8. Higher biases were observed as compared to the non-boundary regions, though variance of the nested training and validation runs were similar (figure 7.10).

The resubstitution loss for the second run at $k = 2$ was high, with a wide standard deviation. Assessment of the inner cross-validation folds revealed that for one (out of ten) of the inner cross-validation runs, only one feature was selected in the inner training process, which was associated with a higher training error rate as compared to the other folds, as well as a corresponding drop in the sensitivity rate (figure 7.10b). This effect of the sequential feature selection process was still seen even when the feature selection was repeated in the absence of the pre-processing feature filtering step.

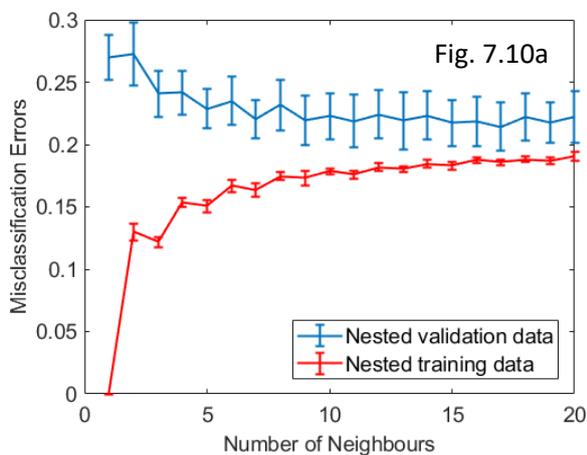
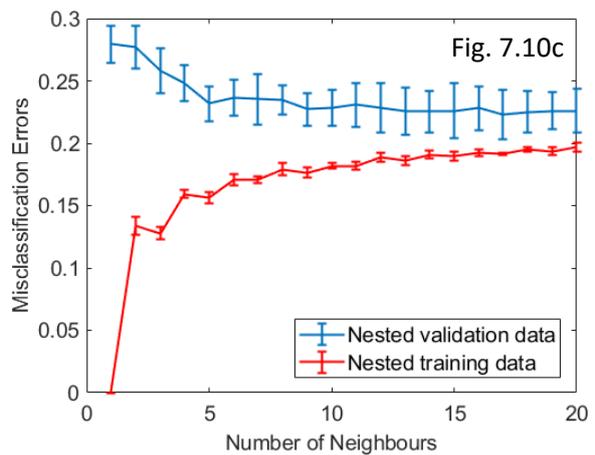
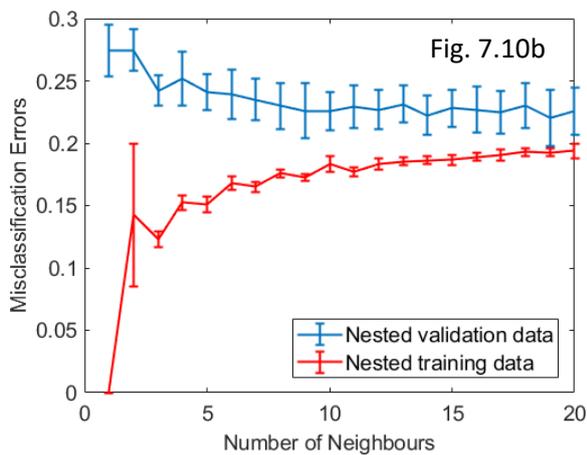


Figure 7.10. Mean misclassification errors for k -NN classification of boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



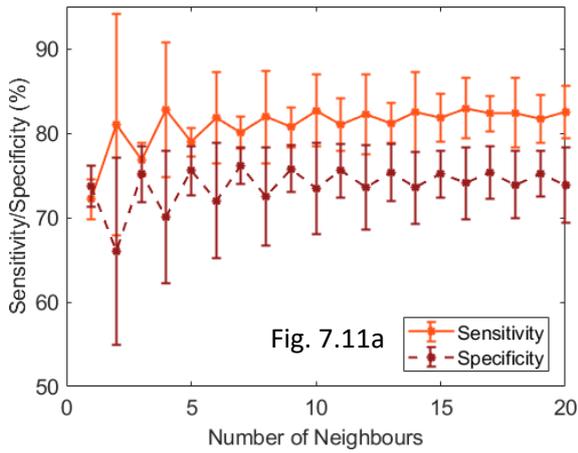
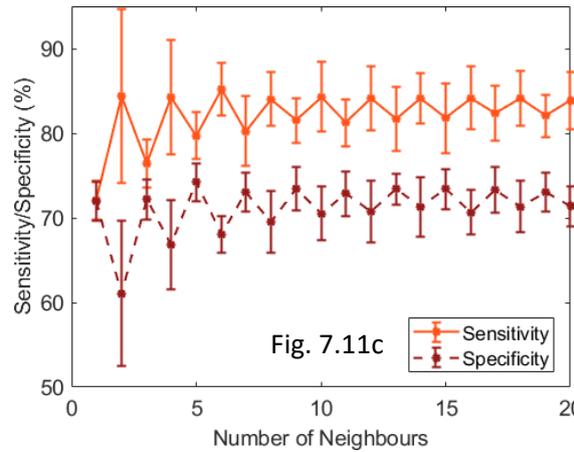
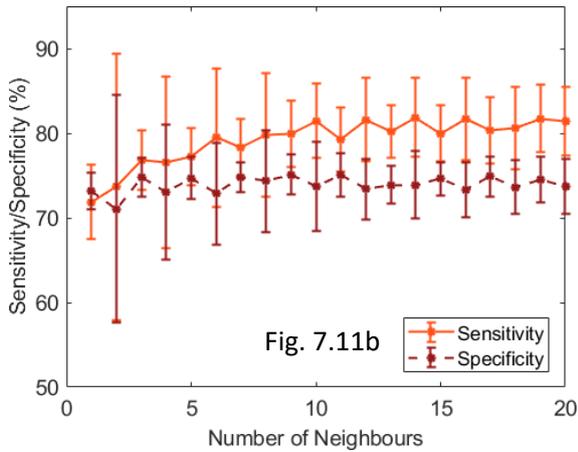


Figure 7.11. Mean sensitivity and specificity plots of nested validation data for k-NN classification of boundary regions with ROI size of 16-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



The trend in number of selected features was similar to that observed for the non-boundary region evaluation with the same 16-pixel ROI size.

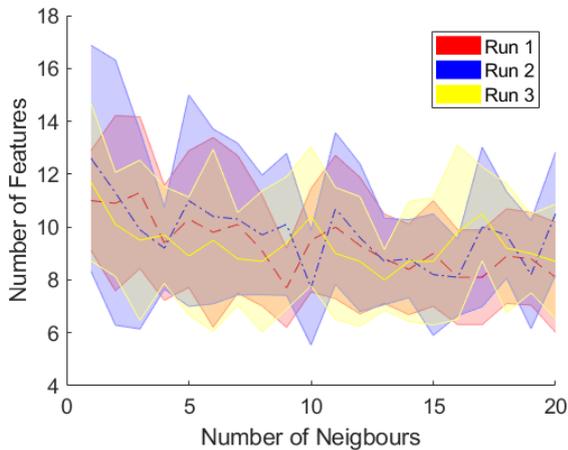


Figure 7.12. Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 16-pixel square). Shaded regions represent standard deviation.

Table 7.4 shows the lowest misclassification error achieved with each of the outer cross-validation training folds and their respective value of k.

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	17	$21.4 \pm 1.9\%$	$82.3 \pm 2.1\%$	$75.3 \pm 3.1\%$
2	19	$22.0 \pm 2.3\%$	$81.8 \pm 3.9\%$	$74.5 \pm 2.7\%$
3	17	$22.3 \pm 2.0\%$	$82.4 \pm 3.3\%$	$73.3 \pm 2.7\%$

Table 7.4. Error rates with optimum parameters on nested validation data (Boundary regions with ROI size of 16-pixel square).

7.7.2.2 ROI size: 8-pixel

The phenomenon observed in the nested training of the boundary regions using 16-pixel ROI size at $k = 2$ was more apparent using an 8-pixel ROI. Across the three training runs, this effected the resubstitution loss between k of 1 and 9 (figure 7.13). In this range, there was variation in the number of inner cross-validation training runs that derived only one single feature in the feature selection process. When averaged with the rest of the inner cross-validation training runs where multiple features were selected, higher inner training loss and larger standard deviations were observed. Because of this, the trends in sensitivity and specificity were different between the three runs at k less than 9 (figure 7.14). Nonetheless, at k greater than 10, this effect was not present and the training curves were following expected patterns.

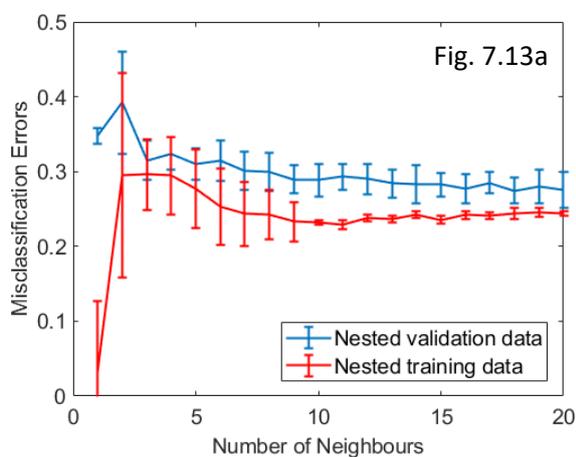
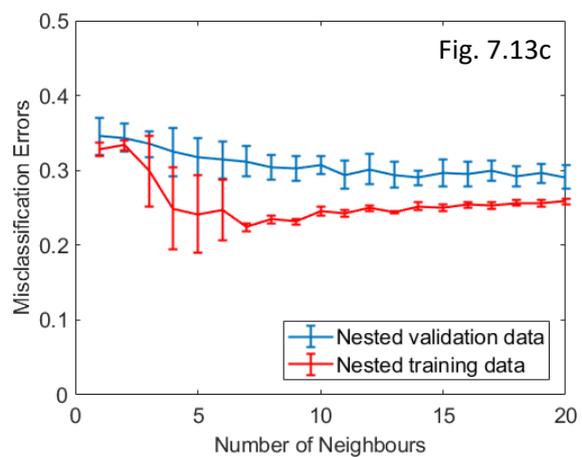
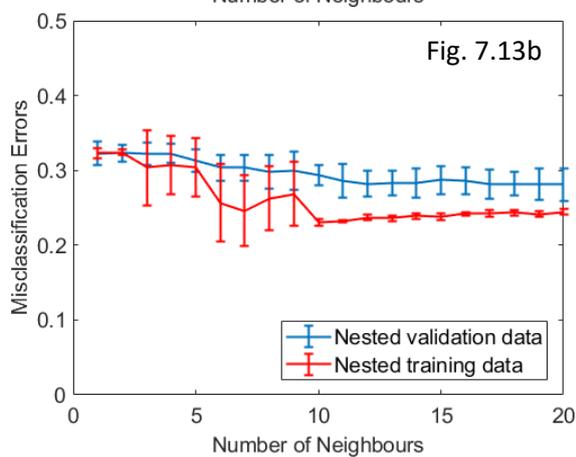


Figure 7.13. Mean misclassification errors for k -NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



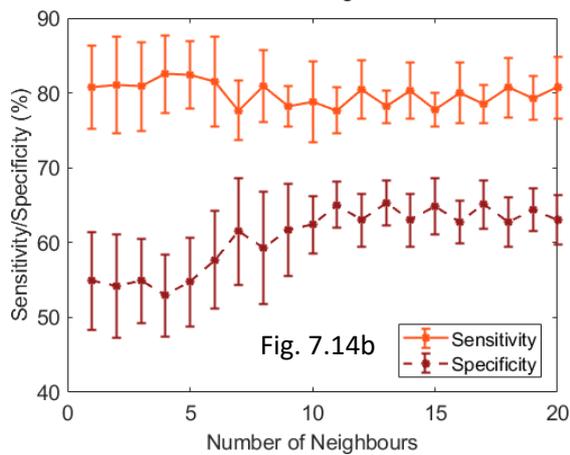
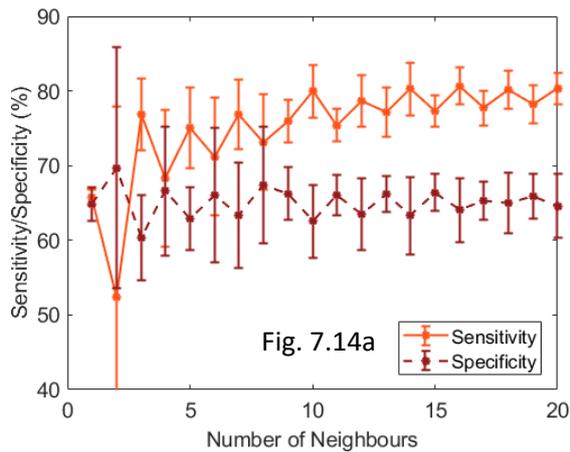
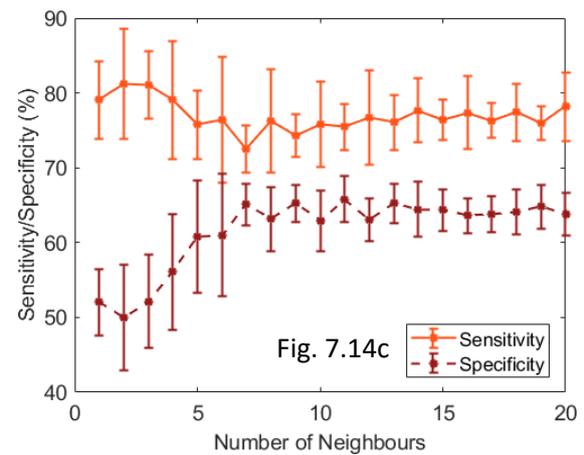


Figure 7.14. Mean sensitivity and specificity plots of nested validation data for k-NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars). a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.



The single feature selection effect can be seen in figure 7.15, where lower average number of features were obtained at k less than 10, beyond which there was less variation in the number of features selected.

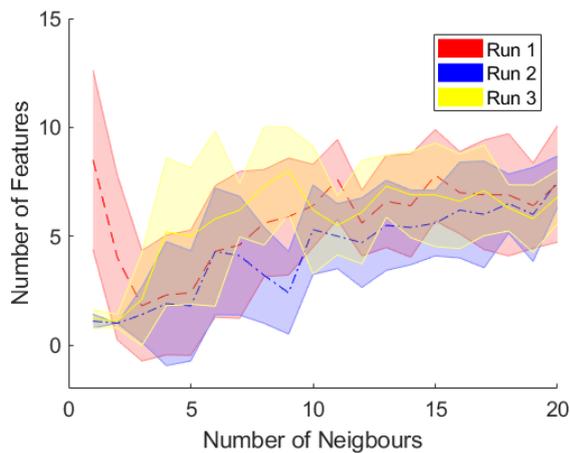


Figure 7.15. Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 8-pixel square). Shaded regions represent standard deviation.

To assess if the pre-filtering process limited the performance of the sequential feature selection, the inner cross-validation runs were repeated with all 130 features in the absence of the AUC-based filter feature selection step. With the application of sequential feature selection alone, similar trends were observed (figures 7.16 and 7.17), where poor model fitting was obtained at k less than 10. Again, acceptable model behaviours were seen at higher numbers of neighbours.

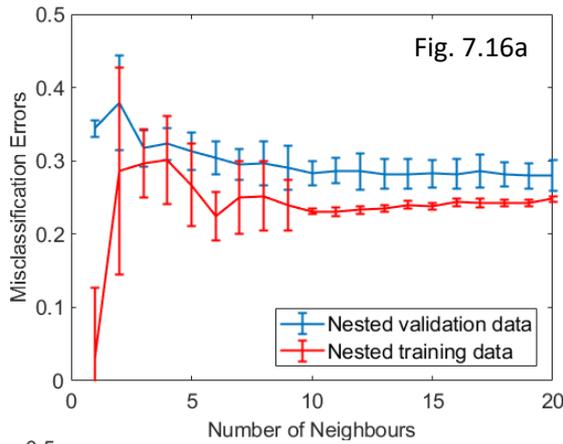


Figure 7.16. Mean misclassification errors for k-NN classification of boundary regions with ROI size of 8-pixel square as a function of neighbourhood size (nested 10-fold cross-validation with standard deviation as error bars), where only sequential feature selection was applied for feature reduction. a) Training run 1; b) Training run 2; c) Training run 3 of outer cross-validation folds.

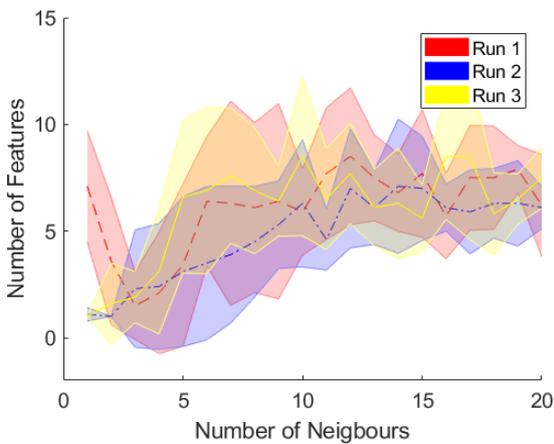
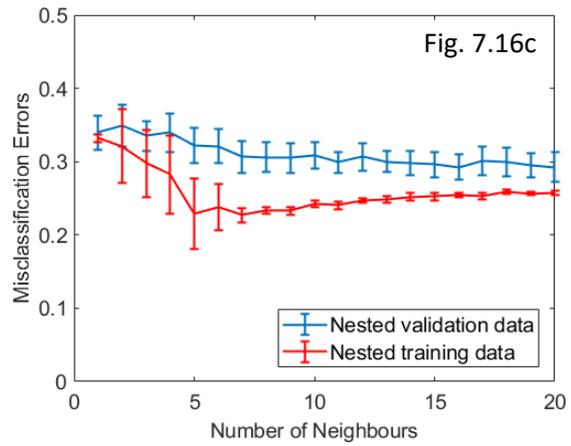
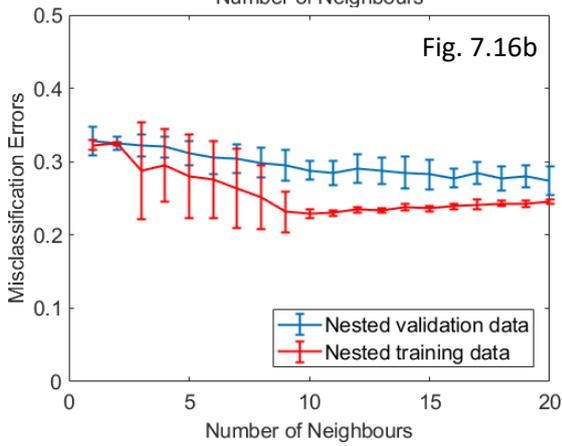


Figure 7.17. Mean number of selected features with varying neighbourhood size for the three training runs of the outer cross-validation folds (Boundary regions with ROI size of 8-pixel square), where only sequential feature selection was applied for feature reduction. Shaded regions represent standard deviation.

Thus, only k values greater than 10 were considered for this group. To maintain methodological consistency, this assessment was based on the former results obtained using both filter and wrapper feature selection processes. Instead of $k = 18$ (misclassification error $27.44 \pm 1.73\%$) for run 1, $k = 20$ (misclassification error $28.10 \pm 2.13\%$) for run 2 and $k = 14$ (misclassification error $28.99 \pm 1.03\%$) for run 3, $k = 19$ were chosen for both runs 1 and 2, whilst $k = 13$ was selected for run 3 (table 7.5).

Training Run	Neighbour size	Misclassification error	Sensitivity	Specificity
1	19	$28.0 \pm 2.2\%$	$78.1 \pm 2.6\%$	$65.9 \pm 3.0\%$
2	19	$28.1 \pm 2.2\%$	$79.3 \pm 2.9\%$	$64.4 \pm 2.9\%$
3	13	$29.4 \pm 1.7\%$	$82.4 \pm 3.3\%$	$73.3 \pm 2.7\%$

Table 7.5. Error rates with optimum parameters on nested validation data (Boundary regions with ROI size of 8-pixel square).

7.8 Estimated performance of classification models with optimised parameters

7.8.1 Estimated classification performance

The average performance of the classification models on the validation runs is shown in table 7.6. Greater classification accuracy was observed for partitioning of tumour versus non-tumour at non-boundary regions, as compared to the tumour boundary where the misclassification errors were almost doubled. For both these groups, lower error rates were associated with the 16-pixel than the 8-pixel ROI size, with greater reduction in specificity than sensitivity.

In the breakdown of the classification accuracy, higher sensitivity than specificity was observed for all four groups, and the associated ROC curves are displayed in figure 7.18.

