IMPERIAL

Hamlyn Centre
for Robotic Surgery

Computer Vision for Multimodal Surgical Guidance in Brain Tumour Resection

Author: Alfie Roddan

Submitted: 3rd October 2025

Supervisors: Dr. S. Giannarou

Prof. Daniel S. Elson

A thesis submitted for the degree of *Doctor of Philosophy*

The Hamlyn Centre for Robotic Surgery

Department of Surgery and Cancer

Imperial College London

The Hamlyn Centre,
Bessemer Building,
Imperial College London,
180 Queens Gate,
South Kensington,
SW7 2AZ

Alfie Roddan © 2025

For Beatrice

Declaration

I hereby certify that the work presented in this thesis is my own original research. All sources and contributions from others have been appropriately acknowledged. Any materials used in the development of this thesis have been cited as required.

The formatting for this thesis is derived from the Thesis Template by R. Robinson at https://github.com/mlnotebook/thesis_template.

Alfie Roddan
October 2025

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EP/T51780X/1) and Intel. Without their support, this research would not have been possible, and I am grateful for the opportunities they have provided me.

To the Intel Geti team: I always looked forward to our catch-ups. Samet, Laurens, Greg and Ollie not only followed my research journey closely but also helped shape it with patience, encouragement, and constructive guidance.

To my supervisors, Matina and Dan: thank you. This PhD has been more than an academic education. I have grown personally over the years, and as our collaboration developed, I believe we truly made an impact with our work.

To my research group, past and present—Chi, Haozheng, Yihang, Alistair, Sam, and João: I have had so much fun working (and not working) with you all. Our discussions, collaborations, and laughs have made this journey far more enjoyable.

To the Hamlyn Centre team—Tobi, Maria, Giulio, and my favourite scientist, Vadzim: I have been lucky to learn from your wisdom and experience, which have guided me throughout this journey.

To the Pecovers and co.: William, Lucy, Matilda, Alex, Raffi, Hugh, Lucy, Rufus, and Phoebe. Some of my most peaceful moments were spent in the garden with you. Long live the commune.

To my friends: you are all my inspiration. I could not have done this without you.

To Beatrice, I like to think parts of this thesis are yours. Maybe some paragraphs you recognise correcting or results you remember me celebrating. We finally did it!

Finally, to my mother, father, and sister: this was just a small stretch of a long road, and you have always had my back along the way. I love you and I am forever grateful.

Abstract

Medical devices play a central role in neurosurgery, where millimetre-scale accuracy is essential for preserving brain function and guiding surgical decisions. Despite the wide use of navigation and imaging tools, few provide real-time, multi-scale tissue characterization. This thesis investigates a multimodal guidance framework that combines macroscopic hyperspectral imaging with probe-based confocal laser endomicroscopy for cellular-level detail. The work advances each modality individually and explores their combination within a broader framework, drawing on photonics, deep learning, and computer vision to improve data acquisition, interpretation, and integration across scales.

Contents

1 Introduction **24**

1.1 Overview . . . 25
1.2 Contributions . . . 33
1.3 Publications . . . 34
1.3.1 Related Publications . . . 34
1.3.2 Other Publications . . . 34

2 Probe-based Confocal Laser Endomicroscopy for Intraoperative Tissue Characterisation **36**

2.1 Overview . . . 37
2.2 Introduction . . . 38
2.3 Related Works . . . 41
2.3.1 Deep Learning Explainability . . . 41
2.3.2 Pixel Attribution . . . 41
2.3.3 Uncertainty . . . 43
2.4 Methodology . . . 47
2.5 Experiments and Analysis . . . 49
2.5.0.1 Datasets . . . 49
2.5.0.2 Implementation . . . 49
2.5.0.3 Evaluation Metrics . . . 50
2.5.0.4 Performance Evaluation . . . 51
2.5.1 T Study . . . 55
2.5.2 Clinical Use . . . 55
2.6 Conclusion . . . 60

3 Hyperspectral Segmentation and Augmentation **61**

3.1 Overview . . . 62
3.2 Introduction . . . 63
3.2.1 Hyperspectral Imaging: Light-Tissue Interactions . . . 63
3.2.1.1 Formation of a Hyperspectral Data Cube . . . 64
3.2.2 Calibration . . . 65
3.2.3 Hyperspectral Segmentation . . . 66
3.2.3.1 Spectral Similarity . . . 66

3.2.3.2 Pixel-wise Spectral Classification . . . 67
3.2.3.3 Automatic Segmentation Models . . . 67
3.3 Related Works . . . 69
3.4 Methodology . . . 72
3.4.1 Geometric Augmentation . . . 73
3.5 Results . . . 77
3.5.1 Dataset . . . 77
3.5.2 Segmentation Models . . . 78
3.5.3 Performance Evaluation Metrics . . . 79
3.5.4 Data Augmentations . . . 80
3.5.5 Comparative Analysis . . . 82
3.6 Discussion . . . 84
3.6.1 Clinical Usage of Hyperspectral Segmentation . . . 85
3.7 Conclusion . . . 87

4 Interactive Hyperspectral Segmentation 88
4.1 Overview . . . 89
4.2 Introduction . . . 90
4.3 Methodology . . . 92
4.3.1 Evaluation Protocol . . . 92
4.3.2 Spectral Comparison Functions . . . 93
4.3.3 RGB Interactive Segmentation . . . 94
4.3.4 RGB and Spectral Fusion . . . 94
4.3.5 SAMSA . . . 95
4.3.6 Training Protocol . . . 96
4.4 Results . . . 98
4.4.1 Datasets . . . 98
4.4.2 Metrics . . . 99
4.4.3 Spectral Comparison Threshold Study . . . 99
4.4.4 Individual Dataset Study . . . 100
4.4.5 Clicks and Shots . . . 104
4.4.6 Generalisation Results . . . 104
4.4.7 Mixed Training . . . 105
4.5 Clinical Usage of SAMSA . . . 108
4.6 Conclusion . . . 109

5 Interactive Hyperspectral Segmentation Extended 110
5.1 Introduction . . . 111
5.2 Methodology . . . 112

5.2.1 Spectral Comparison Functions . . . 112
5.2.2 Spatial Models . . . 112
5.2.3 Spectral-Spatial Models . . . 113
5.2.4 Training Protocol . . . 114
5.2.5 Automatic Segmentation . . . 115
5.3 Results . . . 118
5.3.1 Datasets . . . 118
5.3.2 Individual Dataset Study . . . 118
5.3.3 Mixed Training . . . 119
5.3.4 Automatic Segmentation . . . 125
5.4 Clinical Usage of SAMSA2.0 . . . 127
5.5 Conclusion . . . 128

6 Multimodal Imaging Platform 129
6.1 Overview . . . 130
6.2 Introduction . . . 131
6.3 Multimodal Imaging Platform Design . . . 132
6.3.1 Calibration of Imaging Modalities . . . 138
6.3.2 Spatial Alignment of Multimodal Data . . . 138
6.4 *Ex vivo* Dataset . . . 142
6.5 Multimodal Tumour Identification . . . 145
6.6 Results . . . 146
6.6.1 Macbeth Colour Validation . . . 146
6.6.2 Timing Measurements . . . 146
6.6.3 Spatial Calibration . . . 146
6.6.4 Feasibility Test with Dye Staining . . . 148
6.6.5 Tumour Identification . . . 151
6.7 Discussion & Conclusion . . . 154

7 Conclusions 155
7.1 Main Conclusions . . . 156
7.2 Limitations . . . 159
7.3 Further Work . . . 161

References 163

Appendices 176

A Permission Documents 177

A.0.1 Explainable Image Classification with Improved Trustworthiness for Tissue Characterisation . . . 177
A.0.2 Open Access Springer . . . 182
A.0.3 Wiley Open Access License . . . 182
A.0.4 SAMSA: Segment Anything Model Enhanced with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation . . . 183

B Deep Learning Basics **186**
B.0.1 The Neural Network . . . 186
B.0.2 Convolutional Neural Networks (CNN) . . . 188
B.0.3 Transformer and Attention . . . 189
B.0.4 Vision Transformer . . . 191

C Transformer Explainability **192**

D Bayes' Rule **195**

E Point Estimates **196**
E.1 System of Linear Equations . . . 197
E.2 Ordinary Least Squares . . . 198
E.3 Maximum Likelihood Estimation . . . 200
E.4 Maximum a Posteriori Estimation . . . 203
E.5 Point Estimates . . . 205

F Gaussian Processes **206**
F.1 Posterior Prediction with GPs . . . 208
F.2 Illustrative Example . . . 209

G Extended Chapter 2 Results **211**

List of Figures

1.1 Global heatmap of age-standardized incidence rates for cancers of the brain and central nervous system in 2022. The map displays the age-standardized incidence rate (ASR, World) per 100,000 population for both sexes combined. Data source: GLOBOCAN 2022, International Agency for Research on Cancer, WHO [87]. 25

1.2 Patient Care Lifecycle for Tumour Resection: A patient-centred overview of the surgical pathway, from initial diagnosis and imaging through preoperative planning, surgery with intraoperative assessment, postoperative care and histopathology. . 27

1.3 Side-by-side comparison of brain tumour representations in different modalities. (A) MRI image of a huge meningioma by Fowle, (B) histopathological section of a psammomatous meningioma by Jensflorian, and (C) CT scan showing a pineal gland tumour by Cerevisae. All images licensed under CC BY-SA 4.0 via Wikimedia Commons. Sources: Fowle, Jensflorian, Cerevisae. 29

1.4 Photograph of an operating theatre featuring an operating microscope, surgical bed, overhead surgical lamps, and life support devices. Photo by Scott & White Healthcare, licensed under CC BY-NC-ND 2.0, via . Accessed April 15, 2025. . . . 29

2.1 Example pCLE data a) Astrocytoma b) pCLE probe c) Glioblastoma d) Meningioma. 38

2.2 Multiple pixel attribution methods run on an un-robust model to shown spurious correlation of crack visible in MNG videos. Red annotation shows salient region psammoma body for the MNG class. 42

2.3 Comparison of PA methods on pCLE classification using ResNet18. For each method, the left column shows the softmax predictions for the three classes (MNG, IDC, and GBM); the middle column displays the input image (MNG test sample); and the right column presents the corresponding PA map. 44

2.4 Comparison of ScoreCAM and ReciproCAM PA methods on pCLE classification using ResNet18. Layout follows the same format as fig. 2.3, showing softmax predictions, input images, and PA maps. 45

2.5 Outline of the proposed method. A PA volume is generated using T forward passes of a CNN model with Dropout applied. $\boldsymbol{X}$ represents the input image, $\hat{\boldsymbol{Y}}$ the output logits of the CNN model and $\boldsymbol{S}$ the volume of PA maps. 47

2.6 Each PA map of proposed method pre-interpolation to image size (for each iteration of T) from Layer 4 of ResNet18 for T=10 on pCLE data. Yellow represents the highest value and dark blue is lowest. 48

2.7 PA maps generated using ResNet18 on MNG pCLE images. a) Grad-CAM++ PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of MNG with the salient region highlighted with red bounding box f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest. 53
2.8 PA maps generated using ResNet18 on GBM pCLE images. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of GBM f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest. 53
2.9 PA maps generated using ResNet18 on invasive ductal carcinoma (IDC) pCLE images. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of IDC with the salient region highlighted with red bounding box f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest. 54
2.10 PA maps generated using ResNet50 on ImageNet. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original image f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest. 54
2.11 ADCC vs T parameter study for ResNet-18 and MobileNetV2 on pCLE dataset. . . 55
2.12 Comparison of Pixel Attribution (PA) methods for pCLE classification using ResNet18. Column 1: softmax predictions (box plots after MC dropout); Col 2: input image; Col 3: PA map; Col 4: risk estimation using MC dropout (100 runs). 56
2.13 Comparison of additional Pixel Attribution (PA) methods for pCLE classification using ResNet18. Format follows fig. 2.12, featuring softmax box plots, input images, PA maps, and risk estimation via MC dropout (100 runs). 57

3.1 Light tissue interaction diagram. A)-D) represent hyperspectral sensor modes. . . 63
3.2 Configuration of a hyperspectral imaging system for surgical applications. The left panel illustrates a common acquisition setup, featuring a hyperspectral camera mounted on an articulated support or robotic arm and positioned above the surgical area to capture spectral data. Illumination for the scene is provided by either dedicated hyperspectral lighting systems or standard broadband light sources. The right panel depicts the standard calibration process. Permission to reuse this figure can be found in section A.0.3. 72
3.3 An illustration of the macro types of augmentation from a sample of the Hib dataset. 74

3.4 An illustration of the data loading and augmentation pipeline. The process involves an initial calibration of the hyperspectral data, followed by an augmentation step. Our proposed method introduces a key modification at step 1, where the white reference itself is augmented prior to being used for calibration [159]. Permission to reuse this figure can be found in section A.0.3. 76
3.5 a) Mean reflectance spectra per class of the pre-processed dataset across the channels. b) Standard deviation per class of the spectra in the pre-processed dataset across the channels. 77
3.6 Visualisation of the Calibration-Jitter augmentation on a sample image. (a) The original RGB image with points of interest labelled. (b) The corresponding ground truth annotation mask. (c-e) A comparison of the pseudo-RGB image and the spectra from the labelled points under different jitter coefficients: (c) a positive coefficient of 0.1, (d) a zero coefficient representing standard white-dark calibration, and (e) a negative coefficient of -0.1. The pseudo-RGB is formed using the closest available wavelengths to 630 nm (red), 540 nm (green), and 480 nm (blue). (f-h) The effect of high (0.1), low (-0.1), and zero jitter coefficients on the spectral profiles of different tissue classes, shown alongside the mean and standard deviation for each class across the entire dataset. The classes are: (f) normal tissue, (g) tumour, and (h) blood vessel [159]. Permission to reuse this figure can be found in section A.0.3. 81
3.7 Qualitative comparison of SegFormer outputs showing the effect of data augmentation. Columns: ground truth, RGB image, normal augmentation (without *Calibration-Jitter*), and with *Calibration-Jitter*. The latter reduces false positive tumour predictions, especially under challenging lighting [159]. Reuse details in section A.0.3. Permission for the reuse of this figure is detailed in section A.0.3. . . . 83
3.8 Results from a spectral classification SVM model following the same implementation from [103]. a) Ground truth label. b) RGB image of target. c) SVM prediction of a). 85

4.1 An outline of SAMSA, where a single click within the pseudo-RGB image is utilised to guide both the RGB and the spectral branches of the model [161]. Permission for the reuse of this figure is detailed in section A.0.4. 96
4.2 An example of the similarity map produced by the equalised Spectral Angle ($SA_{Equalized}$) function. The target anatomy is clearly distinguished from the background, highlighting the discriminative power of spectral information. 100
4.3 Qualitative results of SAM2$_{Base}$ on the Hib dataset. Left: pseudo-RGB with click prompt (blue circle). Right: output segmentation map. 103

4.4 Qualitative segmentation results on Hib (top) and Heipor (bottom). (a,g) Input RGB with user click; (b,h) ground truth (red: Tumour, blue: Vasculature, green: Healthy, black: Background); (c,i) SAMSA segmentation; (d-f) probability maps from SAM2$Tuned$, $SAEqualized$, and SAMSA, showing improved localisation [161]. Permission for the reuse of this figure is detailed in section A.0.4. 103
4.5 Performance analysis on the Hib dataset. (a) Macro D@0.5 score as a function of the number of user clicks, comparing SAMSA against the fine-tuned baseline. (b) Single-click Macro D@0.5 score in relation to the number of training examples (shots), illustrating the advantage conferred by spectral information in low-data regimes [161]. Permission for the reuse of this figure is detailed in section A.0.4. . 104
4.6 Qualitative results on the Heipor dataset. Performance gains are modest, as the segmentation task is relatively simple for RGB-based models. 106
4.7 Qualitative results on the Hib dataset. SAMSA successfully uses spectral information to locate the tumour just outside the visually prominent ring, overcoming the limitations of the RGB-only models. 107

5.1 SAMSA2.0 Architecture 113
5.2 Sample segmentation results on the Hib dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity 121
5.3 Sample segmentation results on the SB-X dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity 122
5.4 Sample segmentation results on the SB-H dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity 123
5.5 Sample segmentation results on the Heipor dataset, liver class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity 123
5.6 Example use case on Hib dataset, finding the boundary of tumour and non-tumour. A) Pseudo-RGB image of the target with tumour prompt. B) Ground truth labels. C) SAMSA2.0-Large standard segmentation from a). D) $SA_{Equalized}$ spectral similarity from a). E) Pseudo-RGB image of the target with healthy prompt. F) SAM2-Large-FT using SAM2 predicted similarity 124

5.7 Automatic segmentation results on the Hib dataset. A) Pseudo-RGB image of the target. B) Ground truth labels. C) SAM2-Large standard automatic segmentation. D) SAM2-Large-FT using SAM2 standard automatic segmentation as input. E) SAM2-Large-FT using our light automatic segmentation. F) SAM2-Large-FT using our heavy automatic segmentation. G) SAMSA2.0-Large-FT using our light automatic segmentation. H) SAMSA2.0-Large-FT using our heavy automatic segmentation. 126
5.8 Example of interaction following automatic segmentation. A) Pseudo-RGB image of the target, with a blue click point. B) SAM2-Large-FT using our heavy automatic segmentation; the target-selected region is highlighted in yellow. C) Output segmentation from SAM2-Large-FT using the prompt in (a). D) Ground truth labels. E) SAMSA2.0-Large-FT using our heavy automatic segmentation; the target-selected region is highlighted in yellow. F) Output segmentation from SAMSA2.0-Large-FT using the prompt in (a). 126

6.1 Our multimodal platform integrates RGB, HSI, and pCLE. The lamb brain phantom was imaged macroscopically using HSI and RGB, both integrated into a Leica microscope. The pCLE, operated manually, performs microscale optical biopsies around the tumour margins. By tracking the pCLE, we can correlate its data with the HSI modality using a tracking target attached to the pCLE probe [160]. Permission for the reuse of this figure is detailed in section A.0.2. 132
6.2 Two views of one of the side optical ports on a LEICA M525 OH4 microscope. . . 133
6.3 Schematic of the dust cap adapter that was manufactured to attach to the side dust cap of an optical port on a LEICA M525 OH4. 134
6.4 Two views of one of the side optical ports on a LEICA M525 OH4 microscope with our adapter attached. 134
6.5 An example of the optical pathway for a camera attached to the port using our custom adapter. Measurements are approximate. 135
6.6 The Macbeth ColorChecker chart from calibrite, used as a calibration target. . . . 136
6.7 Example images of the pCLE holder design and its in situ application. 137
6.8 Methodology for calibration and analysis of re-projection error: a) The stereo calibration process between the RGB and HSI cameras, where R and T represent the rotational and translational transformations determined during calibration. b) An illustration of the pivot-based calibration, highlighting the alignment points between the different imaging systems. Here, R_{probe} and T_{probe} denote the rotation and translation from the marker to the probe tip, which are computed during this procedure. c) A depiction of how the re-projection error is calculated, showing the spatial offset between corresponding points in the RGB and hyperspectral images [160]. Permission for the reuse of this figure is detailed in section A.0.2. 139

6.9 Example output images of the pCLE probe tip pose estimation. We have overlayed the estimated probe tip in red, a green cylinder representing the metal tip and the probe world axis on the first keydot. 140
6.10 An overview of the four-stage calibration procedure required before an experiment. The steps are: (1) Stereo calibration to co-register the RGB and HSI cameras; (2) Probe-tip calibration to find the transformation between the tracking marker and the probe tip; (3) White-dark referencing for spectral consistency; and (4) Background subtraction for pCLE measurements. 141
6.11 Preparation of the *Ex vivo* dataset. 143
6.12 The five brain phantoms. Left HSI viewpoint of sample. Right. Annotations in HSI viewpoint. 144
6.13 Macbeth colour checker results (Calibrite). Measured values are compared at each wavelength with reference data from the Munsell Color Science Laboratory, available at https://www.rit.edu/science/munsell-color-science-lab-educational-resources. 147
6.14 Timing measurements between probe and RGB timestamps. Left: global timing differences across all poses in one experiment. Right: local differences showing a mean error of 5.53 ms. 148
6.15 A lamb brain specimen with one half stained with acriflavine. The absorption between 420–500 nm matches the acriflavine absorption spectrum shown in the bottom left. 149
6.16 Live demonstration on a lamb brain specimen, half-stained with acriflavine. The 3D axis (left) shows probe position in world coordinates. The grid shows four synchronized views: top-left the live RGB feed, top-right the pre-scan HSI with probe position, bottom-left the live pCLE image, and bottom-right the spectral data from the probe tip. Colours are consistent across the world coordinates, spectra, and HSI view. 150
6.17 Example results during procedure. a) The macro tumour prediction Softmax logits, giving an indication of tumour location. b) Still frame from video at beginning of scanning. Blue bar on the pCLE frame indicates non-tumour classification c) Still frame at end of scanning with convex hull of pCLE points which are classified as positive. Red bar on pCLE frame indiciates tumour classification 153
6.18 An illustrative overview of the multimodal algorithm for tumour identification. The figure depicts examples of healthy (green) and tumour (red) classifications, with the HSI modality results shown in the upper part and the pCLE results in the lower part. Fusion of the two modalities is achieved through a logical OR operation, followed by a component filtering stage [160]. Permission for the reuse of this figure is detailed in section A.0.2. 153

B.1 Multi-layer perceptron (MLP), comprised of multiple layers of nodes. Each node has a weight associated. Each layer is comprised of these nodes, with a bias term for scaling. The activation functions between the layers have been omitted but should be present after each layer. 186
B.2 Different activation functions and their derivatives over the domain $D : [-2, 2]$. . 187
B.3 Examples of convolutional neural network (CNN) architectures 188
B.4 Left: Scaled dot product attention using query, key, and value vectors. Right: The multi-head attention framework [195]. 189
B.5 Vision Transformer architecture [51] 191

F.1 Setup for Gaussian Process regression 209
F.2 Five function samples from the GP prior using the RBF kernel 209
F.3 Posterior GP predictions with 95% confidence intervals 210

List of Tables

1.1 Classification of central nervous system tumours. Based on origin, WHO grade, and clinical characteristics [13] 26

2.1 ADCC vs Latency Study for ResNet-18 and MobileNetV2 on pCLE Dataset. Latency(ms) is the time to compute one PA map using a batch size of one. 52
2.2 ADCC vs Latency Study for ResNet-50 ImageNet Dataset. Latency(ms) is the time to compute one PA map using a batch size of one 52

3.1 An overview of the different data augmentation pipelines used for training. 80
3.2 Parameter study of the coefficient (Coeff) utilised to manage the effect of *Calibration-Jitter* [159]. Permission to reuse this Table can be found in section A.0.3. 82
3.3 A summary of model performance metrics, all of which are reported as the macro average across the classes. The evaluated metrics include the F1-score, accuracy (Acc), true positive rate (TPR), and false positive rate (FPR) [159]. The license to reproduce this table is located in section A.0.3. 82

4.1 Macro Results: Performance of models with varying input modalities (Mod.) and user clicks. Bold indicates peak performance per metric and dataset [161]. Permission for the reuse of this table is detailed in section A.0.4. 99
4.2 Per-class D@0.5 results for the Hib dataset using a single click. The lower section shows the zero-shot performance on the Tumour class when it is excluded from training [161]. Permission for the reuse of this table is detailed in section A.0.4. . 101
4.3 Model performance (Macro D@0.5) for cross-dataset and mixed-dataset training scenarios. The number of HSI channels and classes for each training configuration are shown [161]. Permission for the reuse of this table is detailed in section A.0.4. 105

5.1 Performance comparison on Heipor, Hib, SB-X, and SB-H Datasets using SAM2 variants. D. denotes DICE; 1c/5c denote user clicks. Model suffixes: T (Tiny), S (Small), B+ (Base+), L (Large), FT (Fine-Tuned), Eq (Equalized), Ad (Adapter). . . 119
5.2 Results across datasets with DICE at 0.5 threshold (D@0.5) at 1 (1c) and 5 (5c) clicks. Model suffixes: T (Tiny), S (Small), B (Base), L (Large), FT (Fine-Tuned). Values are percentages; bold indicates the best performance per column. 120

6.1 Generation parameters for the fiducial marker pattern. 137

6.2 This table presents the volume of injected agar-acriflavine and the size (height and width) of the exposed tumour at the selected key frame. The last 5 columns detail the class distribution of HSI spectra and pCLE images across the dataset [160]. Permission for the reuse of this table is detailed in section A.0.2. 142
6.3 A summary of the spatial and temporal performance statistics for the multimodal platform [160]. Permission for the reuse of this Table is detailed in section A.0.2. 146
6.4 Performance of HSI spectral classification, presenting the DICE coefficient for each class and the average (Avg) score. The table also includes the processing speed in frames per second (FPS) for both the SVM and Neural Network (NN) models [160]. Permission for the reuse of this Table is detailed in section A.0.2. . 151
6.5 Results for the binary classification of tumour tissue using pCLE, detailing the macro accuracy and F1-score. Processing speed in frames per second (FPS) is also provided [160]. Permission for the reuse of this Table is detailed in section A.0.2. 151
6.6 An overview of the results for the multimodality algorithm, presenting DICE and Recall metrics. Performance is shown for the individual modalities as well as their combination. In the table, BG denotes the background class, while Avg signifies the average performance across all classes [160]. Permission for the reuse of this figure is detailed in section A.0.2. 152

G.1 Performance evaluation study on pCLE data based on the ADCC and time metrics. Coh is Coherence, Comp is Complexity, AD is average drop and they are reported for completeness. Time(s) is the average time to compute one PA map using a batch size of one. 211
G.2 Performance evaluation study on ImageNet based on the ADCC and time metrics. Coh is Coherence, Comp is Complexity, AD is average drop and they are reported for completeness. Time(s) is the average time to compute one PA map using a batch size of one. 212
G.3 T vs ADCC Study for Resnet-18 and MobileNetV2 on pCLE data. 212

Acronyms

5-ALA 5-Aminolevulinic Acid. 30, 39

ADCC Average Drop Coherency Complexity. 14, 20, 50–52, 55, 60
AE-CAI Augmented Environments for Computer Assisted Interventions. 61

CAI Computer-aided Intervention. 69
CAM Class Activation Mapping. 42
CEUS Contrast-Enhanced Ultrasound. 28
CMOS Complementary Metal-Oxide-Semiconductor. 136
CNN Convolutional Neural Networks. 13, 39, 42, 43, 47, 51, 58, 60, 67, 70
CNS Brain and Central Nervous System. 25
CT Computed Tomography. 13, 27–29, 90
CV Coefficient of Variation. 14, 48, 51, 53, 54

DICE Dice Similarity Coefficient or Sørensen–Dice coefficient. 20, 21, 97–99, 109, 114, 118–120, 151, 152, 160
DL Deep Learning. 41, 43, 49, 51, 55, 60, 69, 86, 161, 162
DRS Diffuse Reflectance Spectroscopy. 30
DWI Diffusion-Weighted Imaging. 27

ERM Empirical Risk Minimisation. 43

FGS Fluorescence-Guided Surgery. 30, 31

GAP Global Average Pooling. 43
GBM Glioblastoma. 13, 14, 43, 44, 49, 53, 58

Heipor HeiPorSPECTRAL. 16, 20, 98–106, 118–120, 123
Hib Human Imaging Benchmark. 14–17, 20, 74, 77, 83, 85, 87, 98–101, 103–107, 109, 118–121, 124, 126, 127, 156, 158
HSI Hyperspectral Imaging. 17, 18, 20, 21, 30, 31, 33, 62–64, 69–71, 85, 86, 88–91, 94, 98, 99, 105, 108–111, 129, 131, 132, 135, 137–146, 148, 150–154, 156–162

ICG Indocyanine green. 30, 39
iCT Intraoperative Computed Tomography. 28
IDC Invasive Ductal Carcinoma. 13, 14, 44, 49, 54, 58
IHC Immunohistochemistry. 27
iMRI Intraoperative Magnetic Resonance Imaging. 28, 131
IOUS Intraoperative Ultrasound. 28

LCC Largest Connected Component. 145, 152
LDA Linear Discriminant Analysis. 67
LiDAR Light Detection and Ranging. 31

MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 36, 61, 88

ML Machine Learning. 41, 69, 90, 91
MNG Meningioma. 13, 14, 42–44, 49, 53, 55, 58
MRI Magnetic Resonance Imaging. 13, 27–29, 90
MRS Magnetic Resonance Spectroscopy. 27

NHS National Health Service. 27, 39
NMS Non Maximum Supression. 116

OCT Optical Coherence Tomography. 30
OT Operating Theatre. 65

PA Pixel Attribution. 13, 14, 20, 37, 41, 42, 44, 45, 47–58, 60, 156, 157
PCA Principal Component Analysis. 67
PCC Pearson Correlation Coefficient. 90, 93, 94, 118
pCLE Probe-Based Confocal Laser Endomicroscopy. 13, 14, 17, 18, 20, 21, 31, 33, 36–39, 42, 44, 45, 48, 49, 52–57, 129, 131, 132, 136–143, 145, 146, 148, 150–154, 156–159, 161, 162
PpIX Protoporphyrin IX. 39

RF Random Forest. 39, 67
RS Raman Spectroscopy. 31

SA Spectral Angle. 66, 67, 90, 93, 94, 99–101, 112, 113, 118, 120
SAM Segment Anything. 89, 90, 112, 113, 115, 125
SAM2 Segment Anything-2. 15–17, 20, 90, 92, 94, 95, 98–101, 103–106, 109, 111–116, 118–127, 156, 158, 159
SAMSA Segment Anything with Spectral Angles. 15, 16, 95, 96, 98, 99, 101–114, 118, 119, 127, 156, 158, 159
SAMSA2.0 Segment Anything with Spectral Angles 2.0. 16, 17, 110, 111, 113, 114, 118–128, 157–159
SB SLIMBRAIN. 16, 20, 118–120, 122, 123
SCF Spectral Comparison Functions. 90, 91, 93, 95, 96, 112
SOTA State of the Art. 42, 48, 51, 60, 67, 78, 113
SVM Support Vector Machines. 15, 39, 67, 84, 85, 145

WHO World Health Organization. 13, 25

Chapter 1

Introduction

This chapter provides an overview of the research problem and its significance. It outlines the motivation for the research, presents the key objectives, and introduces the central questions and hypotheses guiding the investigation. The chapter also outlines the structure of the thesis, offering the reader a roadmap for the chapters that follow.

1.1 Overview

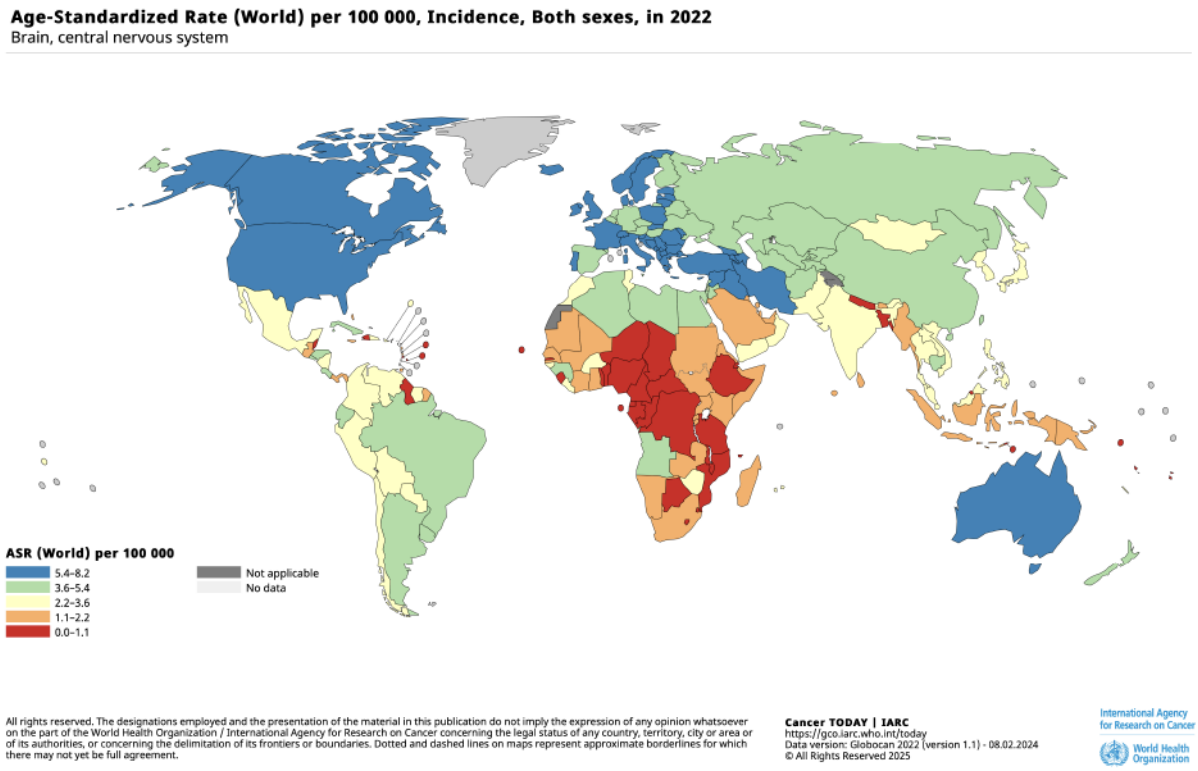


Figure 1.1: Global heatmap of age-standardized incidence rates for cancers of the brain and central nervous system in 2022. The map displays the age-standardized incidence rate (ASR, World) per 100,000 population for both sexes combined. Data source: GLOBOCAN 2022, International Agency for Research on Cancer, World Health Organization (WHO) [87].

Brain and Central Nervous System (CNS) tumours represent a diverse group of neoplasms arising from various cell types within the brain and spinal cord. As illustrated in fig. 1.1, these tumours are prevalent worldwide. Data from 2022 [87] shows that they rank as the ninth leading cause of cancer-related mortality globally. This combination of widespread occurrence and high mortality rates underscores their profound public health significance.

Within the CNS, tumours are classified based on their cellular origin, histopathology, and clinical behaviour. The WHO system is the primary framework used for this classification, categorising tumours by cellular origin and assigning a grade from low-grade (I and II) to high-grade (III and IV) that reflects their malignancy [13, 25, 114]. Some key examples are shown in table 1.1.

Tumour classification is crucial for treatment planning and the management strategies vary widely depending on tumour type. Complete surgical resection yields excellent outcomes for benign lesions such as meningiomas and pilocytic astrocytomas [132, 139], whereas infiltrative gliomas require a treatment approach that combines surgery with radiation and chemotherapy [187, 208]. In contrast, some deep-seated tumours, such as brainstem gliomas, may not benefit from aggressive resection at all [150, 168]. In this thesis, we focus exclusively on surgical cases where resection

Table 1.1: Classification of central nervous system tumours. Based on origin, WHO grade, and clinical characteristics [13]

Tumour Type	Cellular Origin	WHO Grade	Characteristics
Meningiomas	Meninges	I (can be II or III)	Typically benign and slow-growing; may cause symptoms from mass effect on brain tissue [210]
Astrocytomas	Astrocytes (a type of glial cell)	I–IV	Range from slow-growing low-grade to aggressive high-grade; glioblastomas represent Grade IV [44]
Glioblastomas	Astrocytes	IV	Highly aggressive, infiltrative, poor prognosis [213]
Pituitary Adenomas	Pituitary gland	I	Generally benign; may disrupt hormone levels or cause local pressure effects [133]
Schwannomas	Schwann cells	I	Often affect cranial nerves (especially vestibular); usually benign but may impair nerve function [82]
Medulloblastomas	Primitive neuroectodermal origin	IV	Highly malignant; arise in cerebellum; spread via cerebrospinal fluid; common in children [154]

is beneficial, examining the imaging technologies and intraoperative techniques that support safe and effective tumour removal.

In neuro-oncological surgery, achieving a balance between tumour removal extent and preservation of critical brain functions is essential. Surgical guidance systems provide information before, during, and after procedures to help surgeons distinguish between healthy and tumour tissue. An illustration of a patient lifecycle for tumour resection is illustrated in fig. 1.2. The current standard for intraoperative tumour pathological identification is histopathological analysis which requires intraoperative removal of small sections of suspicious tissue for microscopic examination [223]. This process, typically involving rapid frozen section analysis [223], requires temporary surgery cessation while samples are expedited to a pathology lab for evaluation. Following the neuropathologist's preliminary interpretation, the surgeon resumes resection based on these findings. While providing valuable guidance, this method is time-consuming and disrupts surgical workflow, potentially increasing operation time and patient risk. Furthermore, intraoperative histopathology, particularly frozen section analysis, is not the gold standard for definitive tumour diagnosis. The rapid freezing process can introduce cryoartifacts that may obscure cellular details [223], and limited sampling in the intraoperative setting might not capture the tumour's full heterogeneity. As noted in the literature [94], while demonstrating good accuracy, discrepancies between intraoperative and final post-operative diagnoses occur due to limitations of rapid processing and staining compared to more comprehensive post-operative techniques. The definitive diagnosis and complete tumour characterization rely on post-operative histopathology, involving formalin fixation, paraffin embed-

ding, detailed staining, Immunohistochemistry (IHC), and increasingly, molecular analyses [25]. This comprehensive approach allows more accurate assessment of tumour type, grade, margins, and molecular profile, establishing post-operative histopathology as the gold standard for guiding adjuvant therapies and predicting patient outcomes. Therefore, while intraoperative histopathology remains a critical tool for surgical decision-making, its provisional nature and workflow disruptions highlight the need for complementary techniques that can provide continuous, accurate tumour delineation.

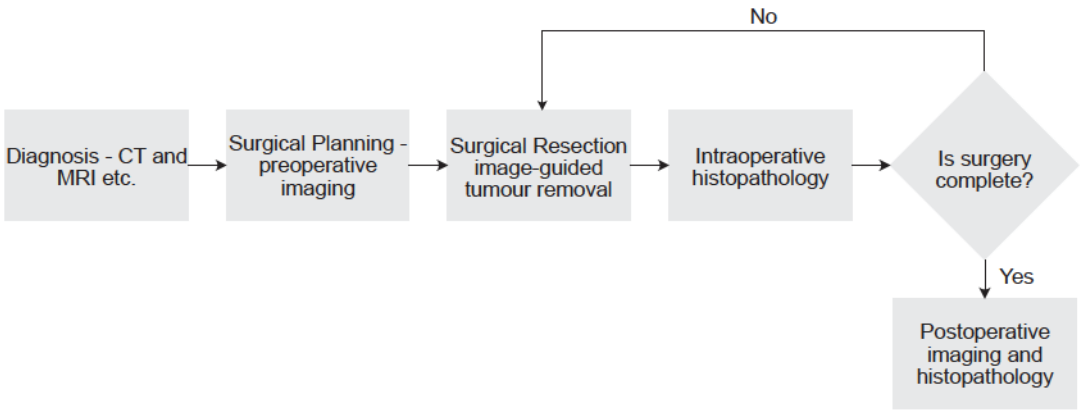


Figure 1.2: Patient Care Lifecycle for Tumour Resection: A patient-centred overview of the surgical pathway, from initial diagnosis and imaging through preoperative planning, surgery with intraoperative assessment, postoperative care and histopathology.

To address these limitations, advanced imaging modalities have emerged as essential tools for enhancing intraoperative decision-making and guiding tumour resection. It is important to note, however, that the routine application of these complementary methods is not currently standard practice within the National Health Service (NHS) and similar healthcare systems, where reliance on the established post-operative histopathology remains the norm. Imaging modalities already play a central role in surgical planning and intraoperative neuronavigation, particularly preoperative Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) [59, 125, 157, 164]. These modalities provide high-resolution anatomical detail that aids in the initial detection, localization, and characterization of brain tumours. MRI, with its superior soft tissue contrast, is particularly valuable for assessing tumour boundaries, edema, involvement of eloquent brain regions, and potential infiltration into surrounding structures [199]. Advanced MRI techniques, such as Diffusion-Weighted Imaging (DWI), perfusion imaging, and Magnetic Resonance Spectroscopy (MRS), can offer additional insights into tumour cellularity, vascularity, and metabolic profiles, contributing to non-invasive tumour classification and grading prior to surgery [130, 169]. CT, while offering lower soft tissue contrast than MRI, is especially useful in identifying calcifications, haemorrhage, or bone involvement, and is often used when MRI is contraindicated. Both imaging modalities are critical for planning the surgical approach [199]. They inform the optimal location for the scalp in-

cision and craniotomy, and guide the safest trajectory to the tumour while minimizing disruption to critical cortical and subcortical structures. Integration of these preoperative images into neuronavigation systems allows the surgeon to map the tumour in three-dimensional space and simulate the procedure in advance, enhancing both precision and confidence during the actual operation [59]. The core modalities of MRI, CT and histopathology are illustrated in fig. 1.3.

However, a significant limitation of preoperative imaging in procedures like a craniotomy is brain shift: the intraoperative deformation of brain tissue caused by factors such as gravity, surgical manipulation, and cerebrospinal fluid loss [70]. This phenomenon reduces the accuracy of static preoperative images as the surgery progresses. To address brain shift, Intraoperative Magnetic Resonance Imaging (iMRI) or Intraoperative Computed Tomography (iCT) scans can be performed, providing updated anatomical information mid-procedure [9, 136, 148]. Unfortunately, the need for specialized equipment and infrastructure makes these approaches cost-prohibitive and inaccessible in many clinical settings. The constant manipulation of brain tissue and the evolving surgical cavity mean that any static registration performed will degrade over time, necessitating continuous or frequent updates to maintain accuracy and relevance for intraoperative guidance. Instead, we must look to other modalities that can offer real-time, or near real-time, feedback on tissue characteristics and anatomical changes as the surgery progresses, providing more dynamic and reliable guidance to the surgeon.

As illustrated in fig. 1.4, the modern surgical room includes surgical lighting, life support systems, and precision tools such as operating microscopes. During resection, surgeons often rely on visual inspection through operating microscopes that offer surface-level RGB imaging [149]. Although RGB imaging is useful for surgical navigation, it does not reliably differentiate tumour from healthy brain tissue without the aid of preoperative imaging or complementary intraoperative modalities [185]. One of these modalities, ultrasound, provides a practical real-time imaging solution during neurosurgical procedures. Intraoperative Ultrasound (IOUS) offers several specific advantages in neurosurgery. It enables rapid visualization of subsurface structures without ionizing radiation, allowing surgeons to identify residual tumour tissue, locate deep-seated lesions, and visualize critical neurovascular structures in real-time [152]. Modern neurosurgical ultrasound systems incorporate Doppler capabilities, enhancing the identification of vascular structures and helping surgeons avoid inadvertent vascular injury [124]. Additionally, Contrast-Enhanced Ultrasound (CEUS) employing microbubble contrast agents has improved the differentiation between tumour and healthy tissue by highlighting areas of disruption in the blood-brain barrier [153]. While IOUS serves as a valuable complement to other neuronavigation tools, its fundamental limitation lies in providing only general subsurface information without precise guidance for margins. This critical gap — determining exactly where to cut at the cortical surface — remains a significant challenge for surgeons relying solely on ultrasound [70]. Given the limitations of subsurface modalities like ultrasound, there is a need for complementary intraoperative imaging that provides real-time, surface-level feedback to assess margins across multiple scales.

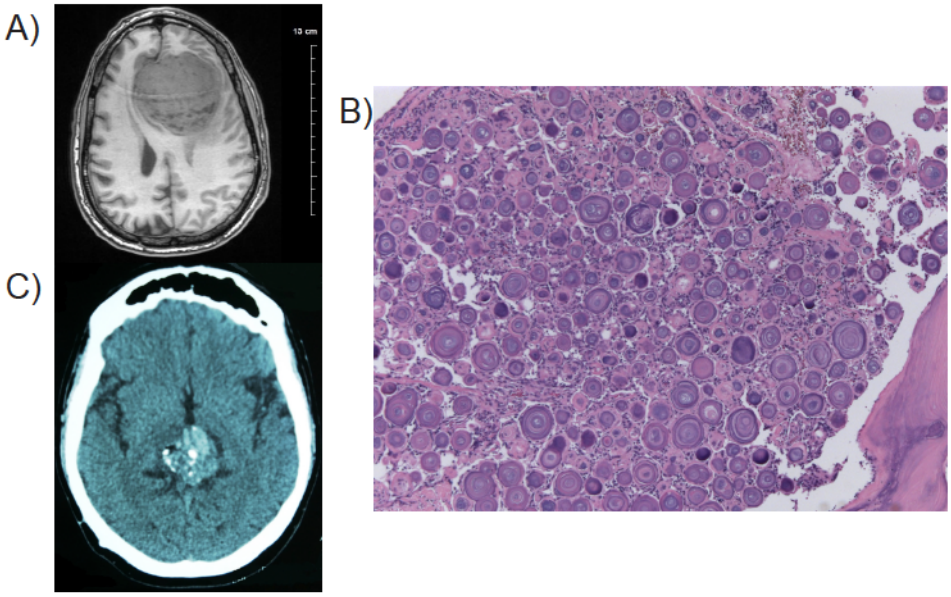


Figure 1.3: Side-by-side comparison of brain tumour representations in different modalities. (A) MRI image of a huge meningioma by Fowle, (B) histopathological section of a psammomatous meningioma by Jensflorian, and (C) CT scan showing a pineal gland tumour by Cerevisae.
All images licensed under CC BY-SA 4.0 via Wikimedia Commons. Sources: Fowle, Jensflorian, Cerevisae.



Figure 1.4: Photograph of an operating theatre featuring an operating microscope, surgical bed, overhead surgical lamps, and life support devices. Photo by Scott & White Healthcare, licensed under CC BY-NC-ND 2.0, via . Accessed April 15, 2025.

Fluorescence-Guided Surgery (FGS) addresses some of these challenges by providing surface-level molecular specificity, which helps in distinguishing cancerous tissue from healthy tissue. This non-ionizing technique involves administering a fluorescent dye — either orally or intravenously — to the patient, allowing the dye to bind to target cells. During the procedure, the dye is excited by light of a specific wavelength, leading it to emit fluorescence; consequently, the operating room lights are dimmed or a filter is added to optics to clearly reveal the fluorescent signal and, thereby, the tumour boundaries [174]. Common example dyes used in FGS include 5-Aminolevulinic Acid (5-ALA) and second-window Indocyanine green (ICG), which highlight the tumour borders based on their selective uptake by cancer cells [14, 29, 36]. The fluorescence signal is often available only for a short period of time. Whilst FGS offers great context of surface-level information to the surgeons, the fluorescence signal is often only available for a short period of time so there is still a need for localized, real-time imaging that provides high-resolution details continuously.

Hyperspectral Imaging (HSI) has emerged as a promising tool for intraoperative imaging by addressing many of the challenges traditionally met by fluorescence-based techniques without relying on exogenous contrast agents. Instead of administering fluorescent dyes, HSI captures data across a wide range of wavelengths—from the visible to the near-infrared—to identify tissue types based on their intrinsic spectral signatures [119, 140]. This technique generates a three-dimensional data cube in which each pixel is associated with a full spectral profile that encodes the underlying biochemical and physical properties of the tissue. The result is a label-free imaging method that eliminates issues such as photobleaching and the temporal limitations often associated with fluorescence signals. In neurosurgical applications, HSI provides macroscopic insights into tumour localization throughout the procedure. By continuously acquiring spectral data, it enables surgeons to differentiate between cancerous and healthy tissue without the need for externally administered markers [37, 190]. The operating environment does not have to be modified by dimming lights or adding optical filters, as the spectral information is captured directly from the tissue's inherent optical properties. This allows for a more seamless integration into the surgical workflow while delivering objective, quantitative assessments of tissue characteristics. This capability enhances the surgeon's ability to discern tumour margins and improves surgical precision by providing detailed, biochemical mapping. However, while HSI offers macro scale spectral information, it lacks the cellular resolution necessary for precise tumour margin delineation, creating a need for complementary approaches.

Probe modalities are built for small scale information, though each has distinct limitations. Optical Coherence Tomography (OCT) offers cross-sectional imaging of tissue microstructure at an impressive resolution, potentially revealing subtle architectural changes indicative of tumour infiltration near the surface. Its rapid acquisition facilitates real-time assessment of tissue layers, albeit with a limited penetration depth and significantly reduced imaging quality in the presence of blood or surgical debris, hampering reliable intraoperative use [21]. Furthermore, Diffuse Reflectance Spectroscopy (DRS) provides label-free biochemical information by analysing the spectral proper-

ties of reflected light, thereby offering insights into tissue composition and oxygenation that can help distinguish tumour from normal brain matter at a localized level; however, it remains relatively novel with limited clinical validation and lacks regulatory approval for widespread surgical applications [115]. Similarly, Raman Spectroscopy (RS) affords label-free molecular characterization of tissue, assisting in differentiating malignant from non-malignant areas with high specificity [47], yet its inherently weak signal when applied to biological samples necessitates lengthy acquisition times, making it impractical as a handheld probe and better suited for benchtop analysis away from the surgical field [64, 91].

To meet the need for microscale precision, Probe-Based Confocal Laser Endomicroscopy (pCLE) offers a particularly effective solution. pCLE enables *in vivo* imaging of tissue microarchitecture at cellular resolution, providing real-time feedback on tumour boundaries without the need for tissue excision. The cellular information is akin to frozen-section analysis but obtained *in-situ*. Similar to FGS, pCLE leverages fluorescent contrast to enhance neoplastic tissue visualization; however, unlike FGS's intermittent, wide-field assessment, pCLE offers continuous, highly localized examination at microscopic resolution. This capability integrates the strengths of the aforementioned techniques while enhancing surgical decision-making and efficiency [23, 30]. The limitation of pCLE lies in its highly focused view, the information remains confined to sampling points that must be integrated across larger areas to provide comprehensive surgical guidance comparable to traditional histopathology. Recent works have focused on combining this modality with robotics and/or vision systems to keep track of these scanning regions [217, 226].

Despite the individual strengths of these imaging techniques, current surgical guidance workflows often suffer from fragmentation — multiple devices operate in isolation, each suited to specific tasks or stages of surgery. This disjointed setup can complicate surgical planning, increase cognitive load, and limit the full potential of image-based decision-making. To overcome this, recent research has explored combining imaging modalities, such as fusing HSI with Light Detection and Ranging (LiDAR) or endoscopy [109, 171], though these efforts have largely focused on offline or diagnostic use rather than real-time surgical guidance [105]. While some combinations of modalities have been investigated, none have effectively bridged the critical gap between microscopic cellular detail and macroscopic contextual visualization — a combination that could revolutionize surgical precision and workflow efficiency.

In this thesis, we present a novel, multi-modal surgical guidance system for neurosurgery that integrates HSI and pCLE to bridge the gap between macro and micro scale visualization. This unique pairing leverages the complementary strengths of both modalities: HSI's ability to provide wide-field, label-free tissue characterization across the surgical field, and pCLE's capacity to deliver cellular-level resolution at specific points of interest. By applying computer vision methods to fuse and interpret this complementary information, we seek to enable continuous, real-time intraoperative guidance that supports tumour localization, tissue differentiation, and precise margin

identification — all within a unified framework. Unlike previous multi-modal approaches that combine techniques with overlapping capabilities, our system exploits the natural synergy between macro and micro scale imaging to provide surgeons with a comprehensive understanding of the surgical environment without workflow disruption. This work aims to simplify surgical decision-making and ultimately contribute to more effective and complete tumour resections through truly complementary multi-scale visualization.

1.2 Contributions

This thesis explores the complementary use of pCLE and HSI for neurosurgical guidance, aiming to improve intraoperative tissue characterization and tumour resection. It presents advances in each modality, introduces novel interactive segmentation methods, and integrates these technologies into a multimodal platform for enhanced tumour detection. The main contributions are:

1. Explainable and Uncertainty-Aware Image Classification

We present a method that combines pixel-level attribution maps with uncertainty estimation to improve the interpretability and trustworthiness of image classification. This enables visualization not only of which image regions the model focuses on, but also how confident it is in those areas. (See Chapter 2.)

2. Calibration-Aware Data Augmentation for HSI Segmentation

We introduce a data augmentation strategy for HSI that simulates realistic sensor calibration variability. This improves model robustness and generalisation, enhancing segmentation performance in neurosurgical scenes. (See Chapter 3.)

3. Interactive HSI Segmentation

We adapt the Segment Anything Model (SAM) to HSI data by incorporating spectral similarity information. The resulting method, SAMSA, enables efficient, interactive segmentation with minimal training data and strong generalization across diverse spectral domains. (See Chapter 4.)

4. Interactive HSI Segmentation Extended

We build upon the previous chapter by introducing SAMSA2.0 which, instead of using a hard-constraint of spectral angles on the output of SAM, prompts using spectral angles as a soft-constraint. Other improvements include gradient memory reduction and automatic segmentation. (See Chapter 5.)

5. Multimodal Imaging Platform for Enhanced Tumour Resection

We develop and evaluate a neurosurgical imaging system that combines pCLE and HSI in a unified platform. By aligning and jointly analysing the data, the system improves tumour detection performance over the use of either modality alone. The approach is validated on *ex vivo* brain tissue. (See Chapter 6.)

1.3 Publications

1.3.1 Related Publications

- Roddan, A., Xu, C., Ajlouni, S., Kakaletri, I., Charalampaki, P. and Giannarou, S. (2023) 'Explainable image classification with improved trustworthiness for tissue characterisation', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer Nature Switzerland, pp. 575–585.
- Roddan, A., Czempiel, T., Elson, D.S. and Giannarou, S. (2024) 'Calibration-jitter: augmentation of hyperspectral data for improved surgical scene segmentation', *Healthcare Technology Letters*, 11(6), pp. 345–354.
- Roddan, A., Czempiel, T., Xu, C., Elson, D.S. and Giannarou, S. (2025) 'SAMSA: segment anything model enhanced with spectral angles for hyperspectral interactive medical image segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*. Cham: Springer Nature Switzerland, pp. 478–488.
- Roddan, A., Czempiel, T., Xu, C., Xu, H., Weld, A., Chalau, V., Anichini, G., Elson, D.S. and Giannarou, S. (2025) 'Multimodal imaging platform for enhanced tumour resection in neurosurgery: integrating hyperspectral and pCLE technologies', *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10.

1.3.2 Other Publications

- Li, X., Keshavarz, M., Kassanos, P., Kidy, Z., Roddan, A., Yeatman, E. and Thompson, A.J. (2023) 'SERS detection of breast cancer-derived exosomes using a nanostructured Pt-black template', *Advanced Sensor Research*, 2(4), p. 2200039.
- Xu, C., Roddan, A., Davids, J., Weld, A., Xu, H. and Giannarou, S. (2022) 'Deep regression with spatial-frequency feature coupling and image synthesis for robot-assisted endomicroscopy', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer Nature Switzerland, pp. 157–166.
- Czempiel, T., Roddan, A., Leiloglou, M., Hu, Z., O'Neill, K., Anichini, G., Stoyanov, D. and Elson, D. (2024) 'RGB to hyperspectral: spectral reconstruction for enhanced surgical imaging', *Healthcare Technology Letters*, 11(6), pp. 307–317.
- Xu, H., Weld, A., Xu, C., Roddan, A., Cartucho, J., Karaoglu, M.A., Ladikos, A. *et al.* (2025) 'SurgRIPE challenge: benchmark of surgical robot instrument pose estimation', *Medical Image Analysis*, 103674.
- DeLorey, C., Davids, J.D., Cartucho, J., Xu, C., Roddan, A., Nimer, A., Ashrafian, H. *et al.* (2023) 'A

cable-driven soft robotic end-effector actuator for probe-based confocal laser endomicroscopy: development and preclinical validation', *Translational Biophotonics*, 5(2), p. e202200015.