Region	ROI size	Misclassification error	Sensitivity	Specificity	AUC
Non-boundary	16-pixel	$13.1 \pm 1.0\%$	$89.6 \pm 4.4\%$	$84.2 \pm 4.1\%$	0.923 ± 0.012
	8-pixel	$16.1 \pm 0.7\%$	$88.2 \pm 4.4\%$	$79.6 \pm 4.4\%$	0.896 ± 0.010
Boundary	16-pixel	$23.1 \pm 0.6\%$	$80.5 \pm 2.2\%$	$73.5 \pm 1.4\%$	0.851 ± 0.007
	8-pixel	$29.0 \pm 0.4\%$	$76.5 \pm 3.1\%$	$65.4 \pm 3.6\%$	0.780 ± 0.005

Table 7.6. Estimated classification performance with optimised classifiers parameters.

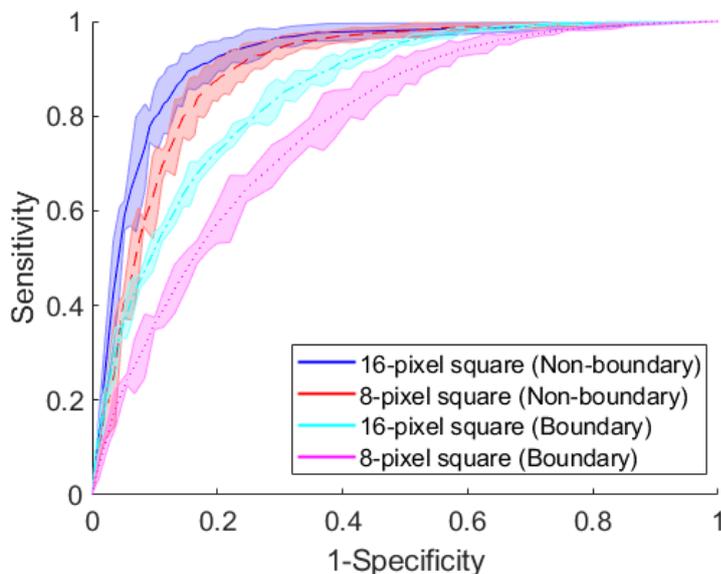


Figure 7.18. Mean ROC curves displaying classification performance for distinguishing between tumour versus non-tumour ROIs at non-boundary and boundary regions, using 16-pixel and 8-pixel ROIs, using outer validation datasets. Shaded regions represent 95% confidence interval.

7.8.2 Feature selection

The number of selected features is shown in figure 7.19, where a maximum of 13 features were chosen across the groups. Evaluation of the feature class revealed that the co-occurrence matrix parameters made up the greatest proportion of the selected features, except for the 8-pixel ROIs at the boundary region where the features were predominantly wavelet-based (figure 7.20). Unlike the patterns seen in chapter 6 with whole tumour/non-tumour ROIs, there was greater contribution from the other feature classes in addition to co-occurrence matrix parameters.

7.8.2.1 Comparison of feature selection between non-boundary and boundary regions

Despite poorer classification results for the boundary ROIs as compared to non-boundary regions, similar numbers of features were obtained for both of these regions when comparing like-for-like ROI sizes. In terms of the constitution of selected features for the 16-pixel size ROIs, similar proportions were obtained for run-length, cooccurrence and wavelet feature classes, with more variation in the contribution from gradient features and autoregressive models (figures 7.20a and 7.20c). This suggests that the poorer classification results may not be due to the lack of features in the selection process, but that even with the optimised selection, the features were poorer at distinguishing between tumour and non-tumour at the boundary compared to non-boundary regions. For the 8-pixel ROIs, although similar feature numbers were obtained, there was more variation in the constitution of the feature classes. Lower classification accuracy was observed the boundary regions despite the inclusion of features from more texture classes.

7.8.2.2 Comparison of feature selection between 16-pixel and 8-pixel sizes

The feature selection process produced more features for the 16-pixel sizes compared to 8-pixel sizes, which was seen for both non-boundary and boundary regions (figure 7.19). Differences were also seen in the constitution of selected features between ROI sizes, even discounting autoregressive features, suggesting that optimisation feature selection process was affected by different ROI sizes.

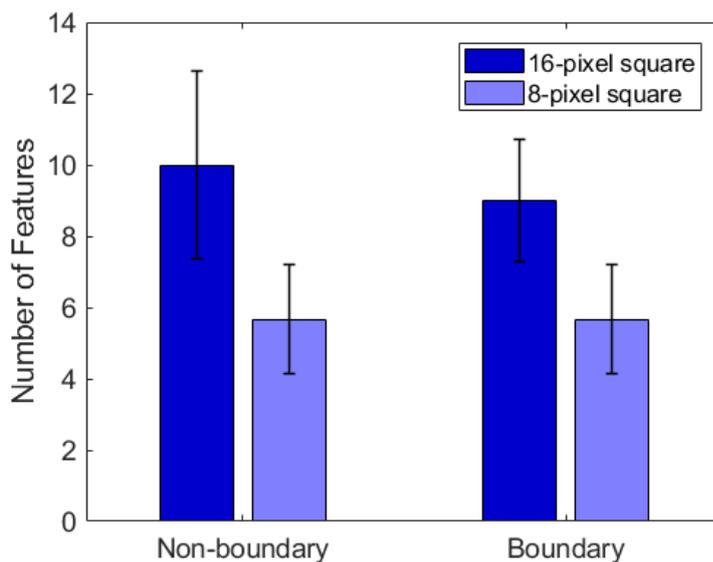


Figure 7.19. Mean number of selected features for optimised classification models (error bars represent standard deviation).

Fig. 7.20a 16-pixel ROI (Non-boundary)

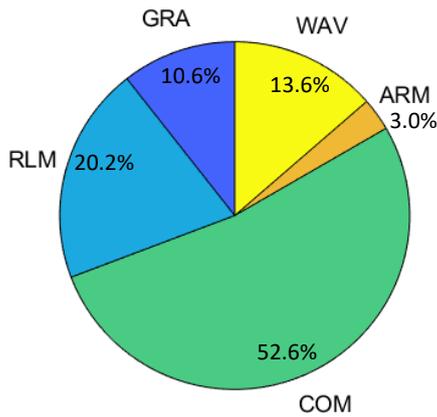


Fig. 7.20b 8-pixel ROI (Non-boundary)

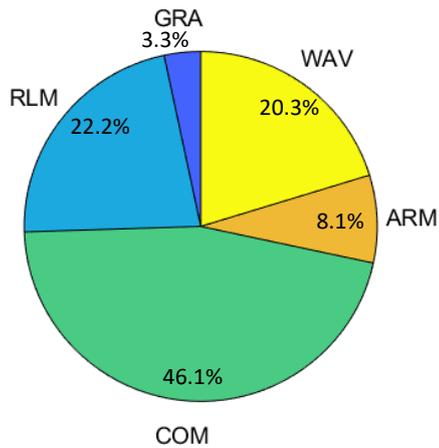
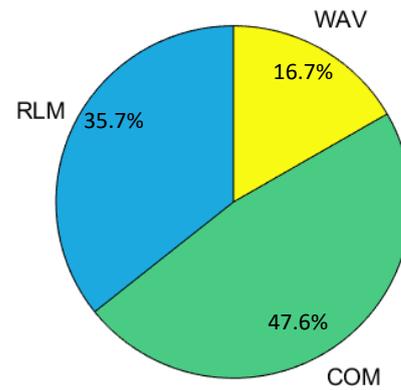


Fig. 7.20c 16-pixel ROI (Boundary)

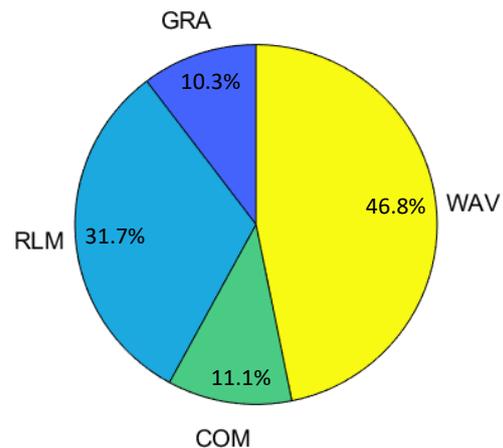


Fig. 7.20d 8-pixel ROI (Boundary)

Figure 7.20. Proportion of selected feature class from optimised classification models across three cross-validation folds.

7.8.2.3 Assessment of individual features

Although there were some features that were common in majority of the runs (table 7.7), there was variation in the other individual features that were selected across the runs for all four groups, which was more apparent for the 8-pixel ROIs than 16-pixel sizes. For 8-pixel ROIs at the boundary region, all three runs produced different feature lists with no overlapping features. In spite of the differences seen at an individual feature level, there was some consistency between the runs at the level of feature class. For example, all the selected co-occurrence matrix features were based on parameters at 4-pixels distance. Interestingly, features based on parameters of the first-order histogram were not selected in any of the groups. Nonetheless, this raises the issue of feature repeatability and stability, especially for the 8-pixel ROIs, which can affect the applicability of the models on new data.

	16-pixel square ROIs			8-pixel square ROIs		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
	Non-boundary regions	45dgr_Fraction 45dgr_LngREmph* GrSkewness* Horzl_GLevNonU* Vertl_Fraction S(0,4)SumOfSqs S(4,4)AngScMom S(4,4)Correlat S(4,4)Entropy S(4,-4)InvDfMom S(4,-4)SumOfSqs** WavEnLL_s-1	GrVariance S(0,4)InvDfMom S(4,-4)DifEntrp S(4,4)SumAverg* S(4,4)SumEntrp S(4,-4)SumOfSqs** WavEnHL_s-1	GrSkewness* Horzl_GLevNonU* Horzl_RLNonUni Vertl_LngREmph* S(0,4)Entropy S(4,-4)Contrast S(4,4)SumAverg* S(4,-4)SumOfSqs** Sigma WavEnLH_s-1 WavEnLL_s-2	45dgr_ShrtREmp* Horzl_LngREmph* Vertl_ShrtREmp S(0,4)DifEntrp S(4,-4)DifEntrp S(4,4)Entropy	S(4,-4)SumAverg S(4,-4)SumEntrp WavEnHH_s-1 WavEnHL_s-1
Boundary regions	45dgr_LngREmph Vertl_LngREmph S(0,4)Contrast* S(4,-4)AngScMom S(4,-4)SumOfSqs S(4,4)SumVarnc* S(4,-4)SumVarnc* WavEnHH_s-2* WavEnLL_s-1 WavEnLL_s-2	45dgr_GLevNonU 45dgr_ShrtREmp Vertl_GLevNonU GrMean S(0,4)Entropy S(4,4)SumAverg S(4,-4)SumEntrp S(4,4)SumVarnc* Teta2 WavEnHL_s-2	Vertl_RLNonUni S(0,4)Contrast* S(4,4)AngScMom S(4,-4)SumVarnc* Teta4 WavEnHH_s-2* WavEnLH_s-2	135dr_Fraction GrNonZeros S(4,4)SumAverg S(4,4)SumEntrp WavEnHH_s-2 WavEnHL_s-1	45dgr_RLNonUni 45dgr_ShrtREmp GrMean WavEnHH_s-2 WavEnHL_s-2 WavEnLH_s-1 WavEnLH_s-2	45dgr_LngREmph Vertl_LngREmph WavEnHL_s-2 WavEnLL_s-2

Table 7.7. List of selected features from optimised classifier models. *Features present in two runs; **Features present in all three runs.

7.9 Re-training of final classification models

Re-training of the data showed similar trends for variances and biases observed in the training dataset across the different groups (figure 7.21) as discussed in section 7.7.2.

For the 8-pixel non-boundary ROI, although the lowest misclassification score was seen at $k = 14$ ($15.5 \pm 0.8\%$), $k = 11$ was selected as it was associated with the next lowest misclassification score. Similarly, $k = 19$ and $k = 15$ were selected rather than $k = 18$ (misclassification error $22.7 \pm 1.0\%$) and $k = 15$ (misclassification error $28.4 \pm 1.5\%$) for the 16- and 8-pixel boundary ROIs respectively.

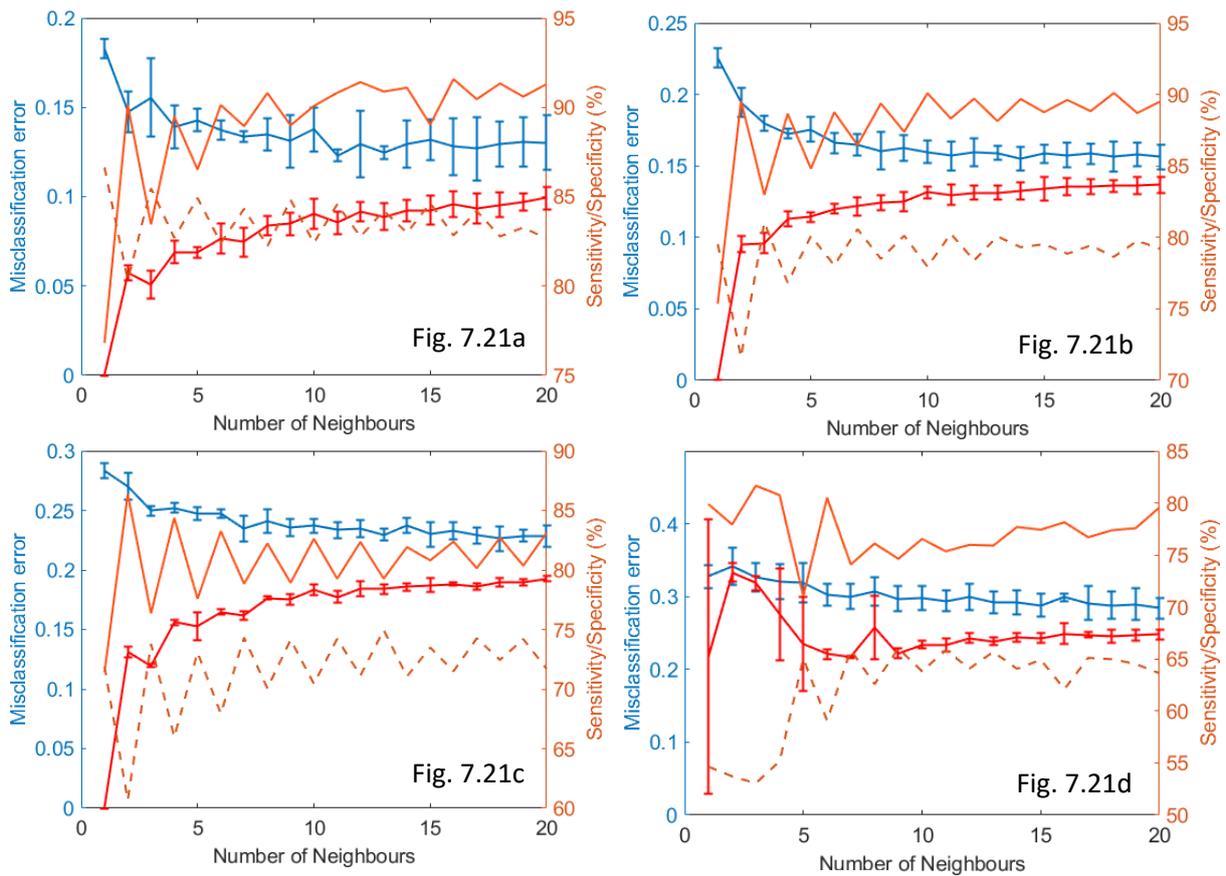
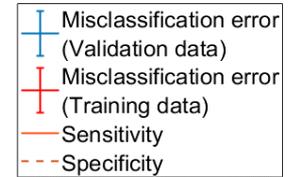


Figure 7.21. Plots for parameter tuning of final k-NN classification models. a) 16-pixel non-boundary ROIs; b) 8-pixel non-boundary ROIs; c) 16-pixel boundary ROIs; d) 8-pixel boundary ROIs.



The selected neighbour size and the corresponding cross-validated error rates are shown in table 7.8. These number of neighbours were used to build the final models, with their error rates displayed in table 7.9.

Region	ROI size	Neighbour size	Misclassification error	Sensitivity	Specificity
Non-boundary	16-pixel	11	$12.3 \pm 0.3\%$	$90.8 \pm 2.5\%$	$84.6 \pm 2.4\%$
	8-pixel	11	$15.7 \pm 1.0\%$	$88.3 \pm 4.2\%$	$80.3 \pm 5.0\%$
Boundary	16-pixel	19	$22.9 \pm 0.1\%$	$80.4 \pm 2.6\%$	$74.2 \pm 1.6\%$
	8-pixel	15	$28.8 \pm 1.5\%$	$77.4 \pm 3.4\%$	$64.9 \pm 4.9\%$

Table 7.8. Error rates with optimum parameters on validation data after re-training.

Region	ROI size	Misclassification error	Sensitivity	Specificity
Non-boundary	16-pixel	8.6 %	95.5 %	87.3 %
	8-pixel	12.9 %	91.7 %	82.6 %
Boundary	16-pixel	18.7 %	86.05 %	77.1 %
	8-pixel	24.7 %	81.6 %	69.1 %

Table 7.9. Error rates of final models based on the total trained data using optimum parameters.

7.10 Independent test set

7.10.1 Classification performance

The results of the application of the independent testing data on the final models is displayed in table 7.10 and figure 7.22, with trends similar to the estimated performance described earlier. There was better accuracy in distinguishing between tumour and non-tumour tissue at non-boundary regions compared to boundary regions, and better classification was observed using 16- than 8-pixel ROIs. This was in spite of achieving slightly better performance than previously estimated with the 8-pixel experiments. As observed earlier, the classification models were associated with higher sensitivity than specificity, indicating that the classification models produced higher false positive rates than false negatives.

Region	ROI size	Misclassification error	Sensitivity	Specificity	AUC
Non-boundary	16-pixel	14.5 %	88.6 %	82.5 %	0.918
	8-pixel	15.5 %	86.5 %	82.5 %	0.905
Boundary	16-pixel	24.7 %	80.2 %	70.8 %	0.841
	8-pixel	26.9 %	83.2 %	63.7 %	0.798

Table 7.10. Performance of final k-nearest neighbour classifier on independent testing data.

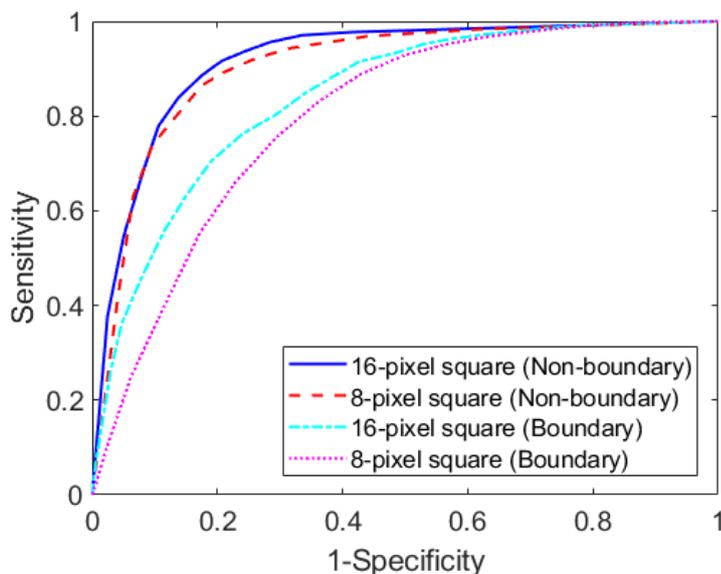


Figure 7.22. ROC curves displaying classification performance for distinguishing between tumour versus non-tumour ROIs at non-boundary and boundary regions, using 16-pixel and 8-pixel ROIs, on independent test data. Shaded regions represent 95% confidence interval.

7.10.2 Feature selection

The number of selected features was similar to that observed in the estimated performance, except for the 16-pixel boundary ROIs, where more features were selected. Like previously, parameters based on first-order histograms were not selected in the process, and co-occurrence matrix parameters predominated the feature list for the 16-pixel ROIs. However, gradient features were not chosen across all the groups. Because of this, the 8-pixel non-boundary ROIs had an equal number of co-occurrence matrix and run-length parameters, whilst the 8-pixel boundary ROIs had the same number of wavelet and run-length parameters.

More features were selected for 16- than 8-pixel ROI sizes. For the non-boundary ROIs, there were two features that were common with the 16- and 8-pixel sizes (long run emphasis at 45° direction and sum average at 4-pixel distance and 135° direction), and twelve remaining different features. All the features between the 16- and 8-pixel ROIs were different. This suggests that variation in ROI sizes can impact on the selection of what is considered to be an important feature set.