- Roddan, A., Yu, Z., Leiloglou, M., Chalau, V., Anichini, G., Giannarou, S. and Elson, D. (2024) 'Towards real-time hyperspectral imaging in neurosurgery', in *Clinical Biophotonics III*. SPIE, vol. 13009, pp. 78–81.
- Xu, C., Roddan, A., Xu, H. and Giannarou, S. (2024) 'FF-ViT: probe orientation regression for robot-assisted endomicroscopy tissue scanning', *International Journal of Computer Assisted Radiology and Surgery*, 19(6), pp. 1137–1145.
- Giannarou, S., Xu, C. and Roddan, A. (2024) 'Endomicroscopy', in *Biophotonics and Biosensing*. Elsevier, pp. 269–284.
- Weld, A., Dixon, L., Roddan, A., Anichini, G., Camp, S. and Giannarou, S. (2025) 'Confidence-based annotation of brain tumours in ultrasound', *arXiv preprint*, arXiv:2502.15484.
- Watson, A.F., Haanaes, N., Chambers, R., Roddan, A., Monfort Sanchez, E., Runciman, M. and Thompson, A.J. (2024) 'Towards label-free flow cytometry for automated cell identification using diffuse reflectance spectroscopy', in *Biophotonics in Point-of-Care III*. SPIE, vol. 13008, pp. 89–102.
- Ezzat, A., Shanthakumar, D., Laskar, N., Cutress, R.I., Hadjiminas, D., Boland, M. *et al.* (2025) 'Impact of resection margin width on local recurrence following breast-conserving surgery and whole breast radiotherapy for pure ductal carcinoma in situ: a systematic review and meta-analysis', *BMJ Oncology*, 4, p. e000633.
- Xu, C., Roddan, A., Kakaletri, I., Charalampaki, P. and Giannarou, S. (2025) 'Interpretable classification of endomicroscopic brain data via saliency consistent contrastive learning', *Medical Image Analysis*.

Chapter 2

Probe-based Confocal Laser Endomicroscopy for Intraoperative Tissue Characterisation

This chapter introduces our first modality, Probe-Based Confocal Laser Endomicroscopy (pCLE). In this chapter we explain how it is used in neurosurgery, current state of the art predictive methods for tissue classification and finally our contribution; a novel explainability method that combines pixel-level attribution maps with uncertainty estimation.

The work in this chapter was adapted from our previously published paper in the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference proceedings in October 2023 [162], the code to reproduce this method can be found here.

2.1 Overview

This chapter investigates real-time and accurate deep learning methods for predicting tumour classes. This work will also provide an overview of pCLE and survey existing classification methods. The integration of machine learning models into surgical practice presents considerable potential for aiding tissue characterisation and facilitating safe tumour resections. By delivering immediate insights into tissue classification, tumour margins, and areas of potential risk, these models can prove invaluable. However, for such systems to achieve clinical acceptance, gaining the trust of surgeons is paramount. This trust is built upon understanding how a model arrives at its conclusions, which allows clinicians to validate its robustness before deployment. Therefore, it is essential that their predictions are not only precise but also interpretable.

Although numerous techniques for interpretable and explainable AI have been developed, most of them encounter difficulties in keeping up with contemporary deep learning architectures. These methods enable the visualisation of image regions which are most influential in a model's prediction. Nevertheless, existing techniques possess notable drawbacks. These include a limited robustness to minor perturbations in the input, which can cause dramatic alterations in attribution maps and thereby diminish a clinician's confidence. Furthermore, a lack of integrated uncertainty measures exists, which are indispensable for assessing the dependability of predictions within high-risk surgical environments. It is also notable that current methodologies rarely provide both salience and uncertainty at the same time, despite the complementary relationship between these two aspects.

To address these challenges, we introduce a novel framework for explainability that merges state-of-the-art Pixel Attribution (PA) with risk estimation derived from Monte Carlo dropout. This approach involves creating a volume of attribution maps via numerous stochastic forward passes and subsequently computing enhanced maps through the estimation of pixel-wise expectation values. The outcome is a unified representation that combines visual salience with uncertainty information, leading to explanations that are both interpretable and quantitatively meaningful.

Our framework is validated using *in vivo* data acquired from pCLE, where it demonstrates marked improvements in the quality and reliability of attributions. The resulting tool is capable of simultaneously highlighting significant image regions while also quantifying the uncertainty associated with the prediction. This represents a step forward in the development of explainable AI systems that can gain the trust of surgeons for decision support before the deployment of models and during operations.

2.2 Introduction

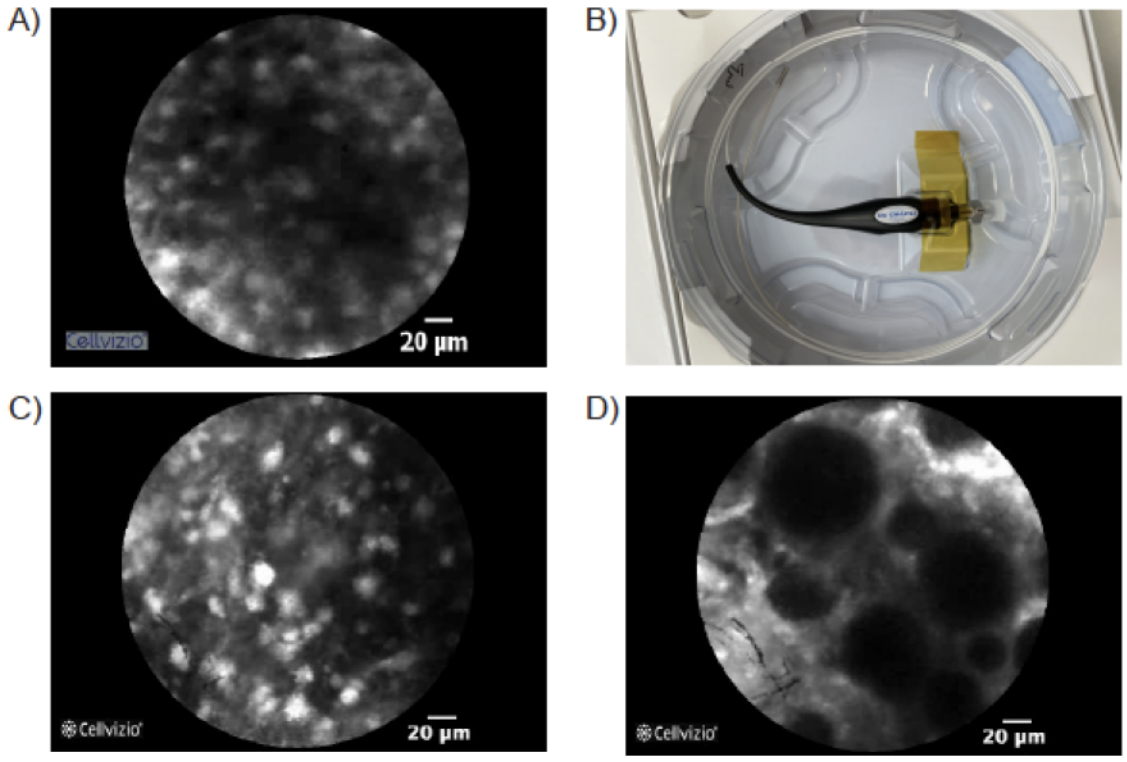


Figure 2.1: Example pCLE data a) Astrocytoma b) pCLE probe c) Glioblastoma d) Meningioma.

pCLE uniquely enables real-time, in situ visualisation of cellular morphology with histology-like detail during procedures [201]. The microscale information gives a trained clinical user insight into the type of tissue being sampled at that point, potentially aiding immediate clinical decision-making. Example pCLE images of cancer tissue, along with a representative probe, are shown in fig. 2.1.

A pCLE camera operates by delivering a low-power laser through a flexible fiber-optic probe, which is passed through the working channel of a standard endoscope [45]. The laser light excites a fluorescent dye that highlights cellular structures that it binds to. Because pCLE is a confocal imaging technique, the system focuses the excitation and detection on a single optical plane, significantly enhancing image resolution and contrast by rejecting out-of-focus light. The fluorescent dye is essential in this process: it emits light only when excited. Its fluorescence lifetime, the short period a molecule remains in an excited state before releasing light, must be long enough to ensure a detectable signal, allowing the system to generate high-resolution, real-time images of the tissue microarchitecture. This enables so-called "optical biopsies," offering a non-destructive way to assess tissue without physical removal.

Clearly, one of the major considerations is the fluorescent dye. The dye must a) bind to the correct target b) have an emission and absorption spectrum with low cross-talk and c) a lifetime long

enough to image [128]. In neurosurgery the most common fluorescent dyes used worldwide are Indocyanine green (ICG) [221], Fluorescein [68] and 5-Aminolevulinic Acid (5-ALA) [186].

ICG is primarily used for visualising blood flow and vascular structures during procedures such as aneurysm clipping or hemangioblastomas resection [128]. Fluorescein is often used to delineate tumour margins, particularly in high-grade gliomas, due to its ability to accumulate in areas with a disrupted blood-brain barrier [128]. 5-ALA, a metabolic precursor that leads to the accumulation of fluorescent Protoporphyrin IX (PpIX) in malignant glioma cells, is widely used for intraoperative identification of tumour tissue in glioblastoma surgery, improving resection accuracy and patient outcomes [128].

In the National Health Service (NHS), the use of pCLE in neurosurgery is usually performed in specialist centres. An example clinical setup is as follows: the patient presents with a suspected high-grade glioma and is scheduled for surgical resection. Approximately three hours prior to the procedure, the patient is administered 5-ALA orally at a dose of 20 mg/kg body weight [15]. Once metabolised, 5-ALA leads to the intracellular accumulation of fluorescent PpIX in malignant glioma cells. PpIX has a peak absorption around 405 nm (blue light) and emits fluorescence in the red spectrum, typically around 635 nm [129]. In this scenario, a pCLE system is employed intraoperatively where the pCLE probe is equipped with a laser source tuned to the absorption peak of PpIX (typically 405 nm), and the detection system includes an emission filter centred around 635 nm to selectively capture the red fluorescence emitted by tumour tissue.

The output images at each "optical biopsy site" are then used for intraoperative tissue assessment. Traditionally, these images are interpreted manually by trained clinicians or neuropathologists. They assess features such as cell density, structure, and fluorescence patterns to make diagnostic judgments. Manual interpretation has shown good accuracy, but it requires specific expertise [55, 145]. Interpreting pCLE images is not standard practice for most neurosurgeons, as it requires specialist training in optical imaging. To reduce this dependency and make real-time interpretation more accessible in the operating room, machine learning models have been introduced. These models can assist or even automate interpretation, enabling consistent and fast tissue classification without needing a specialist at the bedside.

Classical machine learning methods such as Support Vector Machines (SVM) [71] and Random Forest (RF) [20] have shown promising results in differentiating tissue. More recently, deep learning methods such as Convolutional Neural Networks (CNN) have shown great performance on small datasets. These models learn patterns directly from the image data, without needing manual feature selection, and often outperform classical methods in accuracy [88, 107]. Other architectures such as transformers [34, 217] have also shown efficacy and are beginning to be explored in pCLE analysis as dataset sizes grow.

While deep learning models, including CNNs and transformers, show strong performance in tissue classification, their integration into surgical workflows requires more than just accuracy. In intra-

operative settings, it is crucial that surgeons can assess the reliability of a model's predictions [50]. For clinical adoption, these models must provide explanations that are interpretable and trustworthy [5]. In addition to explainability, effective risk estimation is essential. Models must not only provide predictions but also communicate their confidence in those predictions. Existing studies often address prediction uncertainty or explainability in isolation, without combining both into a unified framework that can support clinical use. In this work, we address this critical gap by developing and evaluating a framework that jointly estimates predictive risk and provides interpretable explanations, tailored for surgical tissue characterisation.

2.3 Related Works

2.3.1 Deep Learning Explainability

For a foundational overview of Deep Learning (DL), please refer to Appendix B. Here, a general deep learning model with L layers can be expressed as

$$\begin{aligned} X^0 &= \text{input features}, \\ X^l &= \Phi\left(W^l X^{l-1} + b^l\right), \quad l = 1, 2, \ldots, L, \\ \hat{y} &= X^L, \end{aligned} \tag{2.1}$$

where X^l represents the activations at layer l, W^l are the trainable weights that scale inputs from the previous layer, b^l are the trainable biases that shift the input to each activation, Φ is the non-linear activation function that enables the network to model complex patterns, and $\hat{y}$ are the output logits of the final layer, representing the unnormalized predictions that can be adapted for tasks such as regression, classification, or segmentation depending on the training strategy and objective function used. DL models share many core components like linear layers, convolutional layers, hidden layers, and attention mechanisms. Their strength lies in its ability to model complex, high-dimensional, non-linear relationships, thanks largely to the inclusion of non-linear activation (or transfer) functions. These capabilities, however, often come at the cost of increased model complexity and a large number of parameters. This raises a key question: What exactly are these models learning? Can we meaningfully interpret the internal workings of a deep, highly parameterised, non-linear model?

In the context of intraoperative tissue characterisation, the usefulness of a Machine Learning (ML) model hinges on the surgeon's confidence in its predictions. Without trust, even the most accurate model becomes impractical [50]. To foster this trust, it is essential that the model offers interpretable and transparent outputs [5]. This thesis focuses exclusively on explainability within computer vision applications, as non-image-based DL explainability falls outside the scope of this work, although many techniques do overlap across domains. In evaluating model explainability, we consider five key performance indicators: inference speed, ease of use, generalisability, trustworthiness, and the ability to localise semantically meaningful image features. In image classification, one widely used approach for improving explainability is pixel attribution.

2.3.2 Pixel Attribution

PA maps are a valuable tool for visually assessing whether a model is effectively capturing meaningful semantic features [222], or whether it has instead learned misleading associations due to spurious correlations in the training data [76]. Spurious correlations refer to dataset-specific artifacts that are predictive during training but do not genuinely reflect the target class characteristics.

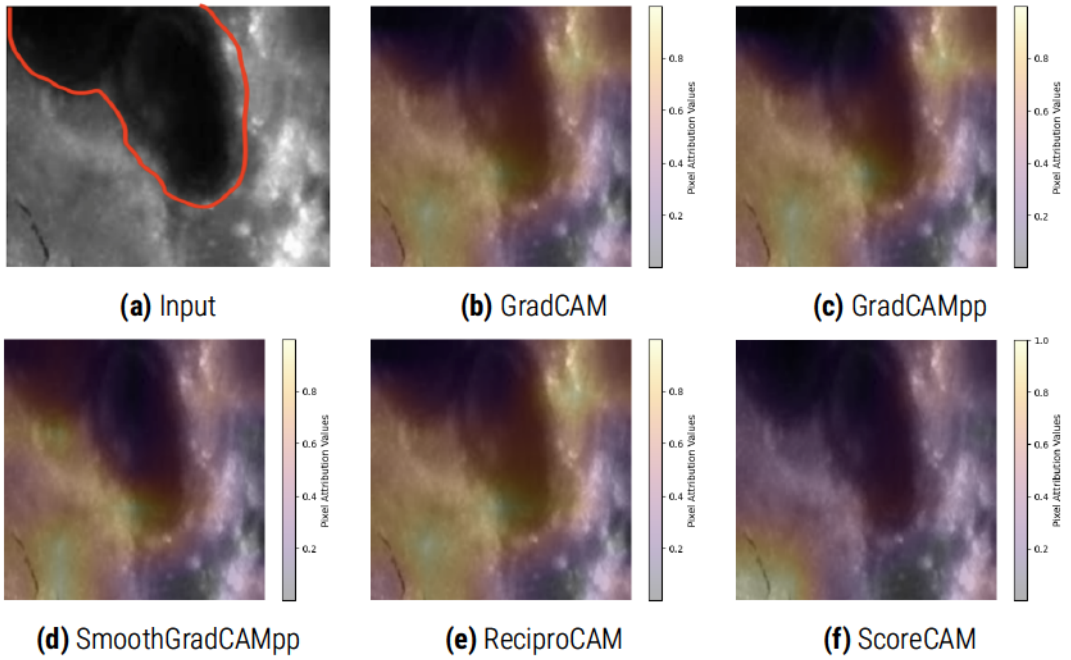


Figure 2.2: Multiple pixel attribution methods run on an un-robust model to shown spurious correlation of crack visible in Meningioma (MNG) videos. Red annotation shows salient region psammoma body for the MNG class.

PA techniques primarily highlight the input regions that the model considers influential on average, making them useful for detecting potential reliance on irrelevant or misleading features. However, there is some caution required when comparing what human experts deem to be salient with the areas that a model identifies as important, because the two perspectives may not align. Human judgments are grounded in domain knowledge and semantic interpretation, whereas model attributions reflect statistical associations learned from the data. Consequently, features highlighted by the model may appear unintuitive, incomplete, or even misleading to experts, complicating the task of determining whether the model is relying on spurious correlations or uncovering subtle but valid cues.

An illustrative case of such a correlation in pCLE data is shown in fig. 2.2, where a small crack in the bottom-left corner of an image present across multiple videos was overly weighted by the model during prediction. This is clearly visible in fig. 2.2f, where the PA method indicates the model has focused on this artifact rather than on medically relevant features. This suggests the model is exploiting dataset-specific "shortcuts" (spurious correlations) rather than learning generalisable representations. Nevertheless, this ability to uncover shortcuts can itself be useful. PA methods serve as a diagnostic tool for generalisability: they reveal when retraining the model with improved strategies, such as augmentations that encourage learning of class-relevant features, is needed.

To process image data effectively, most PA methods build upon CNN, which remain central to achieving state-of-the-art (State of the Art (SOTA)) performance. One of the earliest PA techniques tailored for CNNs was Class Activation Mapping (CAM) [225], which identifies the most influential

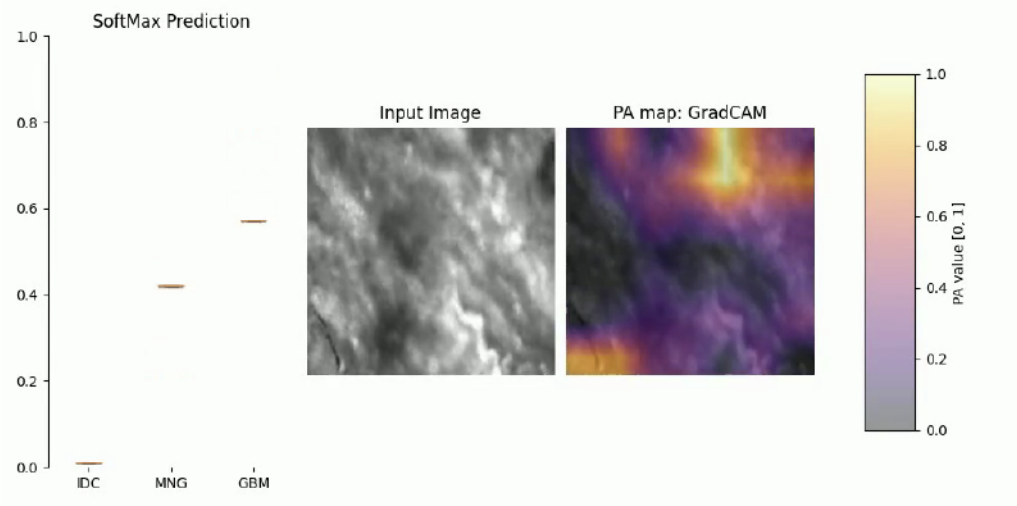
channels in the final convolutional layer through a single forward pass. A limitation of CAM, however, is its dependency on Global Average Pooling (GAP) [110], which restricts the range of architectures it can be applied to. Grad-CAM [176] addresses this by extending compatibility to any CNN architecture that is differentiable from the output logit to the selected intermediate layer. Nonetheless, Grad-CAM tends to produce blurry localisation maps. Improvements such as Grad-CAM++ [31] and SmoothGrad-CAM++ [138] enhance spatial sharpness and semantic accuracy, though their computational demands limit real-time deployment, especially in surgical environments [24].

To address this, methods like Score-CAM [204] and Recipro-CAM [24] offer both architectural flexibility and real-time capability. Score-CAM improves the precision of object localisation while retaining class-specific sensitivity, using a perturbation-based approach that masks parts of the input image and monitors the change in prediction score. This aligns with other perturbation-based explainability methods such as RISE [147], LIME [158], and earlier works [6, 222]. Meanwhile, Recipro-CAM is designed to optimise inference speed without sacrificing attribution quality. It leverages the receptive field properties of CNNs to generate spatial masks efficiently and evaluates their impact on the model's output in a manner akin to Score-CAM.

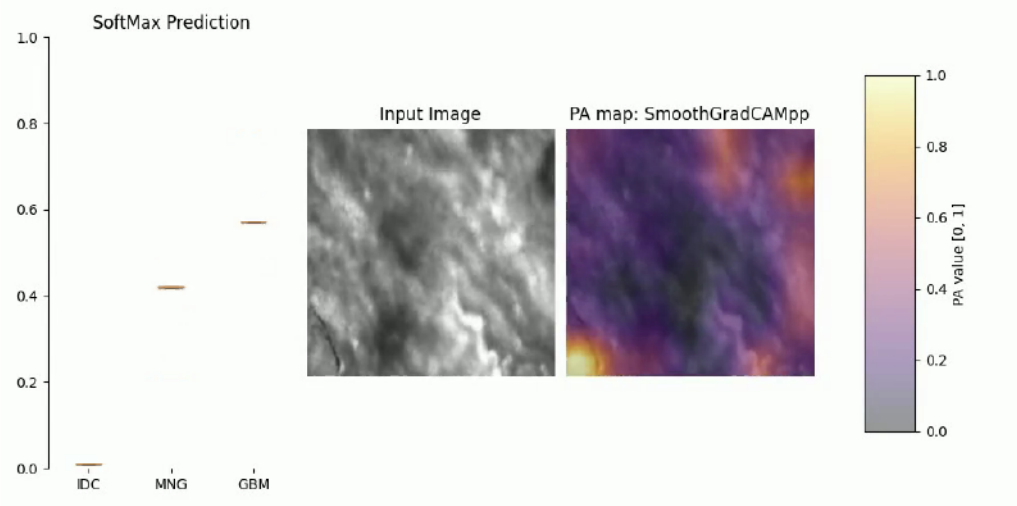
2.3.3 Uncertainty

While the pixel attribution methods discussed earlier are efficient, easy to deploy, and capable of highlighting semantically relevant regions, they often lack trustworthiness, primarily due to the way their underlying models are trained. DL models trained via Empirical Risk Minimisation (ERM) tend to be overconfident in their predictions [67] and susceptible to adversarial inputs [74]. As shown in figs. 2.3 and 2.4, the model assigns a softmax score of roughly 60% to the GBM class when the input is actually a MNG image. Although this may appear to indicate the model is "60% sure," the softmax value is in fact only a point estimate of relative class likelihoods, not a calibrated measure of uncertainty. A clinician viewing these results therefore receives no feedback from the model about how unsure or risky the prediction truly is. Most DL models, particularly those not designed for precision-critical tasks, do not conform strictly to either frequentist or Bayesian paradigms. Instead, they focus on optimising an objective function, typically a differentiable loss function, and are evaluated empirically. For more on point estimates, refer to Appendix E.

ERM involves minimising the average loss over the training data [194], and this approach underpins most modern ML systems. The goal is to reduce generalisation error and indirectly enhance other performance metrics. However, as noted in [73], "...empirical risk minimization is prone to overfitting. Models with high capacity can simply memorize the training set." To counteract this, training strategies are typically designed to mitigate overfitting. This is done commonly through validation based early stopping and regularisation to help produce models that generalise well. Although ERM is broadly applicable, in certain scenarios it overlaps with or mirrors frequentist and Bayesian approaches.

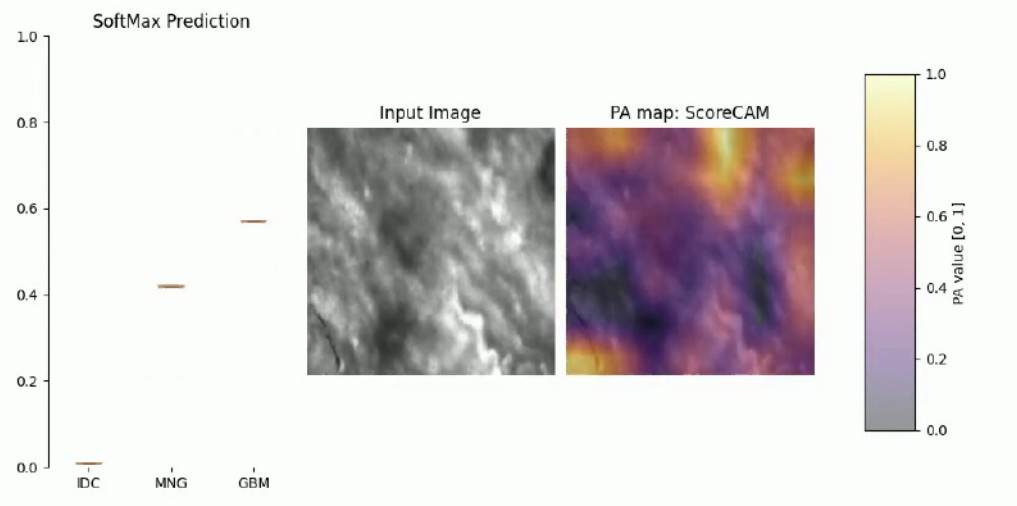


(a) GradCAM PA method.

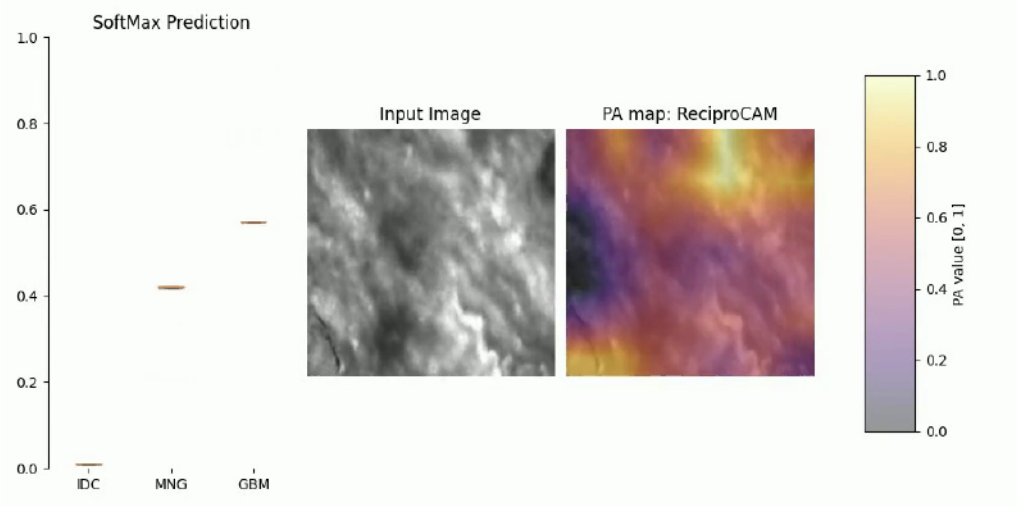


(b) SmoothGradCAM++ PA method.

Figure 2.3: Comparison of PA methods on pCLE classification using ResNet18. For each method, the left column shows the softmax predictions for the three classes (MNG, Invasive Ductal Carcinoma (IDC), and Glioblastoma (GBM)); the middle column displays the input image (MNG test sample); and the right column presents the corresponding PA map.



(a) ScoreCAM PA method.



(b) ReciproCAM PA method.

Figure 2.4: Comparison of ScoreCAM and ReciproCAM PA methods on pCLE classification using ResNet18. Layout follows the same format as fig. 2.3, showing softmax predictions, input images, and PA maps.

Bayesian Neural Networks [135] offer a principled approach to modelling uncertainty and exhibit improved regularisation. However, due to the complexity and non-linearity of neural networks, exact Bayesian inference becomes computationally intractable. To manage this, variational inference methods are employed [75, 83]. One such method, Bayes by Backprop [18], leverages Dropout [84] to approximate Bayesian inference. Dropout, originally introduced as a regularisation technique, works by randomly "dropping" (i.e., setting to zero) a subset of neurons during training. This prevents units from co-adapting too strongly, effectively training an ensemble of subnetworks that share parameters, thereby improving generalisation and reducing overfitting. Crucially, the stochastic nature of Dropout also makes it amenable to uncertainty estimation, linking its role in regularisation to Bayesian-inspired approaches. Beyond these benefits, Dropout has also shown promise in improving feature saliency. While Bayes by Backprop can yield trustworthy models, it often struggles to scale to state-of-the-art architectures due to its computational demands. A more scalable and practical alternative is Monte Carlo Dropout [67]. Monte Carlo Dropout treats a model trained with Dropout as an approximation to a deep Gaussian process [43, 66]. For more about Gaussian processes refer to Appendix F. Under this framework, an output distribution is generated by performing multiple stochastic forward passes with Dropout active at inference time. This distribution provides an estimate of the model's uncertainty, which can be particularly valuable in high-stakes domains like surgery. Surgeons may use this risk signal to gauge the reliability of the model's predictions during decision-making. Importantly, this technique provides a computationally efficient form of model averaging [84]. Although later work has questioned the strength of Monte Carlo Dropout as a true Bayesian inference method [61], its practical utility for estimating prediction uncertainty remains intact. To date, explainability and uncertainty estimation have largely been explored independently as tools for evaluating model suitability in clinical and surgical contexts.

2.4 Methodology

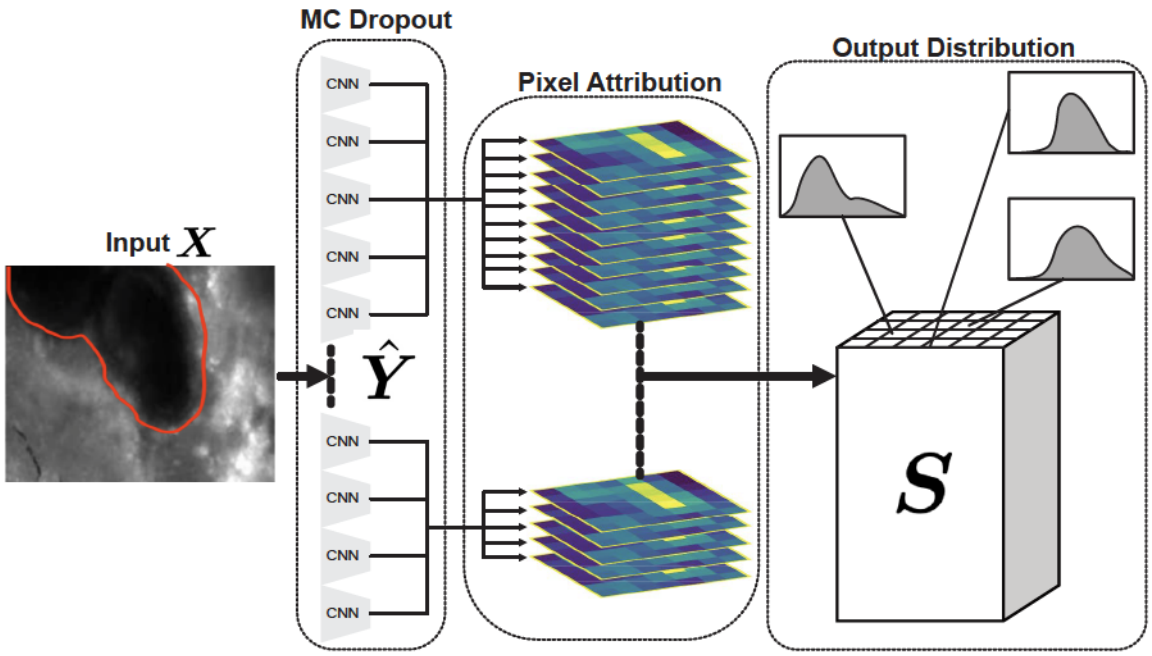


Figure 2.5: Outline of the proposed method. A PA volume is generated using T forward passes of a CNN model with Dropout applied. $\boldsymbol{X}$ represents the input image, $\hat{\boldsymbol{Y}}$ the output logits of the CNN model and $\boldsymbol{S}$ the volume of PA maps.

The aim of the proposed method is to produce an improved PA map of a classification model, while providing risk estimation of the model's explainability. Further aiding the decision making during intraoperative tissue characterisation.

In our method, any CNN classification model trained with Dropout can be used. Let $\hat{\boldsymbol{Y}}$ be the output logits of the CNN model, where Dropout is enabled at test time, with input image $\boldsymbol{X} \in \mathbb{R}^{height \times width \times channels}$. Any PA method can be used to generate a PA map using the output logits $S = f_s(\hat{\boldsymbol{Y}}) \in \mathbb{R}^{height \times width}$ where $f_s(.)$ is the PA method. We propose to repeat the above process for T iterations to create a volume of PA maps $\boldsymbol{S} = \{S_1, ..., S_T\} \in \mathbb{R}^{height \times width \times T}$. We show this visually in fig. 2.6. A visual representation of how the volume is generated is show in fig. 2.5. The aim is to use this volume to generate a pixel-wise distribution of PA values from which we can infer risk. To achieve this, we compute the expectation and variance values of the volume along the third dimension as:

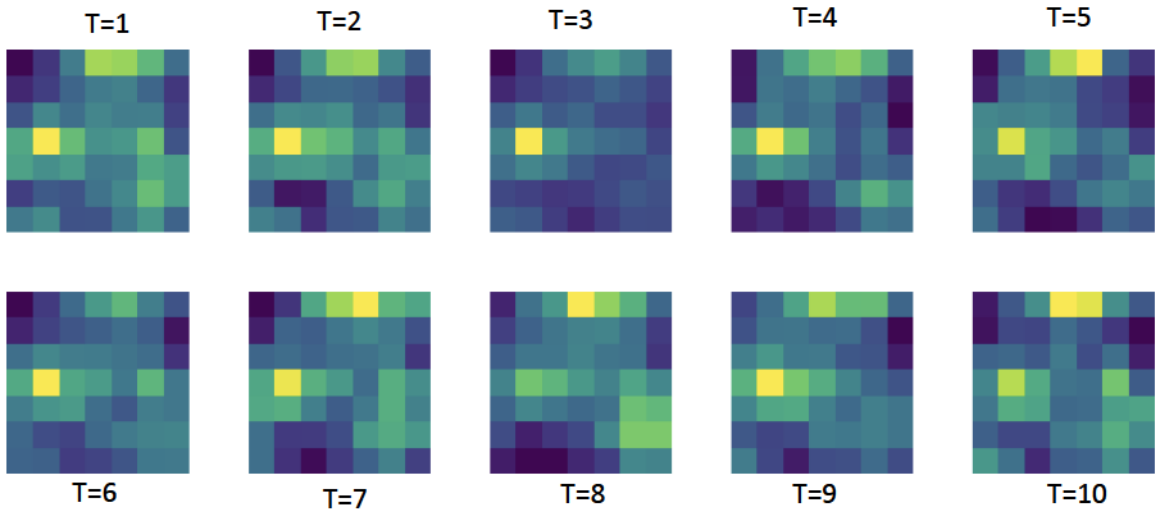


Figure 2.6: Each PA map of proposed method pre-interpolation to image size (for each iteration of T) from Layer 4 of ResNet18 for T=10 on pCLE data. Yellow represents the highest value and dark blue is lowest.

$$\begin{aligned} \mathbb{E}(\boldsymbol{S}_{i,j}) &\approx \frac{1}{T}\sum_{t=1}^{T} f_s(\hat{\boldsymbol{Y}}_t)_{i,j} \\ Var(\boldsymbol{S}_{i,j}) &\approx \frac{1}{T}\sum_{t=1}^{T} f_s(\hat{\boldsymbol{Y}}_t)_{i,j}^T f_s(\hat{\boldsymbol{Y}}_t)_{i,j} - \mathbb{E}(\boldsymbol{S}_{i,j})^T \mathbb{E}(\boldsymbol{S}_{i,j}), \end{aligned} \tag{2.2}$$

where, i, j represent the pixel's row and column coordinates, respectively. The expectation $\mathbb{E}(\boldsymbol{S}_{i,j})$ of each pixel (i, j) is used to generate an enhanced PA map of size $height \times width$. The intuition is that the above distribution of PA values can produce less noisy and overconfident estimation of a pixel's contribution to the final explainability map compared to a single estimate.

Advancing SOTA explainability methods, in our method we also estimate the risk of the enhanced PA map generated above. For the risk estimation, it is important to consider that different pixels in the PA map correspond to different semantic features which contribute differently to the output logits. This makes the pixel-wise distributions (and therefore expectation and variance values) have different scales. For this purpose, the Coefficient of Variation (CV) is used to estimate pixel-wise risk, as it allows us to compare pixel-wise variances despite their different scales. This is mathematically defined as:

$$S_{i,j}^{cv} = \frac{\sqrt{Var(\boldsymbol{S}_{i,j})}}{\mathbb{E}(\boldsymbol{S}_{i,j})} = \frac{std(\boldsymbol{S}_{i,j})}{\mathbb{E}(\boldsymbol{S}_{i,j})}. \tag{2.3}$$

2.5 Experiments and Analysis

2.5.0.1 Datasets

The developed explainability framework has been validated on an in vivo and ex vivo pCLE dataset of MNG, GBM and metastases of an invasive ductal carcinoma (IDC). All studies on human subjects were performed according to the requirements of the local ethic committee and in agreement with the Declaration of Helsinki (No. CLE-001 Nr: 2014480). The Cellvizio© by Mauna Kea Technologies, Paris, France was used in combination with the mini laser probe CystoFlex© UHD-R. The distinguishing characteristic of the MNG is the psammoma body with concentric circles that show various degrees of calcification. Regarding GBMs, the pCLE images allow for the visualization of the characteristic hypercellularity, evidence of irregular nuclei with mitotic activities or multinuclear appearance with irregular cell shape. When examining metastases of an IDC, the tumor presents as egg-shaped cells with uniform evenly spaced nuclei. Our dataset includes 38 MNG videos, 24 GBM and 6 IDC. Each pCLE video represents one tumour type and corresponds to a different patient. The data was curated to remove noisy images and similar frames. This resulted in a training dataset of 2500 frames per class (7500 frames in total) and a testing dataset of the same size. The dataset was split into a training and testing subset, with the division done on the patient level. To show generalisation to other domains, the proposed method was also evaluated on ImageNet [46], a dataset comprised of 1000 classes of natural scenes images.

2.5.0.2 Implementation

To implement the DL models we use the open-source framework PyTorch [142], a NVIDIA Geforce RTX 3090 graphics card for parallel computation and a 12th Gen Intel(R) Core(TM) i9-12900K CPU (using 16 threads for latency experiments). To show our method generalises across domains, we trained and tested on both pCLE data and ImageNet. For the pCLE data we trained two lightweight models; ResNet-18 [79] and MobileNetV2 [172]. A learning rate yielding stable convergence was first identified and fixed across all experiments. This maintained experimental control by isolating the effect of each model modification without confounding it with repeated hyperparameter re-optimisation. ResNet-18 was trained with a learning rate of 0.001, whereas MobileNetV2 had a learning rate of 0.01. Both were trained using the Adam-W [113] optimiser and weight decay of 0.01. For ImageNet, we trained a ResNet-50 with a learning rate of 0.1, using the Stochastic Gradient Descent optimizer [167] and the Step learning rate scheduler with a step size of 30 and gamma of 0.1. For the Resnet-50 model, we trained using distributed training on 3 X NVIDIA RTX A5000 graphics cards and evaluation was done on the NVIDIA Geforce RTX 3090 graphics card using the trained weights. All models were trained from scratch with a Dropout probability of 0.2 and a batch size of 256. At test time, we set $T = 10$ (T study was conducted and shown is analysis). PA methods were implemented with the help of TorchCAM [58] and ReciproCAM was implemented

using the authors' source code.

2.5.0.3 Evaluation Metrics

Evaluating a PA method is not a trivial task as a PA map may not need to be inline with what a human deems "reasonable" [3]. Segmentation scores like intersection over union may be used with caution to compare thresholded PA maps to ground truth maps with annotated salient regions. By doing so, we can measure how informed the model is about a particular class. To quantify how misinformed a model is, we can estimate at its average drop [31]:

$$AverageDrop(f_s, \hat{\boldsymbol{Y}}, \boldsymbol{X}) = 100 \times \frac{max(0, \hat{\boldsymbol{Y}}(\boldsymbol{X}) - \hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}}))}{\hat{\boldsymbol{Y}}(\boldsymbol{X})}, \tag{2.4}$$

where, $\hat{\boldsymbol{X}} = \boldsymbol{X} \odot f_s(\hat{\boldsymbol{Y}}(\boldsymbol{X}))$. The above equation measures the effect on the output score of the classification model if we only include the pixels which the PA method scored highly. A minimum average drop is desired.

As average drop was found to not be sufficient on its own, the unified method Average Drop Coherency Complexity (ADCC) [151] was introduced which is the harmonic mean of average drop, coherency and complexity, defined as:

$$\begin{aligned} ADCC(f_s(\hat{\boldsymbol{Y}})) = (&\frac{1}{Coherency(f_s(\hat{\boldsymbol{Y}}))} \\ &+ \frac{1}{1 - Complexity(f_s(\hat{\boldsymbol{Y}}))} \\ &+ \frac{1}{1 - AverageDrop(f_s, \hat{\boldsymbol{Y}}, \boldsymbol{X})})^{-1}. \end{aligned} \tag{2.5}$$

Coherency is the Pearson Correlation Coefficient which ensures that the remaining pixels after dropping are still important, defined as:

$$Coherency(f_s(\hat{\boldsymbol{Y}})) = 100 \times \frac{Cov(f_s(\hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}})), f_s(\hat{\boldsymbol{Y}})}{\sigma(f_s(\hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}})))\sigma(f_s(\hat{\boldsymbol{Y}}))}, \tag{2.6}$$

where $Cov(.,.)$ is the covariance. A higher coherency is better. Complexity is the L1 norm of the output PA map.

$$Complexity(f_s(\hat{\boldsymbol{Y}})) = 100 \times ||f_s(\hat{\boldsymbol{Y}})||_1. \tag{2.7}$$

Complexity is used to measure how cluttered a PA map is. For a good PA map, complexity should be minimal. As it has been shown in the literature, the metrics in (eq. (2.4)), (eq. (2.6)) and (eq. (2.7)), cannot be used individually to evaluate a PA method [151]. Instead, ADCC combined with computation time gives us a reliable overall metric of how a PA method is performing.

2.5.0.4 Performance Evaluation

The proposed method has been compared to combinations of ResNet-18 and MobileNetV2 with SOTA PA methods on both medical and natural scenes datasets. At test time, Dropout it not enabled for these standard methods (only for the proposed method). In table 2.1, our method outperforms all compared CNN–PA method combinations in ADCC, with the exception of Score-CAM. Similar to our approach, Score-CAM performs multiple forward passes on perturbed inputs, making it less susceptible to overconfidence. Although the effects of our method are similar to the effects achieved by Score-CAM, our method is more parallelisable. By introducing slight perturbations to the model itself, we follow the perturbation principle of Score-CAM while also producing an explicit risk estimation.

In fig. 2.7, we can see our method includes more regions (top part) of the image and slightly sharper in localisation both of which would help the coherency and average drop metrics. Risk estimations from (eq. (2.3)) are also displayed and provide an added visualisation for a surgeon to evaluate both the model and the model's explanation. As it can be seen, areas of low CV match the areas of high PA values which shows the proposed explainability method is precise. During intraoperative surgery a surgeon can visualise an assistive DL model's explanation in order to assess or compare with what pixel's were found to be most relevant to the classification. Whilst a PA method does not need to highlight a salient region, it does need to provide a fair and precise PA map for the surgeon to use. Our PA method removes overconfidence and also provides an added visualisation of relative precision, improving on SOTA PA methods. In figs. 2.7 to 2.9 we can also see that by adding our method on top, the PA methods produce heat maps that more closely resemble those generated by ScoreCAM.

In table 2.2, we show that this method generalises to the natural scenes domain, as also illustrated visually in fig. 2.10. We believe that the better performance of our method is because of the random dropping of features taking place during Dropout at test time which helps to suppress noise in the estimated enhanced PA map. The combination of Recipro-CAM with our proposed method improves performance (increases ADCC) at the expense of increasing the computational complexity. We believe that this could be reduced using a batched implementation of Recipro-CAM. We attribute slow down in SmoothGradCAM++ when Dropout is applied during test time to the perturbations it adds on top of the PA method. Our validation study shows that Grad-CAM, Grad-CAM++ and Recipro-CAM are often leading in terms of speed as expected from the literature.

Table 2.1: ADCC vs Latency Study for ResNet-18 and MobileNetV2 on pCLE Dataset. Latency(ms) is the time to compute one PA map using a batch size of one.

ResNet18				
	Original		Proposed	
PA Method	ADCC ↑	Latency ↓	ADCC ↑	Latency ↓
Grad-CAM	76.6	**75.6**	**77.7**	9.7
Grad-CAM++	76.2	**5.5**	**78.2**	10.5
SmoothGradCAM++	74.8	**70.7**	**75.6**	103.9
Score-CAM	**80.5**	**121.4**	**80.5**	1267.2
Recipro-CAM	66.5	**3.55**	**75.9**	36.4
MobileNetV2				
	Original		Proposed	
PA Method	ADCC ↑	Latency ↓	ADCC ↑	Latency ↓
Grad-CAM	29.3	**8.8**	**48.0**	12.5
Grad-CAM++	37.8	**8.7**	**59.5**	13.4
SmoothGradCAM++	24.5	**71.3**	**37.1**	88.8
Score-CAM	**43.9**	**315.1**	**43.9**	3154.3
Recipro-CAM	33.3	**5.8**	**55.5**	65.6

Table 2.2: ADCC vs Latency Study for ResNet-50 ImageNet Dataset. Latency(ms) is the time to compute one PA map using a batch size of one

ResNet50				
	Original		Proposed	
PA Method	ADCC ↑	Latency ↓	ADCC ↑	Latency ↓
Grad-CAM	67.9	**11.3**	**72.4**	25.4
Grad-CAM++	67.6	**11.3**	**72.1**	26.3
SmoothGradCAM++	64.6	**84.6**	**71.8**	134.0
Score-CAM	**74.3**	**134.0**	**74.3**	14147.9
Recipro-CAM	63.0	**7.5**	**70.9**	86.6

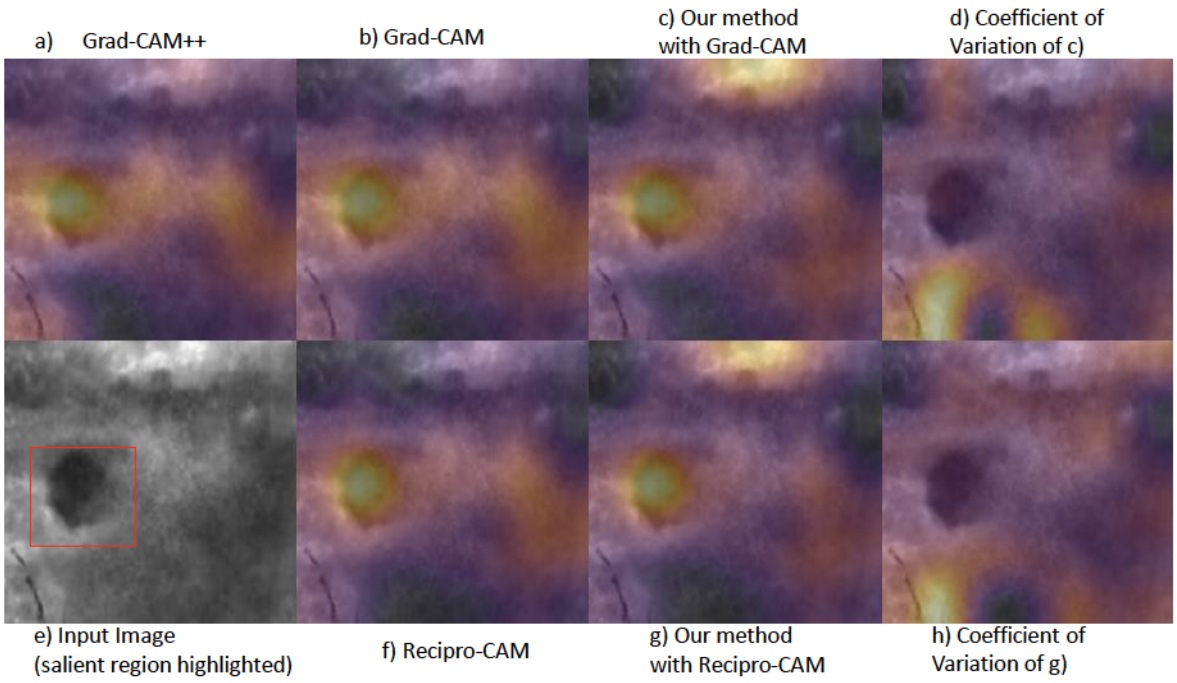


Figure 2.7: PA maps generated using ResNet18 on MNG pCLE images. a) Grad-CAM++ PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of MNG with the salient region highlighted with red bounding box f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest.

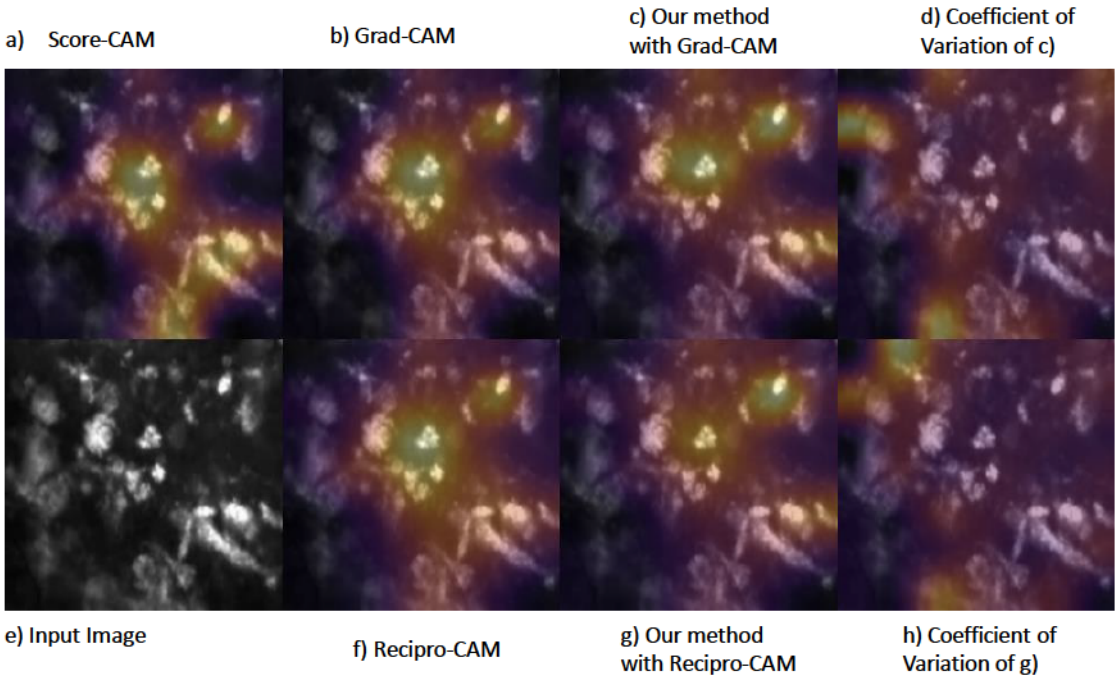


Figure 2.8: PA maps generated using ResNet18 on GBM pCLE images. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of GBM f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest.

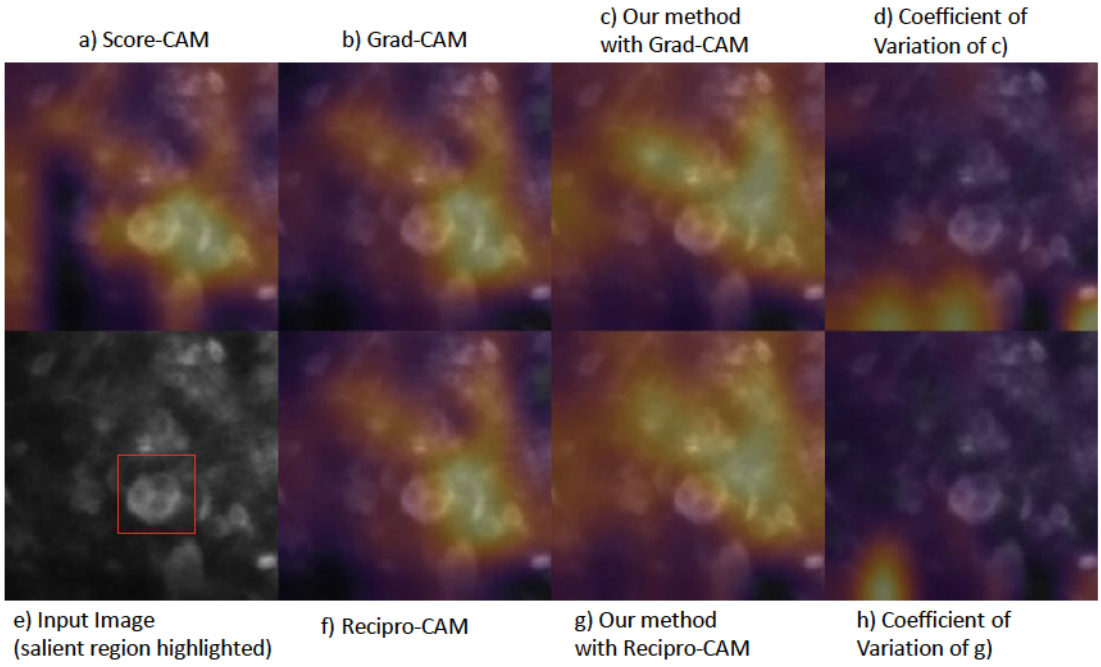


Figure 2.9: PA maps generated using ResNet18 on invasive ductal carcinoma (IDC) pCLE images. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original pCLE image of IDC with the salient region highlighted with red bounding box f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest.

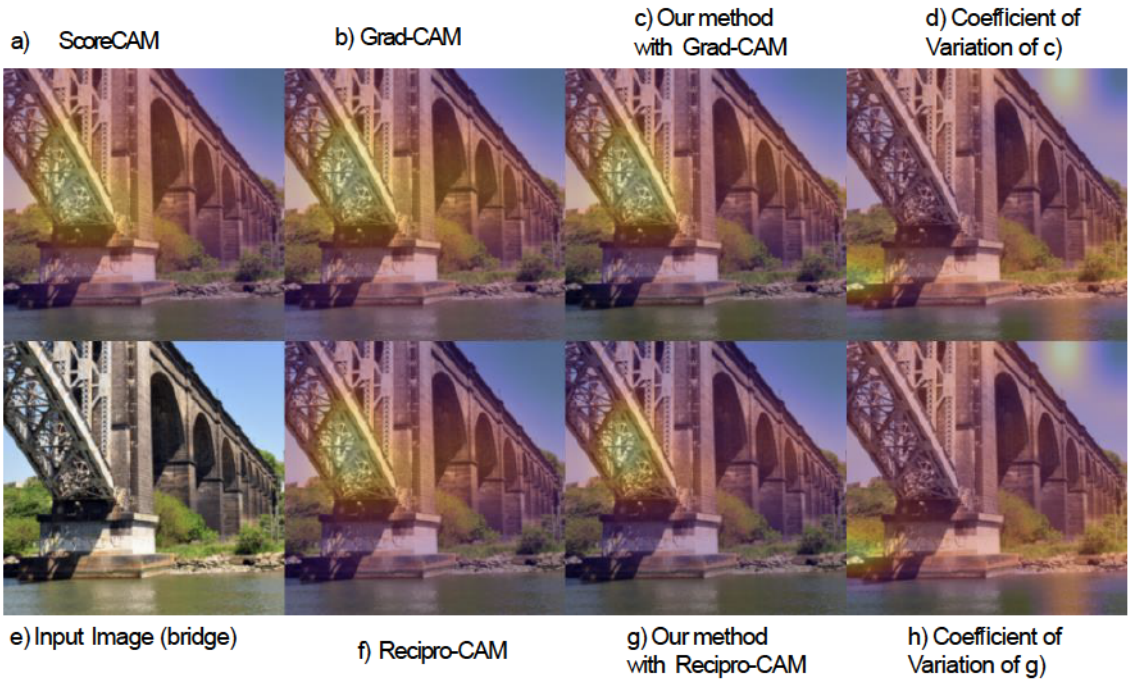


Figure 2.10: PA maps generated using ResNet50 on ImageNet. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) original image f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest.