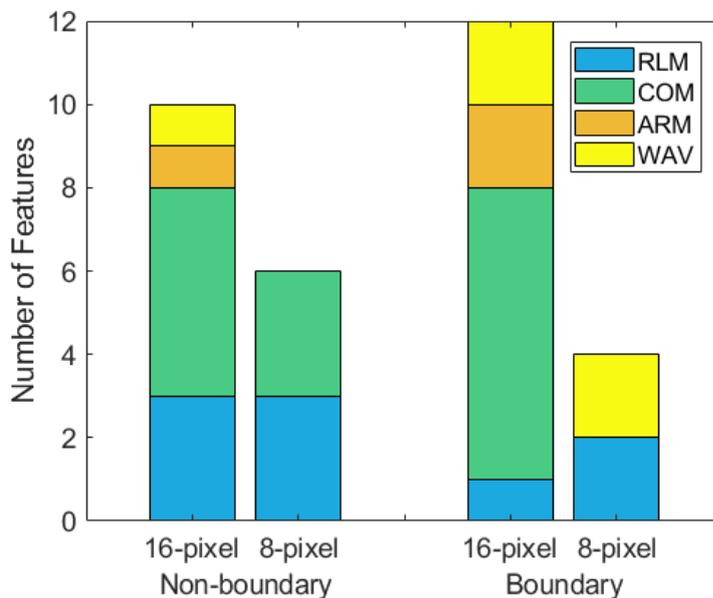


Figure 7.23. Number of selected features for the different texture classes of the optimised final classification models.

Non-boundary regions		Boundary regions	
16-pixel square ROI	8-pixel square ROI	16-pixel square ROI	8-pixel square ROI
135dr_LngREmph	135dr_RLNonUni	S(0,4)DifEntrp	Horzl_Fraction
45dgr_GLevNonU	45dgr_LngREmph	S(0,4)SumOfSqs	Horzl_RLNonUni
45dgr_LngREmph	Horzl_Fraction	S(4,4)DifEntrp	WavEnHH_s-1
S(0,4)DifEntrp	S(4,-4)Contrast	S(4,-4)DifEntrp	WavEnHH_s-2
S(4,4)AngScMom	S(4,4)SumAverg	S(4,-4)Entropy	
S(4,4)Entropy	S(4,-4)SumAverg	S(4,4)SumAverg	
S(4,-4)SumAverg		S(4,4)SumEntrp	
S(4,4)SumEntrp		Sigma	
Teta1		Teta2	
WavEnLL_s-1		Vertl_GLevNonU	
		WavEnHH_s-2	
		WavEnHL_s-2	

Table 7.11. List of selected features from optimised final classifier models.

Discussion

In addition to the observations from chapter 5, the results here show that a combination of texture features can be used to distinguish between tumour and non-tumour tissue effectively. The accuracy of the partitioning is affected by the size of the evaluated region, where excellent classification was achieved when assessing tumour versus surrounding tissue as a whole. The accuracy reduced as the evaluated region decreased in size, although acceptable classification rates were still achieved in the use of either 16-pixel square or 8-pixel square ROIs, with AUC of at least 0.9. When mixed ROI samples at the tumour boundary were introduced, the accuracy dropped as expected, lending itself as a more difficult classification task. Despite the case, acceptable accuracy was achieved for this region at the tumour boundary, with an AUC of 0.84 or just under 0.8 depending on the ROI size used for evaluation (16- and 8-pixel respectively).

In terms of the make-up of the misclassification error, higher sensitivity was achieved than specificity for all the experiments, translating to less tumour regions being labelled as non-tumour than vice versa. This is important in the context of clinical application where high false negatives would result in more regions of the tumour target being missed, which could lead to poorer disease control. Although the accuracy of the classification for normal tissue is also important, poorer tumour accuracy should incur a heavier penalty in the selection of the classification models. As the converse was observed here, there would be value in further pursuing the use of texture features for classification.

One of the other strengths of this work includes the evaluation of a heterogenous mix of cases from multiple centres with different scanning parameters, with no segregation of cases regarding the use of IV contrast. Even in the presence of a diverse mix of cases relatively good accuracy was achieved, demonstrating that this approach can be generalisable to range of cases from different centres. This is crucial in the development of classification models, which should be built based on data representative of the population. Thus, this work supports the applicability of this approach in the general context.

In terms of the methodology, an estimate of the classification performance was achieved through the outer cross-validation folds, which was further validated with an independent test sample. An alternative design would be to use all the data with a nested cross-validation approach, which would provide an estimate of the performance, without the use of an independent test dataset. The latter approach would achieve a good estimation of the classification from the outer cross-validation folds without the need for an independent test, on the assumption that the data is a full representation of the samples in the population. This design has the added advantage of using all the available data for training as well as prediction, which would increase the accuracy of the predictive performance. Although this was considered, the former approach was favoured, as the latter design acts only as a simulation of the test data, and does not take into account sampling variance, which would be provided in an independent test set (437). However, to fully account for sampling variance, it would be best to validate the results on a completely new independent test dataset. Additionally, one could consider increasing the number of the outer cross-validation folds to improve on the estimation of prediction variance. Also, to check for the robustness of the prediction, the cross-validation should be repeated to assess for model instability, where similar predictions would be made in the presence of stable models.

In the absence of a sliding window approach, sample selection bias was reduced with the use of a random number generator to initiate the location of the ROIs. Through the nature of random sampling, for a given region on an image slice, some of the ROIs overlapped with each other, resulting in the assessment of texture features in overlapping areas. The presence of such samples is important in building the classification model, especially if the work is extended to the application of the classification on a pixel-wise basis, i.e. for segmentation.

Despite having achieved promising results with the feature set used here, new features can be assessed for any contribution towards improving the classification rates, especially for the boundary region where the task is more difficult and the bias was seen to be higher. A range of features from different texture classes have been used here, but there is scope to extend this analysis to more parameters. Wavelet-based parameters have been identified as useful features especially for the tumour boundary where segmentation is more difficult. In this work, only the energy of the Harr wavelet was computed, albeit at two scales. The feature list can be expanded by computing other first-order statistics on the wavelet. Moreover, texture features such as the cooccurrence matrix, run-length matrix and gradient can be further computed on the Harr wavelet, as in the approach taken by Aerts et al (430). Examples of other potentially useful parameters include higher-order features, where the statistics is computed based on the first-order features. For instance, neighbourhood grey-tone difference matrices describe properties based on local regions including business, complexity and texture strength (438). Tamura's features are also calculated based on similar principles, which includes contrast and directionality (439). These features are thought to be better at representing the textural cues that are perceived by the human visually, which may lend well to tasks such as segmentation which is based on visual perception. A combination of some of these features were used in the characterisation of PET and CT texture in head and neck (440), as well as in the segmentation of lung cancer (441). As these features describe the texture from a different perspective to those explored in this work, they can potentially provide more information on the image and further complement the feature description, to achieve better partitioning of the groups.

One issue that was not addressed in this study is stability of the selected features. Feature stability is the property where the feature subset is consistently identified, where individual features are true markers and are reliable in their measurement. An ideal feature selection process would reproducibly identify all true markers in repeated tests, which would allow accurate classification across multiple studies. Potential sources of feature instability in this study include uncertainties from feature measurements, as well as the feature selection process that is aimed solely at achieving the best discriminatory performance. To minimise feature instability, the stability rank of all the computed features can be assessed for their reliability within each class through intra-class correlation coefficients (442). Instead of using a feature selection process that constructs a classifier based only on its best predictive accuracy, results from the stability ranking of the features can be incorporated into the process. Ensemble methods where a set of different feature selector is used followed by aggregating the results of the selectors can also be performed to improve feature stability (442).

The efficiency of the workflow needs to be improved for future work. With the procedures that have already been established, this is best performed through the computation of the texture parameters within MATLAB, independent of MaZda.

Potential further areas of work include an evaluation on the spatial context of the classified regions. This would provide some understanding on any anatomical trends in the failures for classification, thereby allowing the performance of texture segmentation to be assessed. Additionally, this knowledge would help in the development of using textural cues in combination with other segmentation techniques.

Following the results in chapter 3 and 4, one potential combination is the application of texture classification following watershed segmentation. As the main limitation seen with the watershed segmentation process is leakage into the mediastinum and chest wall, texture classification can be applied to differentiate between the tumour and these tissue types, to help improve the precision of the segmentation. Additionally, the use of texture classification after the watershed approach would limit the region for evaluation and decrease the computational needs, potentially allowing the exploration of a multi-class classification approach.

Conclusions

Good discrimination of tumour from non-tumour tissue can be achieved based on multiple texture feature sets using 16- and 8-pixel ROI sizes, in the absence of overlapping tissue classes. Texture classification at the tumour boundary is more challenging, with better performance seen with 16-pixel sizes than 8-pixel sizes. Following improvements to the efficiency of the workflow, the integration of spatial information needs to be developed in order to fully establish a working segmentation process.

Chapter 8

Project overview

As target volume delineation contributes to the largest source of uncertainty in the radiotherapy planning and treatment process, contouring assessments are vital to ensure that errors are kept to a minimum. This is particularly important in the setting of clinical trials where differences in the treatment volume can affect trial outcome and lead to erroneous results. Currently, the outlining assessment process is performed manually, which is inefficient and subjective.

This project focuses on the evaluation of different image processing techniques for tumour segmentation, with an aim to produce computer-generated contours that can be used to automate the assessment process. The evaluation was performed in the setting of advanced lung tumours as this is a body site that is being actively investigated in clinical trials. Although there are numerous reports of a wide variety of segmentation techniques used on peripheral lung tumours, their application in advanced lung cancer is limited. Moreover, there are no current available commercial software products that provide solutions for tumour delineation. Thus, the most appropriate method(s) for developing an automated tumour segmentation process was explored in this work.

A heterogeneous dataset comprising of 79 cases (total of 1997 image slices) from a range of centres was used to train and test the segmentation techniques. Moreover, because of the clinical diversity of the cases, the assessment was performed across a variety of tumour locations, reflective of the range of presentations seen typically in day-to-day practice. An efficient workflow within MATLAB was developed to allow the different segmentation techniques to be analysed. The upper and lower HU values were analysed from a subsample of cases to determine appropriate values in the adoption of a thresholding approach, with further adaptation based on the presence of solid or non-solid lesions. Despite the usefulness of the thresholding approach, the selection of the HU values was carried out with a generous margin to minimise the exclusion of false negative regions.

Four image processing techniques were evaluated; marker-controlled watershed segmentation, Chan-Vese active contour approach, edge-based active contour approach, and graph-cut technique with superpixel generation. To initiate the segmentation process, the submitted clinician contours were used as priors as a novel technique, which were used to define the segmentation bounding region. For the watershed and graph-cut approaches, the priors were also used to provide shape information in the form of internal markers. However, to minimise overfitting issues with the active contour approaches, instead of shape information, size information was derived from the priors for initialisation purposes. Parameter selection was conducted systematically using a cross-validation approach with the training dataset for the active contour and graph-cut techniques, with the assessment of the overall performance on the testing dataset. It was found that in the setting of advanced lung tumours, the edge-based active contour approach had the best quantitative performance. Conversely, marker-controlled watershed segmentation achieved the highest DSC with the fastest computational time for isolated lung tumours, indicating that the segmentation performance is affected by the tumour location. However, qualitative analysis revealed inaccuracies with tumour coverage, inclusion

of GGOs and exclusion of normal tissue structures, which supports the need to further improve the segmentation workflow in order for such an application to be applied successfully in the clinic. To this end, a semi-automatic approach with additional user-defined exclusion markers was also evaluated for the watershed method.

Furthermore, in addition to the first order histogram features, the contribution of texture features was also explored as an alternative approach to image segmentation. This comprised of statistical, autoregressive and wavelet-derived parameters. A k-NN classifier was trained using a cross-validation approach to build the models for partitioning ROIs into tumour and non-tumour classes, where the classification performance was assessed using an independent test dataset. In a preliminary study, the discrimination of whole tumour and the surrounding tissues was explored, where excellent partitioning was achieved with the use of a multiple texture feature set over a single discriminatory texture feature and the mean value for the ROIs. The filter feature selection technique was applied in this work, which revealed issues with feature redundancy. Following these results, both filter and sequential feature selection methods were used to extract the multiple texture feature set in the classification of smaller ROIs, where good discrimination was observed in the partitioning of tumour from non-tumour region, although the classification at the tumour boundary was associated with higher error rates. One limitation with this approach is the issue of feature instability, which may limit the applicability of the model on new samples.

Overall conclusions

Automatic workflows for the segmentation of advanced lung tumours from a heterogenous dataset were implemented using clinician submitted contours as a given prior. Four state-of-the-art conventional image segmentation techniques were assessed, where overall, the edge-based active contour approach achieved better performance based on quantitative and qualitative measures as well as computational time. Despite the promising results, as with the graph-cut, Chan-Vese active contour and marker-controlled watershed techniques that were explored, errors in tumour coverage and leakages were observed. The behaviour of the segmentation and patterns of failure varied according to the methods used as well as site of disease. In addition to other techniques for improving the segmentation performance, reproducibility studies need to be carried out to assess the reliability of the segmentations.

Multiple texture feature sets were found to be more successful at discriminating whole tumour from non-tumour regions than a single discriminatory texture feature and the mean value. In the work towards the goal of developing a segmentation process based on multiple feature sets, good discrimination was achieved based on non-overlapping tumour and non-tumour tissue with using smaller 16- and 8-pixel ROI sizes. The classification task was more challenging at the tumour boundary which resulted in comparatively lower accuracy. Further improvement is needed to increase the efficiency of the workflows, in order to develop a segmentation process whereby the texture classification is performed in the context of spatial information.

Proposed future developments

Based on the observations and thoughts from this work, the major limitation to implementing these workflows in the clinical setting is in the accuracy and precision of the derived

segmentation. There are four main areas of future developments that can potentially improve the segmentation performance.

Firstly, there is a need to extend the application of these techniques from the 2D setting to 3D, which has the potential improve their performance in the presence of additional contextual information from the Z-plane. One other benefit to 3D analysis is the ability to estimate the tumour limits at the superior and inferior border, which is not feasible in this work. This is especially important for contouring assessment, as it is known that large delineation errors are seen in the cranial caudal plane (388). For 3D applications higher out of plane resolution would be necessary, which would entail the interpolation of anisotropic voxels into isotropic units with uniform spatial resolution to allow algorithms to work well. Additionally, with the move towards target delineation on 4D CT, the robustness of the segmentation techniques should also be tested in this setting, notwithstanding the potential improvement in performance of the segmentation with information from multiple phases (249).

Secondly, having evaluated the behaviour of the different techniques on CT imaging, at this juncture it would be useful to assess if PET imaging could complement the processes that have been developed. For example, regions of low uptake on the PET images can potentially be used as exclusion structures to improve the watershed performance at the mediastinum. Similarly, these regions can also be applied as additional external markers for the graph-cut method. For the active contour approaches, an initial segmentation on the PET appearances could be used as the initialisation boundary to be applied on the CT images, which would provide additional shape information that would be more in keeping to that of the tumour. These proposals require an approximation of the (non-)tumour component from the PET imaging rather than highly accurate and precise delineations which have been shown to be challenging to obtain, due to the heterogeneity of tracer uptake and blurring of edge boundaries that is typically seen on PET images. Moreover, there would be less impact from the uncertainties inherent with registration in multi-modality approaches, which would be preferred to attempting the segmentation of the target based principally on the PET images. These approaches use PET information to increase the prior knowledge of the segmentation, and has been demonstrated to be useful to define the location of the tumour and to limit the segmentation from encroaching into normal tissues (443).

On a similar note, the third area which would need further investigation is the application of a combination of different segmentation techniques. With the knowledge of the performance associated with each method, it is unlikely that a single automatic segmentation approach would provide the solution to fulfil the delineation task. Processes involving the sequential combination of techniques have been explored, although these have been applied mainly to provide solutions for specific problems, such as vasculature removal or GGO inclusion. As suggested previously, texture-based discrimination could be performed following watershed segmentation, which can then be refined through an active contour application. A different means of pursuit is akin to having multiple experts generate one set of manual gold-standard outlines, where the results of the independent solution to the segmentation task provided by each automatic segmentation approach is converged. This can be performed through an ensemble approach, where a voting rule can be used to select voxels in which majority agreement exists. This has been adopted by Gu et al (221) and Velazquez et al (216) where multiple contours were grown from multiple seed points and the voting scheme was applied on a voxel-wise basis. Another method is through the simultaneous truth and performance level

estimation (STAPLE) algorithm which computes probabilistic estimates for the true segmentations, from which a consensus can be picked based on the accepted probability (444). Huo et al used a similar approach for brain tumour segmentation where the probabilistic maps from each of the different soft segmentation approaches were averaged to compute the final delineation (445). The premise to such approaches is in the acceptance of variations of the estimates provided by each of the segmentation processes, where errors from each process would get diluted when all the results from the different techniques are combined.

Lastly, from the work in texture analysis, it was evident that there were numerous texture features that could potentially be calculated, each of which could contribute in part to the segmentation. In addition to the features that have been discussed, there can be further higher order statistics that can be calculated, across different ROI sizes. With the sheer number of possible permutations, this lends itself to deep learning approaches which would be better suited to solve complex problems. There is great interest in the advances made in the application of deep learning to medical image analysis, which hold promise as a game changer in many fields. There are growing number of studies in its application in medical image segmentation (202), with recent reports in the literature on its application in delineation of OARs (306, 385, 446-453) as well as the segmentation of GTVs and CTVs (198-200, 203, 204, 446, 454-457). Interestingly, some studies have used other image processing techniques either before or after application of the deep learning processes with an aim of improving the performance of the algorithm (385, 451, 452, 455), as well as for contour refinement. Despite the comparatively fewer number of tumour segmentation studies, deep learning methods show great promise and warrants investigation in this setting. Further research into the application of different deep learning architecture needs to be carried out to ascertain if conventional image processing techniques would be required to complement the performance of deep learning techniques for our purposes.

References

1. Cancer Research UK. Achieving a world-class radiotherapy service across the UK. July 2009(Access: http://www.cancerresearchuk.org/cancer-info/prod_consump/groups/cr_common/@nre/@pol/documents/generalcontent/crukmg_1000ast-3360.pdf).
2. Bhide SA, Nutting CM. Recent advances in radiotherapy. *BMC medicine*. 2010;8:25.
3. Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *The Lancet Oncology*. 2011;12(2):127-36.
4. Kam MK, Chau RM, Suen J, Choi PH, Teo PM. Intensity-modulated radiotherapy in nasopharyngeal carcinoma: dosimetric advantage over conventional plans and feasibility of dose escalation. *International journal of radiation oncology, biology, physics*. 2003;56(1):145-57.
5. Nutting CM, Corbishley CM, Sanchez-Nieto B, Cosgrove VP, Webb S, Dearnaley DP. Potential improvements in the therapeutic ratio of prostate cancer irradiation: dose escalation of pathologically identified tumour nodules using intensity modulated radiotherapy. *The British journal of radiology*. 2002;75(890):151-61.
6. Kuban DA, Tucker SL, Dong L, Starkschall G, Huang EH, Cheung MR, et al. Long-term results of the MD Anderson randomized dose-escalation trial for prostate cancer. *International Journal of Radiation Oncology* Biology* Physics*. 2008;70(1):67-74.
7. Wolden SL, Chen WC, Pfister DG, Kraus DH, Berry SL, Zelefsky MJ. Intensity-modulated radiation therapy (IMRT) for nasopharynx cancer: update of the Memorial Sloan-Kettering experience. *International Journal of Radiation Oncology* Biology* Physics*. 2006;64(1):57-62.
8. Kwong DL, Sham JS, Leung LH, Cheng AC, Ng W, Kwong PW, et al. Preliminary results of radiation dose escalation for locally advanced nasopharyngeal carcinoma. *International Journal of Radiation Oncology* Biology* Physics*. 2006;64(2):374-81.
9. ICRU report 50. International commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy. 1993.
10. ICRU report 62. International commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy. Supplement to ICRU report 5. 1999.
11. Van Herk M, editor *Errors and margins in radiotherapy*. Seminars in radiation oncology; 2004: Elsevier.
12. Njeh C. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics/Association of Medical Physicists of India*. 2008;33(4):136.
13. Liang J, Li M, Zhang T, Han W, Chen D, Hui Z, et al. The effect of image-guided radiation therapy on the margin between the clinical target volume and planning target volume in lung cancer. *Journal of medical radiation sciences*. 2014;61(1):30-7.
14. Giraud P, Elles S, Helfre S, De Rycke Y, Servois V, Carette M-F, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. *Radiotherapy and oncology*. 2002;62(1):27-36.
15. Rexilius J, Hahn HK, Schlüter M, Bourquain H, Peitgen H-O. Evaluation of accuracy in MS lesion volumetry using realistic lesion phantoms. *Academic radiology*. 2005;12(1):17-24. e1.