2.5.1 T Study

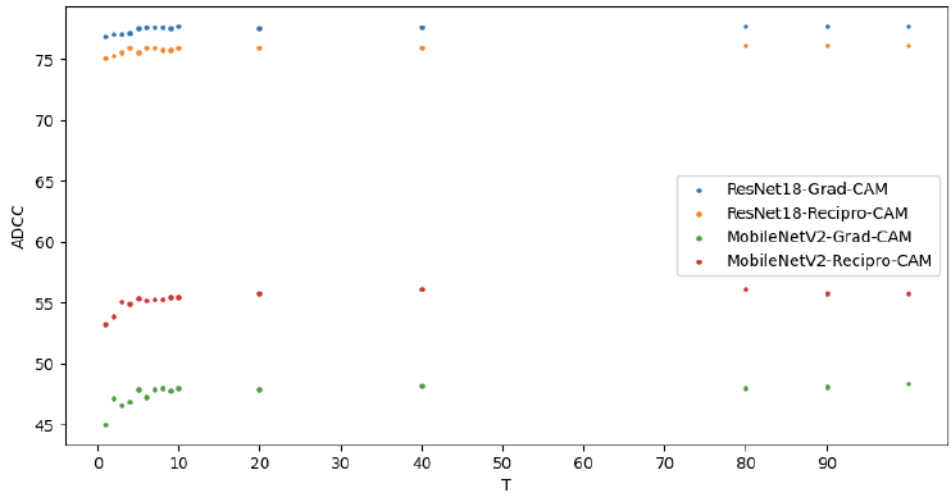


Figure 2.11: ADCC vs T parameter study for ResNet-18 and MobileNetV2 on pCLE dataset.

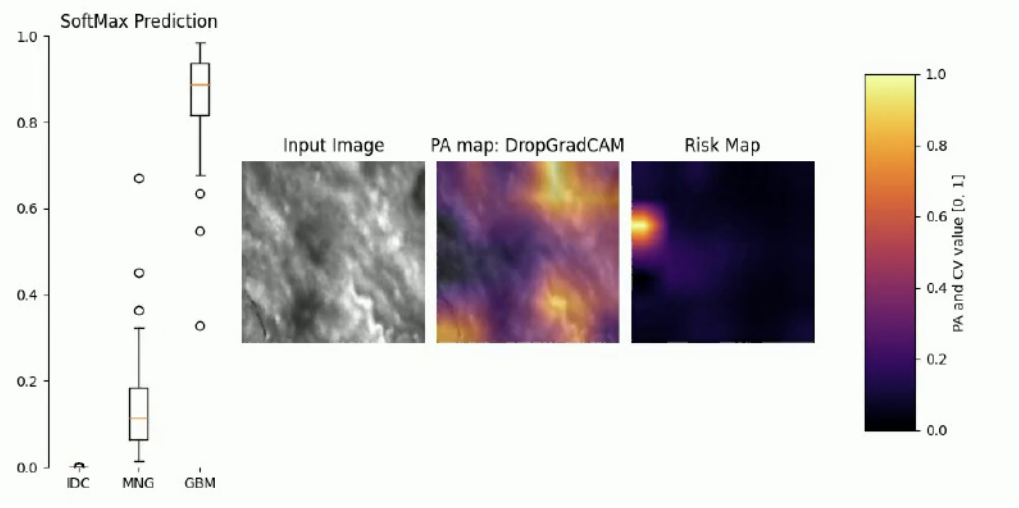
A parameter search was performed to find the optimal value of T. As show in table G.3 there is a positive correlation between ADCC and the value of T. With increase of T, there is an implicit increase in latency. We found the optimal trade-off of ADCC against latency to be $T = 10$. The raw values of ADCC against T can be seen visually in fig. 2.11.

2.5.2 Clinical Use

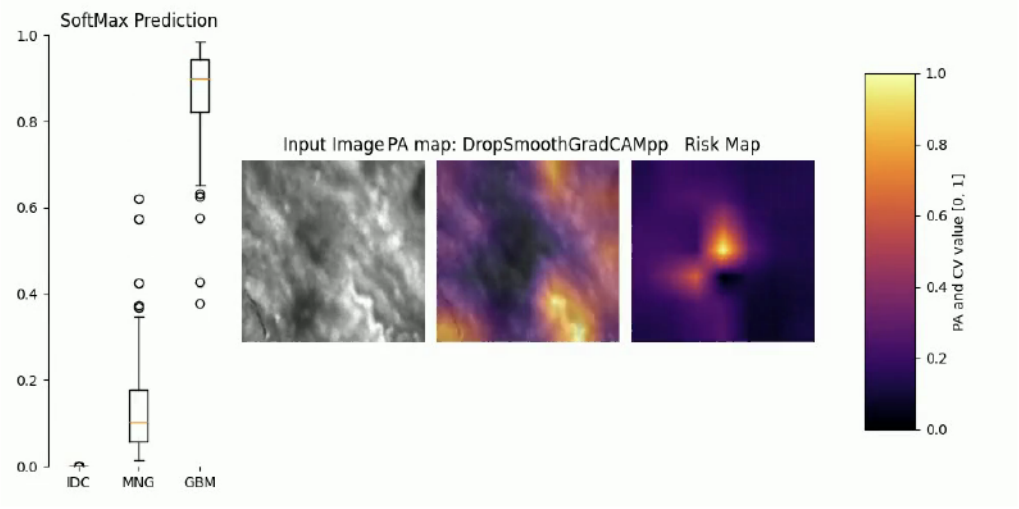
Clinicians are often far removed from the technical details of DL explainability and uncertainty estimation. PA maps, such as Grad-CAM, can highlight relevant regions, but without an indication of reliability they risk being over-interpreted. Expecting clinicians to judge "which heatmap is more statistically meaningful" is impractical. Integrating Monte Carlo Dropout with PA maps addresses this limitation by showing not only where the model focuses, but also how confident it is. Compared to figs. 2.3 and 2.4, which offers a point estimate, context-free attributions, figs. 2.12 and 2.13 displays model variability, allowing clinicians to contrast between stable and uncertain saliency regions.

In the intraoperative environment, the practical utility of explainability methods must be reconsidered to focus on clinical workflow optimization. Rather than requiring surgeons to validate the superiority of Grad-CAM over ReciproCAM, or ResNet over MobileNet architectures, the critical question becomes which complete pipeline provides actionable intraoperative value. The technical minutiae of explainability method selection should be abstracted away from the clinician, who requires clear, interpretable guidance during time-critical surgical procedures.

We hypothesize that intraoperative explainability and uncertainty quantification should be consolidated into a unified framework. An integrated prediction such as "MNG with 80% confidence" provides more actionable clinical value than isolated heatmaps indicating "where" or "why" the

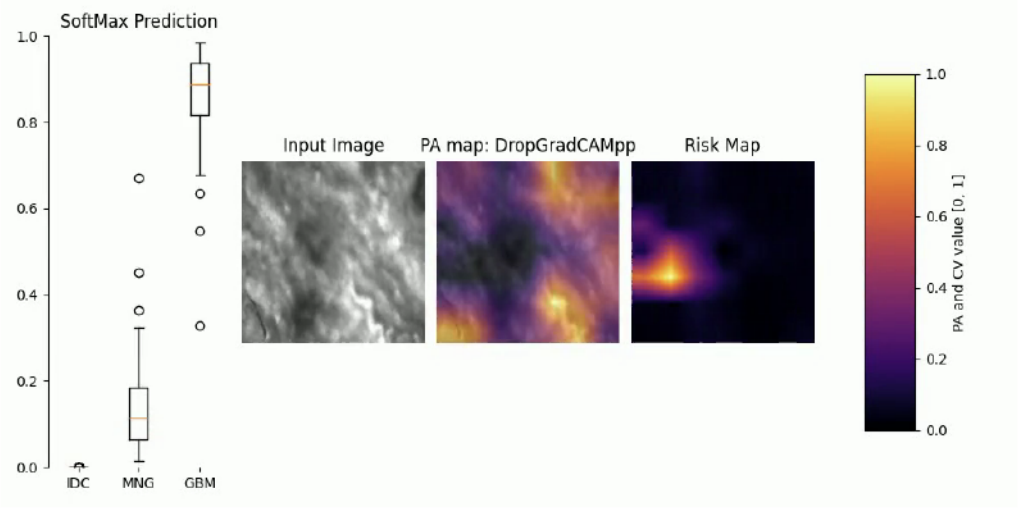


(a) DropGradCAM PA method.

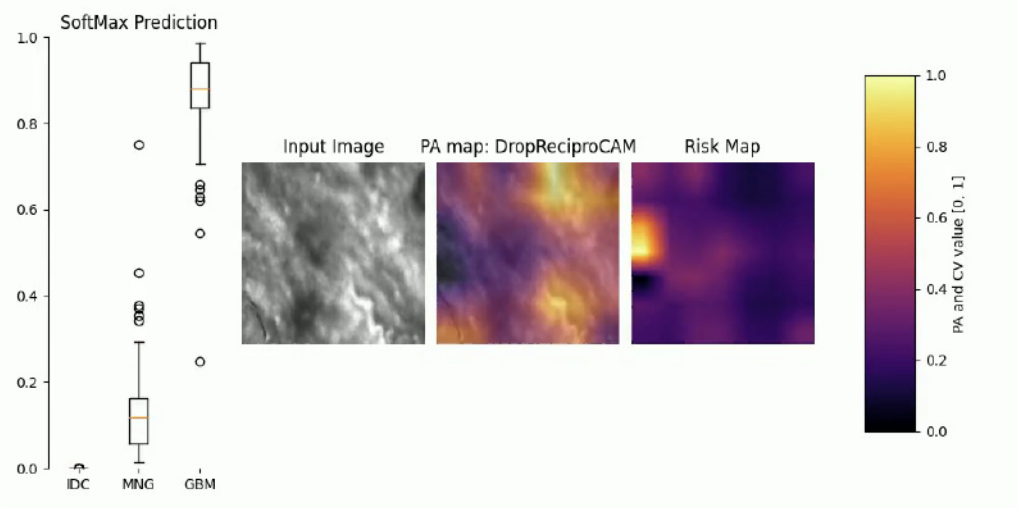


(b) DropSmoothGradCAM++ PA method.

Figure 2.12: Comparison of Pixel Attribution (PA) methods for pCLE classification using ResNet18. Column 1: softmax predictions (box plots after MC dropout); Col 2: input image; Col 3: PA map; Col 4: risk estimation using MC dropout (100 runs).



(a) DropGradCAM++ PA method.



(b) DropReciproCAM PA method.

Figure 2.13: Comparison of additional Pixel Attribution (PA) methods for pCLE classification using ResNet18. Format follows fig. 2.12, featuring softmax box plots, input images, PA maps, and risk estimation via MC dropout (100 runs).

model reached its conclusion, particularly for single-step CNN architectures. While multi-stage model pipelines may benefit from intermediate explanations, the surgical context demands streamlined, confidence-calibrated outputs that directly inform decision-making without cognitive overhead.

The four-class classification framework used in this study (Healthy, MNG, GBM, IDC) provides a broad test environment for evaluating explainability methods but does not mirror intraoperative scenarios, where surgeons usually know the tumor type. This discrepancy highlights the need for deployment-specific validation that goes beyond accuracy to address clinical utility and workflow integration.

Pre-Deployment Clinical Validation Framework. We propose that any model intended for intraoperative deployment must undergo rigorous clinical validation analogous to medical device approval processes. Beyond standard performance metrics of speed and accuracy, models should demonstrate that learned representations correspond to clinically relevant features with minimal uncertainty deviation, rather than spurious correlations that may fail under surgical conditions.

The proposed validation protocol involves surgeons evaluating classification results paired with our Monte Carlo Dropout-enhanced PA maps across representative test cases. This process validates two critical aspects: first, that the model consistently learns class-relevant features with low variance across multiple inference passes; second, that uncertainty estimates provide meaningful bounds that align with clinical expectations of model reliability.

Experimental Validation: Binary MNG Detection. Consider a binary classification model developed for intraoperative MNG detection. The experimental validation setup requires a carefully curated dataset comprising histopathological sections from confirmed MNG cases alongside healthy alternatives.

The required dataset labels include: (1) binary ground truth classification (MNG vs. healthy), (2) confidence scores from the original pathologist review, (3) region-of-interest annotations marking diagnostically relevant areas such as psammoma bodies, and (4) inter-observer agreement metrics from multiple pathologist reviews to establish baseline human uncertainty. Additionally, clinical metadata including patient demographics, size measurements, and pre-operative imaging should be available to enable comprehensive model evaluation across diverse clinical presentations.

During validation, surgeons would assess classification results accompanied by our uncertainty-aware saliency maps, displaying both the model's classification confidence and the spatial consistency of its PA maps. The evaluation protocol would require surgeons to determine whether highlighted regions correspond to known MNG features while assessing whether uncertainty bounds appropriately reflect diagnostic ambiguity in borderline cases. Each prediction would be scored for clinical relevance, with particular attention to cases where model uncertainty aligns with regions

of known diagnostic difficulty.

This validation approach ensures that deployed models not only achieve statistical benchmarks but demonstrate clinically meaningful feature learning with quantified reliability, bridging the gap between algorithmic performance and surgical decision support. By consolidating explainability with uncertainty quantification, we provide clinicians with the interpretable, confidence-calibrated guidance essential for safe intraoperative AI integration.

2.6 Conclusion

In this work we have introduced the first combination of risk in PA methods. Using our proposed framework we not only improve on all the tested non-perturbation based SOTA PA method's ADCC performances but also produce an estimation of risk on the output PA values. From this work we hope to encourage trust between the surgeon and DL models by reducing overconfidence.

While we believe this method represents one of the most effective ways to integrate risk estimation with explainability without compromising CNN performance, it is important to acknowledge that the clinical demand for such methods is not universal. In many routine cases, explainability and uncertainty visualisations are of limited interest to clinicians, who may prioritise speed and accuracy over model interpretability. Furthermore, the rapid emergence of extremely large, high-performing models may reduce the perceived necessity for post-hoc interpretability, as their empirical reliability could overshadow the need for detailed explanations. Nonetheless, for high-stakes or ambiguous cases where model error carries significant consequences, the ability to jointly convey the models focus and confidence remains a valuable safeguard.

Chapter 3

Hyperspectral Segmentation and Augmentation

This chapter explores the application of hyperspectral imaging and segmentation in neurosurgery. We introduce a novel data augmentation technique, calibration-jitter, designed to make models more robust to calibration inconsistencies in hyperspectral imagery.

The work in this chapter was adapted from our previously published paper in Healthcare Technology Letters as part of the Augmented Environments for Computer Assisted Interventions (AE-CAI) Workshop in the satellite events of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2024 [159], the code to reproduce this method can be found here.

3.1 Overview

This chapter provides an introduction to Hyperspectral Imaging (HSI) and reviews its current applications, particularly in the context of deep learning-based segmentation. For a wide array of segmentation tasks, ranging from neurosurgery [7, 184] and remote sensing [17] to the analysis of natural scenes [28], HSI offers a powerful alternative to the conventional RGB imaging. In contrast to conventional imaging, which captures light across a few wide spectral bands, HSI gathers data from hundreds of narrow, contiguous bands, creating a distinct "spectral fingerprint" for every pixel. This detailed spectral information reveals material compositions and properties that are not visible to the naked eye or through standard cameras. This capability has proven to be highly beneficial within the medical domain. For instance, recent research has demonstrated that HSI enables more effective automatic segmentation than standard RGB imaging for tasks such as organ delineation [175].

The detailed spectral data provided by HSI is of critical importance in disciplines like neurosurgery, where it aids the precise identification and delineation of various brain tissues during intricate operations, ultimately aiming to improve surgical outcomes and minimise complications [7]. Furthermore, when integrated with the predictive power of machine and deep learning, HSI can offer significant decision support, especially for applications like tumour resection that demand the accurate definition of boundaries [103]. A challenge arises because traditional data augmentation, which improves model robustness, is often not possible for hyperspectral data due to its unique structure. The limitation of these conventional techniques is their inability to simulate the subtle variations characteristic of hyperspectral acquisition, particularly the fluctuations in lighting conditions and the differing sensitivities of sensors.

To overcome these shortcomings, this work introduces a novel spectral augmentation method termed *Calibration-Jitter*. The proposed technique is designed to enhance the capacity of deep learning models to generalise and extract robust spectral features by emulating the variations that occur within the hyperspectral calibration procedure itself.

Calibration-Jitter leverages the inherent variability within the hyperspectral calibration process to generate a more diverse set of training examples. This is achieved by introducing controlled perturbations, or "jitter," that simulate fluctuations in illumination and sensor sensitivity, thereby producing augmented data more representative of real-world acquisition conditions. Such critical spectral variations cannot be captured by conventional geometric augmentations.

3.2 Introduction

3.2.1 Hyperspectral Imaging: Light-Tissue Interactions

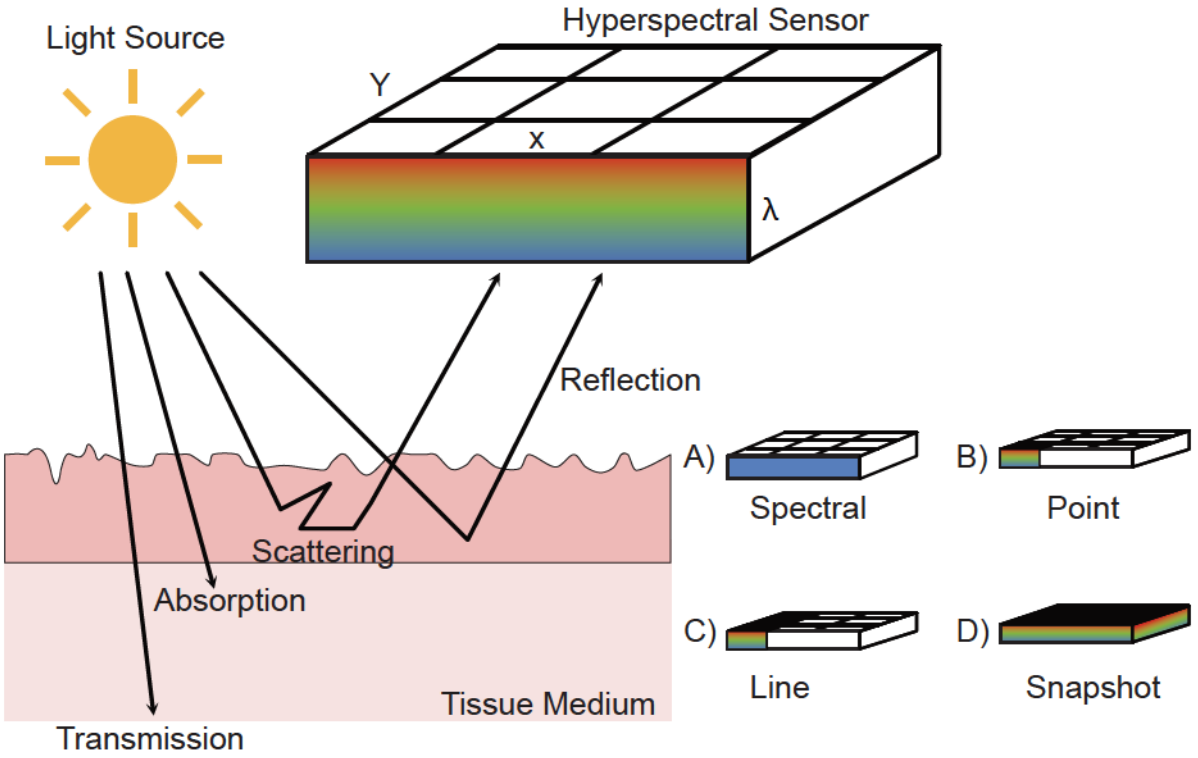


Figure 3.1: Light tissue interaction diagram. A)-D) represent hyperspectral sensor modes.

Conventional imaging modalities, such as standard digital cameras, operate by capturing light within three broad spectral bands: red, green, and blue. This is typically achieved using a Bayer filter array, which places a mosaic of tiny colour filters over the camera sensor's pixels. Each pixel records the intensity of light that passes through its specific red, green, or blue filter, providing only a single colour value at a given location. A complete colour image is then reconstructed through a process of interpolation, where the missing two colour values for each pixel are estimated from its neighbours. While this method is highly effective for capturing human-perceptible colour, it discards a large amount of spectral information, limiting its ability to differentiate materials based on the reflected light.

Multispectral Imaging, a step up from conventional cameras, captures light in a discrete number of bands, typically between 5 and 20, usually in the visible and infrared sections of the electromagnetic spectrum. HSI, in contrast, takes this concept further by collecting information in a much larger number of narrow and contiguous spectral bands (typically more than 20), providing a far richer spectral signature. This capability allows for the better differentiation of tissues and pathologies based on their distinct light-absorption and scattering characteristics.

The fundamental principle underlying HSI is the interaction of light with biological tissue. When a sample is illuminated, the light can be absorbed, scattered, reflected, or transmitted. This is depicted on the left hand side of fig. 3.1. The unique spectral fingerprint of a tissue is primarily dictated by the concentration and composition of its constituent chromophores, such as oxyhaemoglobin, deoxyhaemoglobin, water, lipids, and melanin. The absorption of light by these chromophores follows the well-established Beer-Lambert Law, which relates the attenuation of light to the properties of the material through which the light is travelling. The law is given by:

$$A = \log_{10}\left(\frac{I_0}{I}\right) = \varepsilon c l$$

where A is the absorbance, I_0 is the intensity of the incident light, I is the intensity of the transmitted light, ε is the molar absorptivity coefficient of the chromophore, c is the concentration of the chromophore, and l is the path length of the light through the sample.

In biological tissues, scattering plays a significant role and complicates this relationship. Light photons are redirected by cellular components and nearby tissue, causing them to travel a longer path through the tissue, thereby increasing the probability of absorption. This combination of absorption and scattering results in a unique reflectance spectrum for each tissue type in situ, which serves as the basis for HSI analysis.

3.2.1.1 Formation of a Hyperspectral Data Cube

A hyperspectral imager captures the spectral information at each spatial pixel, generating a three-dimensional data structure known as the "data cube", "hyperspectral stack" or "hypercube." This is depicted at the top of fig. 3.1. This cube contains two spatial dimensions (x, y) and one spectral dimension (λ), with each pixel containing a full spectrum of the reflected light. The intensity value of a pixel at a specific wavelength is a function of the incoming light, the sensor's response, and the tissue's properties. A simplified camera equation can model this relationship:

$$P_\lambda = I_\lambda \cdot R_\lambda \cdot T_\lambda \cdot Q_\lambda$$

where P_λ is the pixel value at wavelength λ, I_λ is the spectral radiance of the incident light, R_λ is the reflectance of the tissue, T_λ is the spectral transmittance of the optics, and Q_λ is the spectral quantum efficiency of the detector.

Different sensor architectures are used to acquire this data cube, each with trade-offs in terms of speed, complexity, and application suitability. This is shown graphically on the right hand side of fig. 3.1.

The most common modes are:

- **Spatial Scanning (Pushbroom):** A line-scan sensor captures one spatial line (x) for all wavelengths (λ) at a time. The second spatial dimension (y) is built by moving the sensor or the sample. This is suitable for fast-moving scenes or conveyor belt applications.
- **Wavelength Scanning (Spectral):** This mode captures a full two-dimensional image (x, y) for a single wavelength (λ). The full spectrum is built by tuning a filter or using a monochromator to scan through different wavelengths.
- **Snapshot Imaging:** This method captures the entire data cube (x, y, λ) in a single exposure. While offering high speed and suitability for dynamic events, these systems often have lower spatial or spectral resolution.

3.2.2 Calibration

Robust performance in clinical settings such as neurosurgery requires careful handling of data variability. One major source is lighting in the Operating Theatre (OT), which can fluctuate with different surgical equipment and light sources. Additional variability arises from the hyperspectral sensor itself, as imaging non-rigid, dynamic tissue such as the brain introduces motion artifacts and subtle spectral distortions. Differences in spectral resolution across sensors further affect sensitivity to lighting conditions. Together, these factors undermine the consistency of spectral signatures across acquisitions.

Temporal and motion-related errors are commonly corrected through stack registration [54], while spectral inconsistencies are addressed via white-dark calibration [12, 178]. White-dark calibration is a crucial step in hyperspectral imaging, converting raw sensor data into meaningful reflectance values and ensuring accurate representation of the tissue's spectral characteristics. Thus, effective calibration is essential for data standardization, enabling machine learning models to capture true tissue properties rather than artifacts introduced by the environment or sensor.

This calibration is dependent on two critical reference measurements. First, the white reference, denoted as $W(\lambda)$, is acquired by imaging a diffuse reflector with near-perfect reflectivity (e.g., a calibrated white panel) under identical illumination conditions as the target object. This captures the combined spectral profile of the illumination source and the wavelength-dependent sensitivity of the sensor. Second, the dark reference, $D(\lambda)$, is measured with the camera lens capped or shutter closed. This step corrects for the sensor's dark current (internal electrical noise) and any stray ambient light.

Subtracting the dark reference from the raw measurement, expressed as $Raw(\lambda) - D(\lambda)$, yields the spectral radiance that reaches the sensor. This intermediate calculation quantifies the light energy reflected or emitted by the sample at each wavelength, having been adjusted for the sensor's inherent noise.

The calibrated reflectance, $R(\lambda)$, for any given wavelength λ is subsequently computed with the following formula:

$$R(\lambda) = \frac{Raw(\lambda) - D(\lambda)}{W(\lambda) - D(\lambda)} \tag{3.1}$$

In this equation, subtracting the dark reference from both the raw and white measurements effectively isolates the signal generated purely by the light source. The subsequent division by the corrected white reference normalizes this signal, scaling the resulting reflectance values to a standard range between 0 and 1. This normalized value signifies the proportion of incident light that the sample reflects at each wavelength, thereby making the spectral data comparable across different sensors and variable lighting conditions.

3.2.3 Hyperspectral Segmentation

The primary objective of hyperspectral segmentation is to divide an image into meaningful regions or objects. This partition is achieved by exploiting the unique spectral signatures of the constituent materials, which are captured in the rich spectral information inherent to hyperspectral data.

3.2.3.1 Spectral Similarity

A foundational approach to hyperspectral segmentation is rooted in the principle that materials with similar properties exhibit similar spectral curves. This gives rise to methods that operate by directly comparing the spectral signature of each pixel against a known reference spectrum.

Among these techniques, Spectral Angle (SA) Mapping is widely used. SA conceptualizes each pixel's spectral signature as a vector within a high-dimensional space. It then computes the angle (θ) between a pixel's vector and a reference vector. A smaller resulting angle signifies greater spectral similarity, implying that the pixel belongs to the same material class as the reference. The popularity of SA stems from its computational efficiency and its inherent insensitivity to variations in illumination.

The formula to calculate the SA (θ) is as follows:

$$\theta = \cos^{-1}\left(\frac{\mathbf{t} \cdot \mathbf{r}}{\|\mathbf{t}\|\|\mathbf{r}\|}\right) \tag{3.2}$$

In this equation, $\mathbf{t}$ represents the target pixel's spectral vector and $\mathbf{r}$ is the reference spectrum's vector. The term $\mathbf{t} \cdot \mathbf{r}$ denotes their dot product, while $\|\mathbf{t}\|$ and $\|\mathbf{r}\|$ are their respective vector magnitudes.

A related technique, Cosine Similarity, is the mathematical equivalent of the cosine of the spectral angle. This metric evaluates the cosine of the angle between two spectral vectors, where a value

of 1 indicates identical spectra and 0 denotes no similarity. Both SA and Cosine Similarity are well-suited for identifying materials by comparing them to a known spectral library, and their application in remote sensing is well-established [19].

3.2.3.2 Pixel-wise Spectral Classification

An alternative to direct spectral comparison is to frame segmentation as a classification task. In this paradigm, each pixel is classified on an individual basis according to its complete spectral signature. The entire spectral curve of a pixel serves as the input to a classifier, which then assigns the pixel to a predefined class [10].

Initial approaches in this domain typically depended on statistical methods and manual feature engineering. For instance, Principal Component Analysis (PCA) was frequently employed to reduce the high dimensionality of hyperspectral data, transforming it into a more manageable set of uncorrelated components to streamline classification. Another traditional technique, Linear Discriminant Analysis (LDA), seeks a linear combination of features that best separates different classes [63]. Subsequently, more advanced machine learning models such as Support Vector Machines (SVM) and Random Forest (RF) became prevalent for pixel-wise classification tasks [103]. By assigning a class to every spectrum, these models generate a comprehensive segmentation map of the observed scene.

3.2.3.3 Automatic Segmentation Models

In recent years, the focus of the field has transitioned to fully automatic segmentation models, with a strong emphasis on deep learning architectures. These models possess the ability to learn intricate spectral and spatial features directly from the input data, moving beyond the limitations of pixel-wise classification. By incorporating spatial context, they produce segmentation maps that are more coherent and accurate [182]. Models such as the U-Net, which is a benchmark architecture in biomedical image segmentation [165], have had their core principles effectively adapted for hyperspectral data analysis [175]. The characteristic encoder-decoder structure of the U-Net enables it to capture both high-level semantic context and precise, fine-grained details, a crucial capability for accurately delineating tissue boundaries in complex biological scenes. Other prominent deep learning models, including SegFormer [216] and various Convolutional Neural Networks (CNN) designs [80], have also achieved State of the Art (SOTA) by effectively integrating both spatial and spectral information. While these end-to-end models mitigate the need for manual feature design, their high performance and resilience to real-world challenges like calibration and spectral shifts are contingent upon access to large annotated training datasets.

This chapter presents a solution to this data-dependency challenge by introducing a novel augmentation technique derived from the principles of hyperspectral calibration. The proposed method simulates realistic spectral variations more effectively than existing techniques, thereby reducing

the reliance on massive, physically acquired datasets. By training on data augmented in this manner, deep learning models can develop a more robust understanding of tissue characteristics, which in turn leads to enhanced segmentation performance and increased reliability in clinical applications.

3.3 Related Works

The intraoperative application of Machine Learning (ML) and particularly Deep Learning (DL) models to analyse hyperspectral data for tissue characterization serves as a critical tool in surgical decision-making. DL architectures are adept at identifying correlations within high-dimensional spaces, a task where traditional statistical models frequently fall short. In recent years, surgical applications of hyperspectral data have leveraged DL for diverse tasks such as detecting tongue tumour [192], segmenting retinal vasculature [69], and performing semantic segmentation in neurosurgical scenes [56]. While DL models efficiently manage the high dimensionality of hyperspectral data [141] and provide strong predictive capabilities, these advantages come at the expense of interpretability and the risk of bias [111]. Often described as "black boxes", their internal decision-making processes are opaque. This lack of transparency can conceal errors or biases inherited from the training data, which may result in undesirable or inequitable predictions.

The development of effective Computer-aided Intervention (CAI) tools necessitates the creation of HSI datasets containing spectral data from a wide variety of surgical procedures for accurate tissue characterization. A crucial step in this process is calibration, which makes these diverse acquisitions comparable by mitigating environmental factors specific to the data capture, such as camera and light source variations [99]. Once calibrated, the HSI data can be processed with statistical or DL methods to generate predictive information that assists surgeons. For these models to be adopted as CAI tools in high-stakes environments, they must be robust, interpretable, and highly accurate [50]. Fulfilling these requirements with DL models demands extensive training data to maximize predictive power and reduce bias [74]. When data is scarce, DL models are prone to overfitting, where they memorize training examples rather than learning generalizable underlying patterns. Consequently, the development of augmentation techniques to expand the volume of training data is essential for enhancing the performance and accuracy of these CAI tools.

Within the RGB domain, a standard practice to mitigate overfitting and enhance the feature learning capabilities of DL models is the routine use of geometric, regularization, and colour augmentations. By expanding the available data, these techniques encourage models to learn features that are more representative of the task at hand [41, 180].

Geometric augmentations involve applying minor perturbations to images through transformations like rotation, flipping, or cropping. By artificially enlarging the size and variety of the training set, these methods provide the DL model with a richer set of examples. Such perturbations discourage the model from merely memorizing the spatial arrangement of the original images, promoting better generalization [180].

Methods classified as regularization augmentations, including Cutout [49] and Cutmix [219], function by deliberately removing portions of image information. This forces the model to learn under more difficult conditions [180]. For example, Cutout randomly excises square regions from an im-

age, whereas Cutmix involves cutting a patch from one image and pasting it onto another. By introducing these types of occlusions and variations, the model is compelled to learn more generalizable representations, which results in improved performance and less overfitting. The effectiveness of these techniques has been demonstrated across a range of computer vision applications, such as image classification, object detection, and semantic segmentation.

The purpose of colour augmentations is to ensure that a model learns the semantic and textural features of objects rather than relying on colour cues alone [180]. Adjustments to brightness, contrast, and hue, along with the swapping of colour channels, contribute to overcoming biases related to lighting, thereby promoting invariance to illumination and colour shifts. This prevents the model from overfitting to the specific colour and lighting characteristics of the training set, enabling it to generalize more effectively to the diverse lighting scenarios, camera sensors, and imaging conditions it is likely to face during deployment [180].

Recent research in the medical field highlights the need for data augmentations to be tailored to the specific modality of the training data [191]. Although standard RGB augmentations are effective for natural images, medical imaging modalities like CT, MRI, or ultrasound possess unique characteristics and artifacts that demand specialized augmentation strategies for effective learning. Moreover, it is critical that any applied augmentation technique reflects the natural variations present in the dataset to prevent the introduction of new biases or the learning of non-representative features. Consequently, augmentations should be designed and parametrized to be invariant to specific datasets and architectures, focusing instead on the intrinsic properties of the data modality.

In medical applications, augmentations developed for the hyperspectral domain are frequently designed with a specific task or architecture in mind. For instance, an 'organ cutmix' technique was proposed for semantic segmentation in surgery, which involves altering the spatial location of organs [175]. Although this method was shown to improve model robustness, it is dependent on dense annotations of the surgical scene, which are often unavailable. In the context of lesion classification, sparse coding algorithms such as K-SVD and A+ have been suggested for augmenting hyperspectral images [220]. These methods learn an overcomplete dictionary from the training data and then generate new image patches by creating sparse combinations of dictionary elements. While effective for data synthesis, sparse coding is limited in its ability to realistically simulate real-world variations like changes in illumination, sensor noise, or occlusions.

Beyond the medical field, the domain of remote sensing provides several HSI-specific augmentation methods. Random occlusion has been introduced as a technique to reduce overfitting in CNNs for HSI classification [78]. This approach involves randomly obscuring rectangular spatial areas within the HSI data during training, which acts as a form of regularization. By injecting this occlusion noise, the model is discouraged from relying on local spatial patterns and is instead prompted to learn more globally relevant features. While this method shows promise, it does not address

other critical aspects of data variation, such as illumination changes, sensor noise, or geometric transforms. Other techniques include the use of superpixel patches [2], pixel-level combinations [106], and label manipulation [77], which modify the ground truth in a controlled way to introduce more variation. However, while these methods are effective at capturing local spatial structures, they risk overlooking important pixel-level characteristics.

To summarise, although several data augmentation strategies exist for HSI, such as occlusion, superpixel methods, and multi-label sampling, a significant gap is their inability to model variations in brightness and colour. Standard colour augmentations common in the RGB domain, including adjustments to brightness, contrast, and hue, are not directly transferable to hyperspectral data. For surgical HSI in particular, applying such methods could introduce physically implausible or unrealistic alterations to the spectral signatures of tissues. Any augmentation technique intended for the surgical domain must carefully preserve the spectral continuity that defines a tissue's spectral fingerprint. This underscores the need for developing augmentation methods specifically for surgical HSI that can simulate changes in illumination and sensor properties without corrupting the inherent spectral response of biological tissue.

3.4 Methodology

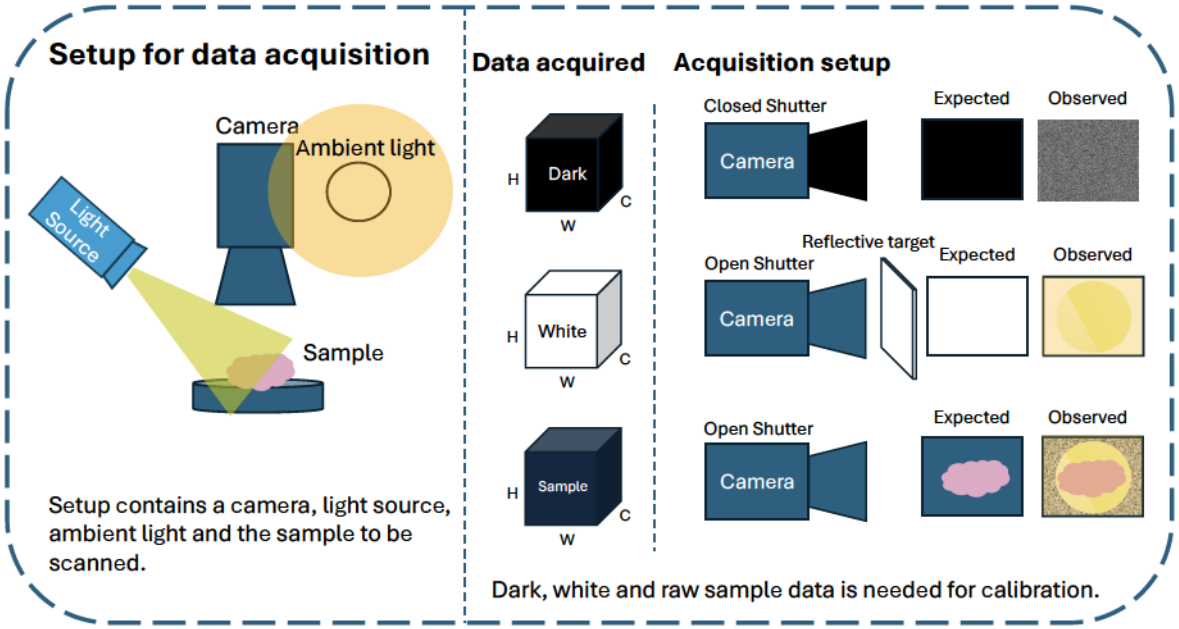


Figure 3.2: Configuration of a hyperspectral imaging system for surgical applications. The left panel illustrates a common acquisition setup, featuring a hyperspectral camera mounted on an articulated support or robotic arm and positioned above the surgical area to capture spectral data. Illumination for the scene is provided by either dedicated hyperspectral lighting systems or standard broadband light sources. The right panel depicts the standard calibration process. Permission to reuse this figure can be found in section A.0.3.

A typical setup for hyperspectral acquisition is depicted in fig. 3.2. This illustration highlights that while a camera records light from a specimen, the resulting data is also influenced by external factors. These include artefacts like light leakage and sensor noise, which can be approximated using a dark reference, as well as variations in illumination, which are compensated for by a white reference. During a surgical procedure, the hyperspectral camera captures a three-dimensional data cube. In this cube, two dimensions represent spatial information (x, y), while the third dimension provides spectral information (λ), generally covering the visible and near-infrared portions of the spectrum. Each hyperspectral image is formally defined as $X \in \mathbb{R}^{C \times H \times W}$, where C denotes the number of spectral channels, and H and W represent the height and width of the image, respectively. For a segmentation task, each hyperspectral image is paired with a corresponding annotation mask of the surgical scene, $y \in \{0, 1, ..., Y_N\}^{H \times W}$, where each pixel's value indicates its class up to the total number of classes, Y_N. The complete dataset is therefore formally expressed as

$$\text{Dataset} = \{(X_i, y_i)\}_{i=1}^{N},$$

with N being the total number of images. As noted in [103], such a dataset often includes ac-

quisitions from multiple cohorts and potentially different locations, which introduces significant variability. The comparability of spectral measurements can be compromised by fluctuations in illumination conditions and camera settings that occur between different acquisition locations or times. The characteristics of the illumination source directly affect the colour and intensity of the captured light. Similarly, differences in camera sensitivity and response across various wavelengths can lead to disparate colour and brightness representations of an identical surgical scene, a problem that can persist even among cameras of the same model. Consequently, calibration is an indispensable step in hyperspectral data acquisition to guarantee that spectral measurements are comparable. The calibrated reflectance, $R(\lambda)$, for a specific wavelength λ is determined by eq. (3.1).

In this formula, $Raw(\lambda)$ is the unprocessed digital value registered by the sensor. The term $W(\lambda)$ represents the white reference measurement, which is acquired from a near-perfect diffuse reflector (such as a highly reflective surface). The term $D(\lambda)$ is the dark reference measurement, captured with the camera's shutter closed or its lens capped. The white reference, $W(\lambda)$, is used to correct for the non-uniform spectral response of both the sensor and the illumination source. In parallel, the dark reference, $D(\lambda)$, compensates for dark current and background light leakage. Following this white-dark calibration process, the reflectance values, $R(\lambda)$, are constrained to a range between 0 and 1, signifying the proportion of incident light reflected by the sample at each wavelength.

In a typical hyperspectral data processing pipeline, the spectra undergo smoothing and are then normalised on a per-pixel basis using a min-max approach after calibration [103]. The purpose of spectral smoothing is to further refine the data by mitigating random noise and minor fluctuations. Subsequently, min-max normalisation addresses shifts in illumination intensity [175], based on the principle that the spectral signature of a specific tissue type should remain consistent regardless of the source brightness or its distance from the camera. This procedure effectively removes the absolute intensity information at each pixel, $R(x, y)$, which in turn precludes the use of standard colour augmentation techniques, such as per-channel brightness or contrast adjustments, for simulating illumination changes in the normalised data.

3.4.1 Geometric Augmentation

Data augmentation is a broad field encompassing various techniques, which can be broadly classified into four major categories: geometric, colour, adversarial, and external knowledge-based methods, as depicted in fig. 3.3. When training datasets are small, they often suffer from limited spatial variation. This can lead to a DL model memorising the spatial locations of particular tissues instead of learning their intrinsic spectral features. To mitigate this, we utilise geometric augmentations to artificially expand the diversity and spatial representation of tissues within the dataset. A standard augmentation pipeline, which we refer to as "normal", incorporates geometric transfor-

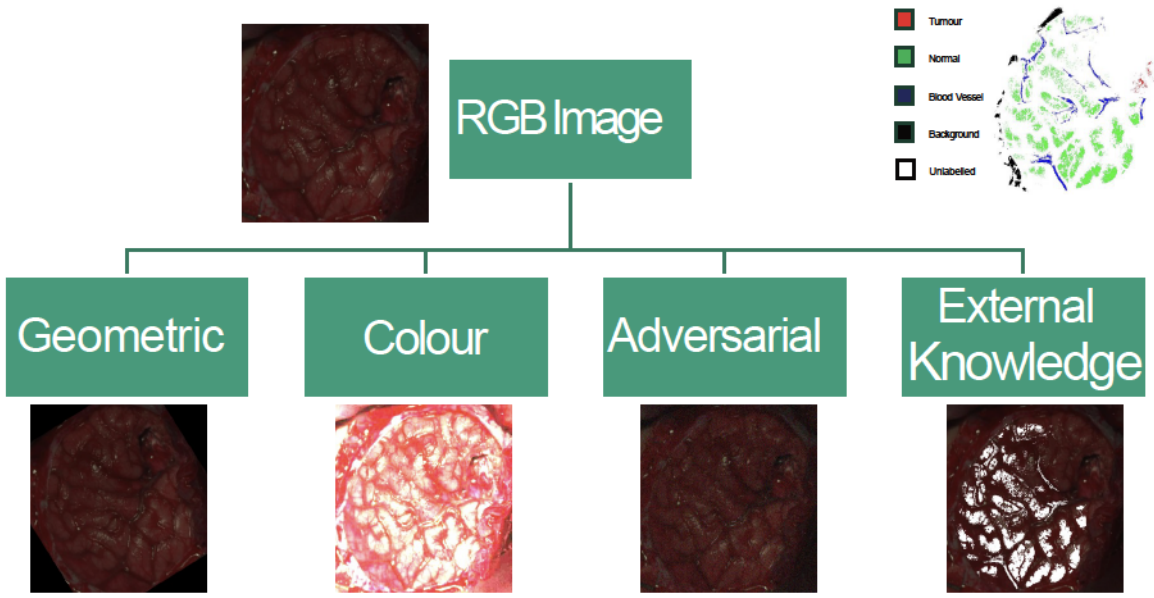


Figure 3.3: An illustration of the macro types of augmentation from a sample of the Human Imaging Benchmark (Hib) dataset.

mations like rotation, cropping, and flipping to simulate varied orientations and perspectives of the surgical field.

The rotation of an image X is defined as:

$$X^{rot} = \text{Rot}_{\theta}(X), \tag{3.3}$$

where Rot_{θ} signifies the rotation operation by a given angle θ, producing the rotated hyperspectral image X^{rot}. The cropping operation is defined by:

$$X^{crop} = \text{Crop}_{x_1:x_2,y_1:y_2}(X), \tag{3.4}$$

where $\text{Crop}_{x_1:x_2,y_1:y_2}$ is the function that extracts a region of interest bounded by the pixel coordinates (x_1, y_1) and (x_2, y_2), yielding the cropped image X^{crop}. The flipping operation is represented as:

$$X^{flip} = \text{Flip}_{axis}(X), \tag{3.5}$$

where Flip_{axis} denotes the flipping operation performed along a designated axis (e.g., horizontal or vertical), resulting in the flipped image X^{flip}. Although the geometric augmentations described above are valuable for introducing spatial variations and diversifying the training data, they do not address the variability caused by different illumination conditions and sensor sensitivities. These factors can have a substantial impact on the spectral signatures recorded in the hyperspectral images.

Our goal is to enrich the hyperspectral dataset by introducing variations that emulate real-world shifts in illumination and sensor noise. Per-pixel min-max normalisation, like L1, standardises intensity values across all channels. This prevents direct replication of illumination or sensor noise per channel, which would disrupt spectral continuity. To overcome this limitation, we introduce *Calibration-Jitter*, a novel data augmentation technique for hyperspectral imagery. This method operates by adjusting the white reference data captured during the calibration phase, crucially before the non-invertible normalisation step is applied. The white reference acts as a standardised proxy for the illumination conditions at the time of data capture. Controlled offsets allow us to simulate varied illumination scenarios and generate augmented samples. Specifically, the white reference is offset by a coefficient that is proportional to its standard deviation, which ensures the augmented samples stay within a plausible intensity range for each image and channel. The new augmented white reference is calculated as follows:

$$W(\lambda)^{\text{augmented}} = W(\lambda) + \text{STD}(\lambda) \cdot \mathbb{U}(\text{low}, \text{high}), \tag{3.6}$$

where $W(\lambda)$ is the original white reference, $\text{STD}(\lambda)$ is the standard deviation of the white reference at each wavelength, and $\mathbb{U}(\text{low}, \text{high})$ is a coefficient sampled from a uniform distribution $\mathbb{U}$ defined by a lower and upper bound (low and high). This coefficient dictates the magnitude of the offset. The resulting augmented white reference, $W(\lambda)^{\text{augmented}}$, is then substituted into eq. (3.1) to calibrate the raw sample data.

The standard calibration pipeline is illustrated in fig. 3.4, where our proposed method is highlighted within the red box. By integrating these augmented samples into the training regimen, the model is exposed to a much wider spectrum of illumination variations. This enhances its capacity to learn features that are invariant to illumination and improves its generalisation to new scenarios with different lighting conditions. This augmentation technique serves as a powerful complement to the "normal" geometric transformations, offering a comprehensive strategy for data augmentation that addresses spatial, sensor, and illumination variability. Ultimately, this leads to an improvement in the performance and generalisation capabilities of models for hyperspectral tissue characterisation. Following the application of these augmentation strategies, a final pre-processing step is performed to standardise the data before it is input to the model. Z-score normalisation is applied across the dataset, a step which is beneficial for ensuring invariance to intensity and enhancing numerical stability during training. Each spectral channel is normalised using the mean and standard deviation derived from the entire training fold, according to the formula:

$$\text{Normalized}_C = \frac{X_C - \mu_C}{\sigma_C}, \tag{3.7}$$

where μ_C and σ_C represent the mean and standard deviation of channel C across the training data, respectively.

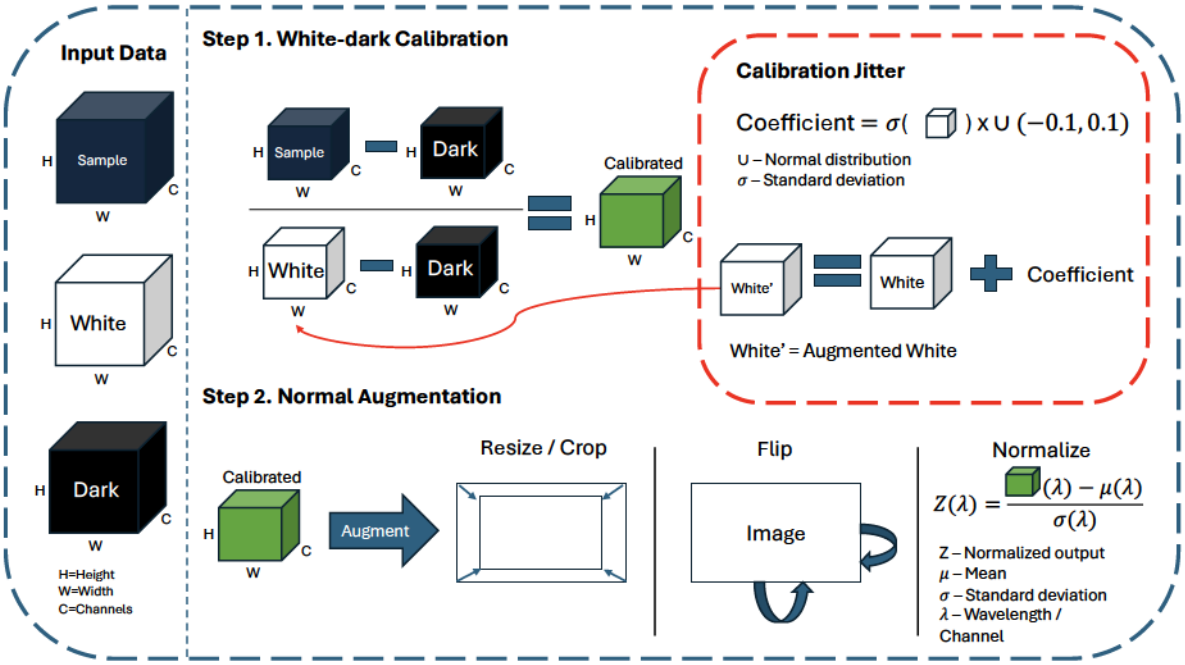


Figure 3.4: An illustration of the data loading and augmentation pipeline. The process involves an initial calibration of the hyperspectral data, followed by an augmentation step. Our proposed method introduces a key modification at step 1, where the white reference itself is augmented prior to being used for calibration [159]. Permission to reuse this figure can be found in section A.0.3.

3.5 Results

3.5.1 Dataset

For this study, we utilise the Hib dataset [103], which consists of 61 images acquired from 34 different patients. At the time of its release, it was the most extensive hyperspectral dataset available for neurosurgery. Nevertheless, its scale is modest when compared to the vast datasets commonly found in the RGB domain, a fact that underscores the scarcity of annotated data in this field and reinforces the critical need for effective data augmentation techniques. In our experiments, we adhere to the per-patient data splits originally defined in the benchmark paper to ensure comparability.

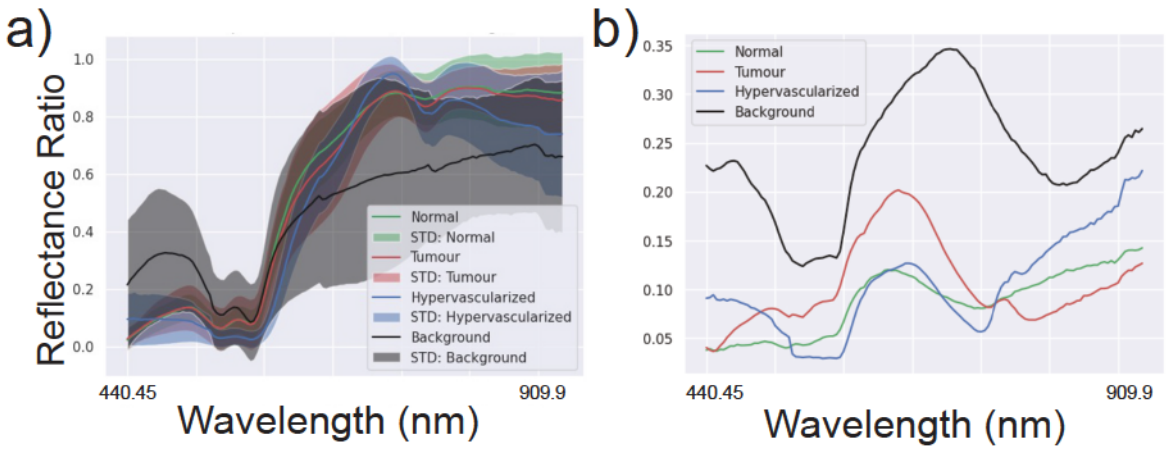


Figure 3.5: a) Mean reflectance spectra per class of the pre-processed dataset across the channels. b) Standard deviation per class of the spectra in the pre-processed dataset across the channels.

The hyperspectral images within the dataset vary in their spatial dimensions (height and width) but are consistent in their spectral resolution, initially containing 826 wavelength bands. These bands span a range from 400.5 nm to 1000.7 nm, with a sampling interval of 0.73 nm. To prepare the raw data for our models, we implemented the same pre-processing pipeline described in the original publication. This procedure involves five key stages: (1) performing a white-dark calibration to normalise the reflectance values; (2) applying a smoothing filter with a kernel size of 5 across the spectral dimension to reduce noise; (3) trimming the spectral range by removing the first 56 and the last 126 bands, which often contain more noise and less useful information; (4) resampling the channel dimension to an optimal sampling interval as recommended by [127]; and finally, (5) applying min-max normalisation along the spectral dimension for each pixel. The mean and standard deviation of the spectra is displayed in fig. 3.5.

Upon completion of this pipeline, the resulting data cubes have a final channel count of 128 bands, covering a spectral range between 440.45 nm and 909.9 nm. The dataset includes sparse annota-

tions that were generated using a custom algorithm detailed in [127]. The data is categorised into five labels: unlabelled, normal, tumour, blood vessel, and background. An analysis of the label distribution across the entire dataset reveals a significant class imbalance. The pixel counts are 308,328 for normal tissue, 37,399 for tumour, 112,407 for blood vessels, and 445,540 for background, while the unlabelled class contains 11,805,660 pixels. This means the unlabelled pixels constitute 92.9% of the dataset, whereas the critical tumour class accounts for merely 0.3%. Consequently, the problem is defined as an imbalanced, sparsely-annotated semantic segmentation task.

3.5.2 Segmentation Models

In contrast to the benchmark paper, which advocates for a two-step process involving per-pixel spectral classification followed by spatial smoothing [103], we adopt a one-step, end-to-end semantic segmentation pipeline. This approach leverages SOTA deep learning networks by processing hyperspectral data in a single forward pass. For this study, we selected two prominent architectures to evaluate: the U-Net, a convolutional network that is a de facto standard for medical image segmentation [165], and SegFormer, a more recent Transformer-based architecture that has achieved state-of-the-art results on many segmentation benchmarks [216]. For the U-Net, we specifically employ a ResNet34 backbone pre-trained on ImageNet [80].

Our implementation was built using the PyTorch framework [143], with the training loop and hardware management handled by PyTorch Lightning [57]. The specific model implementations were sourced from established libraries: Segmentation Models PyTorch (SMP) for the U-Net [86] and the Hugging Face Transformers library for SegFormer [214]. All models were trained on a single Nvidia RTX 3090 GPU using the AdamW optimiser [113]. Instead of exhaustive hyperparameter optimisation, a learning rate demonstrating reliable convergence was selected through preliminary experiments and reused throughout. This standard practice balances computational efficiency with rigorous comparison between model variants. We used a learning rate of 1×10^{-4}, betas of $(0.9, 0.999)$, a weight decay of 1×10^{-2}, and a consistent batch size of 8.

To address the significant class imbalance inherent in the dataset, we employed a partial weighted cross-entropy loss function. The unlabelled class is ignored entirely during the loss computation. The loss is calculated as:

$$L = -\sum_{i=1}^{Y_N} w_c \cdot y_{o,c} \cdot \log(p_{o,c}), \tag{3.8}$$

where L is the final loss value, Y_N is the number of annotated classes, $y_{o,c}$ is the binary indicator if observation o belongs to class c, $p_{o,c}$ is the predicted probability of that observation, and w_c is the class-specific weight. The weight term w_c is calculated using the inverse frequency of each class within the labelled portion of the dataset, ensuring that under-represented classes have a greater

impact on the loss. The formula for this weight is:

$$w_c = \frac{\frac{1}{f_c}}{\sum_{j=1}^{Y_N} \frac{1}{f_j}}, \tag{3.9}$$

where f_c is the frequency of class c in the training data.

3.5.3 Performance Evaluation Metrics

To evaluate the performance of our semantic segmentation models, we employ several standard metrics. It is important to note that, due to the sparse nature of the ground truth, all metrics are calculated exclusively using the pixels that have a valid class label, while unlabelled regions are ignored.

The primary metric for assessing the overall performance, where no single class is considered more important than another, is the macro-averaged F1 score. This score is the mean of the F1 scores for each individual class, providing a balanced measure of a model's effectiveness. The F1 score for a given class i is calculated as the harmonic mean of its precision and recall:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{3.10}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{3.11}$$

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{3.12}$$

The final reported F1 score is then the average over all Y_N classes: $\text{F1} = \frac{1}{Y_N} \sum_{i=1}^{Y_N} \text{F1}_i$. In these formulae, TP_i, FP_i, and FN_i represent the number of true positives, false positives, and false negatives for class i, respectively.

For a more comprehensive evaluation, we also report several supplementary metrics. The overall accuracy (Acc) measures the fraction of correctly classified pixels across all classes:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.13}$$

Additionally, we report the True Positive Rate (TPR), which is equivalent to recall, and the False Positive Rate (FPR) to better understand class-specific performance:

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN} \tag{3.14}$$

For these two metrics, TP, TN, FP, and FN represent the total counts of true positives, true negatives, false positives, and false negatives across all evaluated pixels.

3.5.4 Data Augmentations

To rigorously evaluate the effectiveness of our proposed *Calibration-Jitter* technique, we designed a comparative study against three distinct baseline augmentation pipelines. These baselines consist of: (1) no augmentations, serving as a minimal performance benchmark; (2) a standard pipeline employing only geometric augmentations; and (3) a brightness bias pipeline that adds a spectrally-agnostic intensity shift, a common technique in RGB image processing.

To ensure a fair comparison, all experiments share a common set of foundational augmentation steps and a consistent validation pipeline. The standard geometric augmentation sequence, applied during the training phase of multiple pipelines, is composed of a resize to $(310, 280)$, a random crop to $(256, 256)$, and both random horizontal and vertical flips (each with a 50% probability). All pipelines, for both training and validation, conclude with Z-score normalisation to ensure numerical stability. For the validation set across all experiments, a deterministic pipeline is used which only resizes the image to $(256, 256)$ before applying the Z-score normalisation. The specific components of each training pipeline are detailed in Table 3.1.

Table 3.1: An overview of the different data augmentation pipelines used for training.

Pipeline Name	**Augmentation Steps**
No Augmentation	1. Resize to $(256, 256)$ 2. Z-score normalisation
Standard Geometric	1. Resize to $(310, 280)$ 2. Random Crop to $(256, 256)$ 3. Random Horizontal & Vertical Flips 4. Z-score normalisation
Brightness Bias	1. Standard Geometric steps (1-3) 2. Brightness Bias shift $\mathbb{U}(-0.2, 0.2)$ 3. Z-score normalisation
Calibration-Jitter (Proposed)	1. *Calibration-Jitter* $\mathbb{U}(-0.1, 0.1)$ 2. Standard Geometric steps (1-3) 3. Z-score normalisation

The motivation for the brightness bias pipeline is to ascertain that any performance improvements from our proposed method are not merely due to the regularising effect of an arbitrary spectral bias. This technique adds a coefficient drawn from a uniform distribution, $\mathbb{U}(-0.2, 0.2)$, to the entire image post-calibration. While simple, this can disrupt spectral continuity and result in physically unrealistic data.

In contrast, our *Calibration-Jitter* pipeline is designed to produce physically plausible variations. It is applied before the geometric augmentations. The method computes the standard deviation for each spectral channel, STD(λ), multiplies it by a coefficient drawn from a uniform distribution $\mathbb{U}(-0.1, 0.1)$, and adds this value to the original white reference data before calibration. The pa-

rameters for this uniform distribution were determined to be optimal through an ablation study, the results of which are presented in Table 3.2.

The qualitative impact of *Calibration-Jitter* is demonstrated in fig. 3.6. As the figure illustrates, the technique generates continuous spectra that effectively simulate different lighting conditions, thereby exposing the model to a more diverse and realistic range of illumination variations during training. The pseudo-RGB images particularly highlight this effect, showing that the lower and upper bounds of the jitter coefficient correlate with visibly darker and brighter scene representations, respectively.

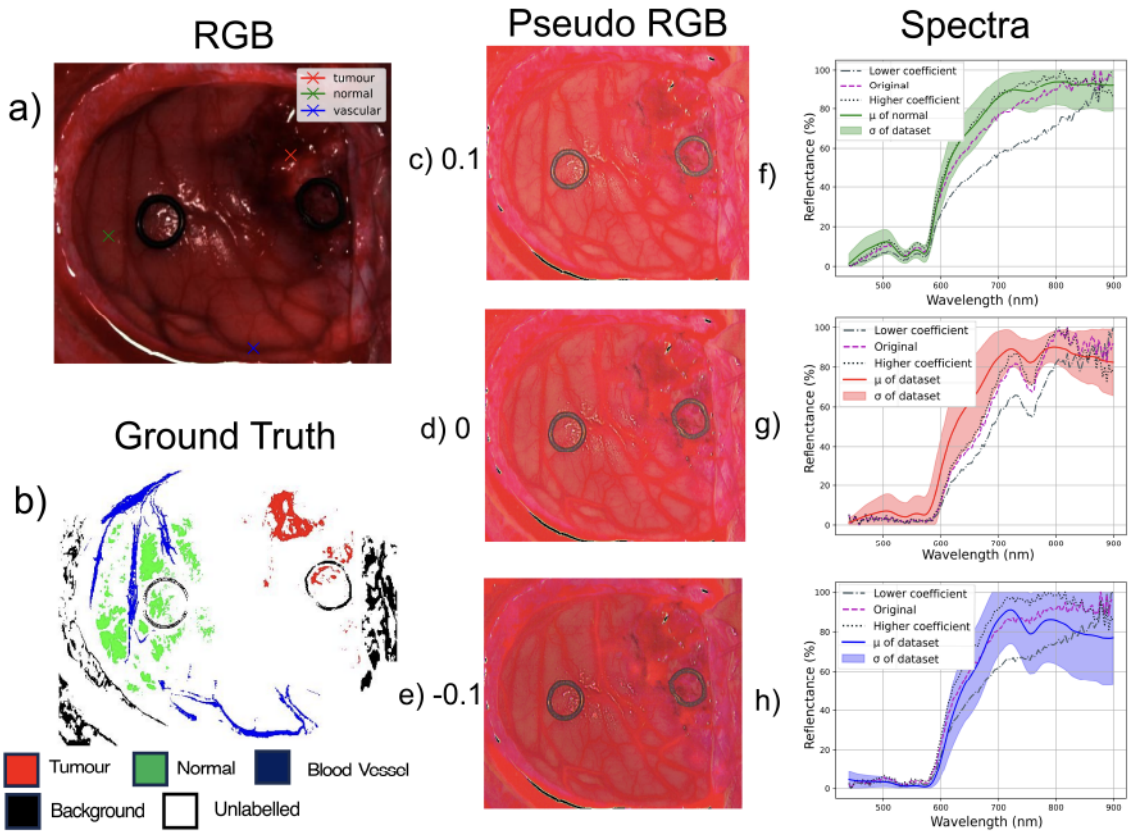


Figure 3.6: Visualisation of the Calibration-Jitter augmentation on a sample image. (a) The original RGB image with points of interest labelled. (b) The corresponding ground truth annotation mask. (c-e) A comparison of the pseudo-RGB image and the spectra from the labelled points under different jitter coefficients: (c) a positive coefficient of 0.1, (d) a zero coefficient representing standard white-dark calibration, and (e) a negative coefficient of -0.1. The pseudo-RGB is formed using the closest available wavelengths to 630 nm (red), 540 nm (green), and 480 nm (blue). (f-h) The effect of high (0.1), low (-0.1), and zero jitter coefficients on the spectral profiles of different tissue classes, shown alongside the mean and standard deviation for each class across the entire dataset. The classes are: (f) normal tissue, (g) tumour, and (h) blood vessel [159]. Permission to reuse this figure can be found in section A.0.3.

3.5.5 Comparative Analysis

Table 3.2: Parameter study of the coefficient (Coeff) utilised to manage the effect of *Calibration-Jitter* [159]. Permission to reuse this Table can be found in section A.0.3.

Model	Coeff	F1 (%)
UNet(ResNet34)	$\mathbb{U}(-0.05, 0.05)$	72.04
	$\mathbb{U}(-0.1, 0.1)$	**73.04**
	$\mathbb{U}(-0.15, 0.15)$	71.39

Table 3.3: A summary of model performance metrics, all of which are reported as the macro average across the classes. The evaluated metrics include the F1-score, accuracy (Acc), true positive rate (TPR), and false positive rate (FPR) [159]. The license to reproduce this table is located in section A.0.3.

Model	Augmentation	F1 (%) ↑	Acc (%) ↑	TPR (%) ↑	FPR (%) ↓
UNet(ResNet34)	None	59.55	87.54	58.91	9.38
	Normal	69.95	91.68	73.00	5.70
	Bias	71.53	92.10	73.40	5.49
	Calibration-Jitter	**73.04**	**92.82**	**73.46**	**5.15**
SegFormer	None	68.79	90.00	72.21	6.17
	Normal	72.78	91.69	75.35	5.87
	Bias	72.32	91.72	76.03	**5.71**
	Calibration-Jitter	**74.35**	**91.77**	**78.46**	**5.71**

The proposed *Calibration-Jitter* pipeline incorporates a randomly chosen coefficient that governs the augmentation's intensity in relation to the source signal. An investigation was conducted to identify the optimal parameter by evaluating several values, as presented in table 3.2. A coefficient that is excessively small may render the augmentation's impact negligible, whereas one that is too large could potentially obscure the original signal. The results from our study suggest that the ideal parameter range for this dataset is $\mathbb{U}(-0.1, 0.1)$. Consequently, this coefficient range was adopted for all subsequent experiments.

The performance without any augmentation, detailed in table 3.3 under 'None', establishes a crucial baseline for comparison. As anticipated, introducing the 'Normal' set of augmentations into the data pipeline yielded a substantial enhancement across all reported metrics for both the UNet and SegFormer models.

Incorporating the 'Bias' augmentation led to a further increase in the F1-score for the UNet model (+1.58%), although it concurrently resulted in a slight decrease for the SegFormer model (-0.46%). Despite this mixed F1 performance, 'Bias' augmentation improved the accuracy, TPR, and FPR for both architectures, which suggests it may act as a regulariser, thereby improving the models' learning process.

When compared directly with the proposed *Calibration-Jitter*, as detailed in table 3.3, the 'Bias' pipeline demonstrated inferior performance for our best model, SegFormer, across the F1-score (-2.03%), Accuracy (-0.05%), and TPR (-2.43%), with a similar trend observed for the UNet model. This outcome suggests that a random intensity bias, which does not account for spectral continuity, is a less effective augmentation strategy than our *Calibration-Jitter* technique. Indeed, the highest Macro F1-score of 74.35% was attained by SegFormer trained with *Calibration-Jitter*, surpassing the 73.04% achieved by UNet with the same augmentation. Crucially, in contrast to the 'Bias' augmentation, *Calibration-Jitter* provides consistent improvements over the 'Normal' baseline across all metrics, highlighting the effectiveness of our proposed method.

Finally, a comparison was made between our top-performing result (SegFormer with *Calibration-Jitter*) and the Hib benchmark base model. The benchmark's highest reported Macro F1-score is 70.2%, achieved using a pixel-wise neural network classification combined with a subsequent clustering step (a spectral/spatial approach). Our proposed method exceeds this benchmark, with an improvement of approximately 4% on the F1-macro score. Furthermore, from a qualitative standpoint, *Calibration-Jitter* contributes to a lower False Positive Rate, as depicted in fig. 3.7, which results in more precise tumour detection than the benchmark method [103].

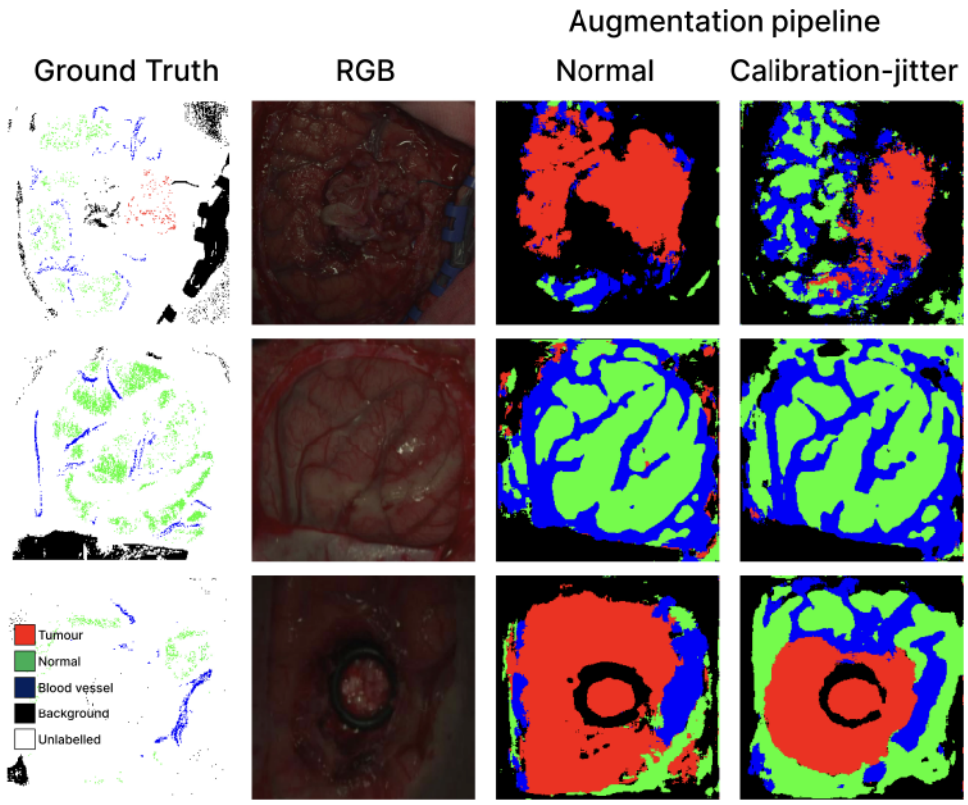


Figure 3.7: Qualitative comparison of SegFormer outputs showing the effect of data augmentation. Columns: ground truth, RGB image, normal augmentation (without *Calibration-Jitter*), and with *Calibration-Jitter*. The latter reduces false positive tumour predictions, especially under challenging lighting [159]. Reuse details in section A.0.3. Permission for the reuse of this figure is detailed in section A.0.3.

3.6 Discussion

The fundamental objective of *Calibration-Jitter* is to compel the model to learn spectral characteristics arising from variations in illumination and sensor sensitivity. Mastering these features is vital for the successful deployment of the model on novel hyperspectral data. Such characteristics are frequently neglected or insufficiently exploited in conventional methodologies.

The efficacy of our proposed illumination augmentation technique is clearly demonstrated by the results in Table 3.3, which show strong performance across various evaluation metrics and segmentation architectures. The 5-fold cross-validation confirms that models trained using *Calibration-Jitter* achieve consistently superior F1-scores, accuracy, TPR, and FPR for both U-Net and SegFormer. This consistent outperformance suggests that the augmentation method is largely independent of the specific model architecture, reinforcing its capacity to boost the generalisation and overall performance of hyperspectral tissue segmentation models.

The qualitative impact of *Calibration-Jitter* is visually evident in fig. 3.7. Models trained with this augmentation generate more precisely localised segmentations and successfully remove noisy artefacts, such as the isolated islands of false predictions observed in the output of the non-augmented model. These visual results show that by simulating varied lighting conditions via controlled white reference offsets, our proposed technique enables the model to learn more robust features. This, in turn, yields enhanced segmentation performance and helps to temper overconfident predictions when encountering unseen data.

To provide a direct comparison, we re-implemented the top-performing model from [103], SVM, and obtained a macro F1-score of 71.23%. Visually, the results from its per-pixel spectral classification method produce a higher resolution output, as seen in fig. 3.8; however, we observed a large number of false positive tumour classifications surrounding the blood vessels, an artefact similar to that seen in automatic segmentation models without sufficient regularisation. In contrast, our automatic segmentation approach significantly reduces implementation complexity by functioning as a single-step process, which obviates the need for iterative loops over the entire image. We foresee that with larger datasets, such end-to-end segmentation models will become increasingly capable of learning the necessary spatial and spectral correlations simultaneously.

In summary, the presented results, alongside the parameter study, offer compelling evidence for the potential of our semantic segmentation approach incorporating *Calibration-Jitter*. By deliberately forcing the model to learn spectral features under varied illumination and sensor sensitivity conditions, we have obtained performance that surpasses that achieved using only geometric augmentations.

Acknowledging the inherent challenges in applying brightness augmentation to hyperspectral data, primarily stemming from the calibration process detailed in fig. 3.6, our technique modifies inten-

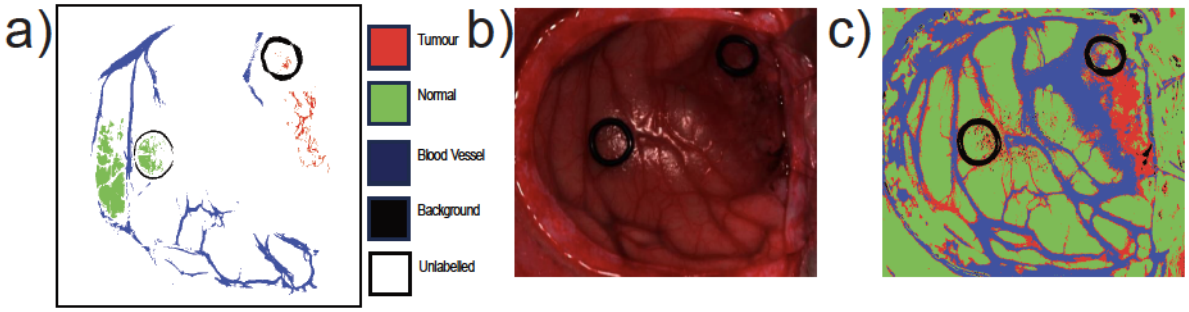


Figure 3.8: Results from a spectral classification SVM model following the same implementation from [103]. a) Ground truth label. b) RGB image of target. c) SVM prediction of a).

sity on a per-channel basis. *Calibration-Jitter* was intentionally designed for simplicity, facilitating its straightforward integration into existing data pipelines as it merely requires the calculation of the white reference's standard deviation. While more complex augmentation methods exist that might offer improved results, they typically necessitate specialised knowledge, such as the spectra of different light sources or tissue-specific colour modifications based on semantic scene understanding. Our approach, despite its straightforward nature, has proven to be remarkably effective, a characteristic we consider to be of significant importance.

3.6.1 Clinical Usage of Hyperspectral Segmentation

Alongside our proposed augmentation method, we evaluated hyperspectral segmentation methods on the Hib dataset. With our proposed augmentation, we achieve a macro F1-score of 74.4%. Even spectral classification methods demonstrate strong performance. In practice, with an appropriate frame-rate HSI camera positioned above the surgical scene, surgeons could visualize an overlay of predicted cancerous tissue regions in real-time. The granularity of segmentation predictions depends critically on label specificity. If a dataset is annotated by tumor grades, visualization could extend beyond binary classification to include infiltration zones, where the model indicates tissue of uncertain malignant potential rather than definitive tumor boundaries.

However, a macro F1-score of 74.4% remains insufficient for clinical translation when evaluated against modern medical computer vision benchmarks. This performance limitation highlights fundamental challenges within HSI methodology that require addressing before practical deployment.

One major limitation of HSI lies in dataset heterogeneity and the resulting barriers to transfer learning. In conventional medical imaging modalities, combining datasets and employing mixed training protocols is standard practice, with single segmentation network architectures generalizing across diverse datasets. HSI presents unique challenges due to hardware variations and inconsistent labeling protocols. While one dataset may capture 25 spectral bands, another may include 128 bands

with entirely different wavelength resolution. Similarly, spectral coverage varies dramatically between datasets, with some limited to the visible spectrum while others extend into near-infrared or beyond.

Traditional DL architectures are designed with RGB inputs in mind, utilizing three distinct bands that, despite potential resolution differences, can be standardized through resizing to produce consistent segmentation outputs. HSI data lacks this standardization, creating incompatibilities between datasets that prevent effective model generalization. Current architectures cannot accommodate the variable dimensionality and spectral range differences inherent in HSI datasets.

Future research should focus on developing HSI-specific architectures capable of mixed dataset training across variable spectral dimensions and wavelength ranges. Additionally, significantly more data collection is required, paired with higher fidelity annotations that enable more clinically meaningful predictions. Ground truth labels must move beyond basic tissue classification to include pathologically relevant gradations that reflect the complexity of surgical decision-making.

Despite these limitations, when contrasted with the gold standard of frozen section analysis, HSI segmentation offers valuable potential as a secondary diagnostic tool. Rather than replacing existing methods, HSI could provide surgeons with an additional perspective to corroborate tissue type boundaries, potentially reducing reliance on time-intensive pathological consultation during critical surgical procedures.

3.7 Conclusion

This chapter introduced a novel augmentation technique named *Calibration-Jitter*. The qualitative evaluation demonstrated its utility in regularising overconfident model predictions and eliminating minor anomalies, thereby enhancing the quality of the final segmentation. These visual improvements are substantiated by quantitative results, which confirm the advantages of the *Calibration-Jitter* approach. Achieving a top macro F1-score of 74.4%, we have surpassed the previously reported best performance on the Hib benchmark of 70.2% [103], which was attained with a spectral/spatial method. In conclusion, the findings reported herein highlight the considerable potential of our proposed augmentation strategy.

Chapter 4

Interactive Hyperspectral Segmentation

This chapter explores a novel interactive segmentation framework that combines RGB foundation models with spectral analysis to address challenges in Hyperspectral Imaging (HSI). A single user click guides both the spatial segmentation and spectral comparison, making the method simple and flexible. The framework works across different datasets and wavelengths, supporting few-shot use and mixed dataset training across hardware ultimately improving HSI's clinical use.

The work presented in this chapter has been accepted for publication in the proceedings of the 2025 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) [161]. The code to reproduce this chapter can be found here.

4.1 Overview

As established in chapter 3, the capacity of HSI to provide enhanced intraoperative guidance stems from its rich spectral information, which allows for the superior differentiation of tissues that are visually indistinct to the human eye [37, 179]. This high-resolution spectral data produces unique 'fingerprints' for different tissue types, revealing subtle physiological and biochemical variations not apparent in standard RGB images [7, 116].

Despite this promise, the widespread clinical adoption of HSI is hindered by several major challenges. The diversity in HSI hardware, with varied spectral ranges and resolutions, leads to significant interoperability issues that obstruct data standardisation [7]. This technological fragmentation, coupled with HSI's limited clinical use to date, has resulted in a scarcity of large-scale, comprehensive datasets. Such data scarcity presents a major barrier to developing robust machine learning models, which rely on extensive, standardised training data [73]. Although recent work has indicated HSI's potential for intraoperative segmentation in neurosurgery [103, 159, 179] and on porcine tissue [175, 184], the development of generalised models that can accommodate hardware differences remains an unresolved problem.

To address these limitations, a novel interactive segmentation framework is proposed. While modern foundation models like Segment Anything (SAM) [97] have revolutionised interactive RGB segmentation, their direct application to HSI is impractical due to fundamental differences in data structure. The common approach of converting other imaging modalities to greyscale for these models is suboptimal, as it discards the unique information that each modality provides [118, 215]. Our method circumvents this by uniting large-scale RGB foundation models with classical spectral analysis. It utilises a single user click to dually guide an RGB-based model and provide a reference for spectral comparison, fusing the spatial context from RGB with the detailed spectral information from HSI to create a flexible and adaptive segmentation tool.

This chapter details the development and validation of this framework. First, the architecture of our proposed interactive method is presented, explaining the mechanism for fusing foundation model outputs with spectral analysis. Following this, the experimental methodology and datasets used to evaluate the system are described. Finally, the results are presented and discussed, demonstrating the effectiveness of our approach in comparison to existing methods and considering its potential for future clinical applications.

4.2 Introduction

As explained in section 3.2, unlike RGB cameras that capture only three broad channels, HSI systems acquire many narrow spectral bands, enabling the detection of subtle tissue differences invisible to conventional imaging [7, 116].

The adoption of this promising technology into routine clinical practice confronts considerable hurdles. The diverse HSI hardware landscape poses a primary challenge, with devices exhibiting significant variations in spectral ranges, resolutions, and sensor types. This heterogeneity leads to major interoperability problems, forming a substantial barrier to data standardisation among different clinical sites and research studies [7]. Such technical fragmentation, together with the relatively limited clinical application of HSI to date, has caused a deficit of large-scale, comprehensive, and annotated datasets. This scarcity of data represents a formidable obstacle for the Machine Learning (ML) field, where robust and generalisable models typically depend on vast and standardised training corpora [73]. Although recent advancements have demonstrated the value of HSI for intraoperative segmentation within fields such as neurosurgery [103, 159, 179] and on porcine organ models [175, 184], the fundamental issue of creating generalised models that can accommodate these inherent hardware differences is yet to be resolved.

The interactive approach is especially beneficial within the medical field, as expert clinical input can markedly enhance performance compared to fully automated systems, which can be fragile or inconsistent [202, 203, 224]. Furthermore, interactivity offers the essential ability to segment novel tissue classes—a critical feature in surgical settings where unforeseen pathological conditions may arise. Within RGB imaging, influential foundation models such as SAM and the subsequent Segment Anything-2 (SAM2) [97, 156] have brought about a revolution in interactive segmentation. Yet, applying these models directly to HSI is impractical owing to fundamental disparities in the data's structure. A common strategy in related work involves converting other imaging modalities, like Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), into greyscale formats before processing them with these models [118, 215]. Such an approach is fundamentally flawed because it discards the distinct diagnostic information inherent to the source modality. Given that current multimodal frameworks lack advanced spectral fusion methods, existing techniques are not easily adaptable for addressing these hyperspectral-specific problems.

To navigate the challenge of limited data availability, classical Spectral Comparison Functions (SCF)s provide a highly flexible alternative for performing interactive HSI image segmentation. Techniques like the Spectral Angle (SA) [19] and Pearson Correlation Coefficient (PCC) [131] are notable for being untrained methods. This characteristic confers a key advantage: the capacity to generalise to novel situations without the need for supplementary data, operating effectively across any spectral range or number of bands. By design, these approaches are interactive, usually functioning by contrasting a user-selected reference spectrum with all other spectra in the

image. However, their effectiveness is limited by two main factors: the 'shading problem', which causes semantic objects to have varied spectral signatures under non-uniform illumination, and the difficulty of setting dependable and consistent segmentation thresholds within and between images.

This work introduces a novel approach to interactive image segmentation, specifically engineered to tackle the inherent limitations of HSI data. This is achieved by synergistically merging the capabilities of large-scale RGB foundation models with traditional spectral analysis techniques. The proposed framework leverages the common interactive paradigm of both fields, allowing a single user click to serve a dual function: guiding the RGB-based model and concurrently establishing a reference for the SCF.

This chapter makes three primary contributions:

1. The introduction of an interactive HSI segmentation framework that utilises a dual-input approach. This system makes efficient use of a single user click in two complementary ways: first, as a prompt for an RGB foundation model, and second, to calculate SCF measurements from the HSI data, leading to improved segmentation outcomes.
2. A demonstration of the framework's efficacy for tumour classification in both few-shot and zero-shot learning scenarios. This validates its robust performance with minimal training examples and on test cases not previously encountered.
3. The presentation of the first ML framework for HSI designed to function irrespective of band count and wavelength variations. This novel capability allows the mixed training of datasets with disparate spectral characteristics and semantic classes into a single and unified methodology, thus tackling the crucial issue of hardware fragmentation.

4.3 Methodology

Our objective is the interactive segmentation of foreground and background elements within a hyperspectral image, denoted as $X \in \mathbb{R}^{H \times W \times C}$. This process is guided by click positions provided by the user. In this notation, H and W correspond to the spatial dimensions, while C signifies the number of spectral channels. Alongside the hyperspectral data, a corresponding pseudo-RGB image $X_{rgb} \in \mathbb{R}^{H \times W \times 3}$ is also available, in addition to a ground truth label map $Y \in [0, ..., N]^{H \times W}$, where N represents the total number of classes. Let $\mathcal{I} = \{I_{i,j}\}$ represent a set of user-provided click coordinates, with each $I_{i,j} = (i, j)$ corresponding to a specific pixel location. The output from each model is a foreground/background similarity map, $\hat{Y} \in [0, 1]^{H \times W}$. In this map, a value of 0 indicates no similarity to the selected pixel(s), whereas a value of 1 signifies a strong similarity.

4.3.1 Evaluation Protocol

The evaluation of each model's performance on foreground/background segmentation, driven by user clicks, follows the procedure established in the SAM2 framework [97]. For every class label c present within the ground truth mask Y, we define a corresponding binary foreground mask as $F = [Y = c]$. To initiate the simulation, a single click is programmatically placed at the centre of the largest connected component within F, simulating the user's first interaction.

At every interaction step, denoted by $k \in \{1, \ldots, K\}$, the model processes the current set of user clicks to produce a probability map $\hat{Y}^{(k)} \in [0, 1]^{H \times W}$. This map is then thresholded to create a binary prediction. The quality of this segmentation is subsequently assessed.

The simulation of user corrections is based on the assumption that a user would click on the area of the object where the model's prediction is least accurate. More specifically, a similarity map is computed between the model's prediction and the ground truth. From this, the pixel within the foreground mask F that possesses the lowest similarity value, and has not previously been selected, is chosen. This newly selected pixel is then appended to the set of clicks for the subsequent interaction. A simplified representation of the evaluation process is provided in listing 4.1.

Listing 4.1: Evaluation Protocol Psuedo Code.

```
for each (image, label, rgb) in test_dataset:
    for each class_id in unique classes of label:
        F = (label == class_id)
        clicks = [centre of largest connected component in F]

        for step in range(max_clicks):
            pred = model(image, rgb, clicks)

            # Mask out unlabelled pixels for fair evaluation
            valid_region = (label != ignore_index)
            dice = compute_dice(pred, F, mask=valid_region)
            log_dice[step].append(dice)

            if step < max_clicks - 1:
                # Compute similarity between prediction and F
                sim = compute_similarity(pred, F)

                # Find least similar pixel in F not already clicked
                next_click = pixel_in_F_with_lowest_similarity(sim, F, clicks)
                clicks.append(next_click)
```

Here, the similarity map (`sim`) serves as a proxy for the model's confidence in its predictions. The subsequent click is strategically placed where the model exhibits the greatest error, thereby guiding it towards rectifying under-segmented regions in progressive steps.

4.3.2 Spectral Comparison Functions

From a collection of user clicks, $\mathcal{I}$, within a region of interest, we extract the corresponding spectra, defined as $S_{i,j} = X(I_{i,j})$. For every pixel in the image, we compute a SCF by comparing its spectrum to all selected spectra, and retain the maximum similarity score. This process results in a similarity map, denoted as $\hat{Y}_{SCF} = SCF(X, \mathcal{I})$.

Two common metrics for this comparison are the SA [19] and the PCC [131]. SA evaluates similarity by measuring the angle between two spectral vectors within the spectral space. In contrast, PCC quantifies the linear relationship between a reference spectrum and a candidate spectrum, with a specific focus on modelling negative correlations to differentiate between positive and negative relationships.

Although both SA and PCC are effective for measuring similarity between spectral samples, they do not inherently establish decision boundaries as they are not learning-based models. To overcome this limitation, histogram equalisation is applied to the similarity maps. This technique enhances the information content by increasing contrast and improving the separability of different regions [72]. This enhanced method is denoted as $SCF_{Equalized}$.

The procedure for histogram equalisation on a given similarity map $\hat{Y}(p) \in [a, b]$ begins with computing the histogram of values across N bins:

$$H(k) = \sum_{p} \delta(k - B(p)), \tag{4.1}$$

where $B(p) = \left\lfloor \frac{N(\hat{Y}(p)-a)}{b-a} \right\rfloor$ assigns each pixel's value to a specific histogram bin, and $\delta(\cdot)$ represents the Kronecker delta function. Subsequently, the cumulative distribution function (CDF) is calculated:

$$C(k) = \frac{1}{P} \sum_{m=0}^{k} H(m), \tag{4.2}$$

where P signifies the total pixel count in the image. The final equalised similarity map is then derived using the CDF:

$$\hat{Y}_{eq}(p) = C(B(p))(b - a) + a. \tag{4.3}$$

We denote the histogram-equalised versions of SA and PCC as $\hat{Y}_{Eq.SA}$ and $\hat{Y}_{Eq.PCC}$, respectively.

4.3.3 RGB Interactive Segmentation

In order to apply powerful, pre-existing RGB foundational models to HSI, we begin by generating pseudo-RGB images. This is achieved through a process of spectral band selection and combination from the original HSI data. For our RGB segmentation backbone, we select SAM2 [156], chosen for its state-of-the-art performance in interactive segmentation and its impressive generalisation capabilities across varied imaging domains, such as medical imaging [134]. From the pseudo-RGB image X_{rgb} and user clicks $\mathcal{I}$, SAM2 produces a confidence map, $\hat{Y}_{SAM2} = \text{SAM2}(X_{rgb}, \mathcal{I})$, which indicates the probability of each pixel belonging to the foreground. As SAM2 operates exclusively on RGB information, its predictions provide a spatial perspective that may diverge from segmentations derived purely from spectral similarity. Throughout this work, we use $\text{SAM2}_{Base}(\cdot)$ to refer to the SAM2-Large model with its weights frozen, whereas $\text{SAM2}_{Tuned}(\cdot)$ signifies the model after it has been fine-tuned.

4.3.4 RGB and Spectral Fusion

Models based on spectral data and those based on RGB data provide complementary information for image segmentation. The RGB-based processing of SAM2 captures spatial and contextual features that may be absent in a pure spectral analysis. Consequently, combining its output with our spectral similarity segmentation offers a path towards enhanced performance. With the goal of improving segmentation quality, we investigate two primary strategies for fusing the spectral and spatial similarity maps.

The first, a straightforward intersection method, involves the direct multiplication of the two similarity maps:

$$\hat{Y}_{\text{SAM2}_{Intersec.}} = \hat{Y}_{SAM2} \cdot \hat{Y}_{SCF}$$

This element-wise product results in high values only within regions where both modalities concur, thereby functioning as a logical AND operation that demands a consensus between spectral and spatial evidence for a pixel's classification.

As a more sophisticated alternative, we employ a UNet architecture [166]. This network receives the similarity maps $\hat{Y}_{SAM2}$ and $\hat{Y}_{SCF}$ as direct inputs and is trained to learn an optimal fusion strategy: $\hat{Y}_{\text{SAM2}_{UNet}} = \text{SAM2}_{UNet}(\hat{Y}_{SAM2}, \hat{Y}_{SCF})$, where $\text{SAM2}_{UNet}(\cdot)$ corresponds to the trained fusion UNet model. In contrast to the deterministic intersection method, this learnable fusion approach is capable of uncovering more nuanced, complementary relationships between the modalities. This could potentially improve segmentation outcomes in scenarios where either the spatial or spectral information alone is insufficient for an accurate classification.

4.3.5 SAMSA

To achieve a further enhancement in segmentation performance, we propose Segment Anything with Spectral Angles (SAMSA), a novel architecture designed to fuse spectral similarity information with the high-resolution spatial features from SAM2. Critically, this model diverges from the previously mentioned fusion methods, which combine the final outputs post-segmentation. Instead, SAMSA integrates the spectral information directly within the upscaling pathway of the SAM2 mask decoder. A conceptual overview of this architecture is provided in fig. 4.1.

When provided with a pseudo-RGB image X_{RGB} and user prompts $\mathcal{I}$, SAMSA executes the standard SAM2 processing pipeline but incorporates an additional step to integrate spectral data. Specifically, the spectral similarity map $\hat{Y}_{SCF}$ is combined with the high-resolution feature maps, S_0, which are extracted from the SAM2 image encoder. This fusion enriches the segmentation decisions with detailed spectral properties. This technique enables our model to exploit spectral characteristics that are imperceptible in the pseudo-RGB representation, all while preserving the high spatial precision inherent to SAM2.

Our training strategy involves freezing the weights of the SAM2 prompt and image encoders, and exclusively fine-tuning the lightweight mask decoder. This approach facilitates the generalisation of SAMSA to new medical datasets using only minimal training data, as the model focuses on learning the optimal method for combining spatial and spectral information. The computational cost of this integration is minimal; the extra steps for spectral comparison and feature map multiplication add an estimated 1.5 to 2 percent to the total FLOPs compared to the original SAM2 model. SAMSA therefore retains its capacity for real-time performance.

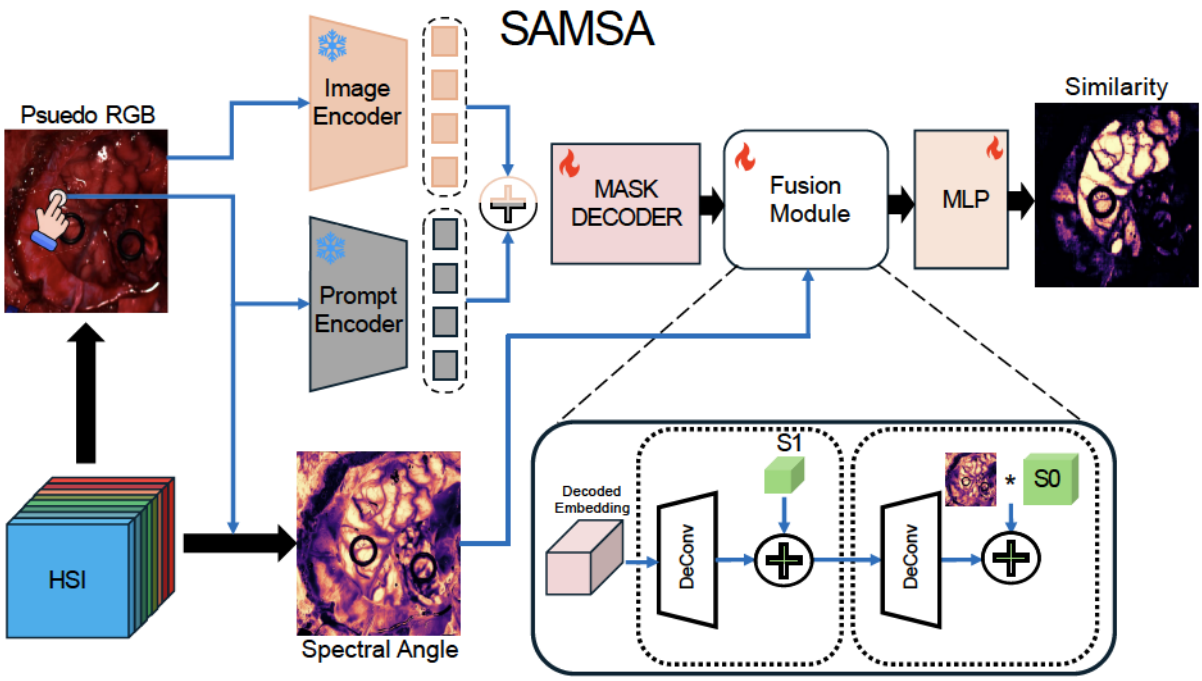


Figure 4.1: An outline of SAMSA, where a single click within the pseudo-RGB image is utilised to guide both the RGB and the spectral branches of the model [161]. Permission for the reuse of this figure is detailed in section A.0.4.

4.3.6 Training Protocol

The training process commences with a given image $X \in \mathbb{R}^{H \times W \times C}$, its corresponding pseudo-RGB projection X_{rgb}, and the ground truth label map Y. Our approach handles each semantic class $c \in \mathcal{C}$ individually, where $\mathcal{C}$ represents the set of all valid, non-ignored classes found within Y.

For every class c, a binary foreground mask is established as $F = [Y = c]$. An initial user interaction is simulated by selecting a starting click (i_0, j_0) located within this foreground area. The sequence of interactions is then built up iteratively:

$$\mathcal{I}^{(k)} = \{(i_0, j_0), (i_1, j_1), \ldots, (i_{k-1}, j_{k-1})\},$$

where subsequent clicks are algorithmically determined by comparing the model's current prediction against the ground truth foreground.

At each interactive step $k \in \{1, \ldots, K\}$, the model generates a foreground confidence map:

$$\hat{Y}^{(k)} = \text{SAMSA}(X_{rgb}, \hat{Y}^{\text{eq}}_{\text{SCF}}, \mathcal{I}^{(k)}).$$

From this prediction, we calculate a combined loss for that click. This per-click loss, denoted $\mathcal{L}_c^{(k)}$ for class c, is a weighted sum of the binary cross-entropy ($\mathcal{L}_{\text{BCE}}$) and soft Dice Similarity Coefficient or Sørensen–Dice coefficient (DICE) ($\mathcal{L}_{\text{DICE}}$) losses:

$$\mathcal{L}_c^{(k)} = \lambda \cdot \mathcal{L}_{\text{DICE}}^{(k)} + (1 - \lambda) \cdot \mathcal{L}_{\text{BCE}}^{(k)},$$

where the weighting factor λ is set to 0.5 throughout our experiments.

The losses are accumulated across all interaction steps and all valid classes within the image. A single backward pass and parameter update is performed for the entire image, guided by a final loss value. This value is computed by summing the per-click losses for all classes and then averaging by the number of clicks, K:

$$\mathcal{L}_{\text{image}} = \frac{1}{K} \sum_{c \in \mathcal{C}} \sum_{k=1}^{K} \mathcal{L}_c^{(k)}.$$

A pseudocode summary of this training loop is provided below in listing 4.2.

Listing 4.2: Training Protocol SAMSA Psuedo Code.

```
for (x, y, rgb) in train_loader:
    optimizer.zero_grad()
    total_loss = 0
    for cls in unique(y):
        if cls in ignore_classes:
            continue

        F = (y == cls)
        clicks = [get_initial_click(F)]

        for k in range(max_clicks):
            pred = model(x, rgb, clicks)
            dice_loss, bce_loss = compute_losses(pred, y, cls)
            loss = 0.5 * dice_loss + 0.5 * bce_loss
            total_loss += loss

            if k < max_clicks - 1:
                sim = model.convert_to_sim(pred)
                next_click = compute_next_click(sim, F, clicks)
                clicks.append(next_click)

    avg_loss = total_loss / max_clicks
    avg_loss.backward()
    optimizer.step()
    scheduler.step()
```

This methodology ensures the model is trained in a setting that mirrors realistic, progressively refined user guidance, thereby enhancing its robustness and responsiveness to interactive click-based inputs.

4.4 Results

For the trainable models—SAM2$_{Tuned}$, SAM2SA$_{Intersec.}$, SAM2SA$_{UNet}$, and $SAMSA$—we follow the same optimization procedure as SAM2 [97]. Much like the evaluation procedure, for every image we create foreground / background masks. Each model is trained using a half DICE and half cross-entropy loss function. As in the evaluation protocol, we mask out regions which are unlabelled. To learn correlations of multiple clicks we follow the same procedure as in evaluation to select the next click pixel location. For our experiments, we employ different optimization strategies depending on the model variant. The learning rate was empirically tuned to identify a stable convergence regime for the baseline model and then held fixed for subsequent experiments. This ensured that performance differences reflected model changes rather than variations in optimisation settings, while avoiding unnecessary grid search. We take latest model weights for the SAM2 architecture (2.1), from there we freeze all parameters unless otherwise stated. For SAM2$_{UNet}$, we use the AdamW optimizer with a weight decay of $1e^{-2}$ and an initial learning rate of $1e^{-3}$, following a Cosine Annealing with Warm Restarts schedule with a minimum learning rate of $1e^{-6}$. For both SAM2$_{Tuned}$ and SAMSA, we utilize the Adam optimizer with a learning rate of 1×10^{-4} for the decoder, implementing a OneCycleLR schedule with a maximum learning rate of 1×10^{-4}, a warm-up percentage of 30%, a cosine annealing strategy, and a division factor of 25.0.

4.4.1 Datasets

We use the Human Imaging Benchmark (Hib) datset, described earlier in section 3.5.1. For our experiments, we adhere to the patient-wise fold splits originally defined in the benchmark study by [103]. Although the spatial dimensions of the images vary, each raw hyperspectral cube initially contains 826 spectral bands, spanning a wavelength range from 400.48 nm to 1000.73 nm with a sampling interval of 0.73 nm. The data is annotated with four primary labelled classes: Background (BG), Tumour, Healthy, and Vasculature, alongside an additional Unlabelled category.

To prepare the data for analysis, a series of pre-processing steps are executed in accordance with the methodology of the original study. This pipeline includes: 1) white-dark calibration; 2) spectral smoothing using a kernel of size 5; and 3) the removal of the initial 56 and final 126 spectral bands. The remaining bands are then resampled along the spectral axis based on the optimal sampling interval identified by [127], followed by a min-max normalisation across the spectral dimension. Upon completion of this pre-processing, the final dataset is composed of 128 spectral bands, covering a reduced wavelength range of 440.45 nm to 909.90 nm.

The second dataset utilised is the HeiPorSPECTRAL (Heipor) dataset, which contains HSI data gathered from porcine subjects at the Heidelberg University Hospital. This dataset is comprised of 20 different organs. Data were acquired using a Tivita®Tissue HSI camera system capturing 100 spectral bands from 500–1000 nm. Each resulting hypercube has a resolution of 640×480

pixels spatially and 100 spectral bands. The associated RGB images for this dataset are generated directly from the HSI data by combining the spectral channels that correspond to the red, green, and blue wavelengths [175].

4.4.2 Metrics

To evaluate the segmentation performance of each model, we utilise two primary metrics based on the DICE (D):

- D@0.5: This represents the standard DICE calculated using a fixed decision boundary of 0.5.
- D@Max: This is the maximum achievable DICE, determined by evaluating all possible thresholds. It indicates the optimal potential performance of a model, independent of a predefined decision boundary.

For our analysis, the results are presented as both macro-averaged and per-class scores, for which we compute the partial DICE coefficient.

Furthermore, we assess the models' interactive performance in a multi-click scenario. This evaluation, extending up to five clicks, simulates a user progressively refining the segmentation by placing additional clicks within the target foreground region. Finally, for any models that require training, we conduct an N-shot evaluation. By training these models with a varying number of examples (specifically 1, 3, 5, 10, and 20), this analysis allows us to investigate the relationship between the quantity of training data and the resulting segmentation quality.

4.4.3 Spectral Comparison Threshold Study

Table 4.1: Macro Results: Performance of models with varying input modalities (Mod.) and user clicks. Bold indicates peak performance per metric and dataset [161]. Permission for the reuse of this table is detailed in section A.0.4.

Mod.	Model	Heipor			Hib		
		1 click		5 clicks	1 click		5 clicks
		D@0.5	D@Max	D@0.5	D@0.5	D@Max	D@0.5
HSI	PCC	0.122	0.472	0.117	$0.373^{\pm 0.019}$	$0.885^{\pm 0.034}$	$0.375^{\pm 0.018}$
	SA	0.117	0.489	0.117	$0.374^{\pm 0.019}$	$0.889^{\pm 0.035}$	$0.374^{\pm 0.019}$
	$SA_{Equalized}$	0.205	0.487	0.137	$0.568^{\pm 0.038}$	$0.885^{\pm 0.034}$	$0.482^{\pm 0.033}$
RGB	$SAM2_{Base}$	0.600	0.773	0.643	$0.523^{\pm 0.056}$	$0.727^{\pm 0.043}$	$0.591^{\pm 0.069}$
	$SAM2_{Tuned}$	0.806	**0.864**	0.886	$0.771^{\pm 0.059}$	$0.905^{\pm 0.036}$	$0.912^{\pm 0.025}$
Fusion	$SAM2SA_{Intersec.}$	0.634	0.755	0.647	$0.605^{\pm 0.048}$	$0.832^{\pm 0.033}$	$0.674^{\pm 0.083}$
	$SAM2SA_{UNet}$	0.692	0.798	0.771	$0.650^{\pm 0.115}$	$0.778^{\pm 0.123}$	$0.673^{\pm 0.096}$
	SAMSA (ours)	**0.811**	0.863	**0.892**	$\mathbf{0.810}^{\pm 0.050}$	$\mathbf{0.929}^{\pm 0.028}$	$\mathbf{0.934}^{\pm 0.031}$

Our initial evaluation of spectral similarity functions revealed that SA yielded superior performance compared to PCC, achieving improvements in the D@Max score of $+0.017$ on the Heipor dataset

and $+0.004$ on the Hib dataset (see table 4.1). To further refine this, we applied histogram equalisation to the SA output ($SA_{Equalized}$), which enhanced the contrast around the 0.5 decision boundary, thereby improving its compatibility with RGB-based models. This equalised version of SA was subsequently adopted as the standard spectral analysis method for all further experiments.

The effectiveness of this spectral comparison approach is visually evident in fig. 4.2. The resulting similarity map clearly delineates the target anatomy from the background, which suggests that an effective segmentation threshold exists. In the absence of a learned decision boundary, the optimal value for such a threshold remains unknown. It is this exact potential that the D@Max metric is designed to quantify, as it evaluates the best possible segmentation by identifying the ideal threshold across the entire range of similarity scores.

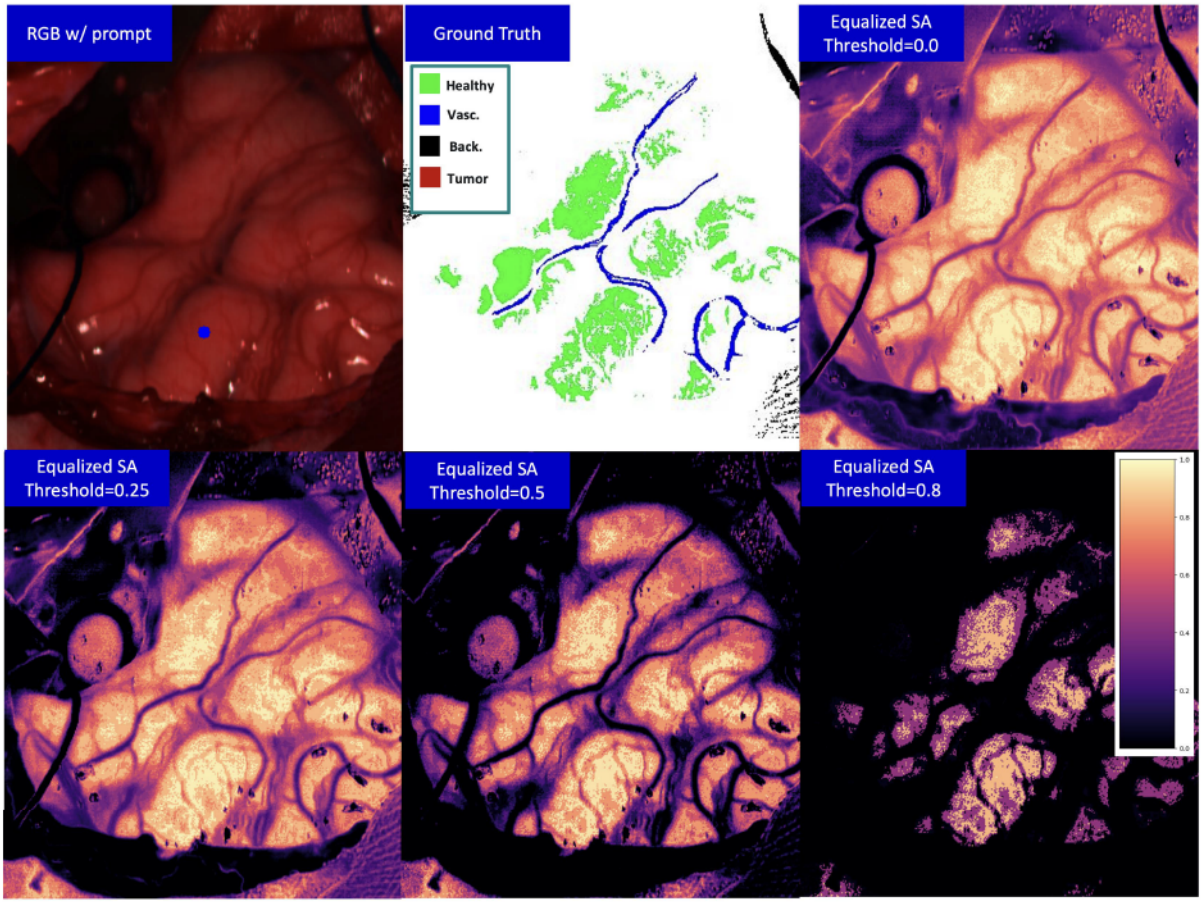


Figure 4.2: An example of the similarity map produced by the equalised Spectral Angle ($SA_{Equalized}$) function. The target anatomy is clearly distinguished from the background, highlighting the discriminative power of spectral information.

4.4.4 Individual Dataset Study

An analysis of the RGB-only model performance (table 4.1) shows that the base SAM2_{Base} model exhibits reasonable out-of-the-box generalisation to medical imagery, scoring a Macro D@0.5 of 0.600 on the Heipor dataset. However, its limitations become apparent on the Hib dataset, where

Table 4.2: Per-class D@0.5 results for the Hib dataset using a single click. The lower section shows the zero-shot performance on the Tumour class when it is excluded from training [161]. Permission for the reuse of this table is detailed in section A.0.4.