16. Tai P, Van Dyk J, Yu E, Battista J, Stitt L, Coad T. Variability of target volume delineation in cervical esophageal cancer. *International Journal of Radiation Oncology* Biology* Physics*. 1998;42(2):277-88.
17. Cooper JS, Mukherji SK, Toledano AY, Beldon C, Schmalfluss IM, Amdur R, et al. An evaluation of the variability of tumor-shape definition derived by experienced observers from CT images of supraglottic carcinomas (ACRIN protocol 6658). *International Journal of Radiation Oncology* Biology* Physics*. 2007;67(4):972-5.
18. Weiss E, Hess CF. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy. *Strahlentherapie und Onkologie*. 2003;179(1):21-30.
19. Lo AC, Liu M, Chan E, Lund C, Truong PT, Loewen S, et al. The Impact of Peer Review of Volume Delineation in Stereotactic Body Radiation Therapy Planning for Primary Lung Cancer: A Multicenter Quality Assurance Study. *Journal of Thoracic Oncology*. 2014;9(4):527-33.
20. Zelefsky MJ, Kollmeier M, Cox B, Fidaleo A, Sperling D, Pei X, et al. Improved clinical outcomes with high-dose image guided radiotherapy compared with non-IGRT for the treatment of clinically localized prostate cancer. *International Journal of Radiation Oncology* Biology* Physics*. 2012;84(1):125-9.
21. Grills IS, Hugo G, Kestin LL, Galerani AP, Chao KK, Wloch J, et al. Image-guided radiotherapy via daily online cone-beam CT substantially reduces margin requirements for stereotactic lung radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*. 2008;70(4):1045-56.
22. del Campo ER, Rivera S, Martínez-Paredes M, Hupé P, Escarpa AS, Borget I, et al. Assessment of the novel online delineation workshop dummy run approach using FALCON within a European multicentre trial in cervical cancer (RAIDs). *Radiotherapy and Oncology*. 2017;124(1):130-8.
23. Eriksen JG, Salembier C, Rivera S, De Bari B, Berger D, Mantello G, et al. Four years with FALCON—an ESTRO educational project: achievements and perspectives. *Radiotherapy and Oncology*. 2014;112(1):145-9.
24. Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Sola AB, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and Oncology*. 2015;114(1):3-10.
25. Goodman KA, Regine WF, Dawson LA, Ben-Josef E, Haustermans K, Bosch WR, et al. Radiation Therapy Oncology Group consensus panel guidelines for the delineation of the clinical target volume in the postoperative treatment of pancreatic head cancer. *International Journal of Radiation Oncology* Biology* Physics*. 2012;83(3):901-8.
26. Wang D, Bosch W, Roberge D, Finkelstein SE, Petersen I, Haddock M, et al. RTOG sarcoma radiation oncologists reach consensus on gross tumor volume and clinical target volume on computed tomographic images for preoperative radiotherapy of primary soft tissue sarcoma of extremity in Radiation Therapy Oncology Group studies. *International Journal of Radiation Oncology* Biology* Physics*. 2011;81(4):e525-e8.
27. Grégoire V, Ang K, Budach W, Grau C, Hamoir M, Langendijk JA, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiotherapy and oncology*. 2014;110(1):172-81.
28. Taylor A, Rockall A, Powell M. An atlas of the pelvic lymph node regions to aid radiotherapy target volume definition. *Clinical Oncology*. 2007;19(7):542-50.

29. Valentini V, Gambacorta MA, Barbaro B, Chiloiro G, Coco C, Das P, et al. International consensus guidelines on Clinical Target Volume delineation in rectal cancer. *Radiotherapy and Oncology*. 2016;120(2):195-201.
30. Ritter T, Quint DJ, Senan S, Gaspar LE, Komaki RU, Hurkmans CW, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *International Journal of Radiation Oncology• Biology• Physics*. 2011;81(5):1442-57.
31. Hall WH, Guiou M, Lee NY, Dublin A, Narayan S, Vijayakumar S, et al. Development and validation of a standardized method for contouring the brachial plexus: preliminary dosimetric analysis among patients treated with IMRT for head-and-neck cancer. *International Journal of Radiation Oncology• Biology• Physics*. 2008;72(5):1362-7.
32. Scoccianti S, Detti B, Gadda D, Greto D, Furfaro I, Meacci F, et al. Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice. *Radiotherapy and Oncology*. 2015;114(2):230-8.
33. Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiotherapy and Oncology*. 2015;117(1):83-90.
34. Gay HA, Barthold HJ, O'Meara E, Bosch WR, El Naqa I, Al-Lozi R, et al. Pelvic normal tissue contouring guidelines for radiation therapy: a Radiation Therapy Oncology Group consensus panel atlas. *International Journal of Radiation Oncology• Biology• Physics*. 2012;83(3):e353-e62.
35. Jabbour SK, Hashem SA, Bosch W, Kim TK, Finkelstein SE, Anderson BM, et al. Upper abdominal normal organ contouring guidelines and atlas: a Radiation Therapy Oncology Group consensus. *Practical radiation oncology*. 2014;4(2):82-9.
36. Duane F, Aznar MC, Bartlett F, Cutter DJ, Darby SC, Jagsi R, et al. A cardiac contouring atlas for radiotherapy. *Radiotherapy and Oncology*. 2017;122(3):416-22.
37. Konert T, Vogel W, MacManus MP, Nestle U, Belderbos J, Grégoire V, et al. PET/CT imaging for target volume delineation in curative intent radiotherapy of non-small cell lung cancer: IAEA consensus report 2014. *Radiotherapy and Oncology*. 2015;116(1):27-34.
38. Khoo V, Joon D. New developments in MRI for target volume delineation in radiotherapy. *The British journal of radiology*. 2006;79(special_issue_1):S2-S15.
39. Caldwell CB, Mah K, Ung YC, Danjoux CE, Balogh JM, Ganguli SN, et al. Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion. *International Journal of Radiation Oncology• Biology• Physics*. 2001;51(4):923-31.
40. Senan S, De Koste JVS, Samson M, Tankink H, Jansen P, Nowak PJ, et al. Evaluation of a target contouring protocol for 3D conformal radiotherapy in non-small cell lung cancer. *Radiotherapy and Oncology*. 1999;53(3):247-55.
41. Vorwerk H, Beckmann G, Bremer M, Degen M, Dietl B, Fietkau R, et al. The delineation of target volumes for radiotherapy of lung cancer patients. *Radiotherapy and Oncology*. 2009;91(3):455-60.
42. Rouette J, Gutierrez E, O'Donnell J, Reddeman L, Hart M, Foxcroft S, et al. Directly Improving the Quality of Radiation Treatment Through Peer Review: A Cross-sectional Analysis of Cancer Centers Across a Provincial Cancer Program. *International Journal of Radiation Oncology• Biology• Physics*. 2017;98(3):521-9.

43. Rooney K, Hanna G, Harney J, Eakin R, Young VL, Dunn M, et al. The Impact of Colleague Peer-review on the Radiotherapy Treatment Planning Process in the Radical Treatment of Lung Cancer. *Clinical Oncology*. 2014;26:S3.
44. Ballo MT, Chronowski GM, Schlembach PJ, Bloom ES, Arzu IY, Kuban DA. Prospective peer review quality assurance for outpatient radiation therapy. *Practical radiation oncology*. 2014;4(5):279-84.
45. Brammer C, Pettit L, Allerton R, Churn M, Joseph M, Koh P, et al. Impact of the introduction of weekly radiotherapy quality assurance meetings at one UK cancer centre. *The British journal of radiology*. 2014;87(1043):20140422.
46. Doll C, Duncker-Rohr V, Rücker G, Mix M, MacManus M, De Ruyscher D, et al. Influence of experience and qualification on PET-based target volume delineation. *Strahlentherapie und Onkologie*. 2014;190(6):555-62.
47. Roques TB, N; Bloomfield, D; Burkill, G; Gaze, M; Gwynne, S; Hanna, G; Illsley, M; Jankowska, P; Mackenzie, J; McAleese, J; Sanghera, P; Simcock, R. Radiotherapy target Volume Definition and Peer Review - RCR guidance. *BFCO*. 2017;17(2).
48. Khalil AA, Bentzen SM, Bernier J, Saunders MI, Horiot J-C, Van Den Bogaert W, et al. Compliance to the prescribed dose and overall treatment time in five randomized clinical trials of altered fractionation in radiotherapy for head-and-neck carcinomas. *International Journal of Radiation Oncology* Biology* Physics*. 2003;55(3):568-75.
49. Peters LJ, O'Sullivan B, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *Journal of clinical oncology*. 2010;28(18):2996-3001.
50. Eisbruch A, Harris J, Garden AS, Chao CK, Straube W, Harari PM, et al. Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22). *International Journal of Radiation Oncology* Biology* Physics*. 2010;76(5):1333-8.
51. Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustig R, et al. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704—a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. *International Journal of Radiation Oncology* Biology* Physics*. 2012;82(2):809-16.
52. Crane CH, Winter K, Regine WF, Safran H, Rich TA, Curran W, et al. Phase II study of bevacizumab with concurrent capecitabine and radiation followed by maintenance gemcitabine and bevacizumab for locally advanced pancreatic cancer: Radiation Therapy Oncology Group RTOG 0411. *Journal of Clinical Oncology*. 2009;27(25):4096-102.
53. Dühmke E, Franklin J, Pfreundschuh M, Sehlen S, Willich N, Rühl U, et al. Low-dose radiation is sufficient for the noninvolved extended-field treatment in favorable early-stage Hodgkin's disease: long-term results of a randomized trial of radiotherapy alone. *Journal of clinical oncology*. 2001;19(11):2905-14.
54. Ibbott GS, Followill DS, Molineu HA, Lowenstein JR, Alvarez PE, Roll JE. Challenges in credentialing institutions and participants in advanced technology multi-institutional clinical trials. *International Journal of Radiation Oncology* Biology* Physics*. 2008;71(1):S71-S5.
55. Cancer Registration Statistics, England. Access: <https://www.gov.uk/people-population-and-community/health-and-social-care/conditions-and-diseases/bulletins/cancer-registration-statistics-england/2015>. 2015.

56. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*. 2015;136(5).
57. National Cancer Registration & Analysis Service and Cancer Research UK: "Chemotherapy, Radiotherapy and Tumour Resections in England: 2013-2014" workbook. London: National Cancer Registration and Analysis Service. 2017.
58. Current Controlled Trials. London: BioMed Central. ISRCTN47674500. ADSCaN: A Randomised Phase II trial of Accelerated, Dose escalated, Sequential Chemo-radiotherapy in Non small cell lung cancer. Access: <https://doi.org/10.1186/ISRCTN47674500>. 2016.
59. Access: <https://radiologypics.com/2013/03/20/ct-of-the-chest-lung-windows-axial-anatomy/>.
60. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology*. 2016;11(1):39-51.
61. Kim SS, Seo JB, Lee HY, Nevrekar DV, Forssen AV, Crapo JD, et al. Chronic obstructive pulmonary disease: lobe-based visual assessment of volumetric CT by Using standard images—comparison with quantitative CT and pulmonary function test in the COPD Gene study. *Radiology*. 2013;266(2):626-35.
62. Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, Engelke C. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. *European radiology*. 2006;16(4):781-90.
63. El-Baz A, Beache GM, Gimel'farb G, Suzuki K, Okada K, Elnakib A, et al. Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *International journal of biomedical imaging*. 2013;2013.
64. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical physics*. 2014;41(5).
65. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*. 2012;48(4):441-6.
66. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*. 2012;30(9):1234-48.
67. Ng H, Ong S, Foong K, Goh P, Nowinski W, editors. Medical image segmentation using k-means clustering and improved watershed algorithm. *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on; 2006: IEEE*.
68. Solomon C, Breckon T. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*: John Wiley & Sons; 2011.
69. Otsu N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*. 1979;9(1):62-6.
70. Kapur JN, Sahoo PK, Wong AK. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer vision, graphics, and image processing*. 1985;29(3):273-85.
71. Ramesh N, Yoo J-H, Sethi I. Thresholding based on histogram approximation. *IEE Proceedings-Vision, Image and Signal Processing*. 1995;142(5):271-9.
72. Sahoo PK, Soltani S, Wong AK. A survey of thresholding techniques. *Computer vision, graphics, and image processing*. 1988;41(2):233-60.

73. Liu D, Yu J, editors. Otsu method and K-means. Hybrid Intelligent Systems, 2009 HIS'09 Ninth International Conference on; 2009: IEEE.
74. Liao P-S, Chen T-S, Chung P-C. A fast algorithm for multilevel thresholding. *J Inf Sci Eng.* 2001;17(5):713-27.
75. Yen J-C, Chang F-J, Chang S. A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing.* 1995;4(3):370-8.
76. Sezgin M, Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging.* 2004;13(1):146-66.
77. Bradley D, Roth G. Adaptive thresholding using the integral image. *Journal of Graphics Tools.* 2007;12(2):13-21.
78. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. *Annual review of biomedical engineering.* 2000;2(1):315-37.
79. Prewitt JM. Object enhancement and extraction. *Picture processing and Psychopictorics.* 1970;10(1):15-9.
80. Sobel I. An isotropic 3×3 image gradient operator. *Machine vision for three-dimensional scenes.* 1990:376-9.
81. Roberts LG. *Machine perception of three-dimensional solids: Massachusetts Institute of Technology;* 1963.
82. Canny J. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence.* 1986(6):679-98.
83. Marr D, Hildreth E. Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences.* 1980;207(1167):187-217.
84. Hsieh J. *Computed tomography: principles, design, artifacts, and recent advances: SPIE press;* 2003.
85. Maini R, Aggarwal H. Study and comparison of various image edge detection techniques. *International journal of image processing (IJIP).* 2009;3(1):1-11.
86. Bhardwaj S, Mittal A. A survey on various edge detector techniques. *Procedia Technology.* 2012;4:220-6.
87. Senthilkumaran N, Rajesh R. Edge detection techniques for image segmentation—a survey of soft computing approaches. *International journal of recent trends in engineering.* 2009;1(2):250-4.
88. Davis LS. A survey of edge detection techniques. *Computer graphics and image processing.* 1975;4(3):248-70.
89. Sharma N, Aggarwal LM. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India.* 2010;35(1):3.
90. Bankman I. *Handbook of medical image processing and analysis: academic press;* 2008.
91. Xu N, Ahuja N, Bansal R, editors. Automated lung nodule segmentation using dynamic programming and EM-based classification. *Medical Imaging 2002: Image Processing;* 2002: International Society for Optics and Photonics.
92. Liu F, Zhao B, Kijewski PK, Wang L, Schwartz LH. Liver segmentation for CT images using GVF snake. *Medical Physics.* 2005;32(12):3699-706.
93. Hsu C-Y, Liu C-Y, Chen C-M. Automatic segmentation of liver PET images. *Computerized Medical Imaging and Graphics.* 2008;32(7):601-10.
94. Adams R, Bischof L. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence.* 1994;16(6):641-7.

95. Hojjatoleslami S, Kittler J. Region growing: a new approach. *IEEE Transactions on Image processing*. 1998;7(7):1079-84.
96. Horowitz SL, Pavlidis T. Picture segmentation by a tree traversal algorithm. *Journal of the ACM (JACM)*. 1976;23(2):368-88.
97. Pan Z, Lu J. A Bayes-based region-growing algorithm for medical image segmentation. *Computing in Science and Engineering*. 2007;9(4):32-8.
98. Pohle R, Toennies KD, editors. Segmentation of medical images using adaptive region growing. *Medical Imaging 2001: Image Processing*; 2001: International Society for Optics and Photonics.
99. Senthilkumar B, Umamaheswari G, Karthik J, editors. A novel region growing segmentation algorithm for the detection of breast cancer. *Computational Intelligence and Computing Research (ICCIC)*, 2010 IEEE International Conference on; 2010: IEEE.
100. Mendez AJ, Tahoces PG, Lado MaJ, Souto M, Vidal JJ. Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. *Medical Physics*. 1998;25(6):957-64.
101. Lin D-T, Lei C-C, Hung S-W. Computer-aided kidney segmentation on abdominal CT images. *IEEE transactions on information technology in biomedicine*. 2006;10(1):59-65.
102. Yan G, Wang B, editors. An automatic kidney segmentation from abdominal CT images. *Intelligent Computing and Intelligent Systems (ICIS)*, 2010 IEEE International Conference on; 2010: IEEE.
103. Kumar S, Moni R, Rajeesh J. Automatic liver and lesion segmentation: a primary step in diagnosis of liver diseases. *Signal, Image and Video Processing*. 2013;7(1):163-72.
104. Keil S, Behrendt FF, Stanzel S, Sühling M, Koch A, Bubenzer J, et al. Semi-automated measurement of hyperdense, hypodense and heterogeneous hepatic metastasis on standard MDCT slices. Comparison of semi-automated and manual measurement of RECIST and WHO criteria. *European radiology*. 2008;18(11):2456-65.
105. Rusko L, Bekes G, Nemeth G, Fidrich M. Fully automatic liver segmentation for contrast-enhanced CT images. *MICCAI Wshp 3D Segmentation in the Clinic: A Grand Challenge*. 2007;2(7).
106. Dinkel J, Khalilzadeh O, Hintze C, Fabel M, Puderbach M, Eichinger M, et al. Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. *Lung Cancer*. 2013;82(1):76-82.
107. Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. *IEEE transactions on medical imaging*. 2008;27(4):467-80.
108. Bornemann L, Dicken V, Kuhnigk J-M, Wormanns D, Shin H-O, Bauknecht H-C, et al. OncoTREAT: a software assistant for cancer therapy monitoring. *International Journal of Computer Assisted Radiology and Surgery*. 2007;1(5):231-42.
109. Beucher S, C L. Use of watersheds in contour detection. *Proc Workshop on Image Processing, CCETT/IRISA*,. 1979:2.1-2.12.
110. Shrimpton P, Hillier M, Lewis M, Dunn M. Doses from computed tomography (CT) examinations in the UK-2003 review: Public Health England, Centre for Radiation, Chemical and Environmental Hazards; 2005.
111. Sethian JA. *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*: Cambridge university press; 1999.

112. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. *International journal of computer vision*. 1988;1(4):321-31.
113. Cohen LD. On active contour models and balloons. *CVGIP: Image understanding*. 1991;53(2):211-8.
114. Xu C, Prince JL. Snakes, shapes, and gradient vector flow. *IEEE Transactions on image processing*. 1998;7(3):359-69.
115. McInerney T, Terzopoulos D. A dynamic finite element surface model for segmentation and tracking in multidimensional medical images with application to cardiac 4D image analysis. *Computerized Medical Imaging and Graphics*. 1995;19(1):69-83.
116. Cootes T. An introduction to active shape models. *Image processing and analysis*. 2000:223-48.
117. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*. 2001;23(6):681-5.
118. Edwards GJ, Cootes TF, Taylor CJ, editors. *Face recognition using active appearance models*. European conference on computer vision; 1998: Springer.
119. Osher S, Sethian JA. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of computational physics*. 1988;79(1):12-49.
120. Suri JS, Liu K, Singh S, Laxminarayan SN, Zeng X, Reden L. Shape recovery algorithms using level sets in 2-D/3-D medical imagery: a state-of-the-art review. *Information Technology in Biomedicine, IEEE Transactions on*. 2002;6(1):8-28.
121. Malladi R, Sethian JA, Vemuri BC, editors. *Topology-independent shape modeling scheme*. *Geometric Methods in Computer Vision II*; 1993: International Society for Optics and Photonics.
122. Malladi R, Sethian JA, editors. *Level set and fast marching methods in image processing and computer vision*. *Image Processing, 1996 Proceedings, International Conference on*; 1996: IEEE.
123. Siddiqi K, Lauziere YB, Tannenbaum A, Zucker SW. Area and length minimizing flows for shape segmentation. *IEEE Transactions on Image Processing*. 1998;7(3):433-43.
124. Kichenassamy S, Kumar A, Olver P, Tannenbaum A, Yezzi A. Conformal curvature flows: from phase transitions to active vision. *Archive for Rational Mechanics and Analysis*. 1996;134(3):275-301.
125. Caselles V, Kimmel R, Sapiro G. Geodesic active contours. *International journal of computer vision*. 1997;22(1):61-79.
126. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116-28.
127. Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical physics*. 2014;41(5):051910.
128. Way TW, Hadjiiski LM, Sahiner B, Chan H-P, Cascade PN, Kazerooni EA, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Medical Physics*. 2006;33(7):2323-37.
129. Suzuki K, Kohlbrenner R, Epstein ML, Obajuluwa AM, Xu J, Hori M. Computer-aided measurement of liver volumes in CT by means of geodesic active contour segmentation coupled with level-set algorithms. *Medical physics*. 2010;37(5):2159-66.

130. Zhao B, Yankelevitz D, Reeves A, Henschke C. Two-dimensional multi-criterion segmentation of pulmonary nodules on helical CT images. *Medical Physics*. 1999;26(6):889-95.
131. Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence*. 2004;26(9):1124-37.
132. Greig DM, Porteous BT, Seheult AH. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society Series B (Methodological)*. 1989:271-9.
133. Zabih R, Kolmogorov V, editors. Spatially coherent clustering using graph cuts. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004*: IEEE.
134. Gridchyn I, Kolmogorov V, editors. Potts model, parametric maxflow and k-submodular functions. *Proceedings of the IEEE International Conference on Computer Vision*; 2013.
135. Kolmogorov V, Blossom V: a new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation*. 2009;1(1):43-67.
136. Kolmogorov V, Shioura A. New algorithms for the dual of the convex cost network flow problem with application to computer vision. *Mathematical Programming*. 2007.
137. Boykov Y, Veksler O, Zabih R, editors. Markov random fields with efficient approximations. *Computer vision and pattern recognition, 1998 Proceedings 1998 IEEE computer society conference on*; 1998: IEEE.
138. Torresani L, Kolmogorov V, Rother C. Feature correspondence via graph matching: Models and global optimization. *Computer Vision—ECCV 2008*. 2008:596-609.
139. Ding CH, He X, Zha H, Gu M, Simon HD, editors. A min-max cut algorithm for graph partitioning and data clustering. *Data Mining, 2001 ICDM 2001, Proceedings IEEE International Conference on*; 2001: IEEE.
140. Stoer M, Wagner F. A simple min-cut algorithm. *Journal of the ACM (JACM)*. 1997;44(4):585-91.
141. Shen Y. A new simple algorithm for enumerating all minimal paths and cuts of a graph. *Microelectronics Reliability*. 1995;35(6):973-6.
142. Kolmogorov V, Zabih R. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence*. 2004;26(2):147-59.
143. Fix A, Gruber A, Boros E, Zabih R, editors. A graph cut algorithm for higher-order Markov random fields. *Computer Vision (ICCV), 2011 IEEE International Conference on*; 2011: IEEE.
144. Kolmogorov V, Rother C. Minimizing nonsubmodular functions with graph cuts—a review. *IEEE transactions on pattern analysis and machine intelligence*. 2007;29(7).
145. Boykov Y, Funka-Lea G. Graph cuts and efficient ND image segmentation. *International journal of computer vision*. 2006;70(2):109-31.
146. Lombaert H, Sun Y, Grady L, Xu C, editors. A multilevel banded graph cuts method for fast image segmentation. *Computer Vision, 2005 ICCV 2005 Tenth IEEE International Conference on*; 2005: IEEE.
147. Kohli P, Torr PH, editors. Efficiently solving dynamic markov random fields using graph cuts. *Computer Vision, 2005 ICCV 2005 Tenth IEEE International Conference on*; 2005: IEEE.
148. Hendrickson B, Leland RW. A Multi-Level Algorithm For Partitioning Graphs. *SC*. 1995;95(28).

149. Kwatra V, Schödl A, Essa I, Turk G, Bobick A, editors. Graphcut textures: image and video synthesis using graph cuts. *ACM Transactions on Graphics (ToG)*; 2003: ACM.
150. Agarwala A, Dontcheva M, Agrawala M, Drucker S, Colburn A, Curless B, et al., editors. Interactive digital photomontage. *ACM Transactions on Graphics (TOG)*; 2004: ACM.
151. Rother C, Kolmogorov V, Blake A, editors. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*; 2004: ACM.
152. Liu J, Sun J, editors. Parallel graph-cuts by adaptive bottom-up merging. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*; 2010: IEEE.
153. Jamriška O, Sýkora D, Hornung A, editors. Cache-efficient graph cuts on structured grids. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*; 2012: IEEE.
154. Recht JM. Method and system for image segmentation. US Patent No 7, 257, 267: Google Patents; 2007.
155. Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*. 2015;24(1):205-19.
156. Walker GV, Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiotherapy and Oncology*. 2014.
157. Conson M, Cella L, Pacelli R, Comerci M, Liuzzi R, Salvatore M, et al. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: From atlas to dose-volume histograms. *Radiotherapy and Oncology*. 2014;112(3):326-31.
158. Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau P-Y, Malandain G, et al. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiotherapy and oncology*. 2008;87(1):93-9.
159. Haas B, Coradi T, Scholz M, Kunz P, Huber M, Oppitz U, et al. Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies. *Physics in Medicine & Biology*. 2008;53(6):1751.
160. Dowling JA, Lambert J, Parker J, Salvado O, Fripp J, Capp A, et al. An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy. *International Journal of Radiation Oncology• Biology• Physics*. 2012;83(1):e5-e11.
161. Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *International Journal of Radiation Oncology• Biology• Physics*. 2010;77(3):959-66.
162. Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. *Radiotherapy and Oncology*. 2012;102(1):68-73.
163. Zikic D, Glocker B, Criminisi A, editors. Classifier-based multi-atlas label propagation with test-specific atlas weighting for correspondence-free scenarios. *International MICCAI Workshop on Medical Computer Vision*; 2014: Springer.
164. Wang H, Yushkevich PA, editors. Multi-atlas segmentation without registration: a supervoxel-based approach. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2013: Springer.
165. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*. 2007;160:3-24.

166. Friedman JH. Regularized discriminant analysis. *Journal of the American statistical association*. 1989;84(405):165-75.
167. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE transactions on information theory*. 1967;13(1):21-7.
168. Breiman L. Random forests. *Machine learning*. 2001;45(1):5-32.
169. Burges CJ. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*. 1998;2(2):121-67.
170. Zhang GP. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 2000;30(4):451-62.
171. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM computing surveys (CSUR)*. 1999;31(3):264-323.
172. Chih-Chin L. A novel image segmentation approach based on particle swarm optimization. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 2006;89(1):324-7.
173. Brink A. Minimum spatial entropy threshold selection. *IEE Proceedings-Vision, Image and Signal Processing*. 1995;142(3):128-32.
174. Frank R, Grabowski T, Damasio H, editors. *Voxelwise percentage tissue segmentation of human brain magnetic resonance images. Abstracts, 25th Annual Meeting, Society for Neuro-Science; 1995.*
175. MacQueen J, editor *Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967.*
176. Ruspini EH. A new approach to clustering. *Information and control*. 1969;15(1):22-32.
177. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*. 1984;10(2-3):191-203.
178. Laws KI. *Textured image segmentation. University of Southern California Los Angeles Image Processing INST, 1980.*
179. Castellano G, Bonilha L, Li L, Cendes F. Texture analysis of medical images. *Clinical radiology*. 2004;59(12):1061-9.
180. Haralick RM. Statistical and structural approaches to texture. *Proceedings of the IEEE*. 1979;67(5):786-804.
181. Tuceryan M, Jain AK. Texture segmentation using Voronoi polygons. *IEEE transactions on pattern analysis and machine intelligence*. 1990;12(2):211-6.
182. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBIImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics*. 2010;26(7):979-81.
183. Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE transactions on biomedical engineering*. 2008;55(7):1822-30.
184. Lennon FE, Cianci GC, Cipriani NA, Hensing TA, Zhang HJ, Chen C-T, et al. Lung cancer—a fractal viewpoint. *Nature reviews Clinical oncology*. 2015;12(11):664.
185. Campbell FW, Robson J. Application of Fourier analysis to the visibility of gratings. *The Journal of physiology*. 1968;197(3):551-66.
186. Daugman JG. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision research*. 1980;20(10):847-56.
187. Turner MR. Texture discrimination by Gabor functions. *Biological cybernetics*. 1986;55(2-3):71-82.
188. Shen L, Bai L. A review on Gabor wavelets for face recognition. *Pattern analysis and applications*. 2006;9(2-3):273-92.

189. Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. *International Journal of Radiation Oncology* Biology* Physics*. 2009;75(2):618-25.
190. De Nunzio G, Pastore G, Donativi M, Castellano A, Falini A. A CAD system for cerebral glioma based on texture features in DT-MR images. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. 2011;648:S100-S2.
191. Bauer S, Nolte L-P, Reyes M. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: Springer*; 2011. p. 354-61.
192. Qurat-UI-Ain GL, Kazmi SB, Jaffar MA, Mirza AM. Classification and segmentation of brain tumor using texture analysis. *Recent Advances In Artificial Intelligence, Knowledge Engineering And Data Bases*. 2010:147-55.
193. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature*. 2015;521(7553):436.
194. Krizhevsky A, Sutskever I, Hinton GE, editors. *Imagenet classification with deep convolutional neural networks*. *Advances in neural information processing systems*; 2012.
195. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-r, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*. 2012;29(6):82-97.
196. Sutskever I, Vinyals O, Le QV, editors. *Sequence to sequence learning with neural networks*. *Advances in neural information processing systems*; 2014.
197. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical image analysis*. 2017;42:60-88.
198. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. *Medical image analysis*. 2017;35:18-31.
199. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*. 2017;36:61-78.
200. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*. 2016;35(5):1240-51.
201. Zhao L, Jia K. Multiscale cnns for brain tumor segmentation and diagnosis. *Computational and mathematical methods in medicine*. 2016;2016.
202. Chu C, De Fauw J, Tomasev N, Paredes BR, Hughes C, Ledsam J, et al. Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans. *F1000Research*. 2016;5.
203. Zhao J, Wang J, Cheng M. Development of a deep neural network derived from contours defined by consensus-based guidelines for automatic target segmentation in hepatocellular carcinoma radiotherapy: A study protocol. *F1000Research*. 2017;6.
204. Men K, Chen X, Zhang Y, Dai J, Yi J, Li Y. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning CT images. *Frontiers in Oncology*. 2017;7:315.
205. Iqbal S, Iqbal K, Arif F, Shaukat A, Khanum A. Potential lung nodules identification for characterization by variable multistep threshold and shape indices from CT images. *Computational and mathematical methods in medicine*. 2014;2014.
206. John J, Mini M. Multilevel Thresholding Based Segmentation and Feature Extraction for Pulmonary Nodule Detection. *Procedia Technology*. 2016;24:957-63.

207. Mullally W, Betke M, Wang J, Ko JP. Segmentation of nodules on chest computed tomography for growth assessment. *Medical Physics*. 2004;31(4):839-48.
208. Taşçı E, Uğur A. Shape and texture based novel features for automated juxtapleural nodule detection in lung CTs. *Journal of medical systems*. 2015;39(5):46.
209. Zhao B, Yankelevitz D, Reeves A, Henschke C. Two-dimensional multi-criterion segmentation of pulmonary nodules on helical CT images. *Medical Physics*. 1999;26(6):889-95.
210. Zhao B, Reeves AP, Yankelevitz D, Henschke CI. Three-dimensional multi-criterion automatic segmentation of pulmonary nodules of helical computed tomography images. *Optical Engineering*. 1999;38(8):1340-8.
211. Fan L, Qian J, Odry BL, Shen H, Naidich D, Kohl G, et al., editors. Automatic segmentation of pulmonary nodules by using dynamic 3D cross-correlation for interactive CAD systems. *Medical Imaging 2002: Image Processing; 2002: International Society for Optics and Photonics*.
212. Fetita CI, Preteux F, Beigelman-Aubry C, Grenier P, editors. 3D automated lung nodule segmentation in HRCT. *International Conference on Medical Image Computing and Computer-Assisted Intervention; 2003: Springer*.
213. Kostis WJ, Reeves AP, Yankelevitz DF, Henschke CI. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. *IEEE transactions on medical imaging*. 2003;22(10):1259-74.
214. Moltz JH, Bornemann L, Kuhnigk J-M, Dicken V, Peitgen E, Meier S, et al. Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans. *IEEE Journal of selected topics in signal processing*. 2009;3(1):122-34.
215. Setio AA, Jacobs C, Gelderblom J, Ginneken B. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical physics*. 2015;42(10):5642-53.
216. Velazquez ER, Aerts HJ, Gu Y, Goldgof DB, De Ruyscher D, Dekker A, et al. A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen. *Radiotherapy and Oncology*. 2012;105(2):167-73.
217. Velazquez ER, Parmar C, Jermoumi M, Mak RH, Van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Scientific reports*. 2013;3:3529.
218. Song J, Yang C, Fan L, Wang K, Yang F, Liu S, et al. Lung lesion extraction using a toboggan based growing automatic segmentation approach. *IEEE transactions on medical imaging*. 2016;35(1):337-53.
219. Namin ST, Moghaddam HA, Jafari R, Esmaeil-Zadeh M, Gity M, editors. Automated detection and classification of pulmonary nodules in 3D thoracic CT images. *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on; 2010: IEEE*.
220. Parveen SS, Kavitha C, editors. Detection of lung cancer nodules using automatic region growing method. *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on; 2013: IEEE*.
221. Gu Y, Kumar V, Hall LO, Goldgof DB, Li C-Y, Korn R, et al. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. *Pattern recognition*. 2013;46(3):692-702.
222. Kubota T, Jerebko AK, Dewan M, Salganicoff M, Krishnan A. Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. *Medical Image Analysis*. 2011;15(1):133-54.

223. Diciotti S, Picozzi G, Falchini M, Mascalchi M, Villari N, Valli G. 3-D segmentation algorithm of small lung nodules in spiral CT images. *IEEE Transactions on Information Technology in Biomedicine*. 2008;12(1):7-19.
224. Kuhnigk J-M, Dicken V, Bornemann L, Bakai A, Wormanns D, Krass S, et al. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE transactions on medical imaging*. 2006;25(4):417-34.
225. Lassen B, Jacobs C, Kuhnigk J, van Ginneken B, van Rikxoort E. Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans. *Physics in Medicine & Biology*. 2015;60(3):1307.
226. Krishnamurthy S, Narasimhan G, Rengasamy U. Three-dimensional lung nodule segmentation and shape variance analysis to detect lung cancer with reduced false positives. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. 2016;230(1):58-70.
227. Diciotti S, Lombardo S, Falchini M, Picozzi G, Mascalchi M. Automated segmentation refinement of small lung nodules in CT scans by local shape analysis. *IEEE Transactions on Biomedical Engineering*. 2011;58(12):3418-28.
228. Santos AM, de Carvalho Filho AO, Silva AC, de Paiva AC, Nunes RA, Gattass M. Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM. *Engineering applications of artificial intelligence*. 2014;36:27-39.
229. Badura P, Pietka E. Soft computing approach to 3D lung nodule segmentation in CT. *Computers in biology and medicine*. 2014;53:230-43.
230. Brown MS, Lo P, Goldin JG, Barnoy E, Kim GHJ, McNitt-Gray MF, et al. Toward clinically usable CAD for lung cancer screening with computed tomography. *European radiology*. 2014;24(11):2719-28.
231. Tan Y, Schwartz LH, Zhao B. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. *Medical physics*. 2013;40(4).
232. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. *American journal of roentgenology*. 2006;186(4):989-94.
233. Vivanti R, Joskowicz L, Karaaslan OA, Sosna J. Automatic lung tumor segmentation with leaks removal in follow-up CT studies. *International journal of computer assisted radiology and surgery*. 2015;10(9):1505-14.
234. Kawata Y, Niki N, Ohmatsu H, Kakinuma R, Eguchi K, Kaneko M, et al. Quantitative surface characterization of pulmonary nodules based on thin-section CT images. *IEEE Transactions on nuclear science*. 1998;45(4):2132-8.
235. Cascio D, Magro R, Fauci F, Iacomi M, Raso G. Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models. *Computers in biology and medicine*. 2012;42(11):1098-109.
236. Farag AA, El-Baz A, Gimel'farb G, Falk R, El-Ghar MA, Eldiasty T, et al., editors. *Appearance models for robust segmentation of pulmonary nodules in 3D LDCT chest images*. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2006: Springer.
237. Farag AA, Abdelmunim H, Graham J, Farag AA, Elshazly S, El-Mogy S, et al., editors. *Variational approach for segmentation of lung nodules*. Image Processing (ICIP), 2011 18th IEEE International Conference on; 2011: IEEE.

238. Farag AA, El Munim HEA, Graham JH, Farag AA. A novel approach for lung nodules segmentation in chest CT using level sets. *IEEE Transactions on Image Processing*. 2013;22(12):5202-13.
239. Soltaninejad S, Keshani M, Tajeripour F, editors. Lung nodule detection by KNN classifier and active contour modelling and 3D visualization. *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*; 2012: IEEE.
240. Van Ginneken B, Armato SG, de Hoop B, van Amelsvoort-van de Vorst S, Duindam T, Niemeijer M, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Medical image analysis*. 2010;14(6):707-22.
241. Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours. *Medical physics*. 2006;33(7Part1):2323-37.
242. Yip SS, Parmar C, Blezek D, Estepar RSJ, Pieper S, Kim J, et al. Application of the 3D slicer chest imaging platform segmentation algorithm for large lung nodule delineation. *PloS one*. 2017;12(6):e0178944.
243. Yoo Y, Shim H, Yun ID, Lee KW, Lee SU, editors. Segmentation of ground glass opacities by asymmetric multi-phase deformable model. *Medical Imaging 2006: Image Processing*; 2006: International Society for Optics and Photonics.
244. Plajer IC, Richter D, editors. A new approach to model based active contours in lung tumor segmentation in 3D CT image data. *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*; 2010: IEEE.
245. Zheng Y, Steiner K, Bauer T, Yu J, Shen D, Kambhamettu C, editors. Lung nodule growth analysis from 3D CT data with a coupled segmentation and registration framework. *Computer Vision, 2007 ICCV 2007 IEEE 11th International Conference on*; 2007: IEEE.
246. Zheng Y, Kambhamettu C, Bauer T, Steiner K, editors. Accurate estimation of pulmonary nodule's growth rate in CT images with nonrigid registration and precise nodule detection and segmentation. *Computer Vision and Pattern Recognition Workshops, 2009 CVPR Workshops 2009 IEEE Computer Society Conference on*; 2009: IEEE.
247. Lermé N, Malgouyres F, Rocchisani J-M. Fast and memory efficient segmentation of lung tumors using graph cuts. 2010;12.
248. Lermé N, Malgouyres F, Létocart L, editors. Reducing graphs in graph cut segmentation. *Image Processing (ICIP), 2010 17th IEEE International Conference on*; 2010: IEEE.
249. Shen Z, Wang H, Xi W, Deng X, Chen J, Zhang Y. Multi-phase simultaneous segmentation of tumor in lung 4D-CT data with context information. *PloS one*. 2017;12(6):e0178411.
250. Browder WA, Reeves AP, Apananosovich TV, Cham MD, Yankelevitz DF, Henschke CI, editors. Automated volumetric segmentation method for growth consistency of nonsolid pulmonary nodules in high-resolution CT. *Medical Imaging 2007: Computer-Aided Diagnosis*; 2007: International Society for Optics and Photonics.
251. Zhang L, Fang M, Naidich DP, Novak CL, editors. Consistent interactive segmentation of pulmonary ground glass nodules identified in CT studies. *Medical Imaging 2004: Image Processing*; 2004: International Society for Optics and Photonics.
252. Van Ginneken B, editor Supervised probabilistic segmentation of pulmonary nodules in CT scans. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2006: Springer.