Model	Macro	Background	Healthy	Vascular	Tumour
$SA_{Equalized}$	$0.568^{\pm 0.038}$	$0.613^{\pm 0.109}$	$0.815^{\pm 0.093}$	$0.506^{\pm 0.130}$	$0.339^{\pm 0.100}$
SAM2_{Base}	$0.523^{\pm 0.056}$	$0.552^{\pm 0.079}$	$0.586^{\pm 0.083}$	$0.335^{\pm 0.095}$	$0.619^{\pm 0.188}$
$\text{SAM2SA}_{Intersec.}$	$0.605^{\pm 0.052}$	$0.564^{\pm 0.079}$	$0.683^{\pm 0.083}$	$0.491^{\pm 0.123}$	$0.680^{\pm 0.164}$
SAM2SA_{UNet}	$0.651^{\pm 0.099}$	$0.673^{\pm 0.074}$	$0.722^{\pm 0.132}$	$0.642^{\pm 0.075}$	$0.565^{\pm 0.312}$
SAM2_{Tuned}	$0.771^{\pm 0.059}$	$0.869^{\pm 0.045}$	$0.778^{\pm 0.081}$	$0.757^{\pm 0.106}$	$0.678^{\pm 0.098}$
SAMSA(ours)	$\mathbf{0.810}^{\pm 0.050}$	$\mathbf{0.881}^{\pm 0.039}$	$\mathbf{0.834}^{\pm 0.066}$	$\mathbf{0.790}^{\pm 0.117}$	$\mathbf{0.738}^{\pm 0.041}$
Zero-shot case - excluded Tumour class from training					
SAM2_{Tuned}	$0.708^{\pm 0.055}$	$0.853^{\pm 0.063}$	$0.735^{\pm 0.080}$	$0.704^{\pm 0.077}$	$0.538^{\pm 0.125}$
SAMSA (ours)	$\mathbf{0.760}^{\pm 0.053}$	$\mathbf{0.881}^{\pm 0.048}$	$\mathbf{0.821}^{\pm 0.081}$	$\mathbf{0.763}^{\pm 0.095}$	$\mathbf{0.576}^{\pm 0.072}$

table 4.2 reveals a significant weakness in segmenting the Vascular class (0.335 D@0.5), indicating poor generalisation to domain-specific structures. We can see this visually in fig. 4.3 where healthy class prompts correspond to a segmentation map covering most of the entire cavity. Performance was substantially improved through fine-tuning, with SAM2_{Tuned} reaching scores of 0.757 on the Vascular class and 0.869 on the Background class.

When evaluating fusion strategies, we found that late fusion approaches, specifically $\text{SAM2SA}_{Intersec.}$ and SAM2SA_{UNet}—while improving upon the base model ultimately underperformed in comparison to the fine-tuned RGB-only SAM2_{Tuned} model. This finding suggests that for spectral information to be maximally effective, it must be integrated at an earlier stage of the segmentation pipeline, a core principle that motivated the architectural design of SAMSA.

SAMSA consistently outperformed the fine-tuned SAM2_{Tuned} baseline across all classes when measured by D@0.5. The improvements were particularly notable on the Hib dataset for the Healthy ($+0.056$) and Tumour ($+0.06$) classes. This translated to an overall Macro D@0.5 increase of $+0.039$ for Hib and $+0.005$ for Heipor. The more modest gains on the Heipor dataset can be attributed to its characteristics, such as annotations that are more aligned with RGB features and a prevalence of large, centrally-located objects (see fig. 4.4). These features are highly favourable for RGB-only models that rely on visual boundaries. This is evidenced by the strong zero-shot performance of SAM2_{Base} on Heipor (0.773), which is only -0.091 D@0.5 behind its fine-tuned counterpart. For this reason, our detailed per-class metric analysis focuses primarily on the Hib dataset, where the contribution of spectral information provides substantial segmentation benefits.

Qualitative results on the Hib dataset are presented in fig. 4.4. When a click is placed on vascular tissue (a), the RGB-only SAM2_{Tuned} model (d) struggles to segment the class effectively. Conversely, the spectral-only SA map (e) successfully identifies the vascular structures but also introduces considerable noise around the nearby tumour region. In contrast, SAMSA (f) leverages the strengths

of both modalities to produce a clean, well-localised probability map for the vascular tissue. On the Heipor dataset, a click on the small bowel (g) demonstrates SAMSA's ability to delineate the class boundaries with high precision, closely matching the ground truth (h). Based on these results, we propose two key applications for our method: 1) as a tool to facilitate more efficient and accurate annotation during the creation of new medical datasets, and 2) for deployment in an intraoperative setting, allowing surgeons to interactively localise and investigate tissue.

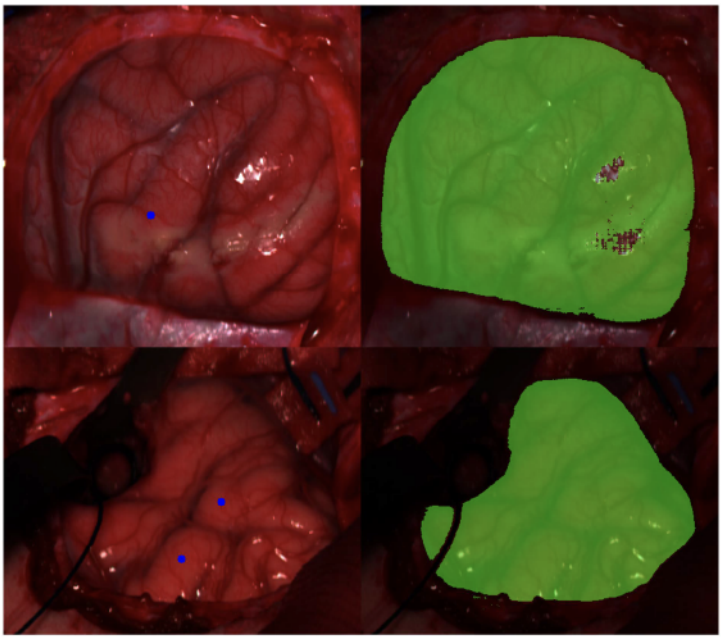


Figure 4.3: Qualitative results of SAM2$_{Base}$ on the Hib dataset. Left: pseudo-RGB with click prompt (blue circle). Right: output segmentation map.

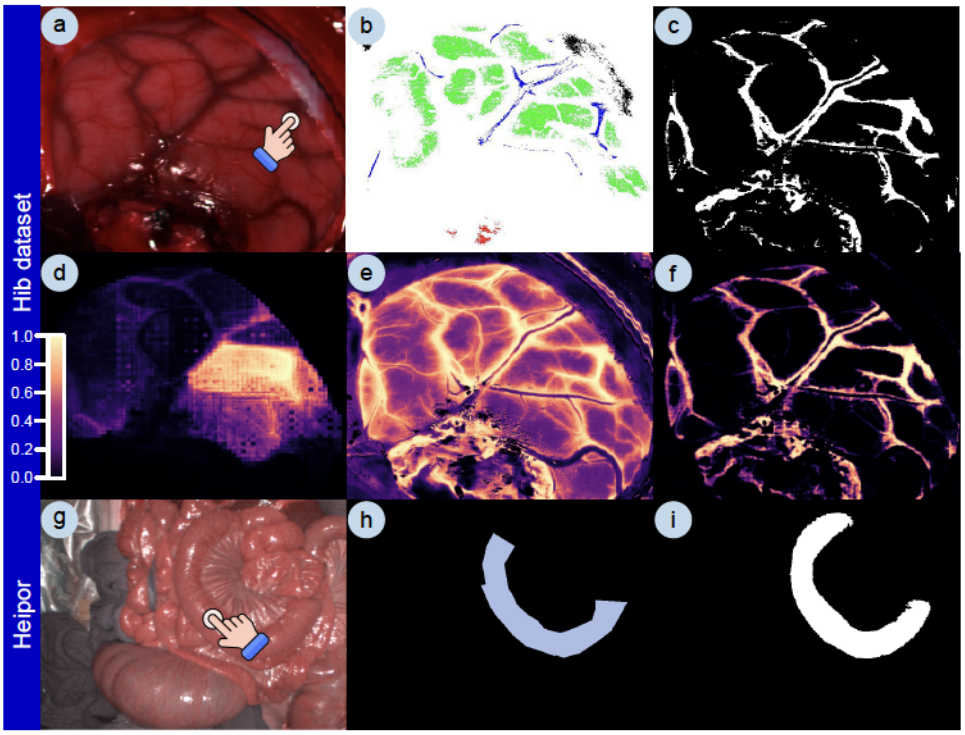


Figure 4.4: Qualitative segmentation results on Hib (top) and Heipor (bottom). (a,g) Input RGB with user click; (b,h) ground truth (red: Tumour, blue: Vasculature, green: Healthy, black: Background); (c,i) SAMSA segmentation; (d-f) probability maps from SAM2$Tuned$, $SAEqualized$, and SAMSA, showing improved localisation [161]. Permission for the reuse of this figure is detailed in section A.0.4.

4.4.5 Clicks and Shots

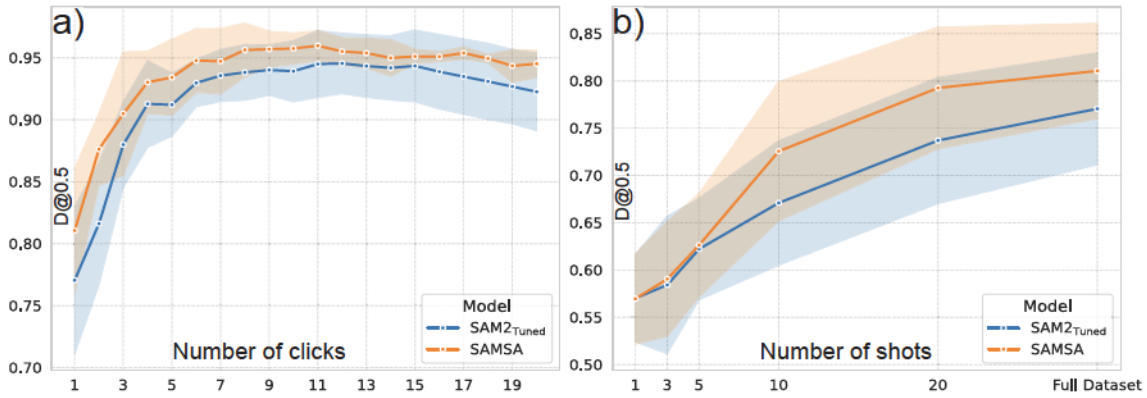


Figure 4.5: Performance analysis on the Hib dataset. (a) Macro D@0.5 score as a function of the number of user clicks, comparing SAMSA against the fine-tuned baseline. (b) Single-click Macro D@0.5 score in relation to the number of training examples (shots), illustrating the advantage conferred by spectral information in low-data regimes [161]. Permission for the reuse of this figure is detailed in section A.0.4.

As would be expected with interactive segmentation models, providing additional user clicks leads to improved performance for all fine-tuned methods. The benefit is particularly significant for SAMSA, where a 5-click input boosts the D@0.5 score by $+0.081$ on the Heipor dataset and $+0.124$ on the Hib dataset. As illustrated in fig. 4.5a, SAMSA consistently outperforms the SAM2_{Tuned} baseline across all click counts on the Hib dataset, ultimately reaching a Macro D@0.5 of 0.95 with five clicks.

In few-shot learning scenarios, both SAMSA and SAM2 perform well due to their foundation model origins, demonstrating robustness in limited-data contexts. Notably, with just 20 training examples, SAMSA achieves a Macro D@0.5 of 0.79 for single-click segmentation (fig. 4.5b). The integration of spectral information consistently enhances the training process; the performance advantage of SAMSA over SAM2_{Tuned} becomes evident with as few as five training examples. This highlights the substantial value of leveraging spectral information, especially within low-data regimes.

4.4.6 Generalisation Results

To assess the generalisation capabilities of the models, we conducted a leave-one-class-out experiment. For this, both the fine-tuned SAM2 and SAMSA models were trained on the Hib dataset with the Tumour class entirely excluded from the training set. The models were then evaluated across all classes, including the unseen Tumour class. This setup is designed to simulate a real-world scenario where a model is required to identify novel structures for which it has no prior supervised training.

The results of this zero-shot experiment are detailed in table 4.2. When the Tumour class was excluded from its training data, the performance of SAM2_{Tuned} on this class dropped by 0.14,

falling below the score achieved by the untrained SAM2$_{Base}$ model. Despite this, its overall Macro performance remained high ($+0.185$ over the base model). SAMSA experienced a similar performance decrease on the unseen Tumour class (-0.17). Critically, however, SAMSA still maintained the highest detection capability for this unseen class. Furthermore, it achieved a superior overall Macro result ($+0.052$ compared to SAM2$_{Tuned}$), which strongly suggests that incorporating spectral information provides a meaningful advantage for generalising to novel classes.

4.4.7 Mixed Training

Table 4.3: Model performance (Macro D@0.5) for cross-dataset and mixed-dataset training scenarios. The number of HSI channels and classes for each training configuration are shown [161]. Permission for the reuse of this table is detailed in section A.0.4.

Training→	**None**	**Heipor**		**Hib**		**Mixed**	
HSI Channels	-	100		128		228	
Num Classes	-	20		4		24	
Test↓	$SAM2_{Base}$ (0-shot)	SAM2 $_{Tuned}$	SAMSA	$SAM2$ $_{Tuned}$	SAMSA	$SAM2$ $_{Tuned}$	SAMSA
Heipor	0.600	0.806	0.811	0.445	0.433	0.807	0.810
Hib	0.523	0.454	0.497	0.771	0.810	0.695	0.765

A key advantage of our approach, unlike traditional hyperspectral models that are constrained to specific spectral configurations, is the ability to train on datasets with heterogeneous spectral properties. This is possible because the spectral similarity function collapses the high-dimensional spectral information into a single-channel similarity map. Consequently, our fusion model is independent of the number or resolution of the input spectral bands, enabling unprecedented flexibility in training. We evaluate this capability by performing cross-dataset and mixed-dataset training experiments.

The results, presented in table 4.3, show that cross-dataset generalisation—training on one dataset and testing on the other—performs poorly, with scores falling even below the zero-shot SAM2$_{Base}$ baseline. However, when the datasets are combined for mixed training, performance improves significantly. For the RGB-only SAM2$_{Tuned}$, the benefits of mixed training are inconsistent; its performance improves on the Heipor dataset but decreases on the Hib dataset. In contrast, SAMSA maintains a balanced and robust performance, outperforming the fine-tuned SAM2 on both test sets (Hib $+0.07$, Heipor $+0.003$). This confirms SAMSA's capacity to effectively learn from and generalise across HSI datasets with varying spectral characteristics and clinical domains.

Qualitative analysis provides further insight into these results. As seen in fig. 4.6 and anticipated from the quantitative scores, the performance improvement on the Heipor dataset is minimal. This is largely because the segmentation task is fairly simple, a fact underscored by the strong zero-

shot performance of the frozen SAM2$_{Base}$ model (as shown previously in table 4.2). In contrast, the more challenging Hib dataset, illustrated in fig. 4.7, reveals the distinct advantages of our approach. Here, we observe that: 1) the frozen SAM2$_{Base}$ model acts as a proficient edge detector, incorrectly identifying the visually distinct ring structure as the tumour; 2) the fine-tuned SAM2$_{Tuned}$ model learns a spurious correlation from the training data, reinforcing the incorrect association between the ring and the tumour; and 3) SAMSA, by integrating spectral information, is able to disregard this misleading visual cue and correctly localise the tumour in the tissue just beyond the ring.

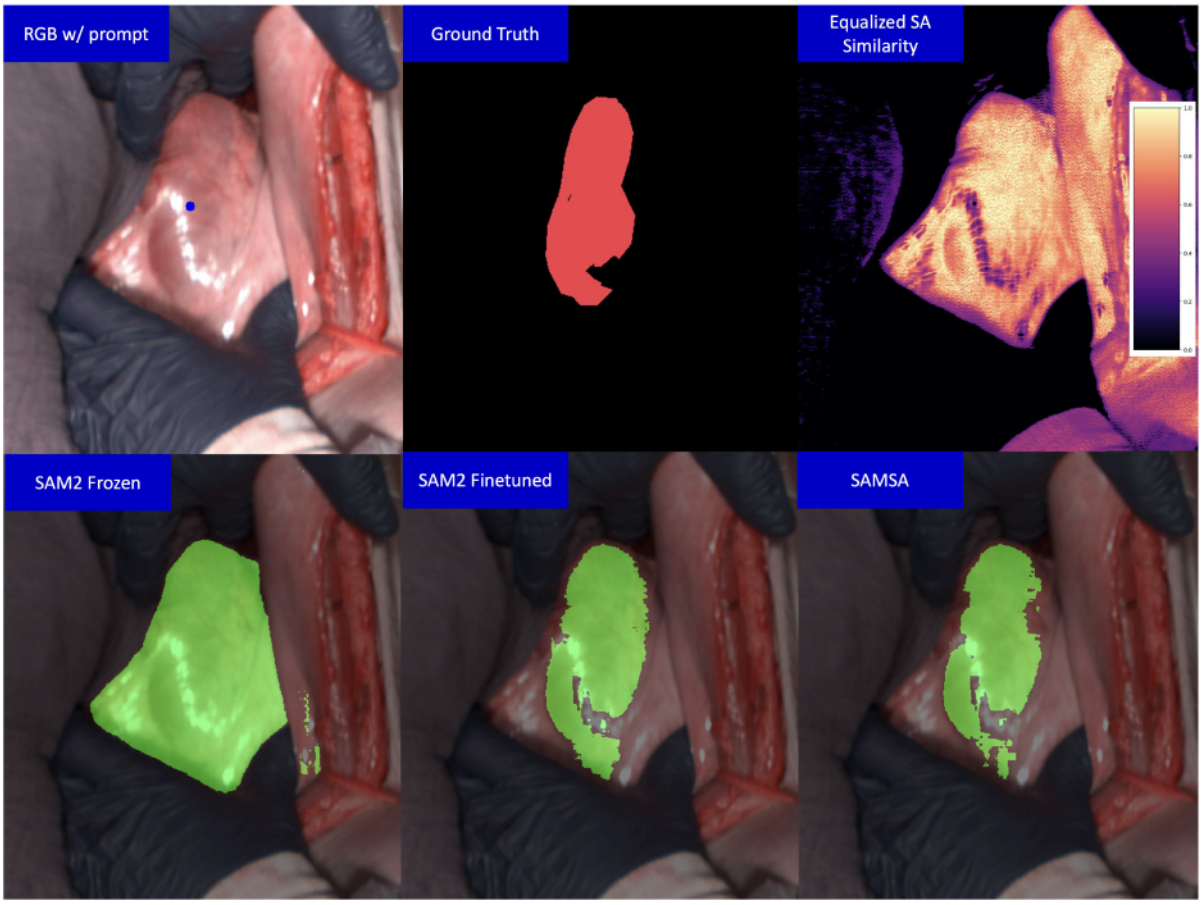


Figure 4.6: Qualitative results on the Heipor dataset. Performance gains are modest, as the segmentation task is relatively simple for RGB-based models.

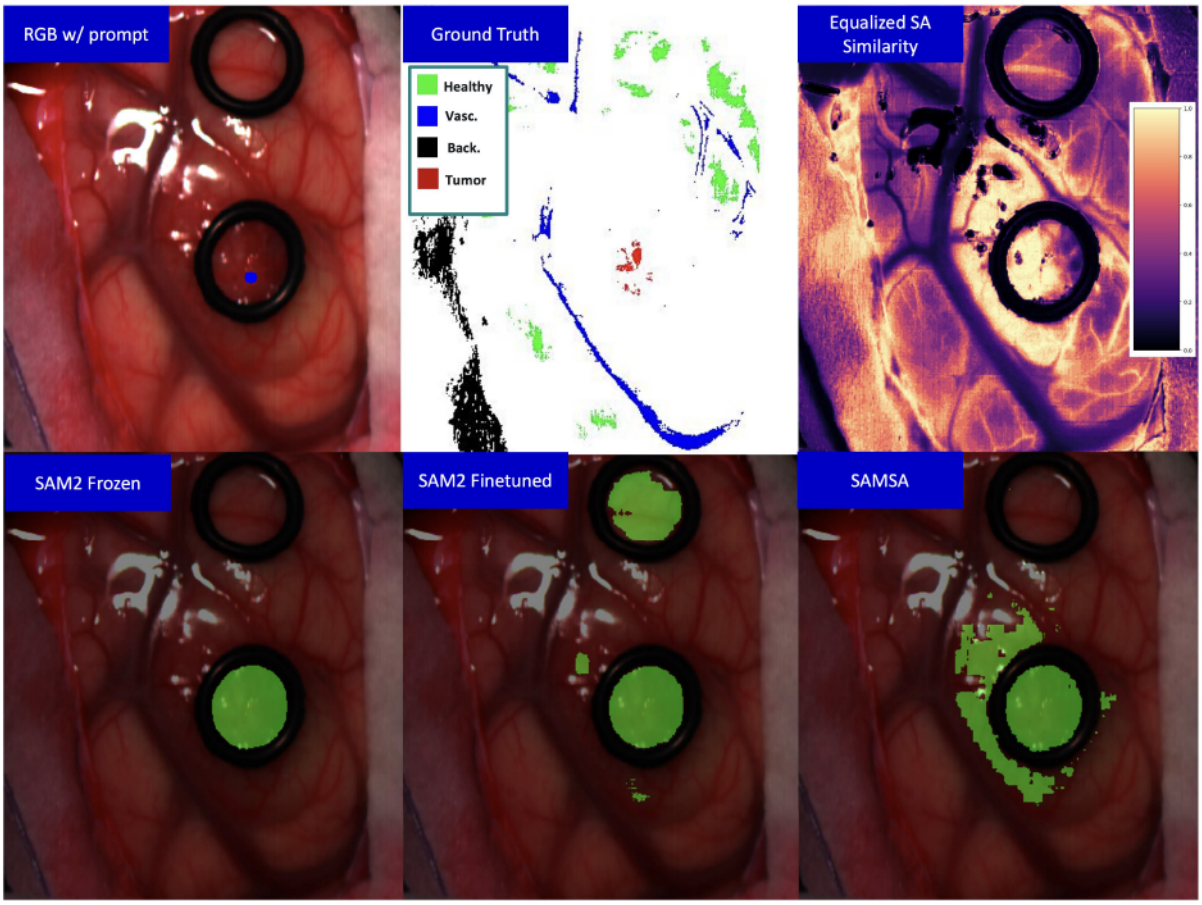


Figure 4.7: Qualitative results on the Hib dataset. SAMSA successfully uses spectral information to locate the tumour just outside the visually prominent ring, overcoming the limitations of the RGB-only models.

4.5 Clinical Usage of SAMSA

As expressed in section 3.6.1, one of the primary disadvantages of semantic HSI segmentation is hardware variation across imaging systems. SAMSA specifically targets this limitation by enabling mixed training across diverse spectral configurations, making cross-dataset compatibility feasible. However, our approach remains constrained by the limited number of available datasets and the requirement that dataset labels maintain consistency without overlap. While foreground-background segmentation of different medical objects is achievable, conflicting or overlapping label definitions across datasets present significant challenges for model training and validation.

Interactive segmentation differs fundamentally from semantic segmentation in its clinical implementation requirements. By requiring positional prompting, two critical intraoperative challenges emerge: first, the practical mechanism for user input during surgery, and second, the management of erroneous or suboptimal prompt placement by the surgeon.

We propose that the most stable and practical platform for HSI camera deployment is attachment to the surgical microscope. Previous work in neurosurgery [120] has demonstrated successful integration through side optical ports. We suggest that combining this stable imaging platform with intuitive user input mechanisms provides an optimal foundation for intraoperative interactive segmentation.

The transition from semantic to interactive segmentation necessarily removes the inherent classification capabilities of the system. While SAMSA can identify spectrally similar regions, it cannot provide diagnostic information such as tissue type or tumor grade classification. Semantic understanding would require integration with additional classification models operating in conjunction with the segmentation framework. This limitation highlights a fundamental trade-off in the interactive approach: while gaining flexibility and cross-dataset compatibility, we sacrifice the autonomous diagnostic capabilities that semantic segmentation could potentially provide. The clinical utility thus shifts from automated tissue classification to surgeon-guided region identification and delineation.

Despite these constraints, the interactive framework enables rapid, accurate segmentation of regions of interest across diverse HSI systems without requiring extensive dataset-specific training. However, our primary objective remains real-time intraoperative medical imaging applications. Future research directions must focus on expanding the available HSI datasets to improve model robustness and developing surgical workflow integration protocols. Additionally, investigation into hybrid approaches that combine the flexibility of interactive segmentation with semantic classification capabilities could potentially address current limitations while maintaining cross-platform compatibility.

4.6 Conclusion

In this work, we introduced SAMSA, a novel framework for interactive hyperspectral image (HSI) segmentation. Our approach successfully integrates the rich descriptive power of spectral data with the robust spatial reasoning of a pre-trained RGB foundation model, SAM2. We operated on the hypothesis that HSI can serve as a form of rich embedding to augment the visual features learned by SAM2. While spectral embeddings alone can be susceptible to issues like shading and require careful thresholding, and RGB models can learn spurious visual correlations, their combination proves highly effective. By fusing the detailed spectral information with the contextual and boundary-aware intelligence of SAM2, our model overcomes the individual limitations of each modality.

Our results demonstrate that SAMSA is a highly effective method, particularly in scenarios with limited training data. On the Hib dataset, our interactive approach achieved a DICE coefficient of 81.0% with just a single user click, increasing to 93.4% with five clicks. This significantly surpasses the performance of previous fully automatic methods. A key innovation of our framework is its ability to train on and generalise across heterogeneous HSI datasets, which possess different spectral and spatial resolutions. This is a crucial advantage over traditional HSI models that are typically constrained to a specific imaging configuration. Furthermore, SAMSA showed a strong capacity to generalise to unseen classes, highlighting its robustness for real-world applications.

The framework's ability to handle mixed datasets opens exciting opportunities, though the performance gains are most pronounced on complex segmentation tasks. On simpler datasets with large, high-contrast objects, the advantage of adding spectral information is less substantial. This suggests that the value of HSI is maximised when visual cues alone are ambiguous or misleading. Building on these promising results, future research will focus on two key directions. The current interactive framework could be extended towards an automatic segmentation model, removing the need for user input and thereby broadening its clinical applicability. Furthermore, to rigorously validate SAMSA's key strength, we plan a more extensive valuation against a wider range of public HSI datasets to confirm its generalisation capabilities across diverse medical domains and imaging systems. Ultimately, SAMSA represents a significant step towards creating more flexible and robust tools for HSI medical image analysis.

Chapter 5

Interactive Hyperspectral Segmentation Extended

This chapter builds upon a prior framework, Segment Anything with Spectral Angles (SAMSA), which addressed hardware fragmentation in Hyperspectral Imaging (HSI) but was limited by a memory-intensive design and late integration of spectral data. We introduce Segment Anything with Spectral Angles 2.0 (SAMSA2.0), a redesigned architecture that resolves these issues through a novel technique called spectral angle prompting. *This method embeds a pixel-wise spectral similarity map, derived from a user's click, directly into the model's input to provide rich spectral context from the earliest processing stage. The resulting framework is not only more accurate and memory-efficient but also incorporates an adapted automatic segmentation mode to generate medically relevant proposals. Through validation on a diverse range of HSI datasets, we demonstrate that SAMSA2.0 offers superior generalisation, accuracy, and clinical utility for HSI interactive medical image segmentation.*

5.1 Introduction

SAMSA [161] was introduced as an interactive framework that combined the spatial reasoning of pretrained RGB segmentation models with the spectral information from HSI. By aligning these complementary feature spaces, SAMSA reduced the need for large annotated datasets and mitigated the effects of scanner variability. Crucially, it addressed hardware fragmentation by demonstrating consistent performance across datasets without requiring retraining.

Despite its success, the original SAMSA framework had clear limitations. The training protocol used a lot of memory, making it difficult train larger Segment Anything-2 (SAM2) models. It was also tested on two datasets, meaning its wider applicability was not fully proven. Architecturally, its main weakness was that it introduced spectral information too late in the network, which limited the influence of the rich spectral cues on the final segmentation. Furthermore, its reliance on manual user clicks, while precise, highlighted a need for a more automated workflow to speed up clinical processes.

This chapter builds on that earlier work by introducing SAMSA2.0, a refined framework designed to fix these weaknesses. The core innovation is a new method we call *spectral angle prompting*. This technique calculates a pixel-wise spectral similarity map based on the user's click and embeds this information directly into the model's input. This 'prompt' acts as a soft constraint, guiding the model with both spatial and spectral context from the very start. This early integration is part of a completely redesigned architecture that is not only more accurate but also more efficient, using significantly less memory in training. To complement this, we have also adapted the automatic segmentation feature from SAM2, changing its 'segment everything' mode to generate more relevant proposals for medical images. Similar scribble and click-based interactive approaches have recently been explored by Wang et al. [205], demonstrating the growing interest in user-guided hyperspectral medical image segmentation.

The main contributions of this chapter are:

1. The introduction of **SAMSA2.0**, a simpler and more powerful model design featuring *spectral angle prompting*, which delivers better segmentation accuracy while using less memory.
2. A thorough test of the new framework on an expanded collection of HSI datasets, providing stronger proof that it can generalise to different tissues, diseases, and imaging systems.
3. The adaptation of SAM-2's automatic segmentation for HSI medical images within the SAMSA2.0 framework, creating a workflow where the system automatically suggests areas that an expert can then check and correct if needed.

5.2 Methodology

As described in chapter 4 interactive segmentation models operate by incorporating both input imagery and spatial cues provided by a user. In the Segment Anything (SAM), such cues may include points, bounding boxes, or masks. Spectral Angle (SA) methods, in contrast, typically rely solely on point-based input to identify spectrally similar regions. A recent hybrid approach, SAMSA, introduced SA-based prompting into SAM by leveraging points, boxes, or masks to guide segmentation based on spectral similarity. In this work, we build upon that idea by integrating spectral angle information earlier in the pipeline. Rather than applying SA post-hoc to refine outputs or prompts, we use the pretrained prompt encoder within SAM to inject spectral similarity directly into the input representation. This early fusion of spectral context allows the model to leverage hyperspectral features more effectively throughout the segmentation process.

To re-iterate section 4.3, the objective is to delineate foreground regions from background using minimal user input, while exploiting the high-dimensional spectral information embedded in the data. Let $X \in \mathbb{R}^{H \times W \times C}$ denote a hyperspectral image with spatial dimensions $H \times W$ and C spectral bands. We assume availability of a pseudo-RGB projection $X_{rgb} \in \mathbb{R}^{H \times W \times 3}$, derived from X, which facilitates spatial reasoning by models pretrained on RGB data. Ground truth segmentation is given as $Y \in 0, \ldots, N^{H \times W}$, where N is the number of semantic classes. User input is represented as a set of spatial cues $\mathcal{I} = I_{i,j}$, where each $I_{i,j} = (i, j)$ denotes a user-provided click at pixel location (i, j). The model's output is a binary probability mask $\hat{Y} \in [0, 1]^{H \times W}$, with values approaching 1 indicating higher confidence of foreground membership relative to the user-provided cues. We follow the same evaluation protocol as is described in section 4.3.1.

5.2.1 Spectral Comparison Functions

We use the same Spectral Comparison Functions (SCF) definition as introduced in section 4.3.2. These models are fully deterministic, interpretable, and operate in a zero-shot setting, making them effective spectral-only baselines for interactive hyperspectral segmentation.

5.2.2 Spatial Models

To encompass a larger range of model sizes, we follow a similar naming scheme for the spatial model as described in section 4.3.3. We use SAM2, a family of models designed for prompt-based segmentation [156], as our spatial backbone. Given spatial prompts $\mathcal{I}$ and pseudo-RGB input X_{rgb}, the model produces a pixel-level confidence map:

$$\hat{Y}_{\text{SAM2}} = \text{SAM2}(X_{rgb}, \mathcal{I}),$$

where $\hat{Y}_{\text{SAM2}} \in [0, 1]^{H \times W}$.

We evaluate several backbone capacities — `Tiny`, `Small`, `Base+`, and `Large` — with two training modes for each:

- SAM2-*Variant*: Frozen model; zero-shot segmentation using pseudo-RGB inputs.
- SAM2-*Variant*-FT: Fine-tuned model; updated on hyperspectral pseudo-RGB data using our interactive protocol.

For example, `SAM2-Tiny` refers to the smallest frozen model, while `SAM2-Tiny-FT` is the same model fine-tuned on our dataset. These serve as RGB-only baselines with varying model capacity and adaptability.

For a comprehensive comparison against other State of the Art (SOTA) methodologies, our evaluation includes the MedSAM-Adapter [117]. The MedSAM-Adapter is specifically engineered to augment the segmentation capabilities of the standard SAM for application in medical imaging. This is accomplished through the implementation of a lightweight adaptation strategy, which distinguishes it from more computationally intensive fine-tuning approaches.

5.2.3 Spectral-Spatial Models

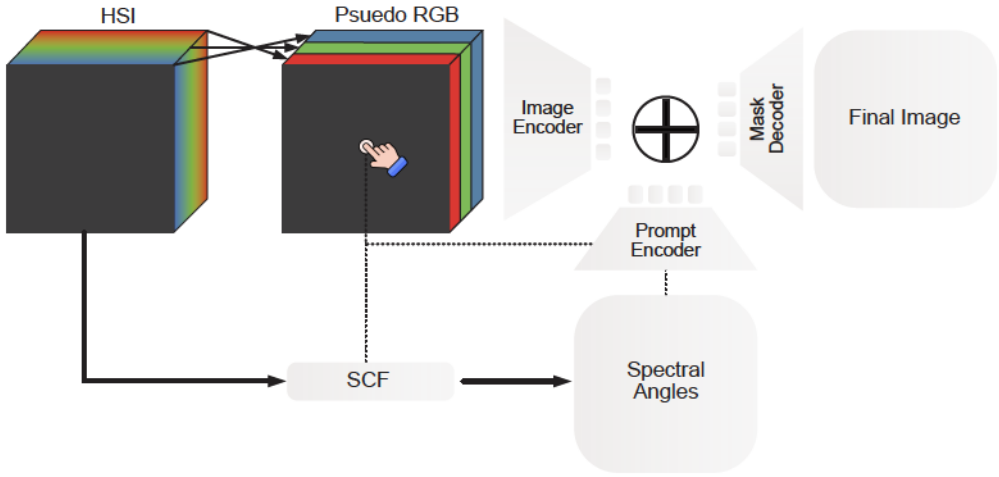


Figure 5.1: SAMSA2.0 Architecture

As with section 5.2.2, we follow a similar but not identical naming scheme to the previous chapter. Unlike the original SAMSA, which incorporates spectral similarity maps only in the later stages, SAMSA2.0 integrates spectral information directly at the input of the prompt encoder.

Given a hyperspectral image $X \in \mathbb{R}^{H \times W \times C}$ and a set of user-provided click locations $\mathcal{I} = \{(x_i, y_i)\}_{i=1}^{N}$, we first compute a spectral similarity map $\hat{Y}_{\text{SCF}}$ as described in Section 5.2.1, using the SA metric.

To enhance regional separability, we apply histogram equalization to the resulting map, producing $\hat{Y}_{\text{SCF}}^{\text{eq}}$. This equalized spectral map is resized to match the resolution of the pseudo-RGB image

X_{rgb} and concatenated channel-wise. The resulting composite input is passed to the SAM2 prompt encoder:

$$\hat{Y}_{\text{SAMSA2.0}} = \text{SAM2}(X_{rgb}, \hat{Y}^{\text{eq}}_{\text{SCF}}, \mathcal{I}),$$

enabling the model to incorporate both spectral and spatial cues from the outset for improved interactive segmentation performance.

As in the original SAMSA, we freeze the image encoder and prompt encoder to preserve generalization from pretraining. Only the mask decoder are updated during training, allowing the model to learn how to effectively fuse spectral guidance with RGB-based spatial context.

5.2.4 Training Protocol

For trainable models, we use a per-class optimization scheme with multi-click gradient accumulation. Note that we adjusted the SAMSA training gradient accumulation (described in section 4.3.6) by only accumulating per class rather than over all classes to reduce memory consumption. For completeness we will re-iterate the training protocol formulae.

Given a training image $X \in \mathbb{R}^{H \times W \times C}$, its pseudo-RGB projection X_{rgb}, and the ground truth label map Y, we process each semantic class $c \in \mathcal{C}$ separately, where $\mathcal{C}$ denotes the set of valid (non-ignored) classes present in Y.

For each class c, we define a binary mask $F = [Y = c]$ and select an initial user click (i_0, j_0) from within the foreground region. This click sequence is extended iteratively to simulate interaction:

$$\mathcal{I}^{(k)} = \{(i_0, j_0), (i_1, j_1), \ldots, (i_{k-1}, j_{k-1})\},$$

where each new click is computed based on the current prediction and the ground truth foreground.

At each step $k \in \{1, \ldots, K\}$, the model predicts a foreground confidence map:

$$\hat{Y}^{(k)} = \text{SAMSA2.0}(X_{rgb}, \hat{Y}^{\text{eq}}_{\text{SCF}}, \mathcal{I}^{(k)}),$$

and we compute the per-click losses using binary cross-entropy $\mathcal{L}_{\text{BCE}}$ and soft Dice Similarity Coefficient or Sørensen–Dice coefficient (DICE) loss $\mathcal{L}_{\text{DICE}}$. These are combined as:

$$\mathcal{L}^{(k)} = \lambda \cdot \mathcal{L}^{(k)}_{\text{DICE}} + (1 - \lambda) \cdot \mathcal{L}^{(k)}_{\text{BCE}},$$

with $\lambda = 0.5$ in our experiments.

Gradients are accumulated over all K interaction steps for each class, and a single backward pass is performed using the average loss:

$$\bar{\mathcal{L}}_c = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}^{(k)}.$$

A pseudocode representation of the training loop is given below in listing 5.1.

Listing 5.1: Training Protocol SAMSA2.0 Psuedo Code.

```
for (x, y, rgb) in train_loader:
    for cls in unique(y):
        if cls in ignore_classes:
            continue

        F = (y == cls)
        clicks = [get_initial_click(F)]

        optimizer.zero_grad()
        total_loss = 0

        for k in range(max_clicks):
            pred = model(x, rgb, clicks)
            dice_loss, bce_loss = compute_losses(pred, y, cls)
            loss = 0.5 * dice_loss + 0.5 * bce_loss
            total_loss += loss

            if k < max_clicks - 1:
                sim = model.convert_to_sim(pred)
                next_click = compute_next_click(sim, F, clicks)
                clicks.append(next_click)

        avg_loss = total_loss / max_clicks
        avg_loss.backward()
        optimizer.step()
        scheduler.step()
```

5.2.5 Automatic Segmentation

SAM and its extension SAM2 follow a very similar automatic segmentation pipeline, with SAM2 introducing several internal optimisations [97, 156].

The purpose of automatic segmentation is to remove the need for user prompts, instead processing the whole image to retrieve all constituent objects.

The automatic mask generation pipeline in SAM2 is based on a hierarchical, crop-based methodology that identifies objects across multiple scales. The image is processed not only at its original resolution but also at several layers of smaller, overlapping crops. This effectively allows the model to 'zoom in' on different regions, improving the detection of smaller objects.

A key principle in this pipeline is the prioritisation of masks generated from smaller, higher-resolution views. During the Non Maximum Supression (NMS) stage, which merges detections from all crops, a confidence score is calculated for each mask. This score is inversely proportional to the area of the crop it originated from, giving higher priority to masks from smaller crops. As a result, the final output favours the more detailed and precise segmentations of smaller objects. A pseudocode representation of the algorithm for SAM2 is given below in algorithm 1.

Algorithm 1: Automatic Segmentation Pipeline (SAM2)

Input: Image I

Output: Set of segmentation masks M

Generate multi-scale crops C of I;

foreach *crop* c *in* C **do**

    Generate candidate masks M_c using SAM2;

    foreach *mask* m *in* M_c **do**

        Assign confidence score $s(m)$ proportional to $1/\text{area}(c)$;

Collect all masks: $M_{\text{all}} \leftarrow \bigcup M_c$ over all crops;

Apply Non-Maximal Suppression (NMS) on M_{all};

    Resolve duplicate masks;

    Prefer higher-scoring masks (from smaller crops);

return M;

Unlike SAM2's training data, medical images pose additional challenges. What constitutes an 'object' is often ambiguous: should tumour grades I and II be treated differently? Ultimately, segmentation depends heavily on the training labels.

Our method simplifies some aspects of SAM2, but this comes at the cost of potential clutter. For example, vessels may span the entire image, with only a few prompts highlighting them, which conflicts with SAM2's stability score. SAM2 prioritises smaller objects, but this does not always align with neuroanatomical structures, such as large brain folds or vessels that cross multiple boundaries. Using smaller crops may lead to inconsistencies in these cases.

We propose two post-processing pipelines to address these challenges: a light and a heavy variation. Pseudocode representations are given below in algorithms 2 and 3.

Algorithm 2: Mask Post-Processing Pipeline Light

Input: Predicted masks
Output: Processed masks
Threshold masks;
    Convert probability masks to binary using a fixed threshold;
Keep largest connected component per mask (optional);
    Remove small disconnected regions;
Compute scores for Non-Maximal Suppression (NMS);
    Use mask confidence or intensity as score;
Apply NMS;
    Remove overlapping duplicate masks;
    Prefer higher-scoring masks;
return final masks;

The light variation is fast and suitable for cases where large structures dominate the image and small, detailed cleanup is less critical.

Algorithm 3: Mask Post-Processing Pipeline Heavy

Input: Predicted masks
Output: Processed masks
Threshold masks;
    Convert probability masks to binary;
Remove tiny noise;
    Remove small connected components;
Morphological closing (optional);
    Close small gaps and smooth mask edges;
Fill holes (optional);
    Fill internal holes within each mask;
Compute scores and apply NMS;
    Use mask confidence for scoring;
    Remove overlapping masks, keeping higher-scoring ones;
Keep largest connected component per mask;
    Ensure each mask is a single contiguous region;
Display masks (optional);
    Overlay masks on original image for visualization;
return final masks;

The heavy variation provides more rigorous cleanup, including noise removal, morphological smoothing, and hole filling, which is particularly useful for small, fragmented structures or highly detailed anatomical masks.

5.3 Results

For optimization of trainable models we followed the same procedure as described in section 4.4. At training time for the mixed training evaluation we employed a weighted random sampling strategy to address the inherent data imbalance across the constituent datasets. Using the sampler, data points were drawn with replacement based on pre-calculated sample weights. This oversampling technique ensures that each training batch contains a balanced representation from all four data sources (Human Imaging Benchmark (Hib), HeiPorSPECTRAL (Heipor), SLIMBRAIN (SB)-X, and SB-H), thereby preventing the model from becoming biased towards the larger datasets during optimisation.

5.3.1 Datasets

For validation we follow SAMSA's datasets Hib and Heipor, which are described in section 4.4.1. In addition, we included the SB database [126], consisting of two subsets: SB-X and SB-H. The SB-X subset was acquired with the Ximea snapshot camera, which captures 25 spectral bands in the 660–950 nm range. In contrast, the SB-H subset was recorded with the Headwall Hyperspec® VNIR E-Series linescan camera, providing 369 effective spectral bands spanning 400–1000 nm. Together, these subsets offer complementary trade-offs between spectral coverage and resolution, enabling us to evaluate models under different imaging conditions. Both SB datasets contain five classes; Healthy, Blood, Tumour, Meninges and Unlabelled.

5.3.2 Individual Dataset Study

Table 5.1 compares segmentation performance across four hyperspectral medical datasets, Heipor and Hib, SB-X, SB-H using spectral baselines, RGB-based models, and our proposed fusion methods. Spectral-only approaches (e.g., Pearson Correlation Coefficient (PCC), SA) show low DICE@0.5 scores due to the lack of spatial modeling but achieve high Max DICE scores, indicating the potential of spectral cues when thresholds are well-tuned. Fine-tuned RGB models (e.g., SAM2-Small-FT) perform better by leveraging spatial features, but still lack spectral awareness. MedSAM-Adapter, in general, has pretty poor performance. We assume that the model has too many trainable weights for the size of the dataset. SAMSA2.0 variants consistently outperform both spectral and RGB models. On average, SAMSA2.0-Small achieves a DICE@0.5 (1-click) score of 84.2%, compared to 81.1% for SAMSA and 80.4% for the best RGB-only model. This corresponds to an average improvement of +3.1% over SAMSA and +3.8% over RGB-only segmentation. In the 5-click setting, SAMSA2.0-Small achieves the highest DICE score of 96.4% on the Hib dataset, demonstrating strong performance with minimal user input. Note that SAMSA results differ from the original paper due to a modified training strategy. To reduce GPU memory, we accumulated gradients over user clicks only, rather than over both clicks and class labels.

Table 5.1: Performance comparison on Heipor, Hib, SB-X, and SB-H Datasets using SAM2 variants. D. denotes DICE; 1c/5c denote user clicks. Model suffixes: T (Tiny), S (Small), B+ (Base+), L (Large), FT (Fine-Tuned), Eq (Equalized), Ad (Adapter).

Model	Heipor			Hib			SB-X			SB-H		
	D.@.5 (1c)	D.@Mx (1c)	D.@.5 (5c)	D.@.5 (1c)	D.@Mx (1c)	D.@.5 (5c)	D.@.5 (1c)	D.@Mx (1c)	D.@.5 (5c)	D.@.5 (1c)	D.@Mx (1c)	D.@.5 (5c)
Spectral												
PCC	12.2	47.2	11.7	37.3	88.5	37.5	31.4	52.7	29.1	37.4	77.5	37.4
SA	11.7	48.9	11.7	37.4	88.9	37.4	28.7	53.5	28.7	37.4	81.0	37.4
SA_{Eq}	20.5	48.7	13.7	56.8	88.5	48.2	37.6	52.5	31.5	48.7	80.4	42.5
RGB												
SAM2-T	64.2	79.3	64.8	54.4	74.8	59.6	47.9	68.5	56.1	67.0	84.6	65.4
SAM2-T-FT	79.6	85.7	88.8	81.2	91.3	92.2	69.5	88.0	88.7	82.5	**92.9**	**93.7**
SAM2-S	62.7	78.6	64.5	55.6	75.4	64.6	49.4	66.7	62.5	65.1	83.1	65.8
SAM2-S-FT	79.9	85.1	88.9	80.8	91.2	93.9	68.9	87.6	89.1	82.6	91.3	92.7
SAM2-B+	61.7	76.6	60.1	53.8	71.2	60.9	43.5	60.3	56.7	57.5	76.4	59.5
SAM2-B+-FT	80.4	85.2	89.2	80.0	91.1	92.5	69.4	87.4	87.4	82.1	92.0	91.8
SAM2-L	60.0	77.3	64.4	52.3	72.7	59.1	53.0	66.7	61.1	62.6	80.9	64.8
SAM2-L-FT	80.2	86.0	89.2	74.9	90.3	91.3	70.7	85.4	90.0	83.4	90.6	91.5
MedSAM-Ad.	50.0	69.1	71.7	55.1	71.8	71.1	55.8	74.0	72.4	70.3	85.1	86.2
Fusion												
SAMSA	80.4	85.9	88.8	81.8	93.2	93.5	71.0	**89.9**	90.1	82.7	91.3	92.9
SAMSA2.0-T	80.5	86.0	89.0	86.0	**94.1**	95.5	**73.4**	88.4	90.1	**85.2**	92.2	92.2
SAMSA2.0-S	**82.2**	**86.5**	**89.7**	86.2	93.6	**96.4**	72.2	89.5	**90.6**	81.8	90.6	91.0
SAMSA2.0-B+	82.0	86.5	89.1	**86.2**	93.7	96.2	71.9	89.5	89.9	84.3	92.6	92.5
SAMSA2.0-L	81.4	86.2	88.7	85.0	93.8	94.6	72.4	88.8	90.2	84.2	91.4	92.2

5.3.3 Mixed Training

The effectiveness of our mixed-training strategy is demonstrated in table 5.2. This table compares the post-fine-tuning performance of a standard RGB model (SAM2) against our hyperspectral-enabled fusion model (SAMSA2.0), both of which were subjected to the same mixed-training protocol.

The results indicate that, in general, mixed-training strategy provides a greater performance uplift to the hyperspectral-enabled SAMSA2.0 architecture. There is much variation in the "best performing model" per dataset, which we believe is due to different sized features for different datasets. This coupled with the sampling strategy could be a source for the variation. For the initial automated delineation from a single click, the SAMSA2.0-Large-FT model achieved the highest average DICE score of 67.1%. During the interactive refinement phase with five clicks, the SAMSA2.0-Small-FT model was most effective, reaching an average score of 82.2%. This suggests that our training approach is particularly successful at leveraging the richer data provided by the hyperspectral model for both initial contouring and subsequent adjustments.

Table 5.2: Results across datasets with DICE at 0.5 threshold (D@0.5) at 1 (1c) and 5 (5c) clicks. Model suffixes: T (Tiny), S (Small), B (Base), L (Large), FT (Fine-Tuned). Values are percentages; bold indicates the best performance per column.

Model	Heipor		Hib		SB-X		SB-H		Average	
	D@.5 (1c)	D@.5 (5c)	D@.5 (1c)	D@.5 (5c)	D@.5 (1c)	D@.5 (5c)	D@.5 (1c)	D@.5 (5c)	D@.5 (1c)	D@.5 (5c)
RGB										
SAM2-T-FT	27.6	**61.2**	84.2	91.6	**77.6**	91.0	70.6	81.3	65.0	81.3
SAM2-S-FT	21.8	50.7	81.4	94.9	72.1	91.3	72.0	82.0	61.8	79.7
SAM2-B-FT	18.1	40.7	78.0	94.8	75.1	87.1	70.9	80.9	60.5	75.9
SAM2-L-FT	**32.3**	58.5	77.5	96.5	70.6	85.3	72.9	81.1	63.3	80.3
Fusion										
SAMSA2.0-T-FT	23.4	54.5	87.0	96.8	73.7	91.8	71.8	82.8	64.0	81.5
SAMSA2.0-S-FT	20.0	57.3	86.6	95.6	72.9	**93.4**	**73.5**	82.5	63.3	**82.2**
SAMSA2.0-B-FT	23.4	48.0	83.9	96.4	73.1	90.4	73.1	**83.2**	63.4	79.5
SAMSA2.0-L-FT	30.7	53.9	**89.9**	**97.8**	74.9	89.8	72.7	80.9	**67.1**	80.6

However, performance gains in the 5-click (5c) scenario are less pronounced than in the 1-click case, with standard SAM2 remaining competitive, especially on Heipor. This likely reflects a limitation of our current multi-click mechanism, which updates prompts by taking the pixel-wise maximum across SA maps—an approach that may not optimally integrate multiple clicks and needs re-visiting.

On the Heipor dataset, the mixed training did not yield a performance advantage for the SAMSA2.0 model. This may indicate that for tissues with characteristics already distinct in the visible spectrum, our training method could not extract additional diagnostic value from the hyperspectral data. However, it is more likely that the sampling strategy is far too strong since the comparative size for Heipor to the other datasets is much larger.

In contrast, the benefits of applying our mixed-training protocol to the hyperspectral model were clearly evident on the Hib, SB-X, and SB-H datasets. The clinical utility was especially notable on the Hib dataset, where the trained SAMSA2.0-Large-FT model achieved a DICE score of 97.8% after five clinician annotations, underscoring the success of our strategy in complex cases.

Qualitative comparisons further highlight the advantages of SAMSA2.0 over fine-tuned RGB baselines. On the Hib dataset (fig. 5.2), both SAMSA2.0-Large and SAM2-Large-FT correctly localise the tumour, but SAMSA2.0 provides a more refined boundary, offering greater clinical interpretability. On the SB-X and SB-H datasets (figs. 5.3 and 5.4), both methods capture the tumour within the marked region, yet SAMSA2.0 extends its prediction beyond the visible margins, suggesting the potential to reveal subsurface infiltration that may not be apparent in RGB-based predictions alone. For the Heipor dataset (fig. 5.5), performance is broadly comparable, though SAMSA2.0

produces fewer false positives, avoiding spurious activations in neighbouring tissue seen with SAM2-Large-FT. Finally, fig. 5.6 illustrates a representative workflow with SAMSA2.0-Large: tumour prompts reveal infiltration extending beyond the visible tumour boundary, while healthy prompts delineate the transition from normal to abnormal tissue. This bidirectional prompting demonstrates how SAMSA2.0 could support decision-making by helping identify resection margins more precisely.

Finally, our mixed-training experiment reveals a nuance regarding model size and its role in the clinical workflow. The training benefited the larger SAMSA2.0-Large model most for the initial segmentation proposal. In contrast, the smaller SAMSA2.0-Small model was more responsive and effective when refining contours with user input, suggesting a potential to investigate model variants further.