253. Netto SMB, Silva AC, Nunes RA, Gattass M. Automatic segmentation of lung nodules with growing neural gas and support vector machine. *Computers in biology and medicine*. 2012;42(11):1110-21.
254. Nie S-d, Li-Hong L, Chen Z-X, editors. A CI feature-based pulmonary nodule segmentation using three-domain mean shift clustering. *Wavelet Analysis and Pattern Recognition, 2007 ICWAPR'07 International Conference on; 2007: IEEE*.
255. Nithila EE, Kumar S. Automatic detection of solitary pulmonary nodules using swarm intelligence optimized neural networks on CT images. *Engineering science and technology, an international journal*. 2017;20(3):1192-202.
256. Nithila EE, Kumar S. Segmentation of lung nodule in CT data using active contour model and Fuzzy C-mean clustering. *Alexandria Engineering Journal*. 2016;55(3):2583-8.
257. Zhou J, Chang S, Metaxas DN, Zhao B, Ginsberg MS, Schwartz LH, editors. An automatic method for ground glass opacity nodule detection and segmentation from CT studies. *Engineering in Medicine and Biology Society, 2006 EMBS'06 28th Annual International Conference of the IEEE; 2006: IEEE*.
258. Zhou J, Chang S, Metaxas DN, Zhao B, Schwartz LH, Ginsberg MS, editors. Automatic detection and segmentation of ground glass opacity nodules. *International Conference on Medical Image Computing and Computer-Assisted Intervention; 2006: Springer*.
259. Kakar M, Olsen DR. Automatic segmentation and recognition of lungs and lesion from CT scans of thorax. *Computerized Medical Imaging and Graphics*. 2009;33(1):72-82.
260. Tao Y, Lu L, Dewan M, Chen AY, Corso J, Xuan J, et al., editors. Multi-level ground glass nodule detection and segmentation in CT lung images. *International Conference on Medical Image Computing and Computer-Assisted Intervention; 2009: Springer*.
261. Hossain MRI, Ahmed I, Kabir MH, editors. Automatic lung tumor detection based on GLCM features. *Asian Conference on Computer Vision; 2014: Springer*.
262. Zinoveva O, Zinovev D, Siena SA, Raicu DS, Furst J, Armato SG, editors. A texture-based probabilistic approach for lung nodule segmentation. *International Conference Image Analysis and Recognition; 2011: Springer*.
263. Jirapatnakul AC, Mulman YD, Reeves AP, Yankelevitz DF, Henschke CI. Segmentation of juxtapleural pulmonary nodules using a robust surface estimate. *Journal of Biomedical Imaging*. 2011;2011:15.
264. Matsumoto S, Ohno Y, Yamagata H, Takenaka D, Sugimura K. Computer-aided detection of lung nodules on multidetector row computed tomography using three-dimensional analysis of nodule candidates and their surroundings. *Radiation medicine*. 2008;26(9):562-9.
265. Yong JR, Qi S, van Triest HJ, Kang Y, Qian W. Automatic segmentation of juxta-pleural tumors from CT images based on morphological feature analysis. *Bio-medical materials and engineering*. 2014;24(6):3137-44.
266. Okada K, Comaniciu D, Krishnan A. Robust anisotropic Gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT. *IEEE Transactions on Medical Imaging*. 2005;24(3):409-23.
267. Okada K, Akdemir U, editors. Blob segmentation using joint space-intensity likelihood ratio test: application to 3D tumor segmentation. *Computer Vision and Pattern Recognition, 2005 CVPR 2005 IEEE Computer Society Conference on; 2005: IEEE*.
268. Okada K, Ramesh V, Krishnan A, Singh M, Akdemir U, editors. Robust pulmonary nodule segmentation in CT: improving performance for juxtapleural cases. *International*

- Conference on Medical Image Computing and Computer-Assisted Intervention; 2005: Springer.
269. Gonçalves L, Novo J, Campilho A. Hessian based approaches for 3D lung nodule segmentation. *Expert Systems with Applications*. 2016;61:1-15.
270. Wang J, Engelmann R, Li Q. Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral-scanning technique. *Medical Physics*. 2007;34(12):4678-89.
271. Wiemker R, Zwartkuis A, editors. Optimal thresholding for 3D segmentation of pulmonary nodules in high resolution CT. *International Congress Series*; 2001: Elsevier.
272. Kubota T, Jerebko A, Salganicoff M, Dewan M, Krishnan A, editors. Robust segmentation of pulmonary nodules of various densities: from ground-glass opacities to solid nodules. *Proceedings of the International Workshop on Pulmonary Image Processing*; 2008.
273. Vezhnevets V, Konouchine V. GrowCut: Interactive multi-label ND image segmentation by cellular automata. *Proc of Graphicon*. 2005;1:150-6.
274. Li X, Wang R. A new efficient 2D combined with 3D CAD system for solitary pulmonary nodule detection in CT images. *International Journal of Image, Graphics and Signal Processing*. 2011;3(4):18.
275. Farag AA, editor Variational approach for small-size lung nodule segmentation. *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*; 2013: IEEE.
276. Gurcan MN, Sahiner B, Petrick N, Chan HP, Kazerooni EA, Cascade PN, et al. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system. *Medical Physics*. 2002;29(11):2552-8.
277. Dornheim J, Seim H, Preim B, Hertel I, Strauss G. Segmentation of neck lymph nodes in CT datasets with stable 3D mass-spring models: Segmentation of neck lymph nodes. *Academic Radiology*. 2007;14(11):1389-99.
278. Lu K, Xue Z, Wong ST, editors. A robust semi-automatic approach for ROI segmentation in 3D CT images. *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*; 2013: IEEE.
279. Honea DM, Ge Y, Snyder WE, Hemler PF, Vining DJ, editors. Lymph node segmentation using active contours. *Medical Imaging 1997: Image Processing*; 1997: International Society for Optics and Photonics.
280. Honea DM, Snyder WE, editors. Three-dimensional active surface approach to lymph node segmentation. *Medical Imaging 1999: Image Processing*; 1999: International Society for Optics and Photonics.
281. Lu K, Higgins WE. Segmentation of the central-chest lymph nodes in 3D MDCT images. *Computers in biology and medicine*. 2011;41(9):780-9.
282. Yu P, Poh CL. Region-based snake with edge constraint for segmentation of lymph nodes on CT images. *Computers in biology and medicine*. 2015;60:86-91.
283. Chen Q, Quan F, Xu J, Rubin DL. Snake model-based lymphoma segmentation for sequential CT images. *Computer methods and programs in biomedicine*. 2013;111(2):366-75.
284. Fabel M, von Tengg-Kobligk H, Giesel F, Bornemann L, Dicken V, Kopp-Schneider A, et al. Semi-automated volumetric analysis of lymph node metastases in patients with malignant melanoma stage III/IV-A feasibility study. *European radiology*. 2008;18(6):1114-22.
285. Fabel M, Bolte H, von Tengg-Kobligk H, Bornemann L, Dicken V, Delorme S, et al. Semi-automated volumetric analysis of lymph node metastases during follow-up—initial results. *European radiology*. 2011;21(4):683-92.

286. Buerke B, Puesken M, Mütter S, Weckesser M, Gerss J, Heindel W, et al. Measurement accuracy and reproducibility of semiautomated metric and volumetric lymph node analysis in MDCT. *American Journal of Roentgenology*. 2010;195(4):979-85.
287. Yan J, Zhao B, Wang L, Zelenetz A, Schwartz LH. Marker-controlled watershed for lymphoma segmentation in sequential CT images. *Medical Physics*. 2006;33(7):2452-60.
288. Yan J, Zhao B, Curran S, Zelenetz A, Schwartz LH. Automated matching and segmentation of lymphoma on serial CT examinations. *Medical physics*. 2007;34(1):55-62.
289. Liu J, Feng C-H, Hua J, Yao J, White JM, Summers RM, editors. Automatic detection and segmentation of abdominopelvic lymph nodes on computed tomography scans. *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on; 2012: IEEE*.
290. Feuerstein M, Deguchi D, Kitasaka T, Iwano S, Imaizumi K, Hasegawa Y, et al., editors. Automatic mediastinal lymph node detection in chest CT. *Medical Imaging 2009: Computer-Aided Diagnosis; 2009: International Society for Optics and Photonics*.
291. Xu J, Greenspan H, Napel S, Rubin DL. Automated temporal tracking and segmentation of lymphoma on serial CT examinations. *Medical physics*. 2011;38(11):5879-86.
292. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D, editors. Automatic detection and segmentation of axillary lymph nodes. *International Conference on Medical Image Computing and Computer-Assisted Intervention; 2010: Springer*.
293. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D. Automatic detection and segmentation of lymph nodes from CT data. *IEEE Transactions on Medical Imaging*. 2012;31(2):240-50.
294. Höink A, Weßling J, Koch R, Schülke C, Kohlhase N, Wassenaar L, et al. Comparison of manual and semi-automatic measuring techniques in MSCT scans of patients with lymphoma: a multicentre study. *European radiology*. 2014;24(11):2709-18.
295. Keil S, Plumhans C, Behrendt FF, Stanzel S, Suehling M, Mühlenbruch G, et al. Automated measurement of lymph nodes: a phantom study. *European radiology*. 2009;19(5):1079-86.
296. Yan J, Zhuang T-g, Zhao B, Schwartz LH. Lymph node segmentation from CT images using fast marching method. *Computerized Medical Imaging and Graphics*. 2004;28(1-2):33-8.
297. Maleike D, Fabel M, Tetzlaff R, von Tengg-Kobligk H, Heimann T, Meinzer H-P, et al., editors. Lymph node segmentation on CT images by a shape model guided deformable surface method. *Medical Imaging 2008: Image Processing; 2008: International Society for Optics and Photonics*.
298. Fischer B, Lassen U, Mortensen J, Larsen S, Loft A, Bertelsen A, et al. Preoperative staging of lung cancer with combined PET-CT. *New England Journal of Medicine*. 2009;361(1):32-9.
299. NICE Clinical Guideline: Lung cancer: The diagnosis and treatment of lung cancer. National Institute for Health and Care Excellence, April 2011. Access: <https://www.nice.org.uk/guidance/cg121>.
300. Fox JL, Rengan R, O'Meara W, Yorke E, Erdi Y, Nehmeh S, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer? *International Journal of Radiation Oncology• Biology• Physics*. 2005;62(1):70-5.
301. De Ruyscher D, Wanders S, Minken A, Lumens A, Schiffelers J, Stultiens C, et al. Effects of radiotherapy planning with a dedicated combined PET-CT-simulator of patients with non-

small cell lung cancer on dose limiting normal tissues and radiation dose-escalation: a planning study. *Radiotherapy and oncology*. 2005;77(1):5-10.

302. van Elmpt W, De Ruyscher D, van der Salm A, Lakeman A, van der Stoep J, Emans D, et al. The PET-boost randomised phase II dose-escalation trial in non-small cell lung cancer. *Radiotherapy and Oncology*. 2012;104(1):67-71.

303. Hatt M, Lee JA, Schmidtlein CR, Naqa IE, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM Task Group No. 211. *Medical physics*. 2017;44(6).

304. Boellaard R, Oyen WJ, Hoekstra CJ, Hoekstra OS, Visser EP, Willemsen AT, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *European journal of nuclear medicine and molecular imaging*. 2008;35(12):2320-33.

305. Access: <http://mirada-medical.com/radiationoncology/efficient/>.

306. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*. 2018;126(2):312-7.

307. Commowick O, Grégoire V, Malandain G. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*. 2008;87(2):281-9.

308. Woo JY, Kim TY, Seok JY, Kim TM, Cho YW, Kim SY, et al. Comparison of Three Commercial Model Based Segmentation Software and Atlas Based Segmentation Software in Contouring of Prostate Cancer and Brain Cancer. *International Journal of Radiation Oncology • Biology • Physics*. 2015;93(3):E605.

309. Hu K, Lin A, Young A, Kubicek G, Piper J, Nelson A, et al. Timesavings for contour generation in head and neck IMRT: Multi-institutional experience with an atlas-based segmentation method. *International Journal of Radiation Oncology* Biology* Physics*. 2008;72(1):S391.

310. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiation Oncology*. 2012;7(1):160.

311. Delpon G, Escande A, Ruef T, Darréon J, Fontaine J, Noblet C, et al. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. *Frontiers in oncology*. 2016;6:178.

312. Roussakis Y, McWilliam A, Hartley A, Sangera P, Benghiat H, Hickman M, et al. PO-0931: Evaluation of multiple auto-segmentation solutions against inter-observer variability. *Radiotherapy and Oncology*. 2015;115:S485-S6.

313. Meillan N, Bibault J-E, Vautier J, Daveau-Bergerault C, Kreps S, Tournat H, et al. Automatic Intracranial Segmentation: Is the Clinician Still Needed? *Technology in cancer research & treatment*. 2018;17:1533034617748839.

314. Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Is accurate contouring of salivary and swallowing structures necessary to spare them in head and neck VMAT plans? *Radiotherapy and Oncology*. 2018.

315. Herraiz Lablanca MD, Paul S, Chiesa M, Grosser KH, Harms W. PO-1006: Evaluation of an auto-segmentation software for definition of organs at risk in radiotherapy. *Radiotherapy and Oncology*.123:S554.

316. Grosu A-L, Lachner R, Wiedenmann N, Stärk S, Thamm R, Kneschaurek P, et al. Validation of a method for automatic image fusion (BrainLAB System) of CT data and 11C-methionine-PET data for stereotactic radiotherapy using a LINAC: first clinical experience. *International Journal of Radiation Oncology• Biology• Physics*. 2003;56(5):1450-63.
317. Daisne J-F, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiation oncology*. 2013;8(1):154.
318. Geraghty JP, Grogan G, Ebert MA. Automatic segmentation of male pelvic anatomy on computed tomography images: a comparison with multiple observers in the context of a multicentre clinical trial. *Radiation Oncology*. 2013;8(1):106.
319. Simmat I, Georg P, Georg D, Birkfellner W, Goldner G, Stock M. Assessment of accuracy and efficiency of atlas-based autosegmentation for prostate radiotherapy in a variety of clinical conditions. *Strahlentherapie Und Onkologie*. 2012;188(9):807-15.
320. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. *Medical physics*. 2011;38(11):6160-70.
321. Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, Yap B, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiation Oncology*. 2014;9(1):173.
322. Mitchell RA, Wai P, Colgan R, Kirby AM, Donovan EM. Improving the efficiency of breast radiotherapy treatment planning using a semi-automated approach. *Journal of applied clinical medical physics*. 2017;18(1):18-24.
323. Stewart J, Lim K, Kelly V, Xie J, Brock KK, Moseley J, et al. Automated weekly replanning for intensity-modulated radiotherapy of cervix cancer. *International Journal of Radiation Oncology• Biology• Physics*. 2010;78(2):350-8.
324. Han X, Hoogeman MS, Levendag PC, Hibbard LS, Teguh DN, Voet P, et al., editors. Atlas-based auto-segmentation of head and neck CT images. *International Conference on Medical Image Computing and Computer-assisted Intervention*; 2008: Springer.
325. Greenham S, Dean J, Fu CCK, Goman J, Mulligan J, Tune D, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. *Journal of medical radiation sciences*. 2014;61(3):151-8.
326. Gooding M, Chu K, Conibear J, Dilling T, Durrant L, Fuss M, et al. Multicenter clinical assessment of DIR atlas-based autocontouring. *International Journal of Radiation Oncology• Biology• Physics*. 2013;87:S714-S5.
327. Pirson A, Nguyen PV, Baiwir M, Coucke PA, Lakosi F, Gulyban A. EP-1196: Atlas-based segmentation for delineating the locoregional node levels during breast radiotherapy. *Radiotherapy and Oncology*. 119:S568.
328. Access: http://www3.gehealthcare.com/en/products/categories/advanced_visualization/applications/lung_vcar.
329. Access: <http://pdf.medicaexpo.com/pdf/mevis-medical-solutions-ag/veolity/101152-140235.html>.
330. Access: <https://www.healthcare.siemens.co.uk/medical-imaging-it/advanced-visualization-solutions/syngovia/use>.
331. Access: http://resources.alleninteractions.com/online/demos/vital_images/course/resources/PDF/CT%20v2.0%20Understanding.pdf.

332. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. *Medical physics*. 2003;30(5):979-85.
333. Boedeker KL, McNitt-Gray MF, Rogers SR, Truong DA, Brown MS, Gjertson DW, et al. Emphysema: effect of reconstruction algorithm on CT imaging measures. *Radiology*. 2004;232(1):295-301.
334. Shrimpton P, Hillier M, Meeson S, Golding S. Doses from Computed Tomography (CT) Examinations in the UK – 2011 Review: Public Health England, Centre for Radiation, Chemical and Environmental Hazards; 2014.
335. Davis AT, Earley J, Edyvean S, Findlay U, Lindsay R, Nisbet A, et al. IPEM topical report 2: the first UK survey of dose indices from radiotherapy treatment planning computed tomography scans for adult patients. *Physics in medicine and biology*. 2018.
336. Davis AT, Palmer AL, Nisbet A. Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review. *The British journal of radiology*. 2017;90(1076):20160406.
337. Maini R, Aggarwal H. A comprehensive review of image enhancement techniques. arXiv preprint arXiv:10034053. 2010.
338. Landau DH, Simon; Laurence, Virginia; Mayles, Philip; Mayles, Helen; Fenwick, John; Miles, Elizabeth; Wilkinson, Dean; Hughes, Laura; Ngai, Yenting; Khan, Iftekhhar. IDEAL-CRT Radiotherapy Planning & Delivery Guidelines Final v4.0 06-09-2012. 2012.
339. Goldman LW. Principles of CT: radiation dose and image quality. *Journal of nuclear medicine technology*. 2007;35(4):213-25.
340. Current Controlled Trials. London: BioMed Central. ISRCTN12155469. IDEAL-CRT: A Phase I/II trial of concurrent chemoradiation with dose-escalated radiotherapy in patients with stage II or stage III non-small cell lung cancer. Access: <https://doi.org/10.1186/ISRCTN12155469>. 2009.
341. Mayer C, Meyer M, Fink C, Schmidt B, Sedlmair M, Schoenberg S, et al. Potential for radiation dose savings in abdominal and chest CT using automatic tube voltage selection in combination with automatic tube current modulation. *AJR American journal of roentgenology*. 2014;203(2):292-9.
342. Schimmöller L, Lanzman R, Dietrich S, Boos J, Heusch P, Miese F, et al. Evaluation of automated attenuation-based tube potential selection in combination with organ-specific dose reduction for contrast-enhanced chest CT examinations. *Clinical radiology*. 2014;69(7):721-6.
343. Hu X, Ding X, Wu R, Zhang M. Radiation dose of non-enhanced chest CT can be reduced 40% by using iterative reconstruction in image space. *Clinical radiology*. 2011;66(11):1023-9.
344. Peng W, Li Z, Xia C, Guo Y, Zhang J, Zhang K, et al. A CONSORT-compliant prospective randomized controlled trial: radiation dose reducing in computed tomography using an additional lateral scout view combined with automatic tube current modulation: Phantom and patient study. *Medicine*. 2017;96(30).
345. Kim H, Park CM, Chae H-D, Lee SM, Goo JM. Impact of radiation dose and iterative reconstruction on pulmonary nodule measurements at chest CT: a phantom study. *Diagnostic and Interventional Radiology*. 2015;21(6):459.
346. Silva JS, Silva A, Santos BS. Image denoising methods for tumor discrimination in high-resolution computed tomography. *Journal of digital imaging*. 2011;24(3):464-9.
347. Wu M-T, Pan H-B, Chiang AA, Hsu H-K, Chang H-C, Peng N-J, et al. Prediction of postoperative lung function in patients with lung cancer: comparison of quantitative CT with perfusion scintigraphy. *American Journal of Roentgenology*. 2002;178(3):667-72.

348. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*. 2011;38(2):915-31.
349. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology*. 2009;252(1):263-72.
350. Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, et al. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PloS one*. 2015;10(3):e0118261.
351. Xu Z, Bagci U, Kubler A, Luna B, Jain S, Bishai WR, et al. Computer-aided detection and quantification of cavitory tuberculosis from CT scans. *Medical physics*. 2013;40(11).
352. van Rikxoort EM, de Hoop B, Viergever MA, Prokop M, van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical physics*. 2009;36(7):2934-47.
353. Leader JK, Zheng B, Rogers RM, Sciruba FC, Perez A, Chapman BE, et al. Automated lung segmentation in X-ray computed tomography: development and evaluation of a heuristic threshold-based scheme. *Academic radiology*. 2003;10(11):1224-36.
354. Zhu Y, Tan Y, Hua Y, Zhang G, Zhang J. Automatic segmentation of ground-glass opacities in lung CT images by using Markov random field-based algorithms. *Journal of digital imaging*. 2012;25(3):409-22.
355. Lim JS. *Two-dimensional signal and image processing*. Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p. 1990.
356. Parker JR. *Algorithms for image processing and computer vision*: John Wiley & Sons; 2010.
357. Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. 1991(6):583-98.
358. Meyer F. Topographic distance and watershed lines. *Signal processing*. 1994;38(1):113-25.
359. Grau V, Mewes A, Alcaniz M, Kikinis R, Warfield SK. Improved watershed transform for medical image segmentation using prior information. *IEEE transactions on medical imaging*. 2004;23(4):447-58.
360. Huang Y-L, Chen D-R. Watershed segmentation for breast tumor in 2-D sonography. *Ultrasound in medicine & biology*. 2004;30(5):625-32.
361. Singh L, Dubey R, Jaffery ZA, Zaheeruddin Z, editors. Segmentation and characterization of brain tumor from MR images. *Advances in Recent Technologies in Communication and Computing, 2009 ARTCom'09 International Conference on*; 2009: IEEE.
362. Xu S, Liu H, Song E. Marker-controlled watershed for lesion segmentation in mammograms. *Journal of digital imaging*. 2011;24(5):754-63.
363. Cui Y, Tan Y, Zhao B, Liberman L, Parbhu R, Kaplan J, et al. Malignant lesion segmentation in contrast-enhanced breast MR images based on the marker-controlled watershed. *Medical physics*. 2009;36(10):4359-69.
364. Bellon E, Feron M, Maes F, Van Hoe L, Delaere D, Haven F, et al. Evaluation of manual vs semi-automated delineation of liver lesions on CT images. *European Radiology*. 1997;7(3):432-8.

365. Yan J, Schwartz LH, Zhao B. Semiautomatic segmentation of liver metastases on volumetric CT images. *Medical physics*. 2015;42(11):6283-93.
366. Chan TF, Vese LA. Active contours without edges. *IEEE Transactions on image processing*. 2001;10(2):266-77.
367. Whitaker RT. A level-set approach to 3D reconstruction from range data. *International journal of computer vision*. 1998;29(3):203-31.
368. Ren X, Malik J, editors. Learning a classification model for segmentation. *Proceedings Ninth IEEE International Conference on Computer Vision*; 2003; Nice, France: IEEE.
369. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*. 2000;22(8):888-905.
370. Malik J, Belongie S, Leung T, Shi J. Contour and texture analysis for image segmentation. *International journal of computer vision*. 2001;43(1):7-27.
371. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*. 2012;34(11):2274-82.
372. Stutz D, editor *Superpixel segmentation: an evaluation*. *German Conference on Pattern Recognition*; 2015: Springer.
373. Neubert P, Protzel P, editors. *Superpixel benchmark and comparison*. *Proc Forum Bildverarbeitung*; 2012.
374. Schick A, Fischer M, Stiefelhagen R, editors. Measuring and evaluating the compactness of superpixels. *Pattern Recognition (ICPR), 2012 21st International Conference on*; 2012: IEEE.
375. Mathieu B, Crouzil A, Puel JB, editors. *Oversegmentation Methods: A New Evaluation*. *Iberian Conference on Pattern Recognition and Image Analysis*; 2017: Springer.
376. Ishikawa M, Ahi ST, Murakami Y, Kimura F, Yamaguchi M, Abe T, et al., editors. *Automatic segmentation of hepatocellular structure from HE-stained liver tissue*. *Proc SPIE*; 2013.
377. Balazsi M, Blanco P, Zoroquiain P, Levine MD, Burnier MN. Invasive ductal breast carcinoma detector that is robust to image magnification in whole digital slides. *Journal of Medical Imaging*. 2016;3(2):027501-.
378. Bejnordi BE, editor *A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images*. *Medical Imaging 2015: Digital Pathology*; 2015: International Society for Optics and Photonics.
379. Zhao L, Li K, Wang M, Yin J, Zhu E, Wu C, et al. Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search MRF. *Computers in biology and medicine*. 2016;71:46-56.
380. Tian Z, Liu L, Zhang Z, Fei B. Superpixel-based segmentation for 3d prostate mr images. *IEEE transactions on medical imaging*. 2016;35(3):791-801.
381. Zhao L, Sarikaya D, Corso JJ. Automatic brain tumor segmentation with MRF on supervoxels. *Multimodal Brain Tumor Segmentation*. 2013;51.
382. Ji S, Wei B, Yu Z, Yang G, Yin Y. A new multistage medical segmentation method based on superpixel and fuzzy clustering. *Computational and mathematical methods in medicine*. 2014;2014.
383. Irving B, Cifor A, Papież BW, Franklin J, Anderson EM, Brady M, et al., editors. *Automated colorectal tumour segmentation in DCE-MRI using supervoxel neighbourhood contrast characteristics*. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2014: Springer.

384. Liao X, Zhao J, Jiao C, Lei L, Qiang Y, Cui Q. A segmentation method for lung parenchyma image sequences based on superpixels and a self-generating neural forest. *PLoS one*. 2016;11(8):e0160556.
385. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. *arXiv preprint arXiv:150403967*. 2015.
386. Chu J, Min H, Liu L, Lu W. A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. *Medical physics*. 2015;42(7):3859-69.
387. Stern EJ, Frank MS, Godwin JD. Chest Computed Tomography Display Preferences: Survey of Thoracic Radiologists. *Investigative radiology*. 1995;30(9):517-21.
388. Van de Steene J, Linthout N, de Mey J, Vinh-Hung V, Claassens C, Noppen M, et al. Definition of gross tumor volume in lung cancer: inter-observer variability. *Radiotherapy and oncology*. 2002;62(1):37-49.
389. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*. 1984(6):721-41.
390. Li Y, Sun J, Tang C-K, Shum H-Y, editors. *Lazy snapping*. *ACM Transactions on Graphics (ToG)*; 2004: ACM.
391. Duda RO, Hart PE, Stork DG. *Pattern classification*: John Wiley & Sons; 2012.
392. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. *Slic superpixels*. 2010.
393. Kou F, Li Z, Wen C, Chen W, editors. *Variance adaptive SLIC*. *Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on*; 2016: IEEE.
394. Liu M-Y, Tuzel O, Ramalingam S, Chellappa R, editors. *Entropy rate superpixel segmentation*. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*; 2011: IEEE.
395. Levinshtein A, Stere A, Kutulakos KN, Fleet DJ, Dickinson SJ, Siddiqi K. *Turbopixels: Fast superpixels using geometric flows*. *IEEE transactions on pattern analysis and machine intelligence*. 2009;31(12):2290-7.
396. Yang S, Zhu Y, Wu X, editors. *An Interactive Segmentation Method Based on Superpixel*. *MATEC Web of Conferences*; 2015: EDP Sciences.
397. Hsu C-Y, Ding J-J, editors. *Efficient image segmentation algorithm using SLIC superpixels and boundary-focused region merging*. *Information, Communications and Signal Processing (ICICSP) 2013 9th International Conference on*; 2013: IEEE.
398. Wang S, Lu H, Yang F, Yang M-H, editors. *Superpixel tracking*. *Computer Vision (ICCV), 2011 IEEE International Conference on*; 2011: IEEE.
399. Kim K-S, Zhang D, Kang M-C, Ko S-J, editors. *Improved simple linear iterative clustering superpixels*. *Consumer Electronics (ISCE), 2013 IEEE 17th International Symposium on*; 2013: IEEE.
400. Cheng J, Liu J, Xu Y, Yin F, Wong DWK, Tan N-M, et al. *Superpixel classification based optic disc and optic cup segmentation for glaucoma screening*. *IEEE Transactions on Medical Imaging*. 2013;32(6):1019-32.
401. Mostajabi M, Yadollahpour P, Shakhnarovich G, editors. *Feedforward semantic segmentation with zoom-out features*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2015.
402. Zhang X, editor *SLIC superpixels for efficient graph-based dimensionality reduction of hyperspectral imagery*. *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXI*; 2015: International Society for Optics and Photonics.

403. Chung H, Lu G, Tian Z, Wang D, Chen ZG, Fei B, editors. Superpixel-based spectral classification for the detection of head and neck cancer with hyperspectral imaging. Proceedings of SPIE--the International Society for Optical Engineering; 2016: NIH Public Access.
404. Chinchor N, Sundheim B, editors. MUC-5 evaluation metrics. Proceedings of the 5th conference on Message understanding; 1993: Association for Computational Linguistics.
405. Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. *Strahlentherapie und Onkologie*. 2012;188(2):160-7.
406. Hanna G, Hounsell A, O'Sullivan J. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. *Clinical oncology*. 2010;22(7):515-25.
407. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology*. 2016;121(2):169-79.
408. Gwynne S, Spezi E, Sebag-Montefiore D, Mukherjee S, Miles E, Conibear J, et al. Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case. *The British journal of radiology*. 2013;86(1024):20120398.
409. de Hoop B, Gietema H, van Ginneken B, Zanen P, Groenewegen G, Prokop M. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. *European radiology*. 2009;19(4):800-8.
410. Dewas S, Bibault J-E, Blanchard P, Vautravers-Dewas C, Pointreau Y, Denis F, et al. Delineation in thoracic oncology: a prospective study of the effect of training on contour variability and dosimetric consequences. *Radiation Oncology*. 2011;6(1):118.
411. Louie AV, Rodrigues G, Olsthoorn J, Palma D, Yu E, Yaremko B, et al. Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiotherapy and Oncology*. 2010;95(2):166-71.
412. Lester JF, Nixon L, Mayles P, Mayles H, Tsang Y, Ionescu A, et al. 156 The I-START trial: ISoToxic Accelerated RadioTherapy in locally advanced non-small cell lung cancer. *Lung Cancer*. 2012;75:S51.
413. Persson GF, Nygaard DE, Hollensen C, Munck af Rosenschöld P, Mouritsen LS, Due AK, et al. Interobserver delineation variation in lung tumour stereotactic body radiotherapy. *The British journal of radiology*. 2012;85(1017):e654-e60.
414. Sluimer IC, van Waes PF, Viergever MA, van Ginneken B. Computer-aided diagnosis in high resolution CT of the lungs. *Medical physics*. 2003;30(12):3081-90.
415. Xu Y, Sonka M, McLennan G, Guo J, Hoffman EA. MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. *IEEE transactions on medical imaging*. 2006;25(4):464-75.
416. Sørensen L, Shaker SB, De Bruijne M, editors. Texture classification in lung CT using local binary patterns. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2008: Springer.
417. Gangeh MJ, Sørensen L, Shaker SB, Kamel MS, De Bruijne M, Loog M, editors. A texton-based approach for the classification of lung parenchyma in CT images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2010: Springer.

418. Hoffman EA, Reinhardt JM, Sonka M, Simon BA, Guo J, Saba O, et al. Characterization of the interstitial lung diseases via density-based and texture-based analysis of computed tomography images of lung structure and function. *Academic radiology*. 2003;10(10):1104-18.
419. Lee Y, Seo JB, Lee JG, Kim SS, Kim N, Kang SH. Performance testing of several classifiers for differentiating obstructive lung diseases based on texture analysis at high-resolution computerized tomography (HRCT). *Computer methods and programs in biomedicine*. 2009;93(2):206-15.
420. Park YS, Seo JB, Kim N, Chae EJ, Oh YM, Do Lee S, et al. Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: comparison with density-based quantification and correlation with pulmonary function test. *Investigative radiology*. 2008;43(6):395-402.
421. Depeursinge A, Sage D, Hidki A, Platon A, Poletti P-A, Unser M, et al., editors. Lung tissue classification using wavelet frames. *Engineering in Medicine and Biology Society, 2007 EMBS 2007 29th Annual International Conference of the IEEE; 2007: IEEE*.
422. Huber MB, Nagarajan MB, Leinsinger G, Eibel R, Ray LA, Wismüller A. Performance of topological texture features to classify fibrotic interstitial lung disease patterns. *Medical Physics*. 2011;38(4):2035-44.
423. Zavaletta VA, Bartholmai BJ, Robb RA. High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. *Academic radiology*. 2007;14(7):772-87.
424. Cunliffe A, Armato SG, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *International Journal of Radiation Oncology* Biology* Physics*. 2015;91(5):1048-56.
425. Korfiatis PD, Karahaliou AN, Kazantzi AD, Kalogeropoulou C, Costaridou LI. Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector CT. *IEEE transactions on information technology in biomedicine*. 2010;14(3):675-80.
426. Phillips I, Ajaz M, Ezhil V, Prakash V, Alobaidli S, McQuaid SJ, et al. Clinical applications of textural analysis in non-small cell lung cancer. *The British journal of radiology*. 2017;91(1081):20170267.
427. Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? *Journal of nuclear medicine*. 2013;54(1):19-26.
428. Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *International Journal of Radiation Oncology• Biology• Physics*. 2014;90(4):834-42.
429. Cook GJ, O'Brien ME, Siddique M, Chicklore S, Loi HY, Sharma B, et al. Non-small cell lung cancer treated with Erlotinib: heterogeneity of 18F-FDG uptake at PET—association with treatment response and prognosis. *Radiology*. 2015;276(3):883-93.
430. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*. 2014;5.
431. Charbonnier J-P, Chung K, Scholten ET, Rikxoort EM, Jacobs C, Sverzellati N, et al. Automatic segmentation of the solid core and enclosed vessels in subsolid pulmonary nodules. *Scientific reports*. 2018;8(1):646.

432. Korfiatis P, Kalogeropoulou C, Karahaliou A, Kazantzi A, Skiadopoulos S, Costaridou L. Texture classification-based segmentation of lung affected by interstitial pneumonia in high-resolution CT. *Medical physics*. 2008;35(12):5290-302.
433. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A. MaZda—a software package for image texture analysis. *Computer methods and programs in biomedicine*. 2009;94(1):66-76.
434. Strzelecki M, Szczybinski P, Materka A, Klepaczko A. A software tool for automatic classification and segmentation of 2D/3D medical images. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. 2013;702:137-40.
435. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*. 2006;7(1):91.
436. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*. 2010;11(Jul):2079-107.
437. Esbensen KH, Geladi P. Principles of proper validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*. 2010;24(3-4):168-87.
438. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*. 1989;19(5):1264-74.
439. Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*. 1978;8(6):460-73.
440. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. *IEEE transactions on medical imaging*. 2009;28(3):374-83.
441. Markel D, Caldwell C, Alasti H, Soliman H, Ung Y, Lee J, et al. Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT. *International journal of molecular imaging*. 2013;2013.
442. He Z, Yu W. Stable feature selection for biomarker discovery. *Computational biology and chemistry*. 2010;34(4):215-25.
443. Cui H, Wang X, Fulham M, Feng DD, editors. Prior knowledge enhanced random walk for lung tumor segmentation from low-contrast CT images. *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE; 2013: IEEE*.
444. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*. 2004;23(7):903-21.
445. Huo J, Okada K, van Rikxoort EM, Kim HJ, Alger JR, Pope WB, et al. Ensemble segmentation for GBM brain tumors on MR images using confidence-based averaging. *Medical physics*. 2013;40(9).
446. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Medical physics*. 2017;44(12):6377-89.
447. Hänsch A, Schwier M, Gass T, Morgas T, Haas B, Klein J, et al., editors. Comparison of different deep learning approaches for parotid gland segmentation from CT images. *Medical Imaging 2018: Computer-Aided Diagnosis; 2018: International Society for Optics and Photonics*.
448. Ibragimov B, Toesca D, Chang D, Koong A, Xing L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. *Physics in Medicine & Biology*. 2017;62(23):8943.

449. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical physics*. 2017;44(2):547-57.
450. Kazemifar S, Balagopal A, Nguyen D, McGuire S, Hannan R, Jiang S, et al. Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning. *arXiv preprint arXiv:180209587*. 2018.
451. Cha KH, Hadjiiski L, Samala RK, Chan HP, Caoili EM, Cohan RH. Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. *Medical physics*. 2016;43(4):1882-96.
452. Fechter T, Adebahr S, Baltas D, Ben Ayed I, Desrosiers C, Dolz J. Esophagus segmentation in CT via 3D fully convolutional neural network and random walk. *Medical physics*. 2017;44(12):6341-52.
453. Tegzes P, Rádics A, Csernai E, Ruskó L. PO-1004: Machine learning methods for automated OAR segmentation. *Radiotherapy and Oncology*. 2017;123:S553-S4.
454. Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori AV, Criminisi A, et al., editors. *DeepMedic for brain tumor segmentation. International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; 2016: Springer.
455. Cardenas CE, McCarroll RE, Elgohari BA, Elhalawani H, Fuller CD, Kamal MJ, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *International Journal of Radiation Oncology* Biology* Physics*. 2018.
456. Li W, Jia F, Hu Q. Automatic segmentation of liver tumor in CT images with deep convolutional neural networks. *Journal of Computer and Communications*. 2015;3(11):146.
457. Wang Y, Zu C, Hu G, Luo Y, Ma Z, He K, et al. Automatic Tumor Segmentation with Deep Convolutional Neural Networks for Radiotherapy Applications. *Neural Processing Letters*. 2018:1-12.

Appendix

Appendix A

Appendix A.1 Plots for k-nearest neighbours classification optimisation using multiple texture feature set

Further to chapter 6 section 6.13.4, the plots for optimisation of the k-NN classification in the respective cross-validation folds using the multiple texture set is shown here. The effect of different neighbour size on the k-NN classification is displayed in figure A.1 where the three runs showed similar trends. For all three folds, better classification was achieved with odd values of neighbourhood size at less than 10, beyond which the error rate increases, which is likely due to increased smoothing effect of the classification boundary.

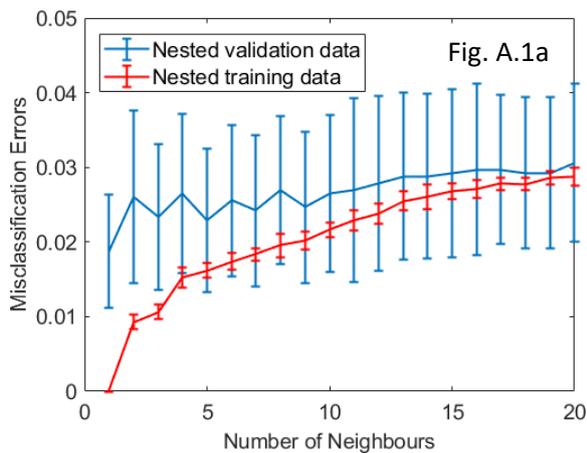
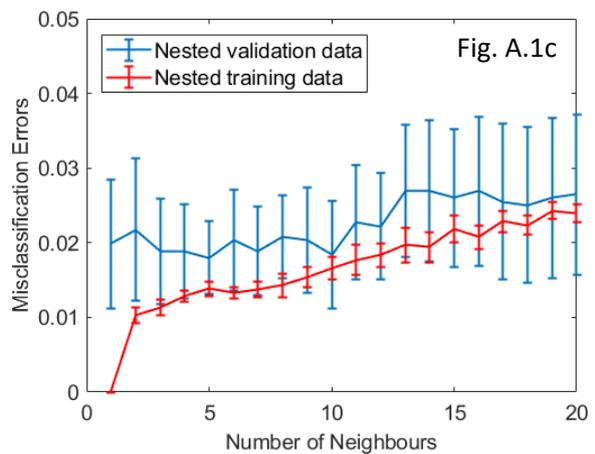
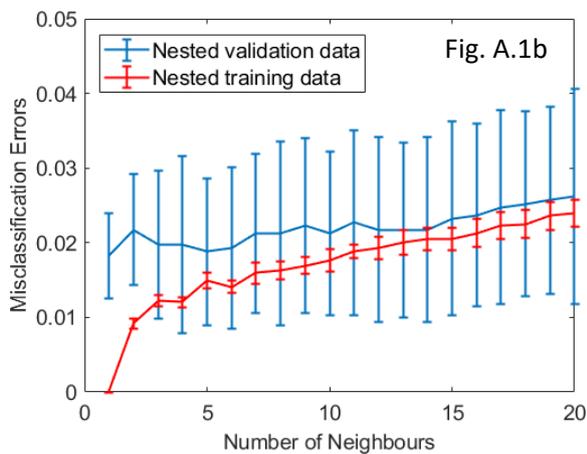


Figure A.1. Mean misclassification errors using k-nearest neighbours classification (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size using multiple texture features for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.



From the sensitivity and specificity plots in figure A.2, the higher error rates at neighbourhood sizes greater than 10 seem to be associated with a corresponding reduction in sensitivity, rather than changes to specificity.

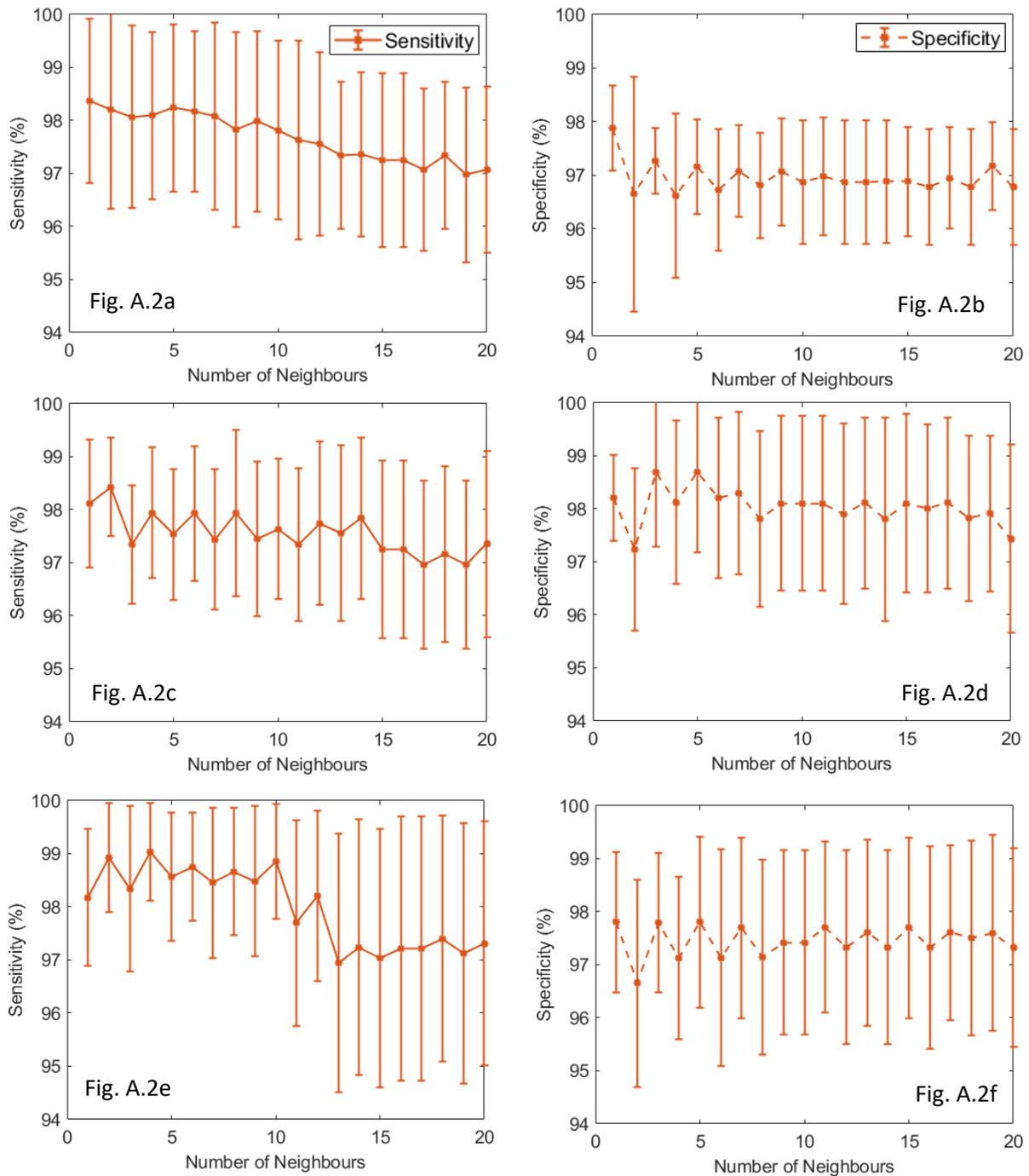


Figure A.2. Mean sensitivity and specificity plots nested validation data using k-nearest neighbours classification (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size with multiple texture features for classification. a, b) Outer cross-validation run 1; c, d) Outer cross-validation run 2; e, f) Outer cross-validation run 3.