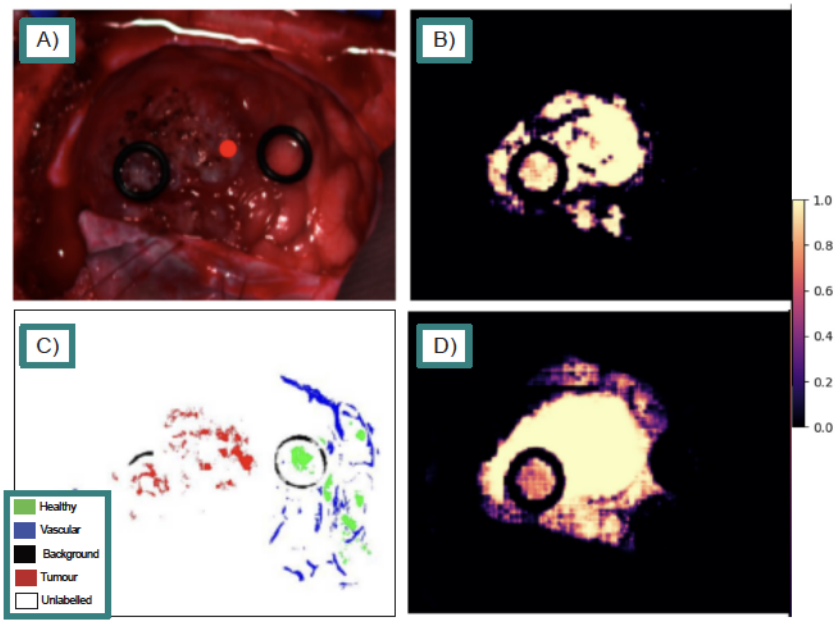


Figure 5.2: Sample segmentation results on the Hib dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity

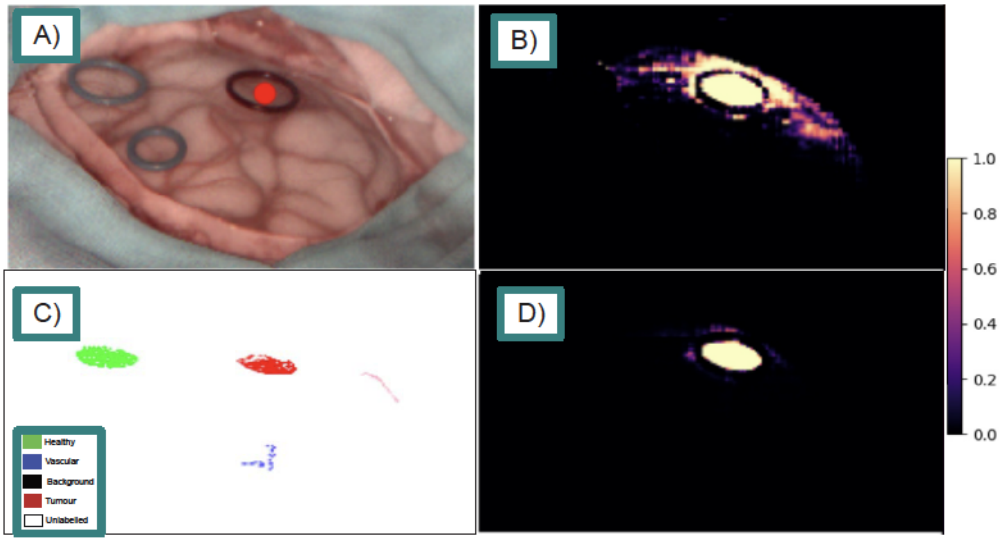


Figure 5.3: Sample segmentation results on the SB-X dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity

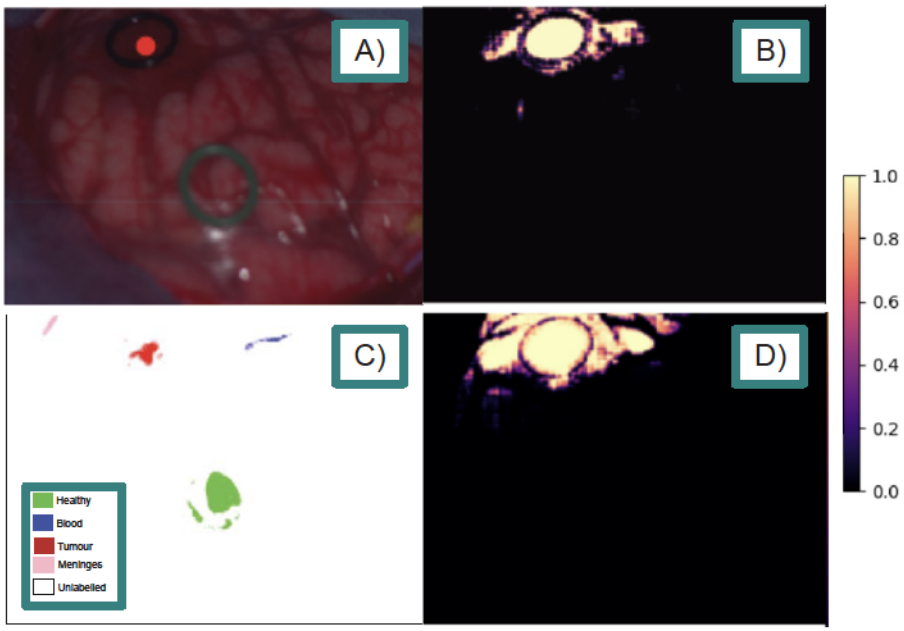


Figure 5.4: Sample segmentation results on the SB-H dataset, tumour class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity

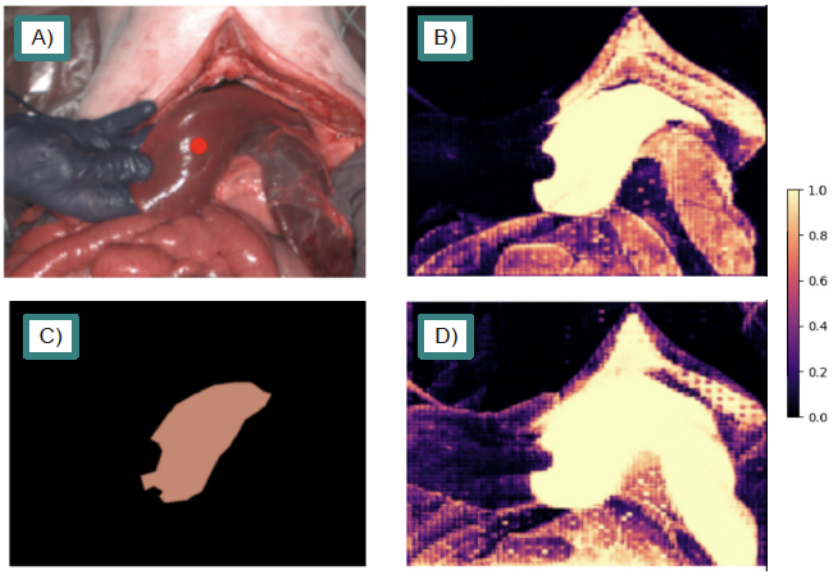


Figure 5.5: Sample segmentation results on the Heipor dataset, liver class. A) Pseudo-RGB image of the target. B) SAMSA2.0-Large standard predicted similarity. C) Ground truth labels. D) SAM2-Large-FT using SAM2 predicted similarity

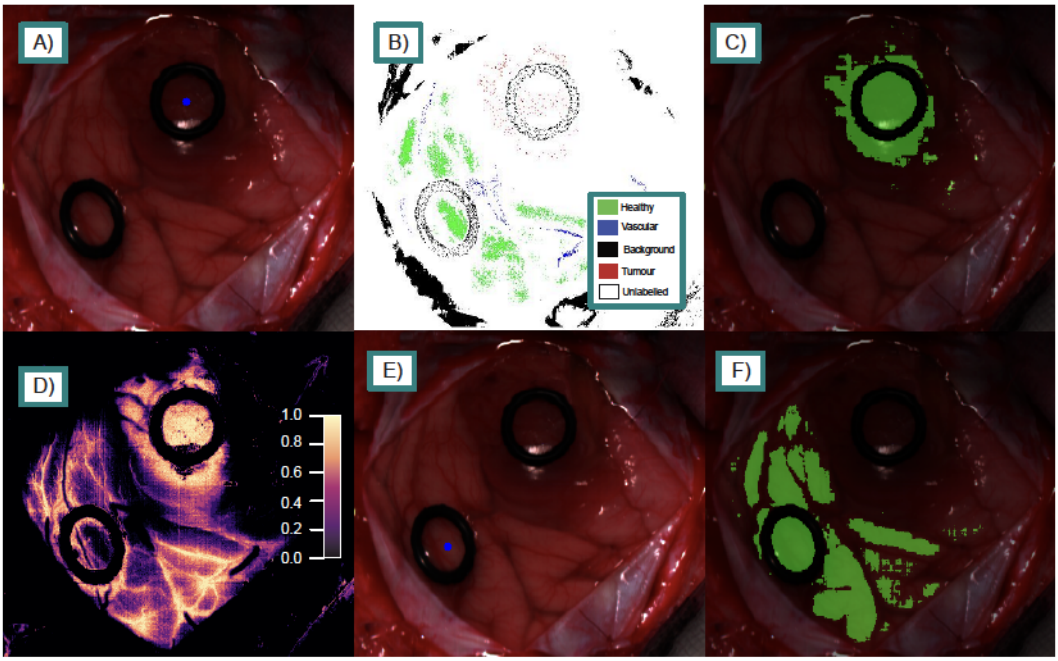


Figure 5.6: Example use case on Hib dataset, finding the boundary of tumour and non-tumour. A) Pseudo-RGB image of the target with tumour prompt. B) Ground truth labels. C) SAMSA2.0-Large standard segmentation from a). D) $SA_{Equalized}$ spectral similarity from a). E) Pseudo-RGB image of the target with healthy prompt. F) SAM2-Large-FT using SAM2 predicted similarity

5.3.4 Automatic Segmentation

Validating automatic segmentations generated by an interactive model remains an experimental task. The authors of SAM [97] rely on expert annotators and qualitative inspection for large-scale model training. We adopt a similar approach in this work. In fig. 5.7, we present the outputs of our automatic segmentation pipelines: SAM2 standard, as well as our light and heavy post-processing variations. The input RGB image is shown in fig. 5.7a, with the corresponding ground truth label in fig. 5.7b. The label reveals a clear hotspot or cluster of tumour annotations located between the two rings, representing the target region for our example surgical workflow. SAM2-Large using the standard automatic segmentation pipeline is shown in fig. 5.7c. While the segmentation is smooth and effectively identifies macro-structures, it lacks clinical relevance. The model captures large brain fold regions, which are disconnected by vessels, and distinguishes between the interior and material of the rings. To demonstrate that these limitations stem from the automatic segmentation pipeline rather than the model itself, fig. 5.7d shows SAMSA2.0-Large-FT with the standard pipeline. The segmentation maps appear noisier, as expected following fine-tuning on medical tissue, and overall performance is similar or slightly worse than SAM2-Large. Applying our light and heavy post-processing pipelines to SAM2-Large-FT results in the segmentations shown in fig. 5.7e and fig. 5.7f. The light variation produces noisier outputs but still highlights patchy regions that the model identifies as distinct. The heavy variation yields smoother, more contiguous masks. A similar trend is observed with SAMSA2.0-Large in fig. 5.7g and fig. 5.7h. These results suggest that simplifying SAM2's complex automatic segmentation pipeline can produce more clinically meaningful masks. However, the outputs remain noisy, and the shapes of segmented regions are often irregular, which may limit direct clinical use without further refinement. To illustrate potential applications, fig. 5.8 presents an example workflow. Suppose a surgeon identifies a patchy region surrounding the tumour as clinically relevant. This region can be selected interactively, as shown in fig. 5.8b and e. A single centre-of-mass click can then be used to generate targeted segmentations, as illustrated in fig. 5.8. The resulting segmentations from SAM2-Large-FT and SAMSA2.0-Large are shown in fig. 5.8c and f, respectively. This workflow has potential to support surgical decision-making by allowing rapid, interactive identification of regions of interest. For example, a surgeon could quickly highlight suspicious tissue during preoperative planning, receive an automatically generated segmentation of the surrounding anatomy, and adjust surgical strategy accordingly. Such a system could also be used intraoperatively to guide resection margins, identify critical structures, or provide real-time visual feedback, ultimately improving both safety and precision in neurosurgical procedures.

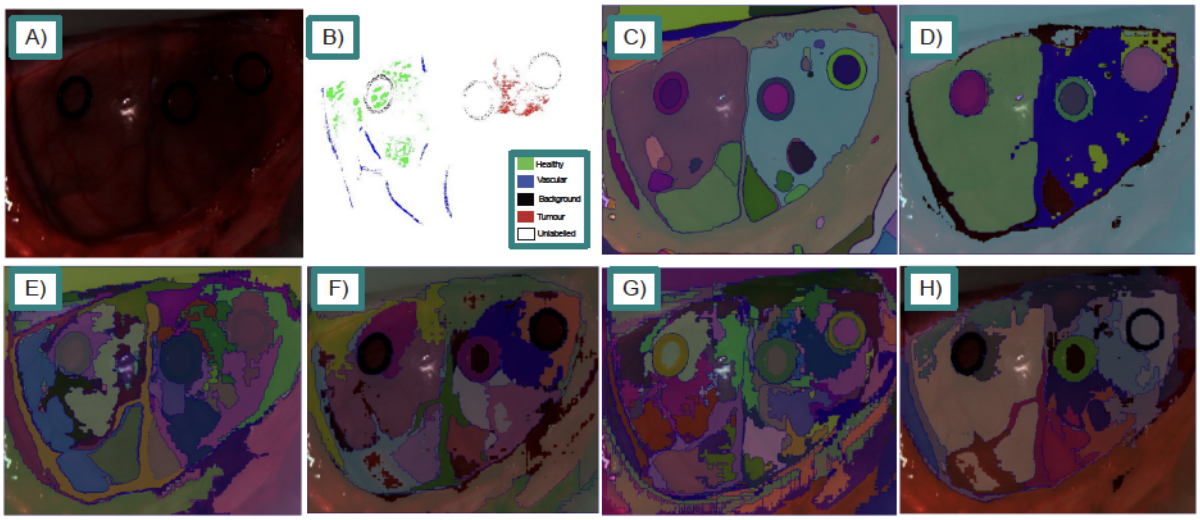


Figure 5.7: Automatic segmentation results on the Hib dataset. A) Pseudo-RGB image of the target. B) Ground truth labels. C) SAM2-Large standard automatic segmentation. D) SAM2-Large-FT using SAM2 standard automatic segmentation as input. E) SAM2-Large-FT using our light automatic segmentation. F) SAM2-Large-FT using our heavy automatic segmentation. G) SAMSA2.0-Large-FT using our light automatic segmentation. H) SAMSA2.0-Large-FT using our heavy automatic segmentation.

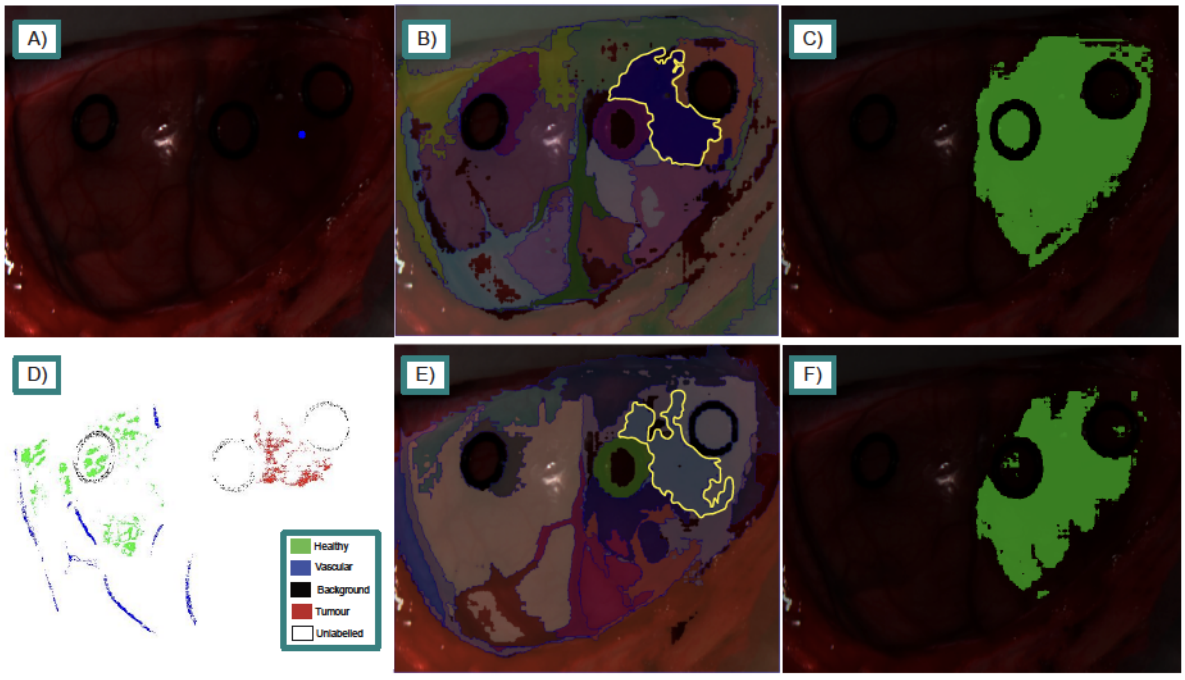


Figure 5.8: Example of interaction following automatic segmentation. A) Pseudo-RGB image of the target, with a blue click point. B) SAM2-Large-FT using our heavy automatic segmentation; the target-selected region is highlighted in yellow. C) Output segmentation from SAM2-Large-FT using the prompt in (a). D) Ground truth labels. E) SAMSA2.0-Large-FT using our heavy automatic segmentation; the target-selected region is highlighted in yellow. F) Output segmentation from SAMSA2.0-Large-FT using the prompt in (a).

5.4 Clinical Usage of SAMSA2.0

The clinical usage follows the main themes described in section 4.5 but we have no enabled automatic mask generation. Automatic mask generation shifts segmentation from fully interactive to hybrid autonomous operation. This reduces reliance on intraoperative input while still allowing surgeon guidance when needed. Automatic segmentation can run continuously, identifying tissue types without disrupting workflow. The dual-mode design supports different surgical needs: surgeons may use interactive mode for precise tumor margin delineation, while routine monitoring can remain automated. Although automation may introduce noise or irregular boundaries, lightweight post-processing produces clinically usable segmentations. This balance between speed and precision reflects surgical practice, where approximate boundaries often suffice for decision-making. The key question is how surgeons adopt the system in practice. A focused clinical study should assess how surgeons integrate SAMSA2.0 into their workflow. Surgeons would perform standardized tasks using both automatic and interactive modes, with data collected on decision time, override frequency, and segmentation accuracy. Usability, workload, and trust surveys, along with post-session interviews, would capture acceptance and preferred mode balance. Such a study requires fully labelled datasets rather than sparsely annotated ones like the Hib dataset to ensure reliable evaluation.

An additional consideration for eventual clinical and commercial deployment is the licensing of the underlying foundation models. The SAM2 model weights are released under the Apache 2.0 license, which permits modification, redistribution, and commercial use with minimal restrictions. This is advantageous for medical applications, where regulatory approval and commercial translation often require the ability to integrate, adapt, and redistribute models within proprietary systems. In contrast, some alternative foundation models impose more restrictive licenses that limit commercial usage or derivative works, potentially hindering clinical adoption. The permissive licensing of SAM2 therefore supports the long-term feasibility of deploying SAMSA in real-world clinical and commercial settings.

5.5 Conclusion

We presented SAMSA2.0, a method that enhances hyperspectral segmentation by embedding spectral angle information as a spatially aware input prompt. This early fusion strategy improves robustness in low-data and noisy conditions, consistently outperforming both spectral-only and RGB-only baselines with minimal user input. In addition, we explored automatic mask generation pipelines, demonstrating that lightweight post-processing can yield clinically meaningful segmentations, though at times with increased noise or irregular boundaries. Looking forward, further gains may come from end-to-end learning with the mask decoder, expanding training on larger and more diverse datasets, and validating performance in prospective clinical workflow studies. Ultimately, integrating SAMSA2.0 into intraoperative environments could provide surgeons with real-time, interactive guidance, supporting precise identification of tumour margins and improving decision-making during procedures.

Chapter 6

Multimodal Imaging Platform

This chapter explores a novel neurosurgical workflow designed to improve the detection and delineation of cancerous tissue. The proposed system integrates a white light RGB microscope with both a Hyperspectral Imaging (HSI) camera and a Probe-Based Confocal Laser Endomicroscopy (pCLE) system. The HSI camera provides a macro-level overview of the surgical field, guiding the surgeon to suspicious areas. The core innovation lies in the co-registration of these broad-field images with cellular-level optical biopsies from the pCLE, aiming to provide a continuous and accurate map for tumour resection.

The work in this chapter was published in the *International Journal of Computer Assisted Radiology and Surgery* [160]. The article is available as an open access publication under the terms of the Creative Commons license (see section A.0.2 for details).

6.1 Overview

Medical device development for open neurosurgery has advanced significantly, beginning in the mid-twentieth century with the introduction of the operating microscope. This instrument revolutionised neurosurgical practice by offering a magnified, three-dimensional view of delicate anatomical details, thereby improving the accuracy and safety of surgical procedures [100]. Subsequent innovation has concentrated on the integration of increasingly sophisticated technologies to further enhance the surgeon's abilities during an operation.

In the modern neurosurgical theatre, a range of technologies are integrated to supply information that extends beyond what the surgeon can directly observe. The rationale behind this is that various modalities provide complementary streams of data for evaluating the surgical site. For instance, whereas the operating microscope delivers high-resolution anatomical details, neuronavigation systems supply a spatial guide derived from pre-operative imaging data [95]. Simultaneously, optical imaging probes can provide real-time data concerning the tissue's biochemical makeup to define the boundaries of pathological tissue [23, 30].

A significant hurdle to the effective application of these systems is the challenge of aligning data from diverse sources into a unified and coherent reference frame for the surgeon. This information fusion needs to be accomplished in a way that does not overwhelm the surgeon, as the ergonomic and effective display of information is crucial for sustaining the flow of the operation and ensuring precision [170, 181].

The approach presented here utilises the primary operating microscope, a standard piece of equipment in all neurosurgical procedures, as a central integration point. Our work involves the development of bespoke computer vision and photonic technologies required to merge and present data from multiple sources directly within the surgeon's field of view via the microscope. The objective is to establish a single, integrated system that provides surgeons with a more comprehensive overview. This framework is intended to facilitate improved intraoperative decision-making without introducing additional complexity to the surgical instrumentation or the operating theatre environment.

6.2 Introduction

The primary objective in neuro-oncological surgery, achieving a complete resection of tumour tissue, is fraught with significant challenges and demands exceptional surgical precision. A successful resection is an important determinant of patient prognosis and of subsequent treatment [207]. The difficulty is particularly severe for infiltrative tumours like gliomas, which often extend into the brain parenchyma far beyond what is visible in conventional imaging [209]. Resecting these margins is difficult, as it risks damaging adjacent cortex, potentially leading to severe neurological disability [52, 81, 200]. Similar difficulties in defining resection boundaries arise with meningiomas, which can invade the surrounding dura or, in rarer malignant cases, the brain itself [16, 38]. While a suite of intra-operative technologies has been developed to improve tumour delineation, including neuronavigation [8], ultrasound [122], fluorescence guidance [29, 36], and Intraoperative Magnetic Resonance Imaging (iMRI) [104], none provide a perfect solution. Their clinical utility is often constrained by factors such as limited accuracy [35], operator dependence [122], prohibitive costs, or complex logistical requirements [163].

Although pCLE offers the capacity for real-time, *in vivo* histological assessment at the cellular level to theoretically guide a more precise resection [14, 30], its practical application is hampered by two major drawbacks: a severely restricted field of view that makes surveying large areas inefficent, and the absence of spatial tracking, which can lead to surgical disorientation [193].

In contrast, HSI is designed for wide-field tissue characterisation. It operates non-invasively by acquiring a complete reflectance spectrum for every pixel in the surgical scene [37]. This spectral information, which correlates with the tissue's biochemical properties, allows for the differentiation between healthy and pathological areas, thereby aiding in diagnosis and margin identification [7, 159].

We hypothesise that a synergistic combination of these two modalities, which has not been previously explored in a neurosurgical context, can overcome their individual limitations. Such an integrated system would leverage HSI's capacity for rapid, wide-area mapping to identify suspicious regions, which could then be subjected to high-resolution cellular investigation using pCLE. This creates a powerful, two-perspective workflow for tumour identification.

This study details the development and initial validation of a system that combines HSI and pCLE to improve brain tumour resection. Specifically, our work aims to develop: (1) a novel hardware and software platform for the co-registration and display of HSI and pCLE data; (2) a navigational framework to address the spatial orientation challenges associated with probe-based imaging (3) perform an initial assessment of the combined system's diagnostic accuracy in differentiating tumour from healthy tissue within an *ex vivo* phantom model, showcasing the complementary nature of the technologies.

6.3 Multimodal Imaging Platform Design

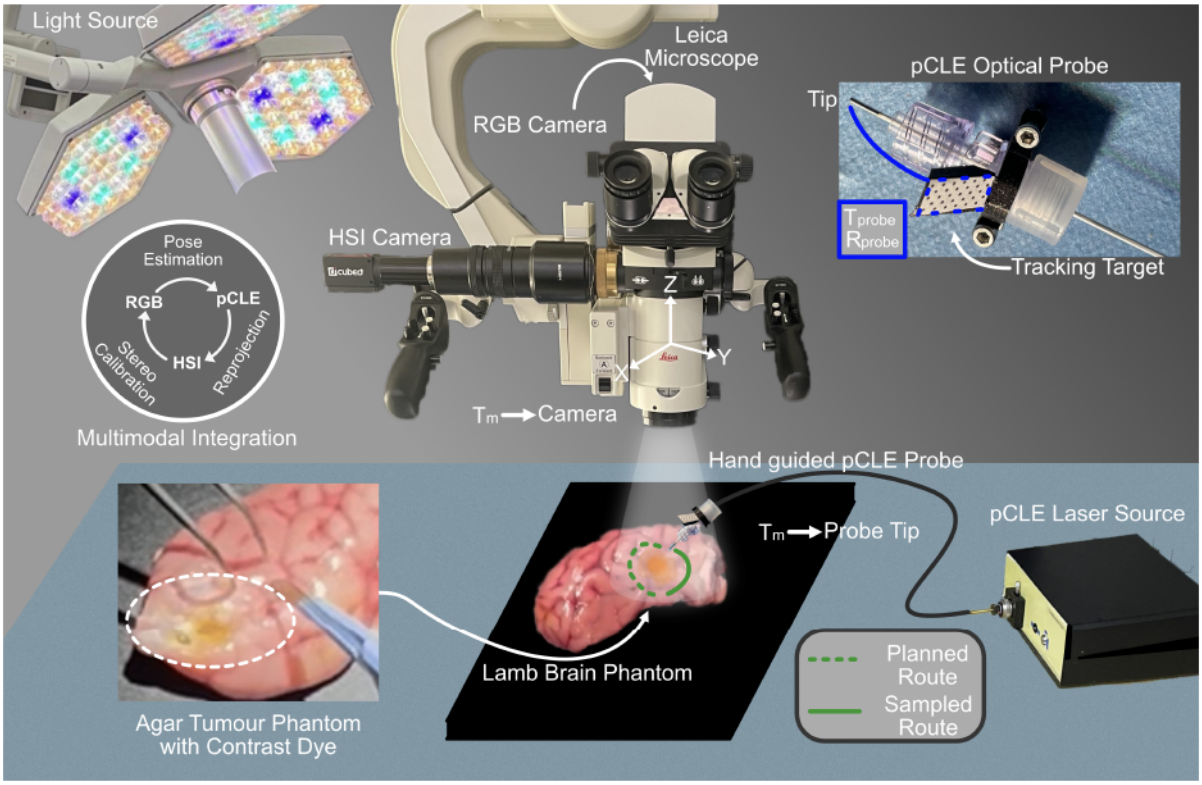


Figure 6.1: Our multimodal platform integrates RGB, HSI, and pCLE. The lamb brain phantom was imaged macroscopically using HSI and RGB, both integrated into a Leica microscope. The pCLE, operated manually, performs microscale optical biopsies around the tumour margins. By tracking the pCLE, we can correlate its data with the HSI modality using a tracking target attached to the pCLE probe [160]. Permission for the reuse of this figure is detailed in section A.0.2.

The multimodal platform described from herein was developed by integrating an operating microscope with both HSI and pCLE for comprehensive tissue analysis. A schematic of the experimental arrangement is presented in fig. 6.1. The core of our platform is a LEICA M525 OH4 operating microscope, which is equipped with an optical beam splitter that directs light to several ports, all sharing a common optical path. Consequently, adjustments to magnification and focus via the objective lens are uniformly applied to the surgeon's eyepieces and any attached imaging hardware. For white-light imaging, a LEICA True Vision TZ+ 3D Surgical Camera was affixed to the rear binoculars. This camera captures stereoscopic images at 120 Hz, and its output serves as the reference coordinate system for the entire platform. It produces composite 1920 x 1080 pixel RGB images, which are divided vertically to represent the left and right views. This configuration enables real-time visualisation of the surgical environment and facilitates the tracking of the pCLE probe.

The HSI component was incorporated by fitting a CUBERT ULTRIS SR5 camera to one of the microscope's optical ports using a custom-designed C-mount adapter. The LEICA M525 OH4 is an

operating microscope equipped with two side optical ports, in addition to front and back binocular ports. Although a standard television and camera mounting solution exists for these side ports, we chose to develop a custom adapter to decrease cost. The side optical port that was utilised is illustrated in fig. 6.2.



(a) Left view of the optical port.

(b) Right view of the optical port.

Figure 6.2: Two views of one of the side optical ports on a LEICA M525 OH4 microscope.

Our initial design approach involved 3D printing an adapter based on estimated dimensions and threads per inch taken from the port's dust cap. Although this method was functional, the plastic construction of the part rendered it suitable for single use only. To create a more robust and reusable solution, we fabricated the final adapter by manually turning it from brass, which ensures a solid fitting. Brass was specifically selected as the material to prevent any potential galling or fusion with the stainless steel components of the microscope's optics. You can find the schematics for the adapter in fig. 6.3 and example of images of the adapter with and without custom optics in fig. 6.4

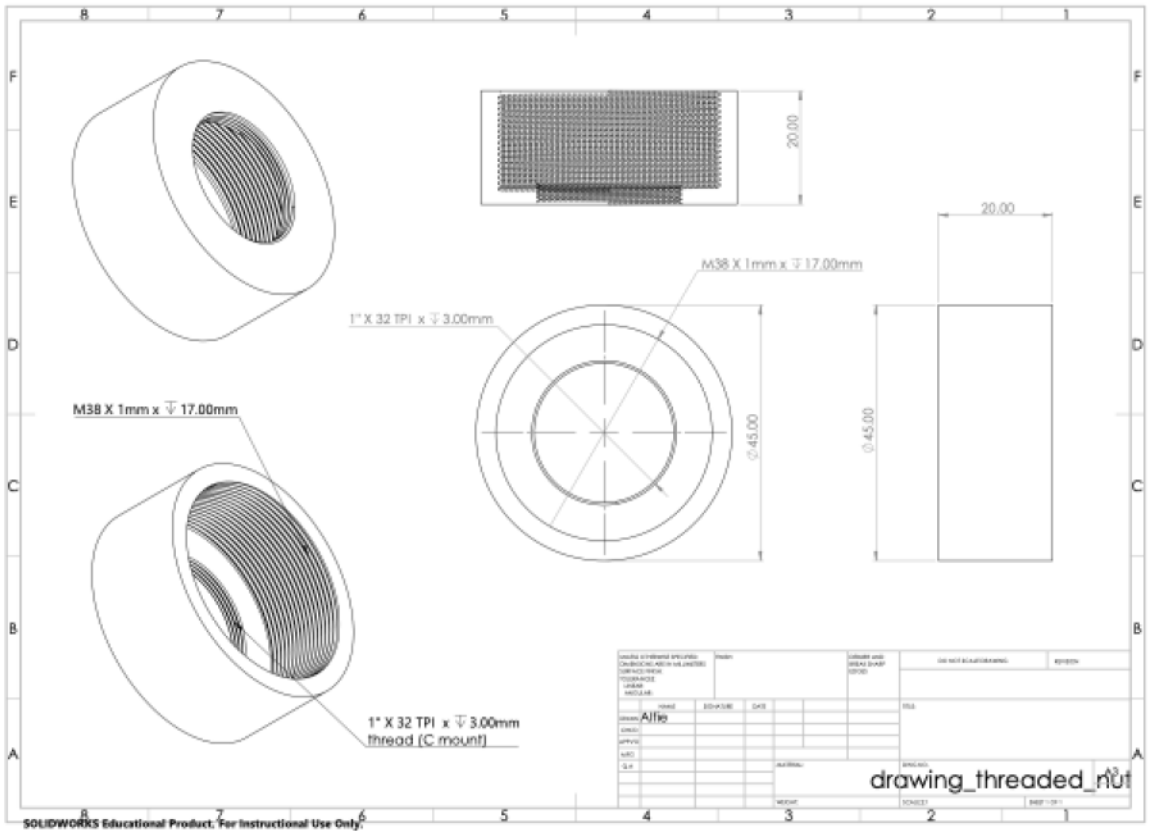
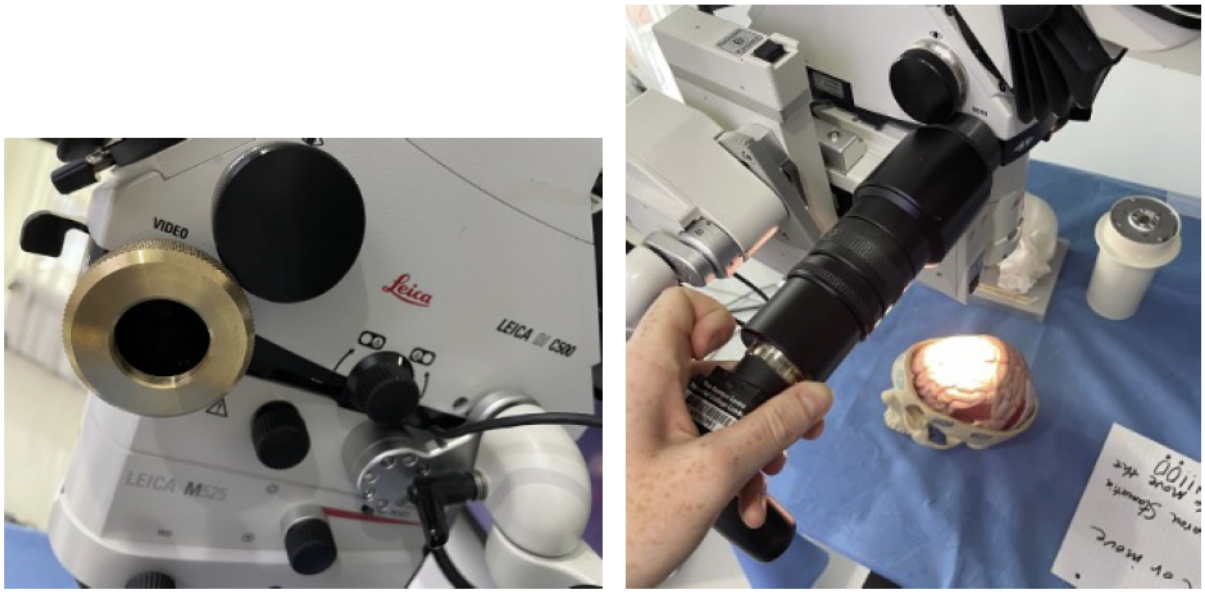


Figure 6.3: Schematic of the dust cap adapter that was manufactured to attach to the side dust cap of an optical port on a LEICA M525 OH4.



(a) Side view of custom adapter. **(b)** Adapter attached with zoom lens and camera.

Figure 6.4: Two views of one of the side optical ports on a LEICA M525 OH4 microscope with our adapter attached.

The optical port is natively designed for optics to be inserted internally, which ensures the camera's focal plane is the same as the view through the binoculars. However, since our custom adapter is based on the external dust cap fitting, the resulting optical setup is altered. This modified arrangement, with its approximate optical pathway, is depicted in fig. 6.5. To rectify this, we utilize a zoom lens to our setup. Any camera attached to our zoom lens can set a reasonable field of view for experimental work. This allows for flexibility in the working distance which a dedicated lens doesn't provide.

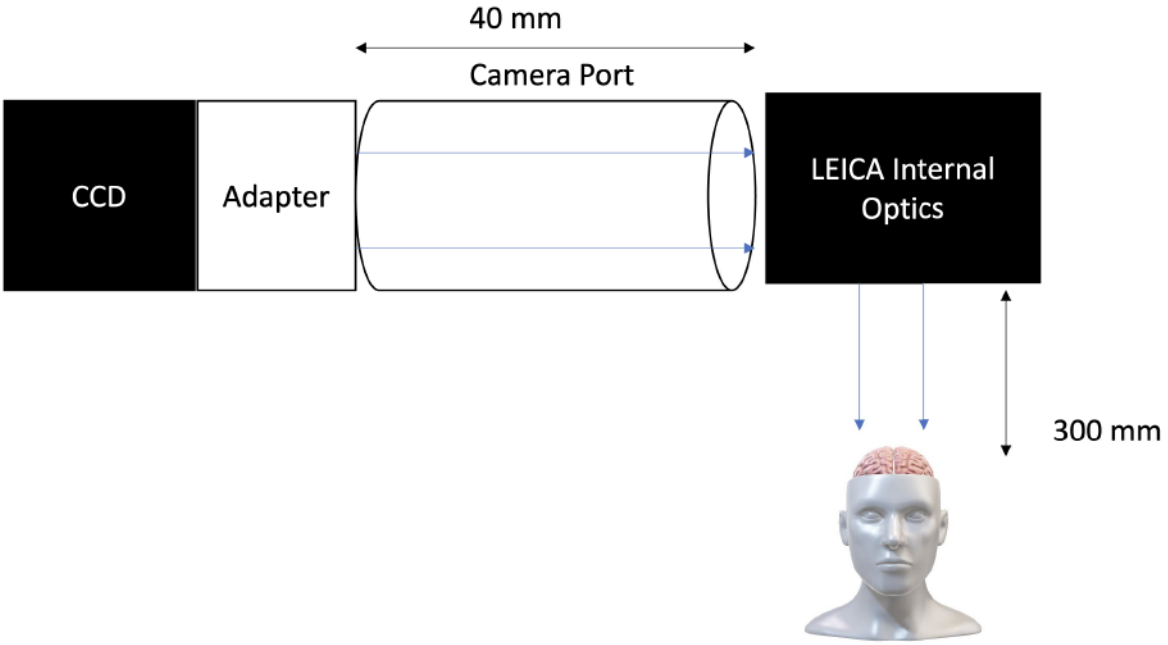


Figure 6.5: An example of the optical pathway for a camera attached to the port using our custom adapter. Measurements are approximate.

This configuration enables the concurrent capture of both white-light and hyperspectral data, designed such that the standard surgical workflow can proceed without interruption. The HSI camera gathers data across 51 spectral bands, spanning a wavelength range from 450 to 850 nm, at a maximal frame rate of 15 Hz and with a spatial resolution of 275 x 290 pixels. However, it was determined that the illumination spectrum from the microscope's light source only extended to approximately 770 nm. Consequently, our analysis was confined to the 40 spectral bands covering the 450–762 nm range. In a similar manner, the exposure duration necessary for adequate signal capture effectively reduced the practical acquisition rate to around 5 Hz. For the validation of the platform's colour accuracy, the spectral response from a Macbeth colour chart was evaluated. The specific chart employed for this validation process is illustrated in fig. 6.6.

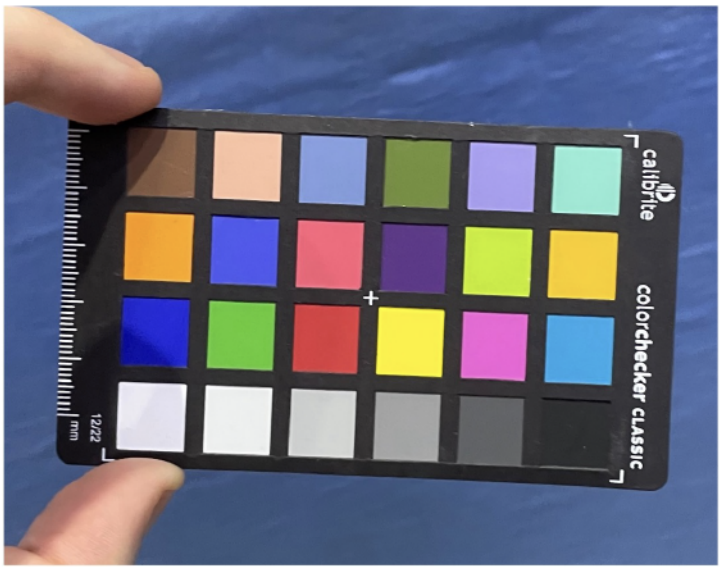


Figure 6.6: The Macbeth ColorChecker chart from calibrite, used as a calibration target.

A custom-built pCLE system was also incorporated into the platform. This system comprises a 488 nm laser for excitation and a Complementary Metal-Oxide-Semiconductor (CMOS) camera (Point Grey Flea®3, Teledyne FLIR LLC, USA) for detection. The optical path includes a 500 nm cut-on longpass fluorescence emission filter (FEL0500, Thorlabs Inc., USA) and a laser line clean-up notch filter (NF488-15, Thorlabs Inc., USA) [85]. For tissue interrogation, a high-resolution, flexible AQ-Flex19 fibre bundle probe (Cellvizio, Mauna Kea Technologies, France) is used. This setup is capable of acquiring images at 120 frames per second, providing real-time cellular-level visualisation over a $325\ \mu m$ field of view, making it suitable for intraoperative use.

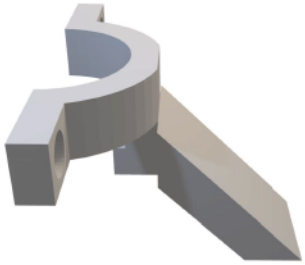
Unlike the other imaging devices, the pCLE system was not physically mounted to the microscope but was instead operated as a freehand tool adjacent to the surgical field. The high temporal and spatial resolution of the RGB camera is leveraged for the precise pose estimation of the pCLE probe. A marker-based approach was selected for tracking the pCLE probe. Although markerless tracking systems exist, they were considered unsuitable for this application due to potential challenges with robustness and the computational demands required in a real-time clinical setting.

Our system utilises a custom-designed marker holder that attaches directly to the standard Cellvizio probe sheath. This sheath, which is similar to a haemostatic valve introducer, incorporates a rubber seal to prevent backflow. It serves as an ideal and stable platform for the marker holder, ensuring that the registration of the probe tip remains constant throughout any procedure. The holder and attached fiducial marker can be seen in fig. 6.7. The marker itself is an asymmetric circle grid pattern generated using the OpenCV library [22]. The asymmetric nature of this pattern is essential as it allows for unambiguous, directional pose estimation. The dimensions of the circles were chosen to balance two competing requirements: they needed to be small enough to avoid obstructing the operative field, yet numerous enough within the camera's working range to yield an accurate and stable pose reading. The pattern specifics are presented in table 6.1.

Table 6.1: Generation parameters for the fiducial marker pattern.

Parameter	Value
Pattern Type	Asymmetric Circles ('acircles')
Grid Dimensions	5 rows × 6 columns
Units	Millimetres (mm)
Grid Element Size	1.0
Circle Radius Rate	5.0

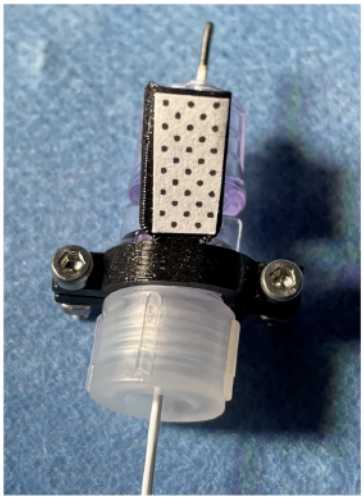
Following a calibration procedure, the determined pose of the probe is projected into the HSI coordinate frame. This final transformation ensures accurate spatial co-registration between the microscopic pCLE data and the macroscopic HSI data.



(a) Side view of holder design.



(b) Side view of pCLE holder in situ.



(c) Front view of pCLE holder in situ.

Figure 6.7: Example images of the pCLE holder design and its in situ application.

6.3.1 Calibration of Imaging Modalities

To ensure data consistency across different acquisitions, a distinct calibration procedure was established for each imaging modality. For this RGB stream, a white-balance calibration was performed to compensate for spectral variations in the illumination provided by the microscope's light source. Executing this step ensures that the captured images have reliable colour fidelity and good contrast.

The HSI modality is also sensitive to illumination conditions and thus required its own calibration process. To normalise the hyperspectral data and remove the influence of the illumination spectrum and sensor noise, a standard white-dark calibration was implemented. This involved acquiring a white reference image (W) of a reflectance target and a dark reference image (D) with the sensor occluded. Each raw hyperspectral image (R) was then calibrated using the expression Calibrated Image $= \frac{R-D}{W-D}$. This conversion results in a hyperspectral data cube where each pixel value represents a unitless reflectance fraction.

The calibration procedure for the pCLE system involved a background subtraction step. A reference frame was first acquired without active fluorescence to capture the background signal from ambient light. This frame was subsequently subtracted from every image captured during the experiment. This process effectively eliminates signal contamination from external light sources, such as the main operating microscope's illuminator, ensuring that the final image accurately represents the true fluorescent signal from the tissue.

6.3.2 Spatial Alignment of Multimodal Data

To co-register data from the various sensors into a common coordinate frame, a spatial calibration of the entire platform was essential. Direct alignment was required between the left stereoscopic view from the surgical microscope and the HSI 'right' view. This alignment was achieved by determining the transformation between the two views via a stereo calibration process, which involved capturing 30 images of a checkerboard pattern, as depicted in fig. 6.8(a). Intrinsic matrices (K_{left}, K_{right}) and extrinsic rotation (R) and translation (T) defining the rigid transformation between sensor frames were then calculated. This transformation maps a 3D point P_{left} in the left camera frame to P_{right} in the right frame: $P_{right} = R \cdot P_{left} + T$. Following this, the point P_{right} can be projected onto the HSI image plane to determine its corresponding pixel coordinates (u_{right}, v_{right}) by applying the camera's intrinsic matrix:

$$\begin{bmatrix} u_{right} \\ v_{right} \\ 1 \end{bmatrix} = K_{right} \cdot \frac{1}{Z_{right}} \cdot P_{right}, \quad P_{right} = \begin{bmatrix} X_{right} \\ Y_{right} \\ Z_{right} \end{bmatrix} \tag{6.1}$$

where (u_{right}, v_{right}) represents the pixel location on the right image.

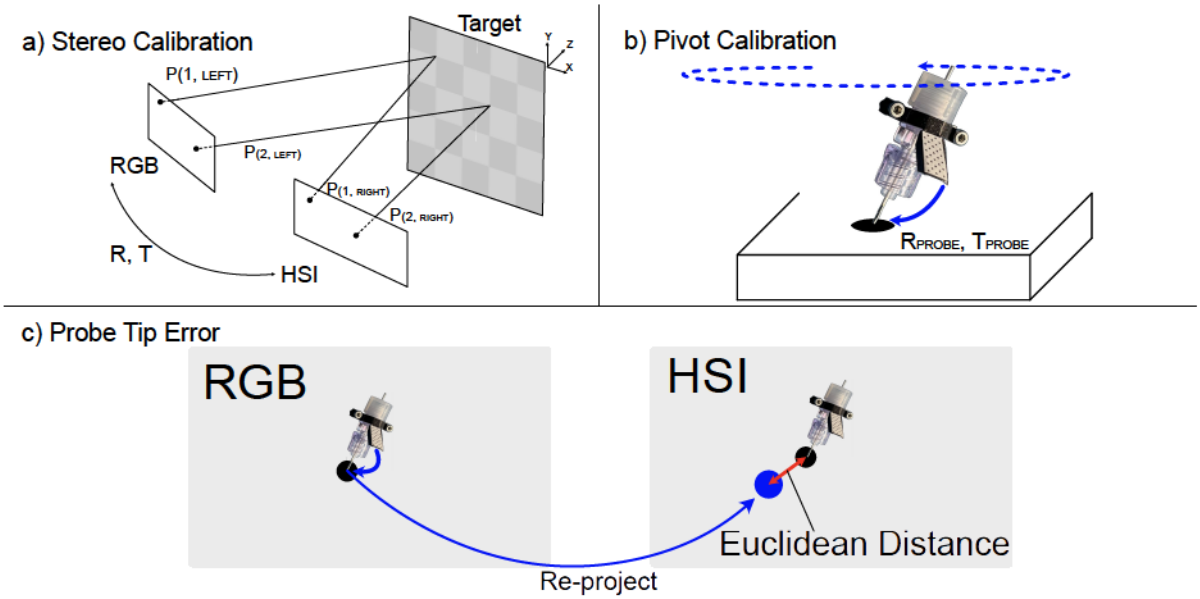


Figure 6.8: Methodology for calibration and analysis of re-projection error: a) The stereo calibration process between the RGB and HSI cameras, where R and T represent the rotational and translational transformations determined during calibration. b) An illustration of the pivot-based calibration, highlighting the alignment points between the different imaging systems. Here, R_{probe} and T_{probe} denote the rotation and translation from the marker to the probe tip, which are computed during this procedure. c) A depiction of how the re-projection error is calculated, showing the spatial offset between corresponding points in the RGB and hyperspectral images [160]. Permission for the reuse of this figure is detailed in section A.0.2.

To map the microscale optical biopsy data from the pCLE probe onto the macroscopic HSI and RGB imagery, it is first necessary to localise the probe's tip. For this purpose, a keydot marker was attached to the probe's shaft, which allowed for its detection within the colour video feed. The detection of this marker was carried out using the OpenCV library [22], after which its six-degree-of-freedom pose was estimated with a RANSAC-based perspective-n-point algorithm. This process yields the pose of the marker itself, not the probe's tip, necessitating an additional calibration step to determine the rigid transformation (comprising rotation R_{probe} and translation T_{probe}) between the two. This transformation was found using pivot calibration [218], a technique shown in fig. 6.8(b). The method involves holding the probe tip stationary at a single point in space while pivoting the probe through a range of different orientations. Since the 3D position of the tip remains fixed as the marker's pose varies, the transformation can be recovered by solving a system of equations. We show some example outputs of the probe tip estimation in fig. 6.9. After calibration, the 3D position of the probe tip, P_{probe}, can be derived from the world coordinates of the marker, P_{marker}, using the following equation:

$$P_{\text{probe}} = R_{\text{probe}} \cdot P_{\text{marker}} + T_{\text{probe}} \tag{6.2}$$

The final step in mapping the location of an optical biopsy involves projecting the computed 3D

position of the pCLE probe tip onto the HSI camera's image plane, utilising the transformation detailed in eq. (6.1).

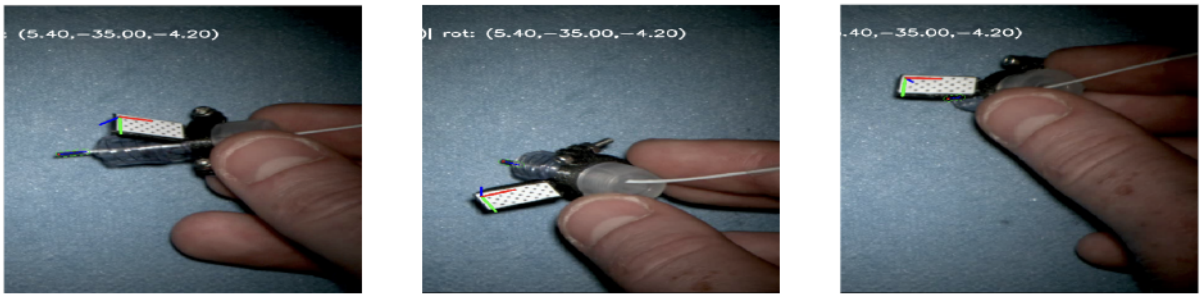


Figure 6.9: Example output images of the pCLE probe tip pose estimation. We have overlayed the estimated probe tip in red, a green cylinder representing the metal tip and the probe world axis on the first keydot.

To assess the accuracy of the probe tip localisation, a systematic protocol was established to measure the re-projection error. This procedure required capturing image data as the pCLE probe tip was sequentially placed on a set of reference markers visible in both the RGB and HSI camera views. The centre of each reference marker, which could be located with high precision in the HSI image plane, acted as a ground truth coordinate. For every ground truth point, the 3D position of the probe tip was calculated via the marker tracking and pivot calibration transformation (as per eq. (6.2)) and subsequently re-projected onto the HSI image plane. The probe tip error for a given measurement was then defined as the Euclidean distance separating this re-projected pixel from the centre of the corresponding ground truth marker. Figure 6.8(c) provides a visual summary of this evaluation process. By averaging these individual errors across all reference points, a robust metric for the system's overall spatial accuracy was determined. This approach, similar to the methodology used in [27], delivers a quantitative evaluation of the geometric transformation's dependability. Given that contemporary clinical guidelines for glioma resection recommend margins of 0.5 to 3.0 cm beyond the gadolinium-enhancing tumour region [123, 197, 198], a probe tip error below 1 mm is considered acceptable for the system's clinical application.

The overall calibration procedure is depicted in fig. 6.10. Once calibrated, the system can be used repeatedly without needing to perform the entire setup process again. Even though the stereo and probe-tip calibrations are a one-time procedure, the white-dark referencing and pCLE background subtraction are performed before each new imaging session to ensure data consistency under varying lighting conditions.

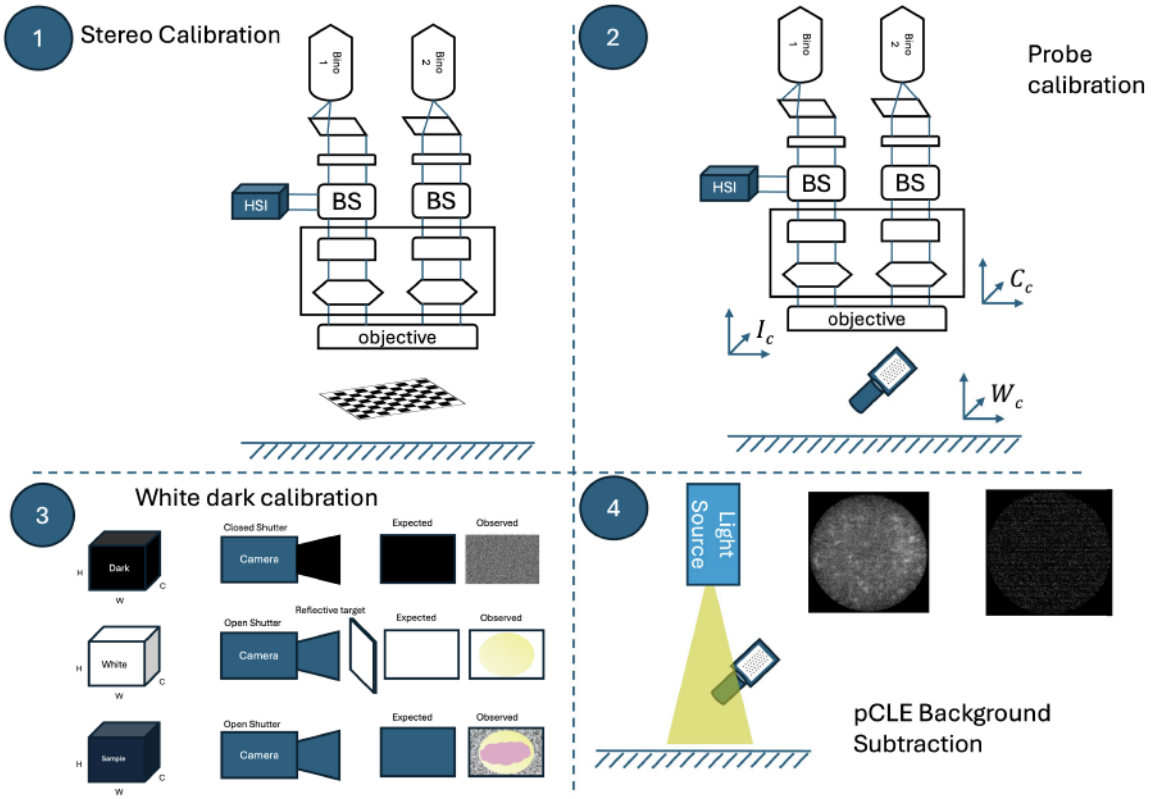


Figure 6.10: An overview of the four-stage calibration procedure required before an experiment. The steps are: (1) Stereo calibration to co-register the RGB and HSI cameras; (2) Probe-tip calibration to find the transformation between the tracking marker and the probe tip; (3) White-dark referencing for spectral consistency; and (4) Background subtraction for pCLE measurements.

6.4 *Ex vivo* Dataset

			Spectra			pCLE Images	
Sample #	Volume (ml)	Size (mm)	BG	Healthy	Tumour	Healthy	Tumour
1	2	15.8 x 17.6	50176	28202	1372	614	852
2	1	6.2 x 5.0	53811	25723	216	484	214
3	2	13.6 x 11.8	43725	34841	1184	572	192
4	2	14.0 x 12.1	53941	24477	1332	218	563
5	1	7.1 x 8.7	53827	25427	496	319	262

Table 6.2: This table presents the volume of injected agar-acriflavine and the size (height and width) of the exposed tumour at the selected key frame. The last 5 columns detail the class distribution of HSI spectra and pCLE images across the dataset [160]. Permission for the reuse of this table is detailed in section A.0.2.

To validate our multimodal platform, a realistic *ex vivo* brain model was developed. Following the methodology of [206], a lamb brain phantom was selected due to its established anatomical similarity to the human brain [92, 93, 137], its cost-effectiveness, and its ready availability. Despite minor spectral deviations, the structural resemblance ensures that the response properties of non-fluorescent tissue to HSI and pCLE are comparable, thereby validating its suitability for this investigation.

To simulate tumour masses for identification experiments, artificial lesions were created within the lamb brains. After evaluating several tumour simulation techniques reported in the literature, such as plastic biomodels [42], 3D printing [189], polyvinyl alcohol [121], and various hydrogels [62, 212], an injectable agar-based model was chosen [206]. Agar was deemed the most suitable material because it closely mimics the texture and light diffusion characteristics of brain tumours [4, 196]. Its injectable nature allows for the creation of subcortical phantom masses without disrupting the overlying surface, simulating a trans-cortical surgical approach with both pre and post-exposure imaging [206]. Crucially, and unlike many alternative phantom materials, agar can be easily mixed with acriflavine, the fluorescent contrast agent required for pCLE. The inclusion of acriflavine results in a bright fluorescent response in agar-containing regions, enabling clear differentiation from non-fluorescent tissue and providing an ideal simulation of clinical tumour visualisation. Example images of agar injection of the lamb brain samples can be seen in fig. 6.11a.