Appendix A.2 Plots for k-nearest neighbours classification optimisation using most discriminatory texture feature Sum Variance (0,4)

The plots for optimisation of the k-NN classification in the respective cross-validation folds using sum variance (0,4) is shown here (see chapter 6 section 6.14.1.1). Training of the k-NN classifier based on the values of the sum variance (0,4) revealed that there was a trend towards lower misclassification errors for increasing neighbourhood size (figure A.3). Run 2 had lower classification errors as compared to runs 1 and 3, suggesting that there might have been some sampling bias despite the use of cross-validation. Despite this, there appears to be stability of the parameter around and beyond $k = 10$, leading to similar error estimates as well as small variance. Higher sensitivity scores were achieved than specificity across the parameter range.

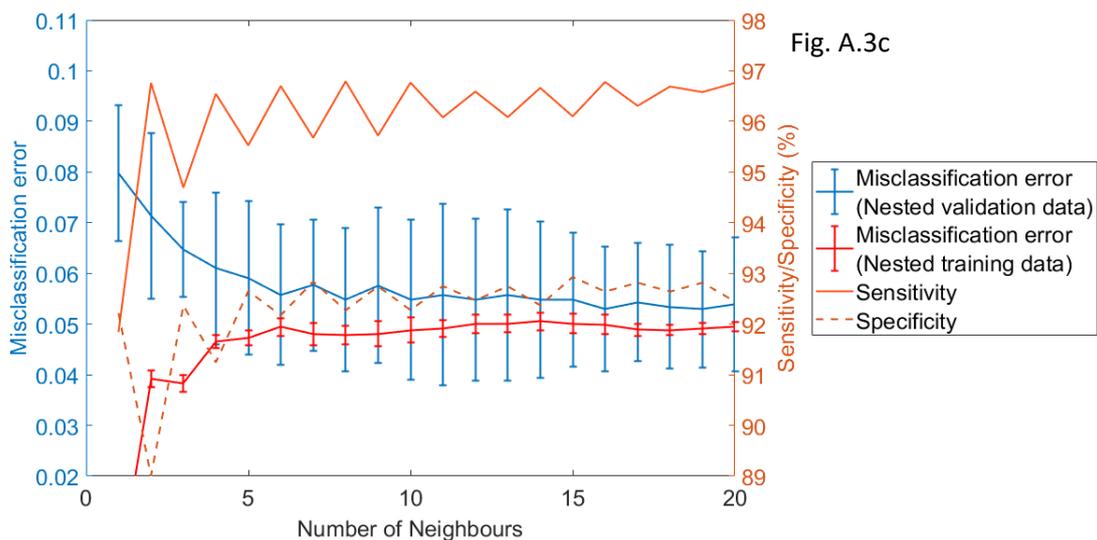
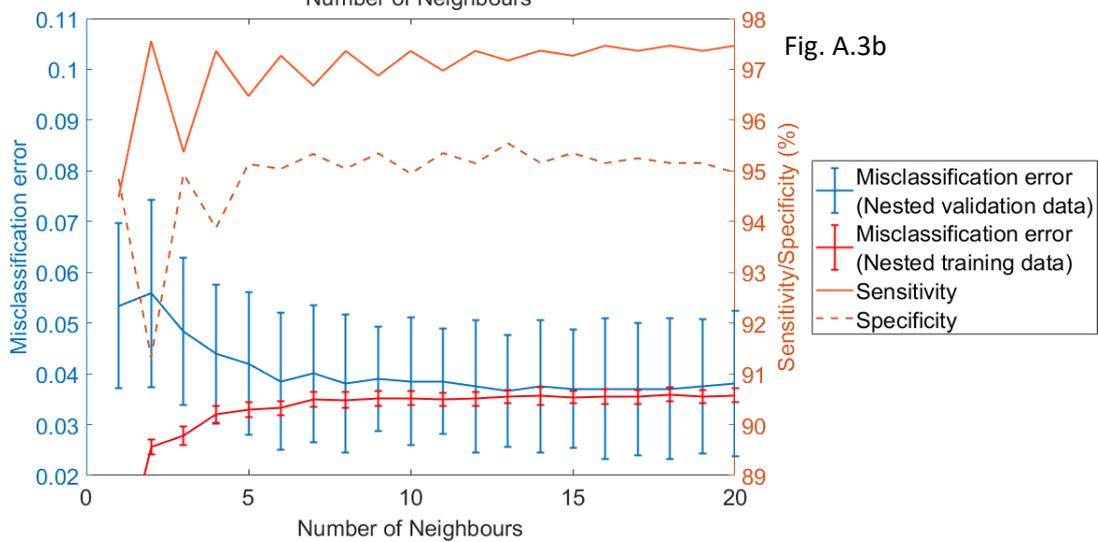
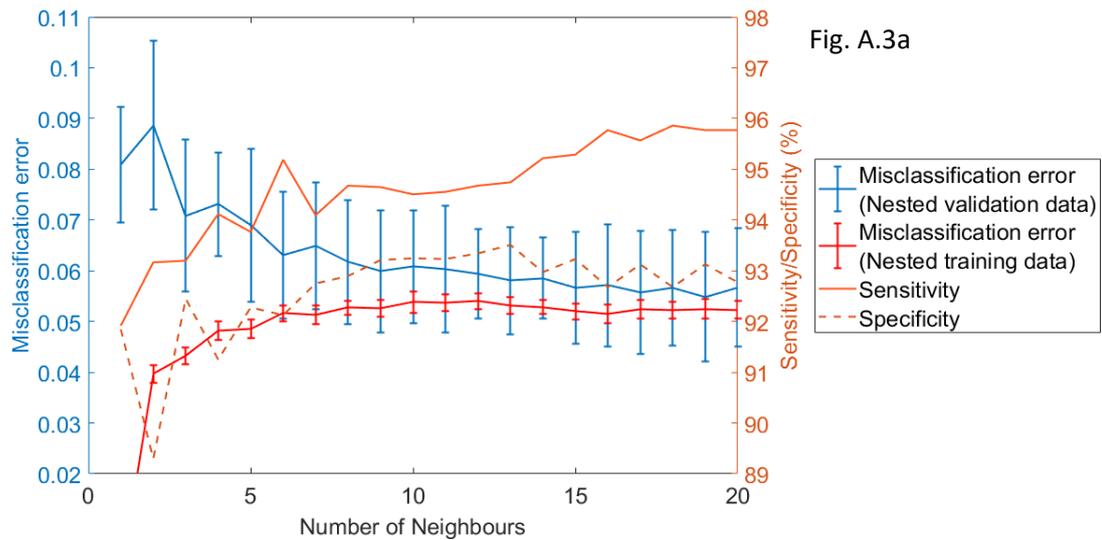


Figure A.3. Performance of k-nearest neighbour classifier (nested 10-fold cross-validation with nested standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both nested training and nested validation data), as well as the mean sensitivity and specificity of the nested validation data using sum variance (4-pixel distance and 45° direction) for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.

Appendix A.3 Plots for k-nearest neighbours classification optimisation using mean as single feature

The plots for optimisation of the k-NN classification in the respective cross-validation folds using mean as the single feature for discrimination is shown here (see chapter 6 section 6.14.2.1). Neighbour sizes above 7 resulted in the lower misclassification errors (figure A.4). Small variance with good error estimates were observed, though the error rates were higher than the previous two classifiers. The nested training and validation curves were seen to converge at higher neighbourhood sizes, though again, parameter stability was present. Converse to sum variance (0,4), specificity scores were higher than sensitivity.

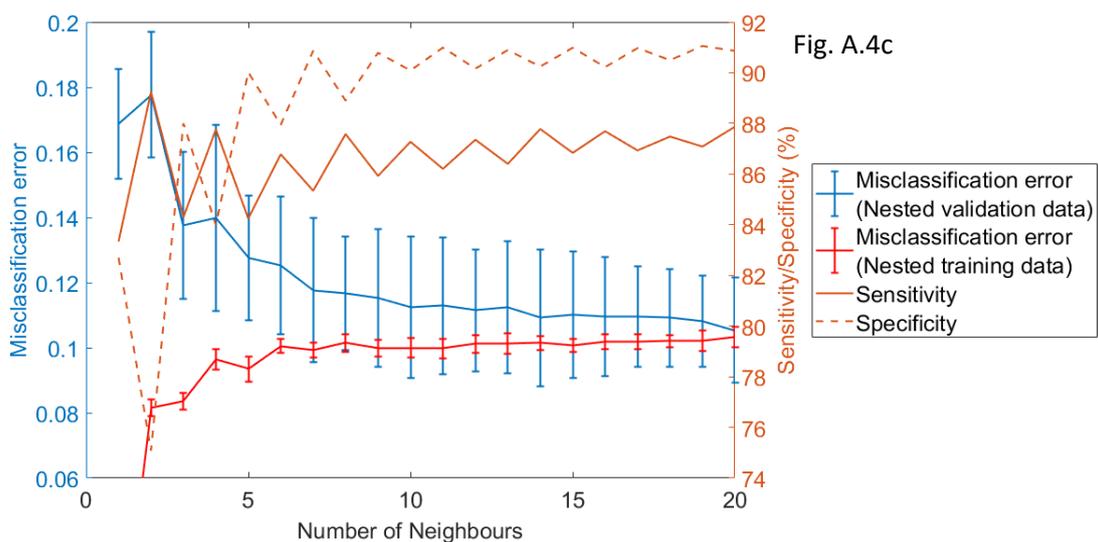
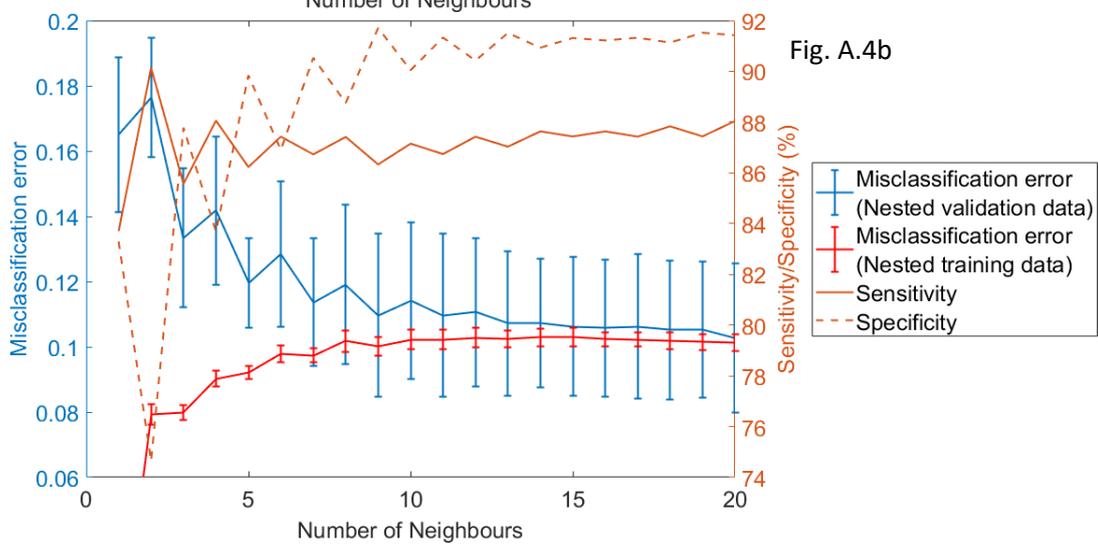
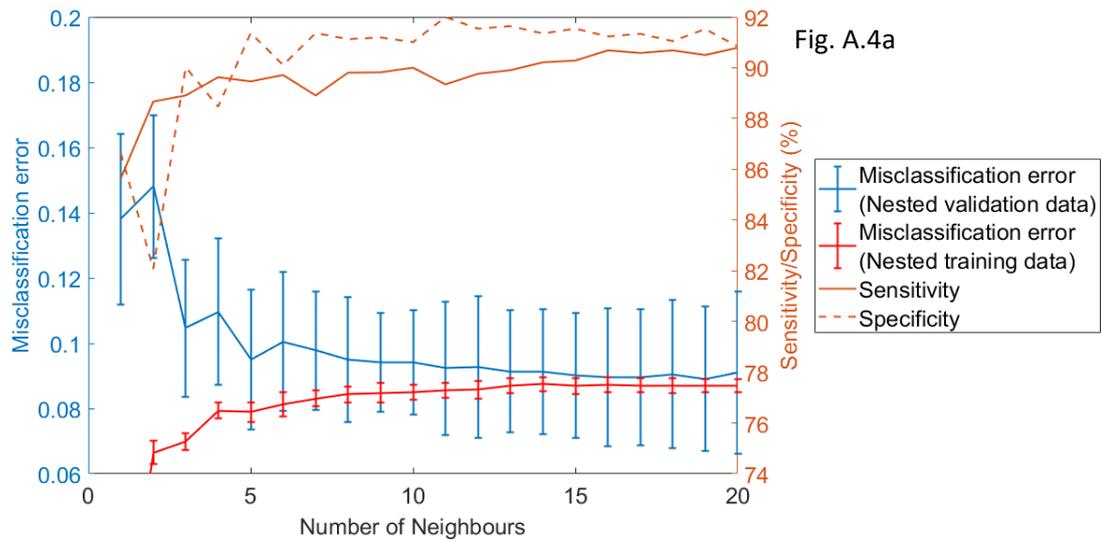


Figure A.4. Performance of k-nearest neighbour classifier (nested 10-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both nested training and nested validation data), as well as the mean sensitivity and specificity of the nested validation data using mean values for classification. a) Outer cross-validation run 1; b) Outer cross-validation run 2; c) Outer cross-validation run 3.

Appendix B

Appendix B.1 Plots for re-optimisation of final k-nearest neighbour classification models

This section pertains the re-optimisation of the final k-NN classification models for chapter 6 section 6.14.3.2.

For the multiple texture feature set, higher variance was seen as compared to the nested cross-validation training results. Although the lowest misclassification error ($2.79 \pm 0.98\%$) was seen at a neighbourhood size of 4, the variance was larger as compared to the next lowest error seen at a size of 5. Additionally, odd number of neighbours is preferred over even numbers with a 2-group classifier, to avoid ties in the classification, especially at smaller neighbourhood sizes.

Stability of the parameters was seen again for sum variance (0,4). Similarly, $k = 17$ was chosen as the final parameter rather than $k = 16$ despite its lowest classification error at $4.91 \pm 1.65\%$). As for classification based on the mean value, $k = 19$ was selected as the parameter used in the final model.

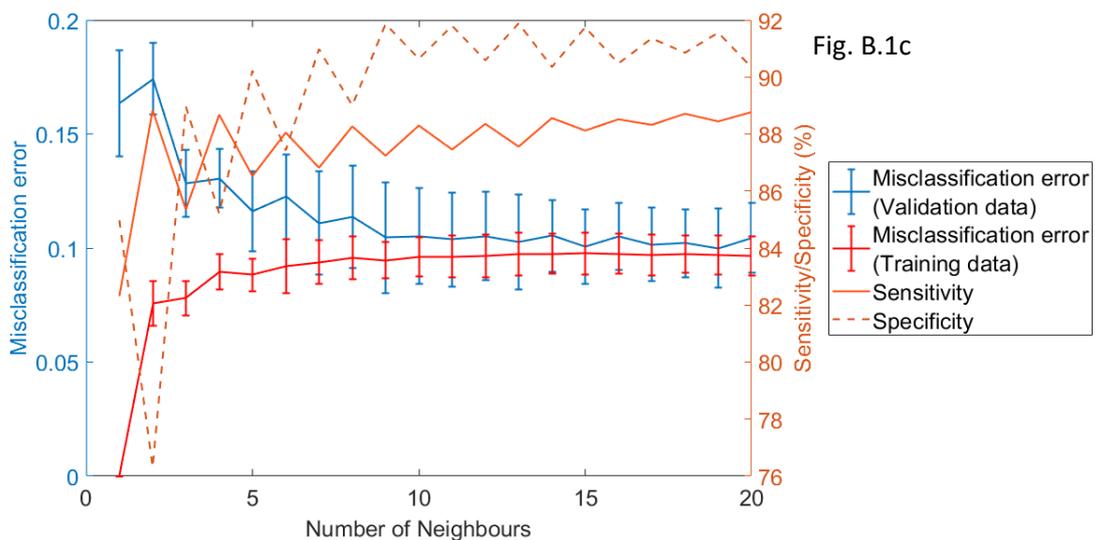
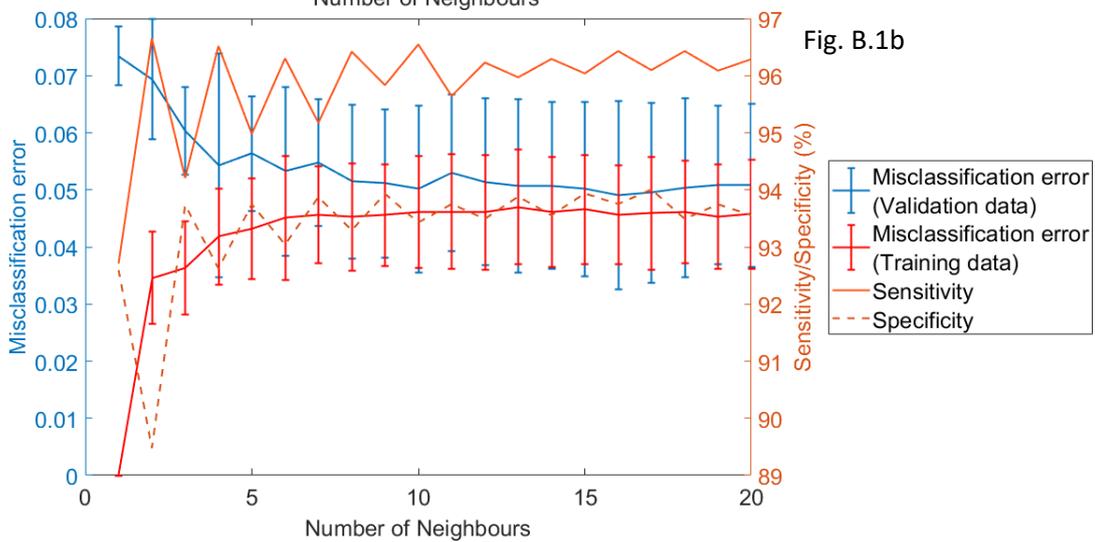
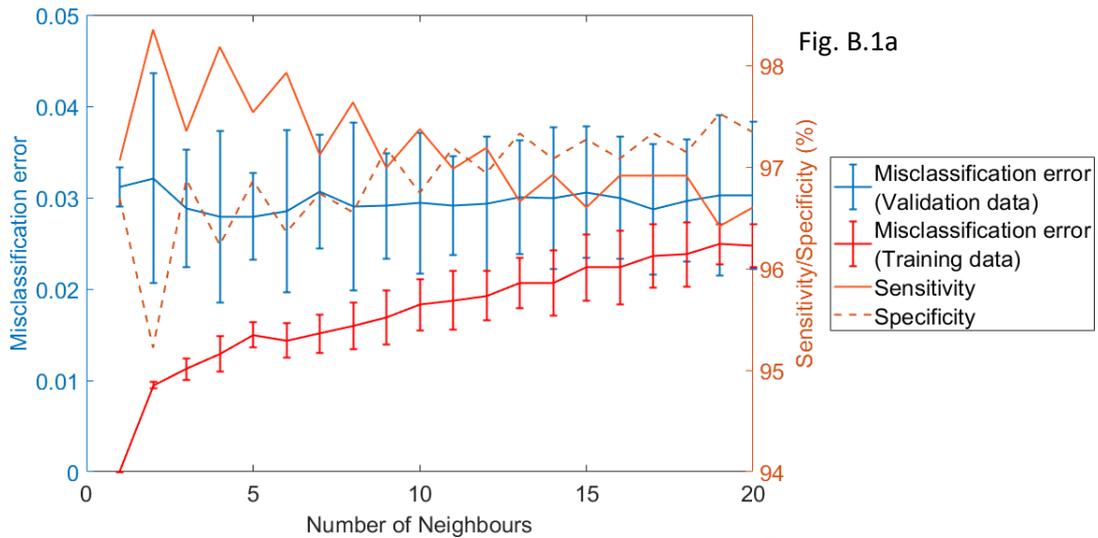


Figure B.1. Performance of k-nearest neighbour classifier (3-fold cross-validation with standard deviation as error bars) as a function of neighbourhood size, displaying results of the mean and standard deviation of the misclassification error (both training and validation data), as well as the mean sensitivity and specificity of the validation data. Feature set used for classification: a) Multiple texture features; b) Sum variance (4-pixel distance and 45° direction); c) Mean values.