The final dataset was constructed from five lamb brain specimens, each approximately 10x7x3.5 cm. The operating microscope was positioned at a working distance of 30 cm from each specimen. At this height, the resulting pixel sizes were 0.12 mm for the RGB camera and 0.31 mm for the HSI camera. The larger pixel size for HSI is a consequence of its lower native resolution. Following [206], a 20% agar solution was prepared and mixed with acriflavine in an approximate 2:1 ratio to achieve a fluorescent dye concentration sufficient for the pCLE system. This mixture, in volumes ranging from 1–2 ml, was injected into each brain and allowed to set for 10 minutes. This phantom



(a) Injection of agar jelly into a lamb brain specimen. **(b)** Neurosurgeon exposing the injection site.

Figure 6.11: Preparation of the *Ex vivo* dataset.

preparation method consistently produced tumour-like masses that were clearly discernible by all three imaging modalities of the platform: white-light RGB, HSI, and pCLE.

The protocol for image acquisition commenced with the surgical exposure of the artificial tumour's core by a neurosurgeon, who resected the overlying brain tissue, example can be seen in fig. 6.11b. In this phantom study, the injection site guided the initial localisation, whereas in a clinical setting, pre-operative imaging would be used. After exposure, HSI was employed for macroscopic tumour localisation, followed by pCLE for a detailed microscopic assessment of the margins. The pCLE sampling strategy was designed to be comprehensive, covering three distinct regions: the tumour core, healthy brain tissue (sampled from sites distant from the tumour), and the tumour margin, where 15–20 discrete optical biopsies were taken for each specimen.

Tumour annotations were created in the RGB images, guided by the known injection parameters and colour differences between agar and brain tissue. These annotations were transferred to the HSI data using the established RGB-HSI co-registration (fig. 6.12). To establish a ground truth for pCLE classification, images from the healthy regions and the tumour core were used to define the two classes for training. Data from the margin scans were held out and used exclusively for testing, a strategy that prevents data contamination and ensures a rigorous evaluation. The dataset was categorised into three classes: healthy brain tissue (healthy), the agar-acriflavine mixture (tumour), and non-tissue areas (background, BG). In the RGB annotations, these classes are visualised as green (healthy), red (tumour), and black (background). A summary of the dataset's composition is provided in table 6.2.

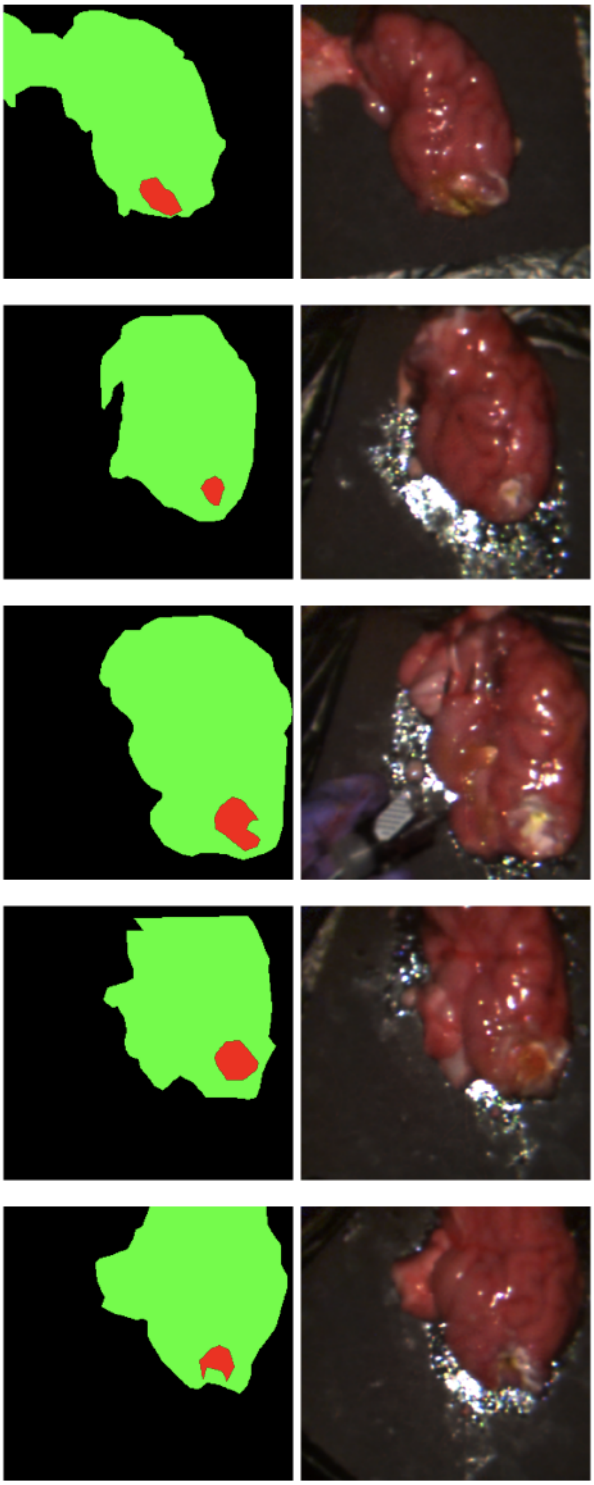


Figure 6.12: The five brain phantoms. Left HSI viewpoint of sample. Right. Annotations in HSI viewpoint.

6.5 Multimodal Tumour Identification

This section details the design of our multimodal algorithm for tumour identification, which incorporates HSI, pCLE, and their combined application. We assessed performance in a data-scarce setting using 5-fold cross-validation, training on four samples and testing on one. Class imbalance was mitigated with undersampling, randomly reducing the majority class to match the minority class size. For HSI segmentation, pixel-wise methods were chosen as they are well-suited to this low-data regime. We selected the two best-performing spectral classification methods identified in [103]: a Support Vector Machines (SVM) [39] and a neural network. The SVM was configured with a radial basis kernel and a regularisation parameter of 1. The neural network features a three-layer architecture, comprising two hidden layers with 128 and 64 nodes respectively, and a final output layer. It was trained using cross-entropy loss and the Adam optimiser [96] with a learning rate of 0.001. In the context of pCLE classification, machine learning has become the predominant approach in recent years [108, 162]. A key priority for our study was the optimisation of inference speed to enable real-time classification during surgical procedures. Consequently, we evaluated two convolutional neural networks renowned for their rapid inference capabilities: MobileNetV2 [173] and EfficientNet [188]. These pCLE models were trained using the identical loss function, optimiser, and learning rate as the HSI neural network. All models supporting GPU acceleration were trained on an NVIDIA 3090 GPU using PyTorch [144], while the Scikit-learn library was utilised for the SVM implementation [146].

To fuse these imaging modalities, a non-machine learning methodology was adopted to improve interpretability and simplify its integration into the surgical workflow. The HSI+pCLE combination began with a logical OR operation between the outputs of the two modalities, thereby highlighting any region identified as a potential tumour area by either technique. Given that pCLE directly targets tumorous tissue, this initial result was refined. The process involved retaining only the Largest Connected Component (LCC) within the HSI segmentation that spatially contained the pCLE-positive points. All other disconnected segments were removed, yielding a pCLE-informed HSI segmentation. Any tumour regions filtered out during this step were replaced with the tissue class corresponding to the next highest confidence score from the HSI models. The complete pipeline is illustrated in fig. 6.18. For a comparative baseline using HSI alone, an analogous filtering approach was implemented. In this HSI-only segmentation, where no pCLE data was available, only the largest connected component (HSI LCC) was retained, and all smaller, disparate segments were discarded, mimicking the filtering process but without the guidance of pCLE information.

This experimental design was structured to facilitate a comprehensive evaluation of each modality, both independently and in combination. Such an approach ensures that the developed algorithm is not only robust but also practical for clinical deployment in surgical applications, especially within data-limited scenarios.

6.6 Results

6.6.1 Macbeth Colour Validation

The Macbeth colour checker results are shown in fig. 6.13, which compares the reflectance spectra of selected patches measured by the hyperspectral system against the reference data.

Overall, the measured spectra follow the same shape as the ground-truth profiles, indicating good agreement in spectral behaviour. In some cases, however, the curves are shifted vertically, with the measured reflectance appearing consistently higher or lower than the reference. This offset is expected and arises from small variations in sensor-to-chart distance during acquisition, which affect the intensity of recorded light. Since this illumination effect scales the spectrum without changing its shape, it does not impact the evaluation of colour fidelity.

6.6.2 Timing Measurements

The timing results are shown in fig. 6.14, which compares probe timestamps with those from the LEICA RGB stream. The plots demonstrate that the two systems remain closely aligned over the course of an experiment. The global timing difference across all poses stays within a narrow range, while the right panel highlights the local behaviour, where the mean offset is approximately 5.53 ms. This level of error is sufficiently small to confirm that the probe and RGB data are effectively synchronized for downstream analysis.

6.6.3 Spatial Calibration

Table 6.3: A summary of the spatial and temporal performance statistics for the multimodal platform [160]. Permission for the reuse of this Table is detailed in section A.0.2.

Reprojection Error (pixels) ↓	**Probe Tip Error** (pixels) ↓	**pCLE Scan Time** (s) ↓
0.34 ± 0.1	1.2 ± 0.2	9.6 ± 4.1

In a multimodal imaging platform such as ours, the precise integration of different modalities is paramount to prevent errors arising from spatial misalignment. To evaluate the quality of this integration, the reprojection error between HSI and RGB was measured using the stereo calibration method described in Section section 6.3. This yielded a reprojection error of 0.34 pixels. In addition, the probe tip localisation error was assessed and found to be 1.2 pixels with a standard deviation of 0.2 pixels. The low magnitude of these errors confirms a high degree of spatial consistency between the modalities. Finally, the average time taken to scan the entire tumour margin with the pCLE probe was recorded as 9.6 seconds, with a standard deviation of 4.1 seconds. A summary of these results is presented in table 6.3.

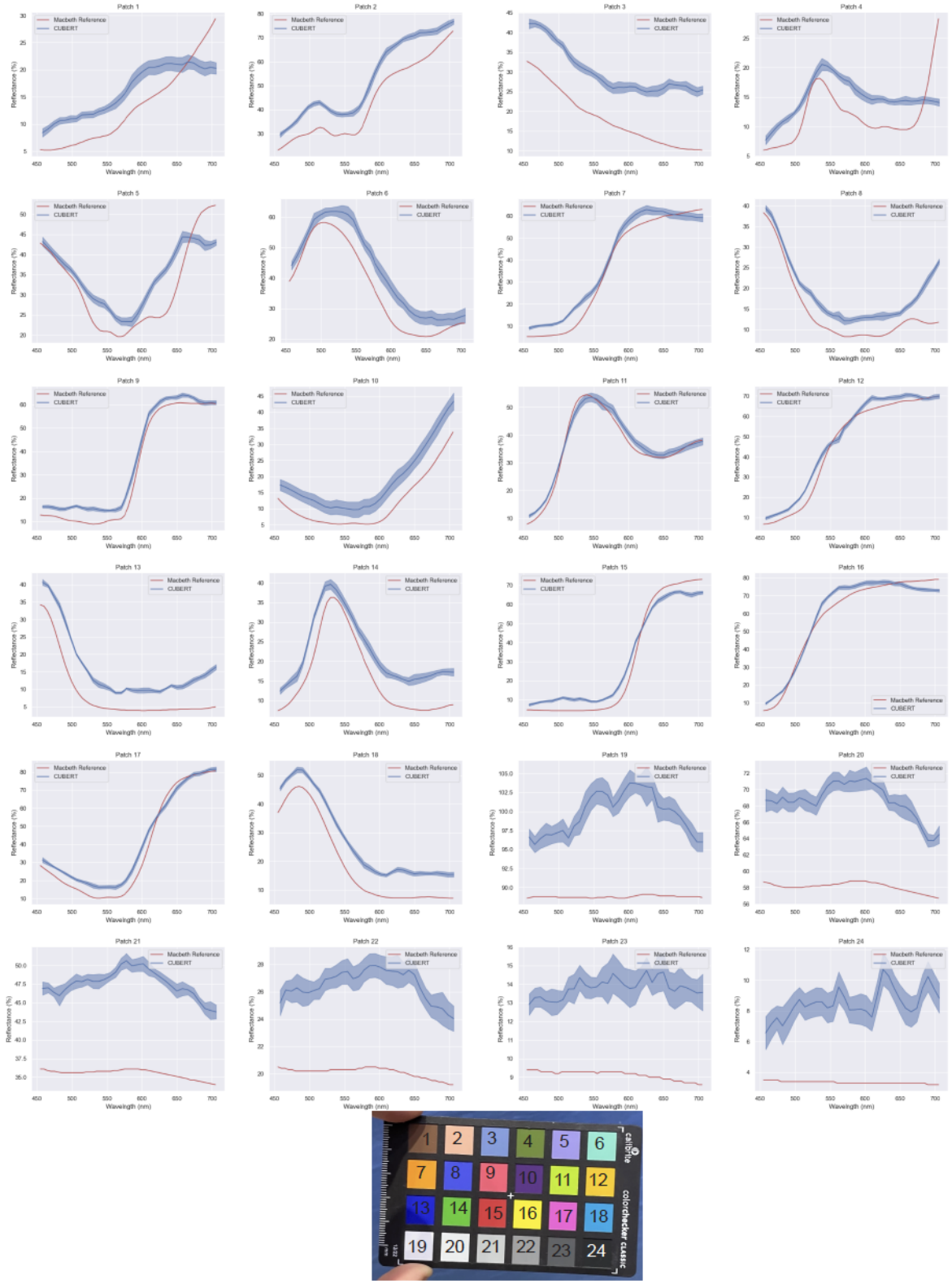


Figure 6.13: Macbeth colour checker results (Calibrite). Measured values are compared at each wavelength with reference data from the Munsell Color Science Laboratory, available at https://www.rit.edu/science/munsell-color-science-lab-educational-resources.

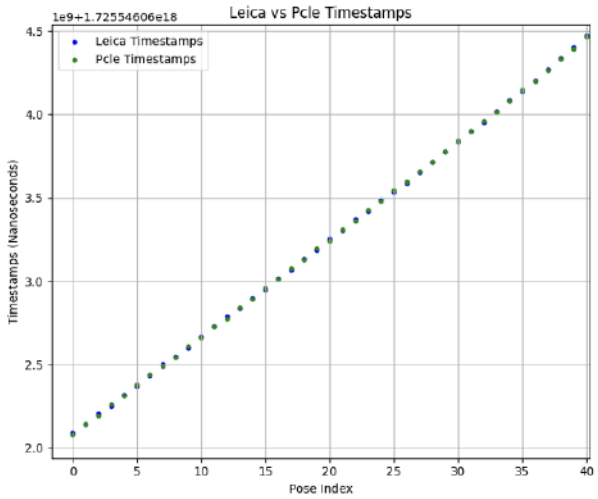


Figure 6.14: Timing measurements between probe and RGB timestamps. Left: global timing differences across all poses in one experiment. Right: local differences showing a mean error of 5.53 ms.

6.6.4 Feasibility Test with Dye Staining

As an initial demonstration, a lamb brain specimen was prepared with one half stained using acriflavine and the other left unstained. The results are shown in fig. 6.15. Both HSI and pCLE clearly distinguish between the two regions. In the hyperspectral data, the stained half shows a distinct absorption feature between 420–500 nm, consistent with the known absorption spectrum of acriflavine. The pCLE scans reinforce this separation: the unstained side shows minimal contrast, while the stained side reveals strong fluorescence. Similarly, fig. 6.16 shows the 3D probe tip position with the corresponding spectra, consistent with the observations above.

Although this experiment uses dye versus no dye, it demonstrates that the two modalities can reliably separate tissue types based on spectral and fluorescence signatures. This forms the basis for later experiments targeting tumour versus non-tumour tissue.

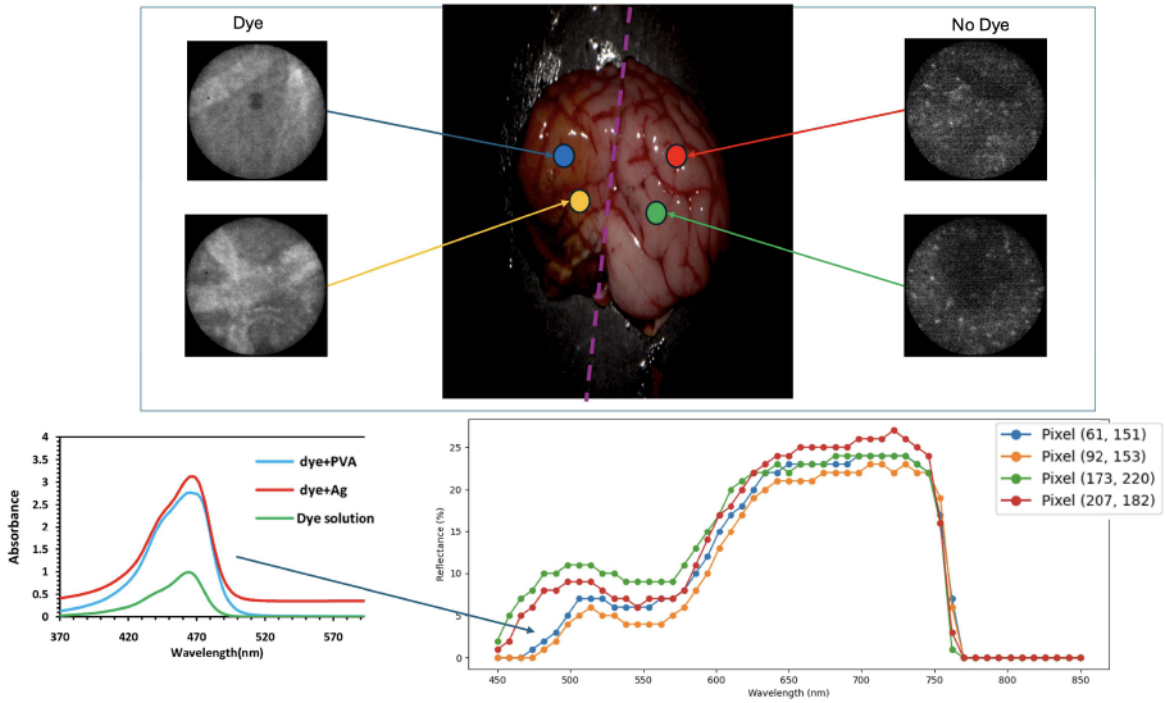
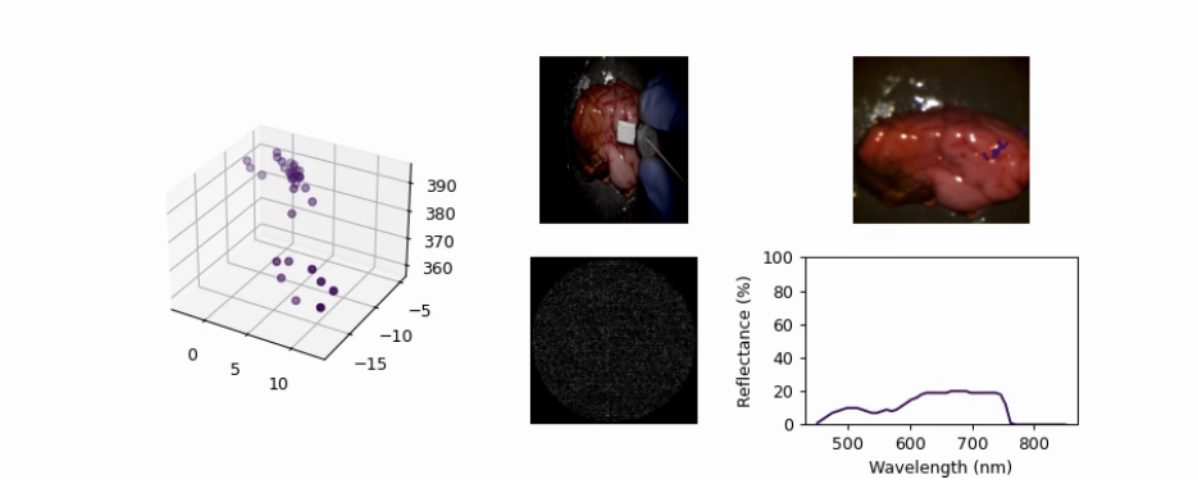
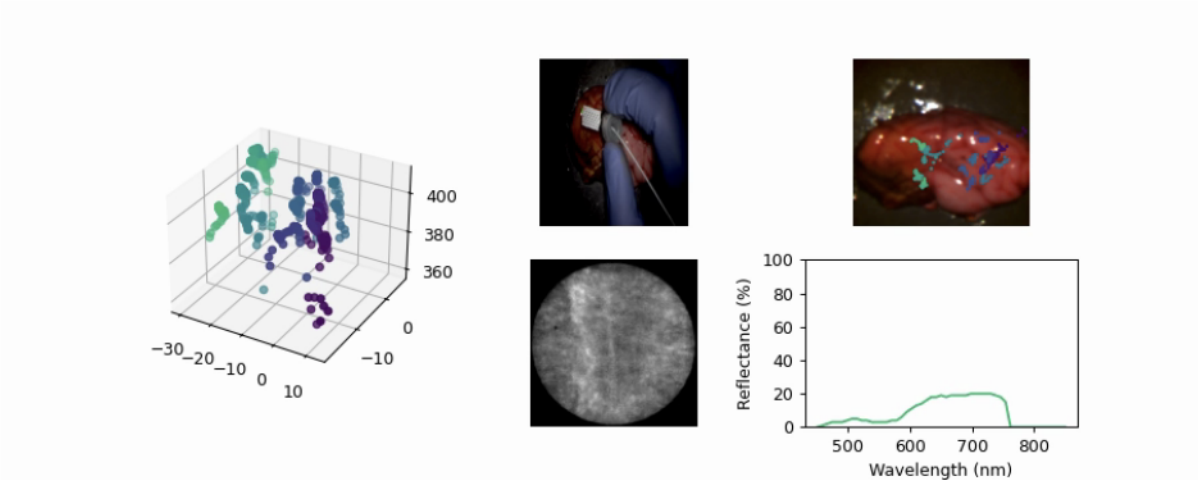


Figure 6.15: A lamb brain specimen with one half stained with acriflavine. The absorption between 420–500 nm matches the acriflavine absorption spectrum shown in the bottom left.



(a) Still frame of video, pCLE scanning side with no dye.



(b) Still frame of video, pCLE scanning side with dye.

Figure 6.16: Live demonstration on a lamb brain specimen, half-stained with acriflavine. The 3D axis (left) shows probe position in world coordinates. The grid shows four synchronized views: top-left the live RGB feed, top-right the pre-scan HSI with probe position, bottom-left the live pCLE image, and bottom-right the spectral data from the probe tip. Colours are consistent across the world coordinates, spectra, and HSI view.

6.6.5 Tumour Identification

Table 6.4: Performance of HSI spectral classification, presenting the Dice Similarity Coefficient or Sørensen–Dice coefficient (DICE) coefficient for each class and the average (Avg) score. The table also includes the processing speed in frames per second (FPS) for both the SVM and Neural Network (NN) models [160]. Permission for the reuse of this Table is detailed in section A.0.2.

	DICE ↑				
Model	**BG**	**Healthy**	**Tumour**	**Avg**	**FPS ↑**
SVM	98.0 ± 1.4	95.7 ± 1.8	55.5 ± 17.6	83.0 ± 6.9	0.014
NN	98.2 ± 0.9	94.7 ± 2.1	53.1 ± 15.0	82.0 ± 5.7	1.47

Table 6.5: Results for the binary classification of tumour tissue using pCLE, detailing the macro accuracy and F1-score. Processing speed in frames per second (FPS) is also provided [160]. Permission for the reuse of this Table is detailed in section A.0.2.

Model	**Accuracy ↑**	**F1 ↑**	**FPS ↑**
EfficientNet	95.7 ± 2.3	95.2 ± 2.1	39.1 ± 2.8
MobileNetV2	95.6 ± 2.0	95.1 ± 1.8	220.5 ± 19.5

Based on the pixel size estimation from section 6.4, the observed probe tip error of 1.2 pixels translates to a physical registration error of approximately 0.14 mm. This result successfully meets our target of achieving sub-millimetre error, as previously stated in section 6.4. Such precision enables accurate margin evaluation and dependable tumour classification, given that the registration error is considerably smaller than the anatomical structures under examination.

The two HSI models, SVM and NN, which were introduced in section 6.5, were evaluated. The results, presented in table 6.4, indicate that both models yielded comparable DICE scores, a finding consistent with existing literature. However, the neural network exhibited an inference speed almost an order of magnitude faster than the SVM. This substantial speed advantage made the NN the clear choice for spectral classification, as it satisfies the real-time requirements of a surgical setting. Consequently, the NN model was selected for integration into our combined algorithm. A similar evaluation was conducted for the pCLE modality. As detailed in table 6.5, the EfficientNet and MobileNetV2 models achieved similar classification performance. Nevertheless, with an inference speed approximately six times greater, MobileNetV2 was chosen for the final combined algorithm.

The constituent components of the combined algorithm, described in section 6.5, were evaluated individually. Their performance, quantified by DICE and recall, is summarised in table 6.6, with particular emphasis on tumour segmentation. The HSI-only model achieved a strong average DICE score and high tumour recall, indicating its ability to reliably identify tumour regions. In contrast, the pCLE-only model attained a higher DICE score for the tumour class but suffered from a marked reduction in recall, leading to missed tumour regions. Owing to its probe-based nature, pCLE can

Table 6.6: An overview of the results for the multimodality algorithm, presenting DICE and Recall metrics. Performance is shown for the individual modalities as well as their combination. In the table, BG denotes the background class, while Avg signifies the average performance across all classes [160]. Permission for the reuse of this figure is detailed in section A.0.2.

	BG	Healthy	Tumour	Avg
DICE ↑				
HSI	98.2 ± 0.9	94.7 ± 2.1	53.1 ± 15.0	82.0 ± 5.7
pCLE	-	-	70.5 ± 5.0	-
HSI LCC	98.2 ± 0.9	95.2 ± 1.9	52.6 ± 30.5	82.0 ± 11.1
HSI+pCLE	98.2 ± 0.9	95.7 ± 0.7	71.7 ± 2.0	88.5 ± 0.7
Recall ↑				
HSI	97.3 ± 1.6	94.5 ± 2.9	82.8 ± 8.3	91.5 ± 3.1
pCLE	-	-	68.8 ± 12.7	-
HSI LCC	97.3 ± 1.6	96.1 ± 1.8	61.9 ± 36.7	85.1 ± 13.2
HSI+pCLE	97.3 ± 1.6	96.1 ± 1.5	94.6 ± 0.9	96.0 ± 0.7

only be applied to the tumour class, as exhaustive sampling of every pixel across the full image is not feasible. An example of the pCLE sampling process is illustrated in fig. 6.17.

In an attempt to reduce false positives from the HSI model and thereby improve its tumour-class DICE score, a filtering step was applied that retained only the largest connected component, aiming to eliminate over-predicted regions. However, this approach did not yield any metric improvements. The standard deviation for this model was nearly double that of the HSI-only model, which suggests that the largest predicted tumour area often did not align with the ground truth.

The combined HSI+pCLE approach yielded substantial performance gains. Specifically, it achieved a markedly higher DICE score for the tumour class compared to the HSI-only model (an increase of 18.6) and, importantly, a significantly improved tumour recall relative to the pCLE-only model (an increase of 25.8). These results highlight the advantages of multimodal fusion in enhancing tumour localisation and reducing the risk of missing critical tumour regions during segmentation. As illustrated in fig. 6.18, the HSI-based prediction initially identifies the tumour along with several false positives. The subsequent pCLE scan, focused on the largest candidate region, enables the combined model to suppress irrelevant areas and refine the segmentation into an accurate and well-defined tumour margin.

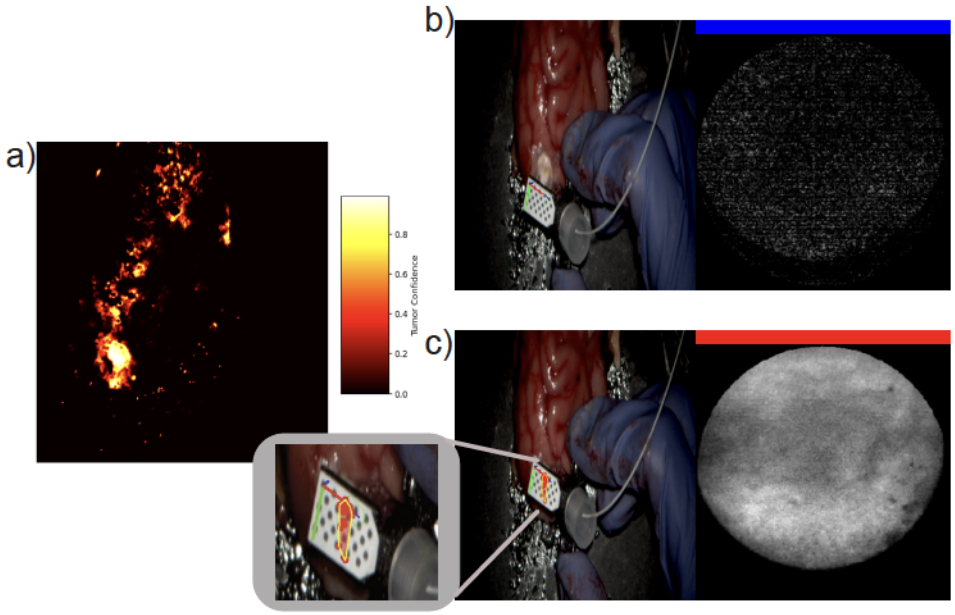


Figure 6.17: Example results during procedure. a) The macro tumour prediction Softmax logits, giving an indication of tumour location. b) Still frame from video at beginning of scanning. Blue bar on the pCLE frame indicates non-tumour classification c) Still frame at end of scanning with convex hull of pCLE points which are classified as positive. Red bar on pCLE frame indiciates tumour classification

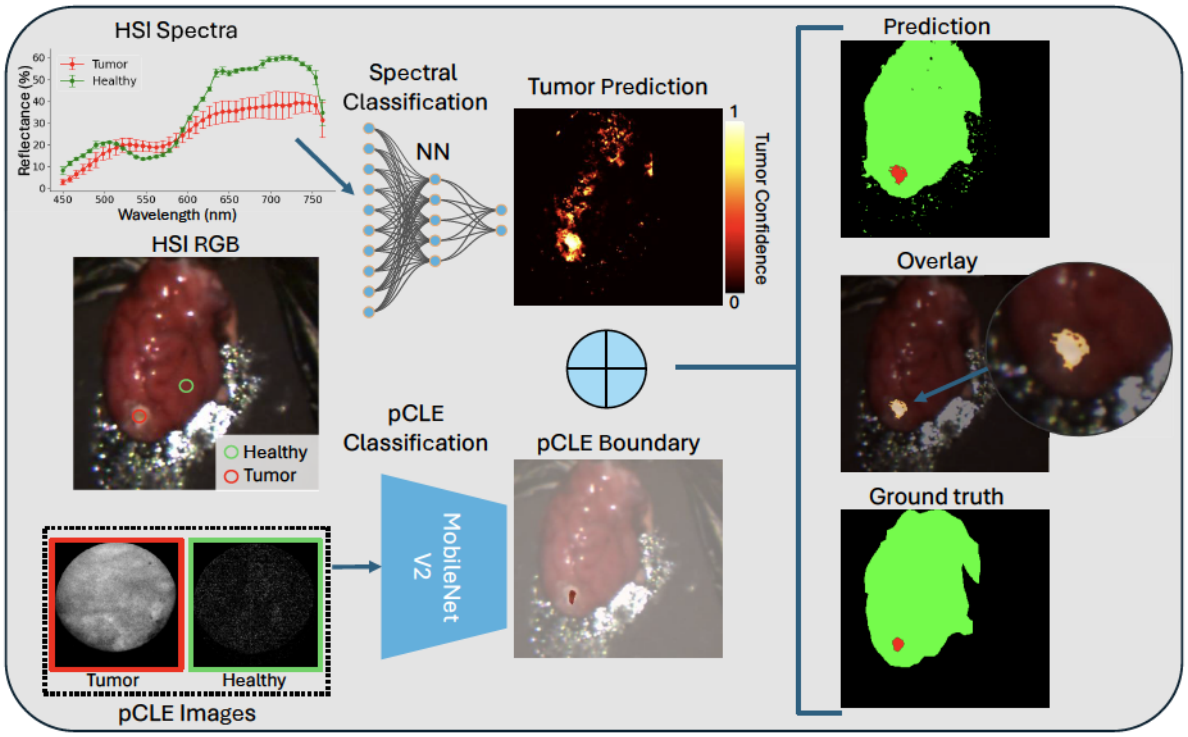


Figure 6.18: An illustrative overview of the multimodal algorithm for tumour identification. The figure depicts examples of healthy (green) and tumour (red) classifications, with the HSI modality results shown in the upper part and the pCLE results in the lower part. Fusion of the two modalities is achieved through a logical OR operation, followed by a component filtering stage [160]. Permission for the reuse of this figure is detailed in section A.0.2.

6.7 Discussion & Conclusion

This work has presented the development of a multimodal imaging platform that combines HSI and pCLE, marking an initial step towards improving tumour localisation in simulated surgical settings. By integrating these two modalities, we have shown that complementary information can be captured: HSI provides a wide-field overview of tissue properties, while pCLE delivers cellular-level detail at targeted sites. Together, they offer a framework for more precise tumour margin assessment.

Importantly, this work should be viewed as a foundational study rather than a complete clinical solution. The primary contribution lies in demonstrating the feasibility and potential value of tightly integrating HSI with high-resolution pCLE within a single platform. This establishes a flexible basis upon which more advanced sensing, learning, and interaction strategies can be developed in future work.

Our experiments demonstrated that while pCLE has the potential to refine tumour boundaries, the current system is constrained by the lack of a feedback mechanism linking the modalities. HSI predictions are currently static, with pCLE points acquired during or afterward. This highlights a disconnect even when imaging from the same viewpoint. A feedback loop where sampled pCLE points update the HSI map in real time would allow surgeons to see an evolving representation of tissue state, while HSI predictions could in turn guide pCLE sampling. This bidirectional integration is expected to substantially improve the accuracy of tumour margin identification.

Beyond technical integration, this work opens several broader research directions. The proposed platform naturally lends itself to the incorporation of learning-based decision support, uncertainty estimation, and adaptive sampling strategies, which could further reduce cognitive load on the surgeon. Additionally, the modular design allows future extensions to alternative optical modalities or foundation models, supporting an open and extensible research framework rather than a task-specific system.

Future work will focus on three directions. First, we will implement this feedback loop to ensure that pCLE measurements can refine the HSI view in real time while remaining resilient to occlusions such as surgical instruments or the surgeon's hand. Second, we will develop flexible, markerless pose estimation for the pCLE probe, simplifying the workflow while preserving accurate spatial tracking. Third, we will expand the dataset to animal cases with real tumours, enabling algorithmic refinement and improving the robustness of tumour boundary detection. Collectively, these advances will strengthen the reliability of the platform and accelerate its path toward clinical translation, where seamless integration, accuracy, and usability are essential for adoption.

Chapter 7

Conclusions

This final chapter draws together the key findings from each preceding chapter, highlighting how they contribute to the central aims of the thesis. It also reflects on the specific limitations of each study and outlines directions for future work to build on these findings.

7.1 Main Conclusions

The main contributions are then laid out in summary below then discussed further afterwards.

Probe-Based Confocal Laser Endomicroscopy (pCLE) for Intraoperative Tissue Characterisation

A Demonstrated that deep learning methods are highly capable of classifying pCLE images for intraoperative tumour identification.

B Highlighted the clinical relevance of balancing accuracy, speed, and interpretability, and explored risk-aware explainability.

C Introduced the first framework that combines risk estimation with Pixel Attribution (PA) methods, enabling both uncertainty quantification and explainability.

D Achieved improved ADCC performance over state-of-the-art non-perturbation PA methods, while also providing risk estimates for attribution outputs.

Hyperspectral Segmentation and Augmentation

A Established hyperspectral segmentation as a valid and clinically relevant approach for application in neurosurgery.

B Proposed a novel augmentation strategy, *Calibration-Jitter*, for hyperspectral segmentation tasks.

C Demonstrated that *Calibration-Jitter* regularises overconfident predictions and mitigates minor anomalies, leading to visually improved segmentations.

D Validated the effectiveness of the approach through quantitative results, surpassing the previous state-of-the-art Human Imaging Benchmark (Hib) benchmark.

E Established the potential of *Calibration-Jitter* as a practical and impactful augmentation method for improving hyperspectral image analysis.

Interactive Hyperspectral Segmentation

A Introduced Segment Anything with Spectral Angles (SAMSA), a novel interactive framework that integrates Hyperspectral Imaging (HSI) data with the spatial reasoning of a pretrained RGB foundation model Segment Anything-2 (SAM2).

B Demonstrated that combining spectral embeddings with SAM2's contextual and boundary-aware features overcomes the limitations of using either modality alone.

C Established the framework's ability to train on and generalise across heterogeneous HSI datasets with varying spectral and spatial resolutions, a key advance over traditional HSI models limited to specific imaging configurations.

D Identified the clinical value of HSI integration in complex or ambiguous segmentation tasks, where RGB-only models are prone to errors or misleading correlations.

Interactive Hyperspectral Segmentation Extended

A Introduced Segment Anything with Spectral Angles 2.0 (SAMSA2.0), an extension of the framework that embeds spectral angle information as a spatially-aware input prompt, enabling early fusion of hyperspectral and spatial cues.

B Demonstrated that this strategy improves segmentation accuracy in low-data and noisy conditions, outperforming both spectral-only and RGB-only baselines with minimal user input.

C Simplified the automatic mask generation pipelines to generate clinically relevant masks.

Multimodal Imaging Platform

A Developed a multimodal imaging platform that integrates HSI and pCLE, representing an initial step toward improved tumour localisation in simulated surgical settings.

B Demonstrated that the two modalities provide complementary information: HSI offers wide-field tissue characterisation, while pCLE delivers targeted cellular-level detail, together enabling more precise tumour margin assessment.

C Outlined future development priorities, including (i) implementation of a real-time feedback loop, (ii) markerless pose estimation for pCLE probe tracking, and (iii) expansion to clinical trials to improve algorithm robustness and clinical reliability.

Defining tumour margins during neurosurgery remains a significant challenge, principally because of the inherent difficulty in clearly distinguishing healthy tissue from tumour within the operative field. A key complication arises when attempting to extrapolate from small, localised biopsy data to guide decisions across the much larger surgical area, particularly when the information is derived from disparate imaging modalities. This thesis has confronted this issue by developing and refining a multimodal strategy, focusing on technologies that provide both cellular-level detail and wide-field context to create a more comprehensive intraoperative map.

The foundation of precise tumour delineation lies in accurate local assessment. To this end, chapter 2 focused on the advanced utility of pCLE, a technique that facilitates rapid 'optical biopsies' at the cellular level. It was demonstrated that deep learning methods are highly effective for classifying pCLE images for intraoperative tumour identification. More significantly, to address the clinical need for trustworthy AI, this research introduced the first framework that combines risk estimation with PA methods. This novel technique provides essential uncertainty quantification and explainability, balancing accuracy with interpretability and thereby strengthening the potential of pCLE to one day supersede frozen section analysis.

While pCLE provides critical microscopic detail, guiding surgery requires a macroscopic view of the entire resection cavity. Chapter 3 established HSI segmentation as a valid and clinically relevant approach for this wide-field characterisation. To overcome primary technical hurdles such as optical calibration, a novel augmentation strategy, *Calibration-Jitter*, was proposed. This method was shown to regularise overconfident predictions and mitigate minor anomalies, leading to visually improved segmentations and quantitatively surpassing the previous state-of-the-art Hib benchmark. This work solidifies HSI as a more robust tool for macro-scale tissue assessment.

Pivoting away from semantic segmentation in low data formats, we introduce SAMSA in chapter 4, a pioneering interactive framework that integrates HSI with the spatial reasoning of a pre-trained RGB foundation model. By leveraging rich spectral information alongside contextual, boundary-aware features, SAMSA bridges wide-field mapping and tissue analysis to enable an interactive 'software optical biopsy': with a single pixel click, the system identifies and segments regions of similar tissue across the image. This integration overcomes the limitations of using either RGB or HSI in isolation, and its ability to train on and generalize across heterogeneous datasets marks a critical advance over traditional HSI semantic segmentation models, offering a flexible and powerful tool for interactive segmentation in complex surgical scenes. In chapter 5, we extend this framework with SAMSA2.0. Here, we relax the reliance on hard constraints, introducing soft constraints, spectral angle prompting, that provide greater flexibility and robustness. Moreover, by modifying the underlying methodology of SAM2, SAMSA2.0 enables not only reduced memory consumption in training but also automatic segmentation for medical objects, further broadening its applicability in surgical workflows.

The ultimate promise of this research is realised in the synergy of these components, culminating in the development of a multimodal imaging platform that physically integrates HSI and pCLE, as described in chapter 6, representing a foundational step toward enhanced tumour localisation. This platform empirically confirms that the two modalities provide complementary information: HSI for the wide-field survey and pCLE for targeted, cellular-level validation. When combined, as demonstrated by the contributions within this thesis, these technologies offer complementary insights, and our work establishes groundwork for further research using these modalities.

7.2 Limitations

While this thesis presents several methodological advances in Probe-Based Confocal Laser Endomicroscopy (pCLE) and Hyperspectral Imaging (HSI) for medical applications, several important limitations must be acknowledged.

Limited Clinical Validation and Adoption Barriers. A fundamental limitation across all proposed methods is the absence of prospective clinical validation. Our risk-aware probe-based confocal laser endomicroscopy framework (chapter 2), while technically sound, has not been evaluated in actual surgical environments where factors such as time pressure, sterile field constraints, and surgeon workflow integration would significantly impact its utility. The clinical demand for explainability and uncertainty visualisation remains questionable, particularly given surgeons' established preference for speed and accuracy over interpretability in routine procedures. The additional cognitive load imposed by risk visualisations may prove counterproductive in practice, especially when weighed against the rapid deployment of high-performing foundation models like Segment Anything-2 (SAM2) that may render post-hoc explainability obsolete. We believe this technique's advantage remains in pre-deployment, since a surgeon and data scientist assessing a model's likelihood of finding salient features over spurious correlations is probably more useful than real-time explainability.

Dataset and Generalisation Constraints. While our work demonstrates the potential of the Calibration-Jitter augmentation technique (chapter 3) and Segment Anything with Spectral Angles (SAMSA) frameworks (chapter 4-chapter 5), evaluation has so far been limited to a relatively narrow set of HSI datasets, focused on specific anatomical regions and imaging conditions. The heterogeneous dataset training approach shows promise for improving generalisation, but training semantic segmentation models solely on limited datasets is not sufficient for immediate clinical deployment. Broader testing across diverse medical imaging scenarios, tissue types, and pathological conditions—and rigorous clinical validation—will be possible as more HSI clinical datasets become available, further supporting reliable translation to real-world applications.

Interactive Segmentation Dependencies. The SAMSA and Segment Anything with Spectral Angles 2.0 (SAMSA2.0) frameworks fundamentally rely on user interaction, creating a potential bottleneck in clinical workflows. While we demonstrate improved performance with minimal clicks, the requirement for expert annotation limits scalability and introduces operator-dependent variability. The transition from interactive to fully automatic segmentation remains unresolved, and our automatic mask generation pipelines often produce irregular boundaries and increased noise that may be clinically unacceptable. The assumption that clinicians will engage with interactive segmentation tools during time-critical procedures may be overly optimistic.

Technical Integration Challenges. The multimodal platform combining HSI and pCLE (chapter 6) exposes significant technical limitations that extend beyond the scope of this thesis. The absence

of real-time feedback mechanisms between modalities represents a critical gap that undermines the practical utility of the system. Current static prediction overlays and temporal disconnects between imaging modes create a fragmented user experience that would likely impede surgical decision-making. The requirement for marker-based pose estimation introduces additional complexity and potential failure points in sterile environments.

Limited Scope of Medical Applications. This work focuses primarily on tumour margin detection and segmentation, representing only a subset of potential HSI applications in surgery. The methods have not been evaluated for other critical surgical tasks such as perfusion assessment, nerve identification, or real-time tissue viability monitoring. The generalisability of our approaches to these broader surgical applications remains speculative.

Statistical and Methodological Limitations. Our quantitative evaluations, while showing promising improvements, are conducted on relatively small datasets with limited statistical power for detecting clinically meaningful differences. The lack of multi-centre validation, inter-observer agreement studies, and comparison with current clinical gold standards limits the strength of our conclusions. Furthermore, the metrics used (F1-score, Dice Similarity Coefficient or Sørensen–Dice coefficient (DICE)) may not adequately capture the clinical utility of the proposed methods, particularly regarding their impact on surgical outcomes and patient safety. In comparison, in the RGB imaging domain, foundation models are often trained on tens of thousands of diverse images, highlighting the current scale limitations of HSI datasets and the need for larger, clinically representative data to robustly validate these approaches.

7.3 Further Work

This research serves primarily as a proof of concept for our proposed medical device: a multimodal combination of Probe-Based Confocal Laser Endomicroscopy (pCLE) and Hyperspectral Imaging (HSI) for neurosurgery. Building on this foundation, several immediate and longer-term directions are clear. The following recommendations represent our assessment of the most critical next steps to advance this multimodal platform towards clinical viability.

Clinical validation. The first step should be preclinical validation using animal models. Conducting safe trials in mice bearing gliomas or meningiomas, followed by validation with histopathology, would provide more reliable estimates of the diagnostic utility of the proposed workflow.

Independent modality improvements. Advances in pCLE and HSI individually will directly strengthen their combined use. For pCLE, deployment-focused research is now critical. While intraoperative feasibility has been demonstrated, larger-scale trials with regulatory-grade machine learning models and devices are required. In addition, research into fluorescent dyes with higher cancer specificity remains an active and promising area. For HSI, clinical adoption is growing, but dataset limitations hinder competitiveness with large-scale RGB models. Since this work was completed, promising methods have been introduced for training across multiple HSI datasets [98], including spectral vector tokenisation and wavelength-informed position encoding. These approaches may unify HSI data sources into single models, though their true clinical potential depends on the scale of available data.

Towards seamless surgical integration. To be clinically viable, the system must function and remain robust to lighting variability. Achieving this robustness will require larger datasets for both HSI and pCLE. In parallel, moving away from rigid setups will be essential. HSI predictions must be registered and dynamically transformed using high-frame-rate RGB guidance to account for tissue deformation, with frequent updates conditioned on prior pCLE samples. Such a fluid, feedback-driven prediction system would greatly enhance intraoperative usability.

Outlook. Taken together, these directions chart a clear path forward: immediate validation, technical refinement, and integration into real-world surgical workflows. Improvements in both pCLE and HSI will reinforce the strength of their combination, while advances in multimodal learning may accelerate clinical adoption. Ultimately, this work lays the foundation for a new generation of multi-modal intraoperative imaging tools.

Wider Context. The specific recommendations outlined above must be considered within the context of the broader transformation that has occurred in medical imaging since the inception of this thesis work. The field has evolved dramatically from small-scale Deep Learning (DL) projects on limited datasets to foundational models providing comprehensive insight into surgical visualisation. This evolution raises important questions about the fundamental approaches we have taken

in this work.

Rethinking HSI validation and calibration. We have applied a biophotonic modality that requires extensive calibration procedures even under controlled laboratory conditions to "see the unseeable" in clinical contexts. However, the necessity of current calibration processes merits careful investigation. While we calibrate "true colour" for film cameras, this is not a prerequisite for RGB foundational models to achieve accurate segmentation. As established scientific principle dictates, without validation against ground truth, observed patterns remain merely correlational. Future work must focus not only on validating HSI with larger datasets, but crucially, with better-labelled ones. Current approaches that generate labels using RGB information arguably undermine the fundamental value proposition of HSI, as DL methods are ultimately constrained to emulating their training labels.

Accelerating pCLE clinical adoption. For pCLE, the path forward appears more defined, though the wider research community faces similar validation challenges. We must subject our models to the same rigorous clinical validation standards applied to medical devices themselves. Regulatory frameworks exist for sound reasons, and pCLE represents an established tool requiring deeper integration rather than fundamental redevelopment. The focus should therefore be on advancing clinical validation protocols and regulatory compliance to facilitate broader adoption of this promising technology.

Overall. Despite these challenges, the convergence of advanced imaging modalities with sophisticated machine learning approaches represents an opportunity to enhance surgical precision and patient outcomes. As foundational models continue to mature and datasets expand, the multimodal platform proposed in this work stands positioned to contribute meaningfully to the next generation of intelligent surgical assistance systems. The journey from laboratory proof-of-concept to clinical reality remains demanding, but the potential to transform neurosurgical practice makes this pursuit both scientifically compelling and clinically imperative.

References

[1] Abnar, S. and Zuidema, W. *Quantifying Attention Flow in Transformers*. May 2020. DOI: 10.48550/arxiv.2005.00928.

[2] Acción, Á., Argüello, F., and Heras, D. B. "Dual-Window Superpixel Data Augmentation for Hyperspectral Image Classification". In: *Applied Sciences* 10.24 (2020), p. 8833.

[3] Adebayo, J., Gilmer, J., Muelly, M., et al. "Sanity Checks for Saliency Maps". In: *arXiv preprint arXiv:1810.03292* (2018).

[4] Albor-Ramírez, E., Reyes-Alberto, M., Vidal-Flores, L., et al. "Agarose Gel Characterization for the Fabrication of Brain Tissue Phantoms for Infrared Multispectral Vision Systems". In: *Gels* 9.12 (Nov. 2023), p. 944. DOI: 10.3390/gels9120944.

[5] Amann, J., Blasimme, A., Vayena, E., et al. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC* 20.1 (Dec. 2020). ISSN: 1472-6947.

[6] Ancona, M., Ceolini, E., Öztireli, C., et al. *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*. Nov. 2017. DOI: 10.48550/arxiv.1711.06104.

[7] Anichini, G., Leiloglou, M., Hu, Z., et al. "Hyperspectral and multispectral imaging in neurosurgery: a systematic literature review and meta-analysis". In: *European Journal of Surgical Oncology* 51.1 (2025), p. 108293.

[8] Asfaw, Z., Young, T., Brown, C., et al. "Charting the success of neuronavigation in brain tumor surgery: from inception to adoption and evolution". In: *Journal of Neuro-Oncology* 170.1 (Oct. 2024), pp. 1–10. DOI: 10.1007/s11060-024-04778-0.

[9] Ashraf, M., Choudhary, N., Hussain, S., et al. "Role of intraoperative computed tomography scanner in modern neurosurgery - An early experience". In: *Surgical Neurology International* 11 (2020), p. 247. DOI: 10.25259/SNI_303_2020.

[10] Audebert, N., Le Saux, B., and Lefèvre, S. "Deep learning for classification of hyperspectral data: A comparative review". In: *IEEE geoscience and remote sensing magazine* 7.2 (2019), pp. 159–173.

[11] Bach, S., Binder, A., Montavon, G., et al. *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation*. 2015. DOI: 10.1371/journal.pone.0130140.

[12] Bahl, A., Horgan, C. C., Janatka, M., et al. "Synthetic white balancing for intra-operative hyperspectral imaging". In: *Journal of Medical Imaging* 10.4 (2023), pp. 046001–046001.

[13] Banan, R. and Hartmann, C. "The new WHO 2016 classification of brain tumors—what neurosurgeons need to know". In: *Acta Neurochirurgica* 159.3 (2017), pp. 403–418. ISSN: 0942-0940. DOI: 10.1007/s00701-016-3062-3.

[14] Belykh, E., Cavallo, C., Gandhi, S., et al. "Utilization of intraoperative confocal laser endomicroscopy in brain tumor surgery". In: *J Neurosurg Sci* 62.6 (2018). Epub 2018 Aug 28, pp. 704–717. DOI: 10.23736/S0390-5616.18.04553-8.

[15] Belykh, E., Martirosyan, N. L., Yagmurlu, K., et al. "Intraoperative fluorescence imaging for personalized brain tumor resection: current state and future directions". In: *Frontiers in surgery* 3 (2016), p. 55.

[16] Beutler, B., Lee, J., Edminster, S., et al. "Intracranial meningioma: A review of recent and emerging data on the utility of preoperative imaging for management". In: *Journal of Neuroimaging* 34.5 (Sept. 2024), pp. 527–547. DOI: 10.1111/jon.13227.

[17] Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., et al. "Hyperspectral remote sensing data analysis and future challenges". In: *IEEE Geoscience and remote sensing magazine* 1.2 (2013), pp. 6–36.
[18] Blundell, C., Cornebise, J., Kavukcuoglu, K., et al. "Weight Uncertainty in Neural Network". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Bach, F. and Blei, D. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1613–1622.
[19] Boardman, J. "Spectral angle mapping: a rapid measure of spectral similarity". In: *AVIRIS. Delivered by Ingenta* (1993).
[20] Bondesson, D., Schneider, M. J., Silbernagel, E., et al. "Automated evaluation of probe-based confocal laser endomicroscopy in the lung". In: *PLoS One* 15.5 (2020), e0232847.
[21] Boppart, S. A., Brezinski, M. E., Pitris, C., et al. "Optical coherence tomography for neurosurgical imaging of human intracortical melanoma". In: *Neurosurgery* 43.4 (1998), pp. 834–841.
[22] Bradski, G. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).
[23] Breuskin, D., Szczygielski, J., Urbschat, S., et al. "Confocal Laser Endomicroscopy in Neurosurgery—An Alternative to Instantaneous Sections?" In: *World Neurosurgery* 100 (2017), pp. 180–185. ISSN: 1878-8750. DOI: https://doi.org/10.1016/j.wneu.2016.12.128.
[24] Byun, S.-Y. and Lee, W. *Recipro-CAM: Gradient-free reciprocal class activation map*. Sept. 2022. DOI: 10.48550/arxiv.2209.14074.
[25] Cano-Valdez, A. M. and Sevilla-Lizcano, D. B. "Pathological Classification of Brain Tumors". In: *Principles of Neuro-Oncology: Brain & Skull Base*. Ed. by Monroy-Sosa, A., Chakravarthi, S. S., Garza-Salazar, J. G. de la, et al. Cham: Springer International Publishing, 2021, pp. 75–105. ISBN: 978-3-030-54879-7. DOI: 10.1007/978-3-030-54879-7_5.
[26] Caron, M., Touvron, H., Misra, I., et al. *Emerging Properties in Self-Supervised Vision Transformers*. Apr. 2021.
[27] Cartucho, J., Wang, C., Huang, B., et al. "An enhanced marker pattern that achieves improved accuracy in surgical tool tracking". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 10.4 (2021), pp. 400–408. DOI: 10.1080/21681163.2021.1997647.
[28] Chakrabarti, A. and Zickler, T. "Statistics of real-world hyperspectral images". In: *CVPR 2011*. IEEE. 2011, pp. 193–200.
[29] Charalampaki, P., Nakamura, M., Athanasopoulos, D., et al. "Confocal-Assisted Multispectral Fluorescent Microscopy for Brain Tumor Surgery". In: *Front Oncol* 9 (2019), p. 583. DOI: 10.3389/fonc.2019.00583.
[30] Charalampaki, P., Javed, M., Daali, S., et al. "Confocal Laser Endomicroscopy for Real-time Histomorphological Diagnosis: Our Clinical Experience With 150 Brain and Spinal Tumor Cases". In: *Neurosurgery* 62 (2015). ISSN: 0148-396X.
[31] Chattopadhyay, A., Sarkar, A., Howlader, P., et al. *Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks*. Oct. 2017. DOI: 10.1109/WACV.2018.00097.
[32] Chefer, H., Gur, S., and Wolf, L. *Transformer Interpretability Beyond Attention Visualization*. Dec. 2020. DOI: 10.48550/arxiv.2012.09838.
[33] Chefer, H., Schwartz, I., and Wolf, L. *Optimizing Relevance Maps of Vision Transformers Improves Robustness*. June 2022. DOI: 10.48550/arxiv.2206.01161.
[34] Chen, W., Liao, D., Deng, Y., et al. "Development of a transformer-based deep learning algorithm for diabetic peripheral neuropathy classification using corneal confocal microscopy images". In: *Frontiers in Cell and Developmental Biology* 12 (2024), p. 1484329.

[35] Chiurillo, I., Sha, R. M., Robertson, F. C., et al. "High-Accuracy Neuro-Navigation with Computer Vision for Frameless Registration and Real-Time Tracking". In: *Bioengineering* 10.12 (2023). ISSN: 2306-5354. DOI: 10.3390/bioengineering10121401.

[36] Cho, S., Salinas, R., and Lee, J. "Indocyanine-Green for Fluorescence-Guided Surgery of Brain Tumors: Evidence, Techniques, and Practical Experience". In: *Frontiers in Surgery* 6 (2019), p. 11. DOI: 10.3389/fsurg.2019.00011.

[37] Clancy, N. T., Jones, G., Maier-Hein, L., et al. "Surgical spectral imaging". In: *Medical Image Analysis* 63 (2020), p. 101699. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101699.

[38] Corona, A., Di, L., Shah, A., et al. "Current experimental therapies for atypical and malignant meningiomas". In: *Journal of Neuro-Oncology* 153.2 (June 2021), pp. 203–210. DOI: 10.1007/s11060-021-03759-x.

[39] Cortes, C. and Vapnik, V. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[40] Cover, T. M. and Thomas, J. A. "Elements of Information Theory". In: *Elements of Information Theory* (Apr. 2005), pp. 1–748. DOI: 10.1002/047174882X.

[41] Cubuk, E. D., Zoph, B., Shlens, J., et al. "RandAugment: Practical Automated Data Augmentation With a Reduced Search Space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[42] D'Urso, P., Hall, B., Atkinson, R., et al. "Biomodel-guided stereotaxy". In: *Neurosurgery* 44.5 (May 1999), pp. 1084–1093. DOI: 10.1097/00006123-199905000-00080.

[43] Damianou, A. C. and Lawrence, N. D. *Deep Gaussian Processes*. Nov. 2012. DOI: 10.48550/arxiv.1211.0358.

[44] Daumas-Duport, C., Scheithauer, B., O'Fallon, J., et al. "Grading of astrocytomas: a simple and reproducible method". In: *Cancer* 62.10 (1988), pp. 2152–2165.

[45] Delaney, P. and Harris, M. "Fiberoptics in confocal microscopy". In: *Handbook of biological confocal microscopy*. Springer, 1995, pp. 515–523.

[46] Deng, J., Dong, W., Socher, R., et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[47] Desroches, J., Jermyn, M., Mok, K., et al. "Characterization of a Raman spectroscopy probe system for intraoperative brain tissue classification". In: *Biomed. Opt. Express* 6.7 (July 2015), pp. 2380–2397. DOI: 10.1364/BOE.6.002380.

[48] Devlin, J., Chang, M.-W., Lee, K., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Oct. 2018. DOI: 10.48550/arxiv.1810.04805.

[49] Devries, T. and Taylor, G. W. "Improved Regularization of Convolutional Neural Networks with Cutout". In: *arXiv preprint arXiv:1708.04552* (2017).

[50] Diprose, W. K., Buist, N., Hua, N., et al. "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator". In: *Journal of the American Medical Informatics Association* 27.4 (Apr. 2020). ISSN: 1527974X.

[51] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Oct. 2020.

[52] Duffau, H. "Acute functional reorganisation of the human motor cortex during resection of central lesions: a study using intraoperative brain mapping". In: *Journal of Neurology, Neurosurgery & Psychiatry* 70.4 (Apr. 2001), pp. 506–513. DOI: 10.1136/jnnp.70.4.506.

[53] Enyinna Nwankpa, C., Ijomah, W., Gachagan, A., et al. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. Nov. 2018. DOI: 10.48550/arxiv.1811.03378.

[54] Erives, H. and Fitzgerald, G. "Automatic subpixel registration for a tunable hyperspectral imaging system". In: *IEEE Geoscience and Remote Sensing Letters* 3.3 (2006), pp. 397–400. DOI: 10.1109/LGRS.2006.873346.

[55] Eschbacher, J., Martirosyan, N. L., Nakaji, P., et al. "In vivo intraoperative confocal microscopy for real-time histopathological imaging of brain tumors". In: *Journal of neurosurgery* 116.4 (2012), pp. 854–860.

[56] Fabelo, H., Ortega, S., Kabwama, S., et al. "HELICoiD project: A new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations". In: *Hyperspectral Imaging Sensors: Innovative Applications and Sensor Standards 2016*. Vol. 9860. SPIE. 2016, p. 986002.

[57] Falcon, W. and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. Mar. 2019. DOI: 10.5281/zenodo.3828935.

[58] Fernandez, F.-G. *TorchCAM: Class Activation Explorer*. https://github.com/frgfm/torch-cam. Accessed: 2025-07-04. 2020.

[59] Filler, A. "The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI". In: *Nature precedings* (2009), pp. 1–1.

[60] Fisher, R. A. "Statistical Methods for Research Workers". In: *Breakthroughs in Statistics: Methodology and Distribution*. Ed. by Kotz, S. and Johnson, N. L. New York, NY: Springer New York, 1992, pp. 66–70. ISBN: 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_6.

[61] Folgoc, L. L., Baltatzis, V., Desai, S., et al. *Is MC Dropout Bayesian?* Oct. 2021. DOI: 10.48550/arxiv.2110.04286.

[62] Forte, A. E., Galvan, S., Manieri, F., et al. "A composite hydrogel for brain tissue phantoms". In: *Materials & Design* 112 (2016), pp. 227–238. ISSN: 0264-1275. DOI: 10.1016/j.matdes.2016.09.063.

[63] Furlanetto, R. H., Moriwaki, T., Falcioni, R., et al. "Hyperspectral reflectance imaging to classify lettuce varieties by optimum selected wavelengths and linear discriminant analysis". In: *Remote Sensing Applications: Society and Environment* 20 (2020), p. 100400.

[64] Gajjar, K., Heppenstall, L. D., Pang, W., et al. "Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis". In: *Analytical Methods* 5.1 (2013), pp. 89–102.

[65] Gal, Y. "Uncertainty in Deep Learning". PhD thesis. University of Cambridge, 2016.

[66] Gal, Y. and Ghahramani, Z. *Dropout as a Bayesian Approximation: Appendix*. June 2015. DOI: 10.48550/arxiv.1506.02157.

[67] Gal, Y. and Ghahramani, Z. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. June 2015.

[68] Gandhi, S., Tayebi Meybodi, A., Belykh, E., et al. "Survival Outcomes Among Patients With High-Grade Glioma Treated With 5-Aminolevulinic Acid–Guided Surgery: A Systematic Review and Meta-Analysis". In: *Frontiers in Oncology* Volume 9 - 2019 (2019). ISSN: 2234-943X. DOI: 10.3389/fonc.2019.00620.

[69] Garifullin, A., Kööbi, P., Ylitepsa, P., et al. "Hyperspectral image segmentation of retinal vasculature, optic disc and macula". In: *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. 2018, pp. 1–5.

[70] Gerard, I., Kersten-Oertel, M., Hall, J., et al. "Brain Shift in Neuronavigation of Brain Tumors: An Updated Review of Intra-Operative Ultrasound Applications". In: *Frontiers in Oncology* 10 (2021), p. 618837. DOI: 10.3389/fonc.2020.618837.

[71] Ghatwary, N., Ahmed, A., Grisan, E., et al. "In-vivo Barrett's esophagus digital pathology stage classification through feature enhancement of confocal laser endomicroscopy". In: *Journal of Medical Imaging* 6.1 (2019), pp. 014502–014502.

[72] Gonzalez, R. C. and Woods, R. E. *Digital Image Processing*. Fourth edition, Global edition. New York, NY: Pearson, 2018.
[73] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.
[74] Goodfellow, I. J., Shlens, J., and Szegedy, C. "Explaining and Harnessing Adversarial Examples". In: *CoRR* abs/1412.6572 (2014).
[75] Graves, A. "Practical Variational Inference for Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Shawe-Taylor, J., Zemel, R., Bartlett, P., et al. Vol. 24. Curran Associates, Inc., 2011.
[76] Hagos, M. T., Curran, K. M., and Mac Namee, B. *Identifying Spurious Correlations and Correcting them with an Explanation-based Learning*. Nov. 2022. DOI: 10.48550/arxiv.2211.08285.
[77] Hao, Q., Li, S., and Kang, X. "Multilabel Sample Augmentation-Based Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.6 (2020), pp. 4263–4278.
[78] Haut, J. M., Paoletti, M. E., Plaza, J., et al. "Hyperspectral Image Classification Using Random Occlusion Data Augmentation". In: *IEEE Geoscience and Remote Sensing Letters* 16.11 (2019), pp. 1751–1755.
[79] He, K., Zhang, X., Ren, S., et al. *Deep Residual Learning for Image Recognition*. Dec. 2015. DOI: 10.48550/arxiv.1512.03385.
[80] He, K., Zhang, X., Ren, S., et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.
[81] Hendricks, B., Scherschinski, L., Jubran, J., et al. "Eloquent noneloquence: redefinition of cortical eloquence based on outcomes of superficial cerebral cavernous malformation resection". In: *Journal of Neurosurgery* 141.2 (Mar. 2024), pp. 291–305. DOI: 10.3171/2023.12.JNS232588.
[82] Hilton, D. A. and Hanemann, C. O. "Schwannomas and their pathogenesis". In: *Brain Pathol.* 24.3 (2014), pp. 205–220.
[83] Hinton, G. E. and Camp, D. van. *Keeping neural networks simple by minimizing the description length of the weights*. 1993. DOI: 10.1145/168304.168306.
[84] Hinton, G. E., Srivastava, N., Krizhevsky, A., et al. *Improving neural networks by preventing co-adaptation of feature detectors*. 2012. arXiv: 1207.0580 [cs.NE].
[85] Hughes, M. and Yang, G.-Z. "Line-scanning fiber bundle endomicroscopy with a virtual detector slit". In: *Biomed. Opt. Express* 7.6 (June 2016), pp. 2257–2268. DOI: 10.1364/BOE.7.002257.
[86] Iakubovskii, P. *Segmentation Models Pytorch*. 2019.
[87] International Agency for Research on Cancer. *IARC*. https://gco.iarc.fr/today/en. Accessed: 2025-04-10. 2022.
[88] Izadyyazdanabadi, M., Belykh, E., Mooney, M. A., et al. "Prospects for theranostics in neurosurgical imaging: empowering confocal laser endomicroscopy diagnostics via deep learning". In: *Frontiers in oncology* 8 (2018), p. 240.
[89] Jain, S. and Wallace, B. C. *Attention is not Explanation*. Feb. 2019. DOI: 10.48550/arxiv.1902.10186.
[90] Janocha, K. and Czarnecki, W. M. "On Loss Functions for Deep Neural Networks in Classification". In: *Schedae Informaticae* 25 (Feb. 2017), pp. 49–59. ISSN: 20838476. DOI: 10.48550/arxiv.1702.05659.

[91] Jermyn, M., Desroches, J., Aubertin, K., et al. "A review of Raman spectroscopy advances with an emphasis on clinical translation challenges in oncology". In: *Physics in Medicine & Biology* 61.23 (2016), R370.

[92] John, S. E., Lovell, T. J., Opie, N. L., et al. "The ovine motor cortex: A review of functional mapping and cytoarchitecture". In: *Neuroscience & Biobehavioral Reviews* 80 (2017), pp. 306–315. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2017.06.002.

[93] Johnson, J. I., Sudheimer, K. D., Davis, K. K., et al. *The sheep brain atlas*. https://brains.anatomy.msu.edu/brains/sheep/index.html. Accessed: 2025-01-02. 2024.

[94] Khan, S. A., Ujjan, B. U., Akhunzada, N. Z., et al. "Diagnostic accuracy of frozen section in detecting malignant brain tumors taking histopathology as gold standard." In: *The Professional Medical Journal* 27.09 (2020), pp. 1814–1817.

[95] Khoshnevisan, A. and Allahabadi, N. S. "Neuronavigation: principles, clinical applications and potential pitfalls". In: *Iranian journal of psychiatry* 7.2 (2012), p. 97.

[96] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[97] Kirillov, A., Mintun, E., Ravi, N., et al. "Segment Anything". In: *arXiv:2304.02643* (2023).

[98] Kong, W., Liu, B., Bi, X., et al. "HyperSL: A Spectral Foundation Model for Hyperspectral Image Interpretation". In: *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[99] Koprowski, R. *Processing of hyperspectral medical images*. en. Studies in Computational Intelligence. Cham, Switzerland: Springer International Publishing, July 2018.

[100] Kriss, T. C. and Kriss, V. M. "History of the operating microscope: from magnifying glass to microneurosurgery". In: *Neurosurgery* 42.4 (1998), pp. 899–907.

[101] Krizhevsky, A., Sutskever, I., and Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks*.

[102] LeCun, Y., Bottou, L., Bengio, Y., et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2323. ISSN: 00189219. DOI: 10.1109/5.726791.

[103] Leon, R., Fabelo, H., Ortega, S., et al. "Hyperspectral imaging benchmark based on machine learning for intraoperative brain tumour detection". In: *NPJ Precision Oncology* 7.1 (2023), p. 119.

[104] Li, P., Qian, R., Niu, C., et al. "Impact of intraoperative MRI-guided resection on resection and survival in patients with gliomas: a meta-analysis". In: *Current Medical Research and Opinion* 33.4 (Apr. 2017), pp. 621–630. DOI: 10.1080/03007995.2016.1275935.

[105] Li, T., Hui, H., Hu, C., et al. "Multiscale imaging of colitis in mice using confocal laser endomicroscopy, light-sheet fluorescence microscopy, and magnetic resonance imaging". In: *Journal of Biomedical Optics* 24.1 (2019), p. 016003. DOI: 10.1117/1.JBO.24.1.016003.

[106] Li, W., Chen, C., Zhang, M., et al. "Data Augmentation for Hyperspectral Image Classification With Deep CNN". In: *IEEE Geoscience and Remote Sensing Letters* 16.4 (2019), pp. 593–597.

[107] Li, Y., Charalampaki, P., Liu, Y., et al. "Context aware decision support in neurosurgical oncology based on an efficient classification of endomicroscopic data". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 1187–1199.

[108] Li, Y., Charalampaki, P., Liu, Y., et al. "Context aware decision support in neurosurgical oncology based on an efficient classification of endomicroscopic data". In: *Int J Comput Assist Radiol Surg* 13.8 (Sept. 2018), pp. 1187–1199. DOI: 10.1007/s11548-018-1806-7.

[109] Lin, J., Clancy, N. T., Qi, J., et al. "Dual-modality endoscopic probe for tissue surface shape reconstruction and hyperspectral imaging enabled by deep neural networks". In: *Medical Image Analysis* 48 (2018), pp. 162–176. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2018.06.004.

[110] Lin, M., Chen, Q., and Yan, S. *Network In Network*. Dec. 2013. DOI: 10.48550/arxiv.1312.4400.
[111] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2021). ISSN: 1099-4300. DOI: 10.3390/e23010018.
[112] Liu, Y., Li, H., Guo, Y., et al. *Rethinking Attention-Model Explainability through Faithfulness Violation Test*. Jan. 2022. DOI: 10.48550/arxiv.2201.12114.
[113] Loshchilov, I. and Hutter, F. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)*. 2019.
[114] Louis, D. N., Perry, A., Reifenberger, G., et al. "The 2016 World Health Organization Classification of tumors of the Central Nervous System: A summary". en. In: *Acta Neuropathol.* 131.6 (2016), pp. 803–820.
[115] Lu, H., Grygoryev, K., Bermingham, N., et al. "Combined autofluorescence and diffuse reflectance for brain tumour surgical guidance: initial ex vivo study results". In: *Biomedical Optics Express* 12.4 (2021), pp. 2432–2446.
[116] Lu, R. and Fei, B. "Medical hyperspectral imaging: a review". In: *Journal of Biomedical Optics* 19.1 (2014), p. 010901. DOI: 10.1117/1.JBO.19.1.010901.
[117] Ma, J., He, Y., Li, F., et al. "Segment Anything in Medical Images". In: *Nature Communications* 15 (2024), p. 654.
[118] Ma, J., He, Y., Li, F., et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (Jan. 2024), p. 654. ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z.
[119] MacCormac, O., Noonan, P., Janatka, M., et al. "Lightfield hyperspectral imaging in neuro-oncology surgery: an IDEAL 0 and 1 study". In: *Frontiers in Neuroscience* 17 (2023). ISSN: 1662-453X. DOI: 10.3389/fnins.2023.1239764.
[120] MacCormac, O., Noonan, P., Janatka, M., et al. "Lightfield hyperspectral imaging in neuro-oncology surgery: an IDEAL 0 and 1 study". In: *Frontiers in Neuroscience* 17 (2023), p. 1239764.
[121] Mackle, E., Shapey, J., Maneas, E., et al. "Patient-Specific Polyvinyl Alcohol Phantom Fabrication with Ultrasound and X-Ray Contrast for Brain Tumor Surgery Planning". In: *Journal of Visualized Experiments* 161 (July 2020), e61344. DOI: 10.3791/61344.
[122] Mahboob, S., McPhillips, R., Qiu, Z., et al. "Intraoperative Ultrasound-Guided Resection of Gliomas: A Meta-Analysis and Review of the Literature". In: *World Neurosurgery* 92 (Aug. 2016), pp. 255–263. DOI: 10.1016/j.wneu.2016.05.007.
[123] Mansouri, A., Lai, C., Scales, D., et al. "A phase II pilot randomized controlled trial to assess the feasibility of the "supra-marginal" surgical resection of malignant glioma (G-SUMIT: Glioma supra marginal incision trial) study protocol". In: *Pilot and Feasibility Studies* 8 (2022), p. 138. DOI: 10.1186/s40814-022-01104-1.
[124] Marchese, E., Della Pepa, G. M., La Rocca, G., et al. "Real-time Imaging for Intraoperative Ultrasound-Guided Doppler Sonography in Neurosurgery". In: *Operative Neurosurgery* 18.5 (2020), pp. 553–562. DOI: 10.1093/ons/opz215.
[125] Marguth, F. "The Effect of CT on Neurosurgery". In: *Cranial Computerized Tomography*. Ed. by Lanksch, W. and Kazner, E. Berlin, Heidelberg: Springer Berlin Heidelberg, 1976, pp. 453–455. ISBN: 978-3-642-66494-6. DOI: 10.1007/978-3-642-66494-6_60.
[126] Martín-Pérez, A., Villa, M., Rosa Olmeda, G., et al. "SLIMBRAIN database: A multimodal image database of in vivo human brains for tumour detection". In: *Scientific Data* 12.1 (2025), p. 836.
[127] Martinez, B., Leon, R., Fabelo, H., et al. "Most relevant spectral bands identification for brain cancer detection using hyperspectral imaging". In: *Sensors* 19.24 (2019), p. 5481.

[128] Mazurek, M., Kulesza, B., Stoma, F., et al. "Characteristics of fluorescent intraoperative dyes helpful in gross total resection of high-grade gliomas—A systematic review". In: *Diagnostics* 10.12 (2020), p. 1100.

[129] Mckechnie, T., Jahan, A., Tait, I., et al. "An Endoscopic System for the Early Detection of Cancers of the Gastrointestinal Tract". In: *Review of Scientific Instruments* 69 (June 1998), pp. 2521–2523. DOI: 10.1063/1.1148952.

[130] McLean, M. A. and Cross, J. J. "Magnetic resonance spectroscopy: principles and applications in neurosurgery". In: *British journal of neurosurgery* 23.1 (2009), pp. 5–13.

[131] Meneses, P. R. "Spectral Correlation Mapper (SCM): An Improvement on the Spectral Angle Mapper (SAM)". In: *Proceedings of the AVIRIS 2000 Workshop*. JPL Publication 00-18. Jet Propulsion Laboratory, California Institute of Technology, 2000.

[132] Moazzam, A. A., Wagle, N., and Zada, G. "Recent developments in chemotherapy for meningiomas: a review". In: *Neurosurgical Focus* 35.6 (2013), E18. DOI: 10.3171/2013.10.FOCUS13341.

[133] Molitch, M. E. "Diagnosis and treatment of pituitary adenomas: a review". In: *Jama* 317.5 (2017), pp. 516–524.

[134] Murali, A., Mascagni, P., Mutter, D., et al. *CycleSAM: One-Shot Surgical Scene Segmentation using Cycle-Consistent Feature Matching to Prompt SAM*. 2024. arXiv: 2407.06795 [cs.CV].

[135] Neal, R. M. "Monte Carlo Implementation". In: *Bayesian Learning for Neural Networks*. New York, NY: Springer New York, 1996, pp. 55–98. ISBN: 978-1-4612-0745-0. DOI: 10.1007/978-1-4612-0745-0_3.

[136] Nimsky, C., Ganslandt, O., Keller, B. V., et al. "Intraoperative high-field-strength MR imaging: implementation and experience in 200 patients". In: *Radiology* 233.1 (2004). Epub 2004 Aug 18, pp. 67–78. DOI: 10.1148/radiol.2331031352.

[137] Nitzsche, B., Frey, S., Collins, L. D., et al. "A stereotaxic, population-averaged T1w ovine brain atlas including cerebral morphology and tissue volumes". In: *Frontiers in Neuroanatomy* 9 (June 2015), p. 69. DOI: 10.3389/fnana.2015.00069.

[138] Omeiza, D., Speakman, S., Cintas, C., et al. *Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models*. Aug. 2019. DOI: 10.48550/arxiv.1908.01224.

[139] Ostrom, Q. T., Cioffi, G., Gittleman, H., et al. "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016". In: *Neuro-Oncology* 21.Supplement_5 (2019), pp. v1–v100. DOI: 10.1093/neuonc/noz150.

[140] Panasyuk, S. V., Yang, S., Faller, D. V., et al. "Medical hyperspectral imaging to facilitate residual tumor identification during surgery". In: *Cancer Biology & Therapy* 6.3 (2007). PMID: 17374984, pp. 439–446. DOI: 10.4161/cbt.6.3.4018. eprint: https://doi.org/10.4161/cbt.6.3.4018.

[141] Paoletti, M., Haut, J., Plaza, J., et al. "Deep learning classifiers for hyperspectral imaging: A review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), pp. 279–317. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2019.09.006.

[142] Paszke, A., Gross, S., Massa, F., et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[143] Paszke, A., Gross, S., Massa, F., et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[144] Paszke, A., Gross, S., Massa, F., et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[145] Pavlov, V., Meyronet, D., Meyer-Bisch, V., et al. "Intraoperative probe-based confocal laser endomicroscopy in surgery and stereotactic biopsy of low-grade and high-grade gliomas: a feasibility study in humans". In: *Neurosurgery* 79.4 (2016), pp. 604–612.

[146] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[147] Petsiuk, V., Das, A., and Saenko, K. *RISE: Randomized Input Sampling for Explanation of Black-box Models*. June 2018. DOI: 10.48550/arxiv.1806.07421.

[148] Pino, M., Imperato, A., Musca, I., et al. "New Hope in Brain Glioma Surgery: The Role of Intraoperative Ultrasound. A Review". In: *Brain Sciences* 8.11 (2018), p. 202. DOI: 10.3390/brainsci8110202.

[149] Pojskić, M., Bopp, M., Saß, B., et al. "Microscope-Based Augmented Reality with Intraoperative Computed Tomography-Based Navigation for Resection of Skull Base Meningiomas in Consecutive Series of 39 Patients". In: *Cancers (Basel)* 14.9 (2022), p. 2302. DOI: 10.3390/cancers14092302.

[150] Pollack, I. F., Agnihotri, S., and Broniscer, A. "Childhood brain tumors: current management, biological insights, and future directions". In: *Journal of Neurosurgery: Pediatrics* 23.3 (2019), pp. 261–273. DOI: 10.3171/2018.10.PEDS18377.

[151] Poppi, S., Cornia, M., Baraldi, L., et al. "Revisiting the Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2021, pp. 2299–2304.

[152] Prada, F., Del Bene, M., Mattei, L., et al. "Intraoperative navigated angiosonography for skull base tumor surgery". In: *World Neurosurgery* 82.6 (2014), e593–e598. DOI: 10.1016/j.wneu.2014.05.025.

[153] Prada, F., Perin, A., Martegani, A., et al. "Intraoperative Contrast-Enhanced Ultrasound for Brain Tumor Surgery". In: *Neurosurgery* 78.4 (2016), pp. 542–552. DOI: 10.1227/NEU.0000000000001173.

[154] Ramaswamy, V. and Taylor, M. D. "Medulloblastoma: From Myth to Molecular". In: *Journal of Clinical Oncology* 35.21 (2017). PMID: 28640708, pp. 2355–2363. DOI: 10.1200/JCO.2017.72.7842. eprint: https://ascopubs.org/doi/pdf/10.1200/JCO.2017.72.7842.

[155] Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning*. 2006.

[156] Ravi, N., Gabeur, V., Hu, Y.-T., et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv preprint arXiv:2408.00714* (2024).

[157] Riahi, F., Kiani, P., Golabbakhsh, A., et al. "Comparison of PET/CT and PET/MRI in central nervous system tumors, a narrative review". In: *International Journal of Physiology, Pathophysiology, and Pharmacology* 16.4 (2024), pp. 89–95. DOI: 10.62347/UMDS1961.

[158] Ribeiro, M. T., Singh, S., and Guestrin, C. *Why Should I Trust You?Explaining the Predictions of Any Classifier*. Feb. 2016.

[159] Roddan, A., Czempiel, T., Elson, D. S., et al. "Calibration-Jitter: Augmentation of hyperspectral data for improved surgical scene segmentation". In: *Healthcare Technology Letters* 11.6 (2024), pp. 345–354. DOI: https://doi.org/10.1049/htl2.12102. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/htl2.12102.

[160] Roddan, A., Czempiel, T., Xu, C., et al. "Multimodal imaging platform for enhanced tumor resection in neurosurgery: integrating hyperspectral and pCLE technologies". In: *International Journal of Computer Assisted Radiology and Surgery* (2025), pp. 1–10.

[161] Roddan, A., Czempiel, T., Xu, C., et al. "SAMSA: Segment Anything Model Enhanced with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Ed. by Gee, J. C.,

Alexander, D. C., Hong, J., et al. Cham: Springer Nature Switzerland, 2026, pp. 478–488. ISBN: 978-3-032-05114-1.

[162] Roddan, A., Xu, C., Ajlouni, S., et al. "Explainable Image Classification with Improved Trustworthiness for Tissue Characterisation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Greenspan, H., Madabhushi, A., Mousavi, P., et al. Cham: Springer Nature Switzerland, 2023, pp. 575–585. ISBN: 978-3-031-43895-0.

[163] Rogers, C., Jones, P., and Weinberg, J. "Intraoperative MRI for Brain Tumors". In: *Journal of Neuro-Oncology* 151.3 (Feb. 2021), pp. 479–490. DOI: 10.1007/s11060-020-03667-6.

[164] Romano, A., Ferrante, M., Cipriani, V., et al. "Role of magnetic resonance tractography in the preoperative planning and intraoperative assessment of patients with intra-axial brain tumours". In: *La radiologia medica* 112.6 (2007), pp. 906–920.

[165] Ronneberger, O., Fischer, P., and Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Navab, N., Hornegger, J., Wells, W. M., et al. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[166] Ronneberger, O., Fischer, P., and Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Navab, N., Hornegger, J., Wells, W. M., et al. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[167] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[168] Ryall, S., Zapotocky, M., Fukuoka, K., et al. "Integrated Molecular and Clinical Analysis of 1,000 Pediatric Low-Grade Gliomas". In: *Cancer Cell* 37.4 (2020), 569–583.e5. DOI: 10.1016/j.ccell.2020.03.011.

[169] Salama, G. R., Heier, L. A., Patel, P., et al. "Diffusion weighted/tensor imaging, functional MRI and perfusion weighted imaging in glioblastoma—foundations and future". In: *Frontiers in neurology* 8 (2018), p. 660.

[170] Salma, A. and Ammirati, M. "Real time parallel intraoperative integration of endoscopic, microscopic, and navigation images: a proof of concept based on laboratory dissections". In: *Journal of Neurological Surgery Part B: Skull Base* 73.01 (2012), pp. 036–041.

[171] Sancho, J., Villa, M., Chavarrías, M., et al. "SLIMBRAIN: Augmented reality real-time acquisition and processing system for hyperspectral classification mapping with depth information for in-vivo surgical procedures". In: *Journal of Systems Architecture* 140 (2023), p. 102893. ISSN: 1383-7621. DOI: https://doi.org/10.1016/j.sysarc.2023.102893.

[172] Sandler, M., Howard, A., Zhu, M., et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520.

[173] Sandler, M., Howard, A. G., Zhu, M., et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 4510–4520.

[174] Schupper, A. J., Rao, M., Mohammadi, N., et al. "Fluorescence-guided surgery: a review on timing and use in brain tumor surgery". In: *Frontiers in Neurology* 12 (2021), p. 682151.

[175] Seidlitz, S., Sellner, J., Odenthal, J., et al. "Robust deep learning-based semantic organ segmentation in hyperspectral images". In: *Medical Image Analysis* 80 (2022), p. 102488.

[176] Selvaraju, R. R., Cogswell, M., Das, A., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Con-*

ference on Computer Vision 2017-October (Dec. 2017), pp. 618–626. ISSN: 15505499. DOI: 10.1109/ICCV.2017.74.

[177] Serrano, S. and Smith, N. A. *Is Attention Interpretable?* 2019. arXiv: 1906.03731 [cs.CL].

[178] Shaikh, M. S., Jaferzadeh, K., Thörnberg, B., et al. "Calibration of a hyper-spectral imaging system using a low-cost reference". In: *Sensors* 21.11 (2021), p. 3738.

[179] Shapey, J., Xie, Y., Nabavi, E., et al. "Intraoperative multispectral and hyperspectral label-free imaging: A systematic review of in vivo clinical studies". In: *Journal of Biophotonics* 12.9 (Sept. 2019). Epub 2019 Apr 29, e201800455. DOI: 10.1002/jbio.201800455.

[180] Shorten, C. and Khoshgoftaar, T. M. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[181] Shuhaiber, J. H. "Augmented reality in surgery". In: *Archives of surgery* 139.2 (2004), pp. 170–174.

[182] Signoroni, A., Savardi, M., Baronio, A., et al. "Deep learning meets hyperspectral image analysis: A multidisciplinary review". In: *Journal of imaging* 5.5 (2019), p. 52.

[183] Simonyan, K. and Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Sept. 2014. DOI: 10.48550/arxiv.1409.1556.

[184] Studier-Fischer, A., Seidlitz, S., Sellner, J., et al. "Heiporspectral-the heidelberg porcine hyperspectral imaging dataset of 20 physiological organs". In: *Scientific Data* 10.1 (2023), p. 414.

[185] Stummer, W., Pichlmeier, U., Meinel, T., et al. "Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial". In: *Lancet Oncology* 7.5 (2006), pp. 392–401. DOI: 10.1016/S1470-2045(06)70665-9.

[186] Stummer, W., Novotny, A., Stepp, H., et al. "Fluorescence-guided resection of glioblastoma multiforme utilizing 5-ALA-induced porphyrins: a prospective study in 52 consecutive patients". In: *Journal of neurosurgery* 93.6 (2000), pp. 1003–1013.

[187] Stupp, R., Mason, W. P., Bent, M. J. van den, et al. "Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma". In: *New England Journal of Medicine* 352.10 (2005), pp. 987–996. DOI: 10.1056/NEJMoa043330.

[188] Tan, M. and Le, Q. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 6105–6114.

[189] Thawani, J. P., Singh, N., Pisapia, J. M., et al. "Three-Dimensional Printed Modeling of Diffuse Low-Grade Gliomas and Associated White Matter Tract Anatomy". In: *Neurosurgery* 80.4 (Apr. 2017), pp. 635–645. DOI: 10.1093/neuros/nyx009.

[190] Thomaßen, M. T., Köhler, H., Pfahl, A., et al. "In vivo evaluation of a hyperspectral imaging system for minimally invasive surgery (HSI-MIS)". In: *Surgical Endoscopy* 37.5 (2023), pp. 3691–3700. ISSN: 1432-2218. DOI: 10.1007/s00464-023-09874-2.

[191] Tirindelli, M., Eilers, C., Simson, W., et al. "Rethinking Ultrasound Augmentation: A Physics-Inspired Approach". In: *arXiv preprint arXiv:2105.02188* (2021).

[192] Trajanovski, S., Shan, C., Weijtmans, P. J., et al. "Tongue tumor detection in hyperspectral images using deep learning semantic segmentation". In: *IEEE transactions on biomedical engineering* 68.4 (2020), pp. 1330–1340.

[193] Ullah, Z., Roy, S., Muhammad, S., et al. "Fluorescence imaging-guided surgery: current status and future directions". In: *Biomater. Sci.* 12 (15 2024), pp. 3765–3804. DOI: 10.1039/D4BM00410H.

[194] Vapnik, V. N. "An overview of statistical learning theory". In: *IEEE Transactions on Neural Networks* 10.5 (1999), pp. 988–999. DOI: 10.1109/72.788640.

[195] Vaswani, A., Shazeer, N., Parmar, N., et al. *Attention Is All You Need*. June 2017. DOI: 10.48550/arxiv.1706.03762.

[196] Vidal-Flores, L., Reyes-Alberto, M., Albor-Ramírez, E., et al. "Fabrication and Characterization of Brain Tissue Phantoms Using Agarose Gels for Ultraviolet Vision Systems". In: *Gels* 10.8 (Sept. 2024), p. 540. DOI: 10.3390/gels10080540.

[197] Vivas-Buitrago, T., Domingo, R. A., Tripathi, S., et al. "Influence of supramarginal resection on survival outcomes after gross-total resection of IDH–wild-type glioblastoma". In: *Journal of Neurosurgery* 136.1 (2022), pp. 1–8. DOI: 10.3171/2020.10.JNS203366.

[198] Wach, J., Vychopen, M., Kühnapfel, A., et al. "A Systematic Review and Meta-Analysis of Supramarginal Resection versus Gross Total Resection in Glioblastoma: Can We Enhance Progression-Free Survival Time and Preserve Postoperative Safety?" In: *Cancers (Basel)* 15.6 (Mar. 2023), p. 1772. DOI: 10.3390/cancers15061772.

[199] Wadley, J., Dorward, N., Kitchen, N., et al. "Pre-operative planning and intra-operative guidance in modern neurosurgery: a review of 300 cases". In: *Annals of the Royal College of Surgeons of England* 81.4 (1999), p. 217.

[200] Wager, M., Rigoard, P., Bouyer, C., et al. "Operating environment for awake brain surgery–Choice of tests". In: *Neurochirurgie* 63.3 (2017), pp. 150–157. DOI: 10.1016/j.neuchi.2016.10.002.

[201] Wallace, M. B. and Fockens, P. "Probe-based confocal laser endomicroscopy". In: *Gastroenterology* 136.5 (2009), pp. 1509–1513.

[202] Wang, G., Li, W., Zuluaga, M. A., et al. "Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning". In: *IEEE Transactions on Medical Imaging* 37.7 (2018), pp. 1562–1573. DOI: 10.1109/TMI.2018.2791721.

[203] Wang, G., Zuluaga, M. A., Li, W., et al. "DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (July 2019), pp. 1559–1572. ISSN: 1939-3539. DOI: 10.1109/tpami.2018.2840695.

[204] Wang, H., Wang, Z., Du, M., et al. *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks*. Oct. 2019. DOI: 10.48550/arxiv.1910.01279.

[205] Wang, Z., Wang, J., Budd, C., et al. *Scribble-Based Interactive Segmentation of Medical Hyperspectral Images*. 2024. arXiv: 2408.02708 [eess.IV].

[206] Weld, A., Dixon, L., Anichini, G., et al. "A Method For Mimicking Tumour Tissue In Brain Ex-Vivo Ultrasound For Research Application And Clinical Training". In: *medRxiv* (2024). DOI: 10.1101/2024.09.30.24314649.

[207] Weller, M., Bent, M. van den, Preusser, M., et al. "EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood". In: *Nature reviews Clinical oncology* 18.3 (2021), pp. 170–186. DOI: 10.1038/s41571-020-00447-z.

[208] Weller, M., Bent, M. van den, Tonn, J. C., et al. "European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas". In: *The Lancet Oncology* 18.6 (2017), e315–e329. DOI: 10.1016/S1470-2045(17)30194-8.

[209] Wen, P. Y., Chang, S. M., Van den Bent, M. J., et al. "Response assessment in neuro-oncology clinical trials". In: *Journal of Clinical Oncology* 35.21 (2017), pp. 2439–2449. DOI: 10.1200/JCO.2017.72.7511.

[210] Whittle, I. R., Smith, C., Navoo, P., et al. "Meningiomas". In: *The Lancet* 363.9420 (2004), pp. 1535–1543.

[211] Wiegreffe, S. and Pinter, Y. *Attention is not not Explanation*. 2019. arXiv: 1908.04626 [cs.CL].

[212] Winkler-Schwartz, A., Yilmaz, R., Tran, D. H., et al. "Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery". In: *World Neurosurgery* 144 (Dec. 2020), e62–e71. DOI: 10.1016/j.wneu.2020.07.209.

[213] Wirsching, H.-G. and Weller, M. "Glioblastoma". In: *Malignant Brain Tumors: State-of-the-Art Treatment* (2016), pp. 265–288.

[214] Wolf, T., Debut, L., Sanh, V., et al. *Transformers: State-of-the-Art Natural Language Processing*. Version v4.10.2. Sept. 2021. DOI: 10.5281/zenodo.5500478.

[215] Wu, J., Wang, Z., Hong, M., et al. "Medical SAM adapter: Adapting segment anything model for medical image segmentation". In: *Medical Image Analysis* 102 (2025), p. 103547. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2025.103547.

[216] Xie, E., Wang, W., Yu, Z., et al. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers". In: *Advances in Neural Information Processing Systems*. Ed. by Ranzato, M., Beygelzimer, A., Dauphin, Y., et al. Vol. 34. Curran Associates, Inc., 2021, pp. 12077–12090.

[217] Xu, C., Roddan, A., Xu, H., et al. "FF-ViT: probe orientation regression for robot-assisted endomicroscopy tissue scanning". In: *International Journal of Computer Assisted Radiology and Surgery* 19.6 (2024), pp. 1137–1145.

[218] Yaniv, Z. "Which pivot calibration?" In: *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by III, R. J. W. and Yaniv, Z. R. Vol. 9415. International Society for Optics and Photonics. Orlando, Florida, USA: SPIE, 2015, p. 941527. DOI: 10.1117/12.2081348.

[219] Yun, S., Han, D., Oh, S. J., et al. "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[220] Zandi, R. "Sparse Coding for Data Augmentation of Hyperspectral Medical Images". PhD thesis. San Jose State University, 2021.

[221] Zeh, R., Sheikh, S., Xia, L., et al. "The second window ICG technique demonstrates a broad plateau period for near infrared fluorescence tumor contrast in glioblastoma". In: *PLOS ONE* 12.7 (July 2017), pp. 1–16. DOI: 10.1371/journal.pone.0182034.

[222] Zeiler, M. D. and Fergus, R. *Visualizing and Understanding Convolutional Networks*. Nov. 2013. DOI: 10.48550/arxiv.1311.2901.

[223] Zhang, J. and Xiong, H. "Brain Tissue Preparation, Sectioning, and Staining". In: *Current Laboratory Methods in Neuroscience Research*. Ed. by Xiong, H. and Gendelman, H. E. New York, NY: Springer New York, 2014, pp. 3–30. ISBN: 978-1-4614-8794-4. DOI: 10.1007/978-1-4614-8794-4_1.

[224] Zhao, F. and Xie, X. "An overview on interactive medical image segmentation". In: *The Annals of the BMVA* 2013.7 (Aug. 2013).

[225] Zhou, B., Khosla, A., Lapedriza, A., et al. *Learning Deep Features for Discriminative Localization*. 2015. arXiv: 1512.04150 [cs.CV].

[226] Zuo, S. and Yang, G.-Z. "Endomicroscopy for computer and robot assisted intervention". In: *IEEE Reviews in Biomedical Engineering* 10 (2017), pp. 12–25.

Appendices

Chapter A

Permission Documents

A.0.1 Explainable Image Classification with Improved Trustworthiness for Tissue Characterisation

SPRINGER NATURE LICENSE

TERMS AND CONDITIONS

Sep 08, 2025

This Agreement between Alfie Roddan ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number 6104280701632

License Date Sep 08, 2025

Licensed Content Publisher Springer Nature

Licensed Content Publication Springer eBook

Licensed Content Title Explainable Image Classification with Improved Trustworthiness for Tissue Characterisation

Licensed Content Author Alfie Roddan, Chi Xu, Serine Ajlouni et al

Licensed Content Date Jan 1, 2023

Type of Use Thesis/Dissertation

Requestor Type academic/university or research institute

Format print and electronic

Portion figures/tables/illustrations

Number of figures/tables/illustrations 4

Will you be translating? no

Circulation/distribution 1 - 29

Author of this Springer Nature content yes

Title of new work Computer Vision for Multimodal Surgical Guidance in Brain Tumour Resection

Institution name Imperial College London

Expected presentation date Jan 2026

Portions Figure 1, Figure 2, Figure 3 and Table 1.

The Requesting Person / Organization to Appear on the License Alfie Roddan

Requestor Location Mr. Alfie Roddan
Imperial College
London, SW72BX
United Kingdom

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

Grant and Scope of License The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

Reservation of Rights Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

Restrictions on use Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

You must not use any Licensed Material as part of any design or trademark.

Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

STM Permission Guidelines An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@

springernature.com or bookpermissions@springernature.com for these rights.

Duration of License Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project Reuse up to distribution or time period indicated in License

Reuse in a dissertation/thesis Lifetime of thesis

Reuse in a journal/magazine Lifetime of journal/magazine

Reuse in a book/textbook Lifetime of edition

Reuse on a website 1 year unless otherwise specified in the License. If you wish to reuse the content on your website for longer than 1 year, please make this clear in the 'additional information' field and the License will include a 'Special Term' to reflect your duration choice.

Reuse in a presentation/slide kit/poster Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.

Reuse in conference proceedings Lifetime of conference proceedings

Reuse in an annual report Lifetime of annual report

Reuse in training/CME materials Reuse up to distribution or time period indicated in License

Reuse in newsmedia Lifetime of newsmedia

Reuse in coursepack/classroom materials Reuse up to distribution and/or time period indicated in license

Acknowledgement The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

Reuse in a dissertation or thesis Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: 'Reproduced with permission from Springer Nature'.

License Fee You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

Warranty The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version). If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

Termination and Cancellation The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

General The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions: Version 1.5 - June 2025

A.0.2 Open Access Springer

Materials reproduced from open access articles published by Springer Nature are distributed under the Creative Commons Attribution (CC BY) license. The publisher provides the following statement regarding the terms of use:

> This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A.0.3 Wiley Open Access License

Materials reproduced from open access articles published by Wiley are distributed under the Creative Commons Attribution (CC BY) license. The publisher provides the following description of the license terms:

> This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

A.0.4 SAMSA: Segment Anything Model Enhanced with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation

Date: Sep 25, 2025

This Agreement between **Alfie Roddan, Imperial College London** ("You") and **Springer Nature** consists of your order details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Order Information

Order Number	502017424
Order Date	Sep 24, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	*SAMSA: Segment Anything Model Enhanced with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation*
Licensed Content Author	Alfie Roddan, Tobias Czempiel, Chi Xu et al
Licensed Content Date	Jan 1, 2026
Type of Use	Thesis/Dissertation
Requestor Type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Translation	no
Circulation/Distribution	1–29
Author of Licensed Content	yes
Title of New Work	*Computer Vision for Multimodal Surgical Guidance in Brain Tumour Resection*
Institution Name	Imperial College London
Expected Presentation Date	Jan 2026
Licensee Name	Alfie Roddan, Imperial College London
Requestor Location	Mr. Alfie Roddan, Imperial College, London SW7 2BX, United Kingdom

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your RightsLink constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By completing the transaction, you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

Grant and Scope of License

The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, worldwide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only.

You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws. A separate license may be required for additional use.

Reservation of Rights

Licensor reserves all rights not expressly granted under this License. Nothing in this License conveys ownership rights to you.

Restrictions on Use

- Minor formatting edits are allowed, provided the meaning is not altered.
- Cropping, adapting, or omitting material that changes meaning is prohibited.
- Licensed Material may not be used as part of a trademark.
- For Open Access Publications, reuse must include clear acknowledgment.

STM Permission Guidelines

Alternative scope of license may apply to signatories of the STM Permissions Guidelines. See `https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/`.

Duration of License

Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year (unless otherwise specified)
Reuse in a presentation/poster	Lifetime of the work

Acknowledgement

Permission must be acknowledged with:
"Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature."

For theses, the following acknowledgement is required:
"Reproduced with permission from Springer Nature."

Warranty

The Licensor warrants that it has rights to license reuse. Licensed Material is provided "as is" without further warranty. Liability for damages is excluded to the extent permitted by law.

Termination and Cancellation

The License is valid for the term specified. Licensor may terminate in case of breach or non-payment.

General

This License is governed by German law. Disputes shall be resolved in Heidelberg, Germany, as far as legally permissible.

Other Conditions

Version 1.5 – June 2025

Chapter B

Deep Learning Basics

B.0.1 The Neural Network

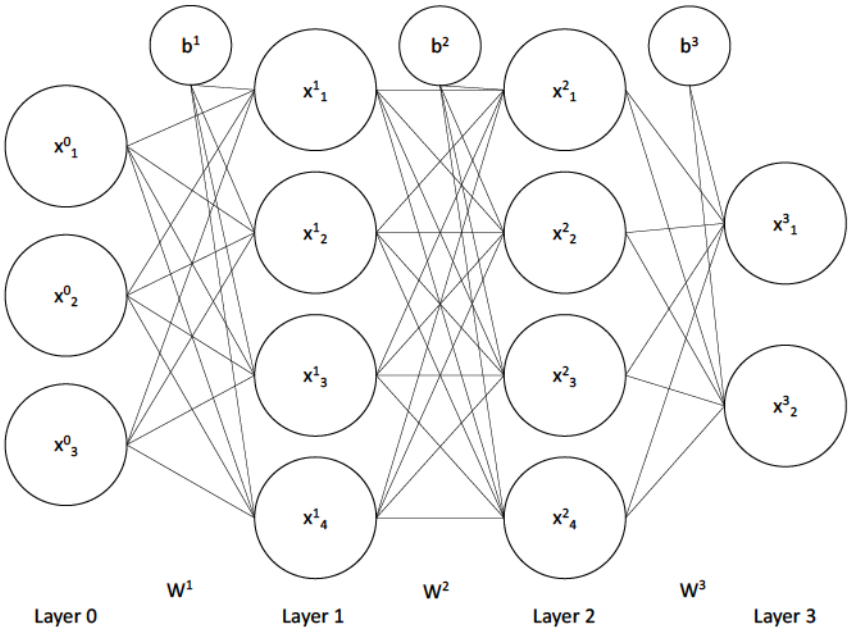


Figure B.1: Multi-layer perceptron (MLP), comprised of multiple layers of nodes. Each node has a weight associated. Each layer is comprised of these nodes, with a bias term for scaling. The activation functions between the layers have been omitted but should be present after each layer.

While an in-depth discussion lies beyond the scope of this work, it is worth briefly highlighting some foundational components of neural and deep learning networks that will become relevant in later sections. For a comprehensive explanation of neural networks and deep learning, I recommend Ian Goodfellow's Deep Learning book [73].

Figure B.1 presents a basic example of a multilayer perceptron (MLP) with input features of shape 3×1, two hidden layers (Layer 1 and 2), and an output layer of shape 2×1. A model of this structure could be applied to tasks such as multi-class or multi-label classification, or even multi-output regression. The output of the final layer, denoted as $\hat{y} = \Phi([x_1^3, x_2^3])$, represents the final

layer logits after being processed by the output activation function Φ. The output layer is calculated as follows:

$$\hat{y} = \Phi(W^L X^{L-1} + b^L), \tag{B.1}$$

where L denotes the final (output) layer, W is the weight matrix, X is the activation matrix from the previous layer (note that the activation functions used in the hidden layers may differ from the final layer's activation), b represents the bias terms, and Φ is the activation function. Activation functions are essential for introducing non-linearity in hidden layers. The computation of logits for each layer (except the input layer) involves the activation from the previous layer, weights, and bias terms, and is given by:

$$X^l = \phi(W^l X^{l-1} + b^l), \tag{B.2}$$

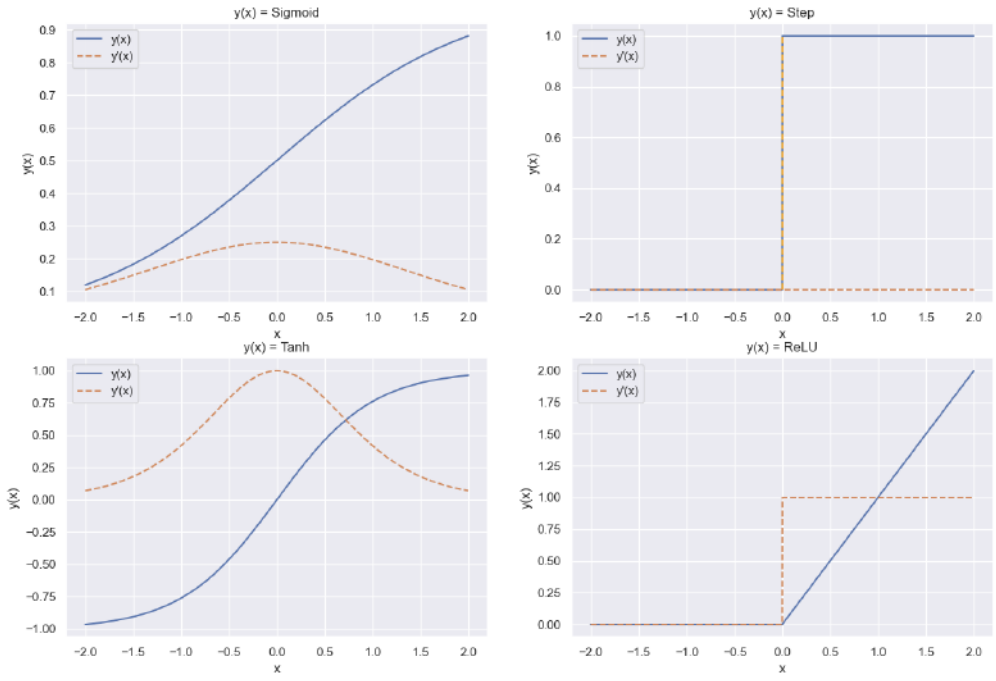
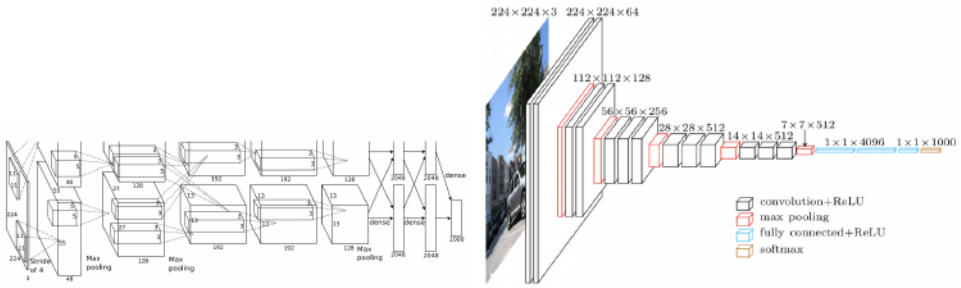


Figure B.2: Different activation functions and their derivatives over the domain $D : [-2, 2]$.

Here, X is the activation matrix of the previous layer and ϕ denotes the activation function for the hidden layers. The choice of the final layer's activation function depends on the network's task, while hidden layer activations serve primarily to introduce non-linearity [53]. Figure B.2 illustrates several widely used activation functions. Additionally, the activation function of the final layer is closely tied to the selected loss function or criterion, which forms the basis for optimization (typi-



(a) AlexNet [101] **(b)** VGG [183]

Figure B.3: Examples of convolutional neural network (CNN) architectures

cally via gradient descent) [90].

B.0.2 Convolutional Neural Networks (CNN)

An image can be represented as a matrix where the number of rows equals the image height, the number of columns corresponds to the image width, and the depth (third dimension) represents the number of channels (commonly red, green, and blue—RGB). These dimensions are integer values indicating the number of pixels along each axis. For instance, a 256×256 image contains 65,536 pixels per channel, totaling 196,608 pixels for an RGB image. This represents a substantial increase in complexity compared to the three-feature input used in Figure B.1. It is also worth noting that a 256×256 RGB image is relatively small by current standards.

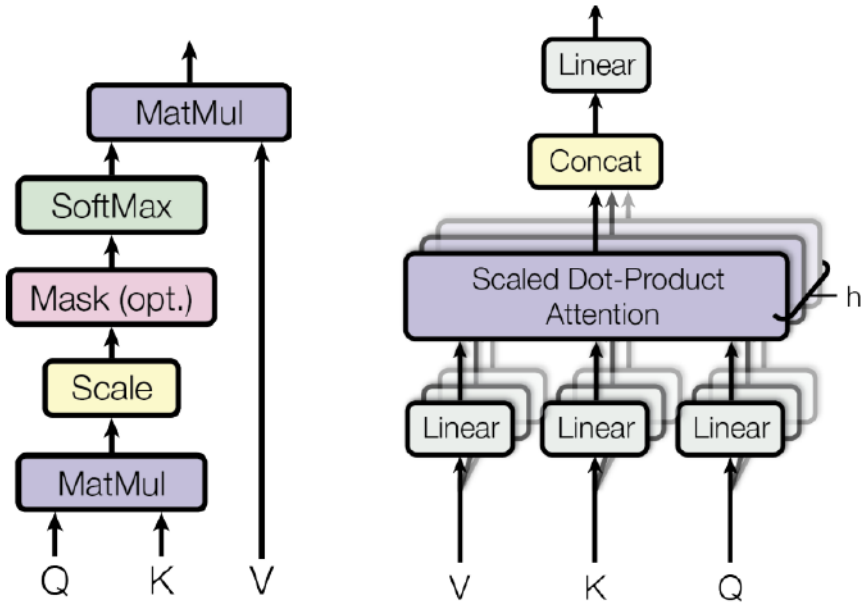
If we were to apply the same architecture as shown in Figure B.1 to image data, we would need to flatten the image, resulting in the loss of spatial relationships between adjacent pixels. For example, the first pixel in the flattened array might correspond to the pixel directly below it in the original image, which would now be separated in the feature vector by the image width. Additionally, this would only represent one channel; the pixel at position $1 + \text{Width} \times \text{Height}$ would correspond to the same location in real space but in a different color channel.

Convolutions address these issues effectively. Traditional neural networks are not well-equipped to handle the structural complexity of RGB image data, but convolutional layers dramatically reduce the computational demands of training. This efficiency is achieved through mechanisms such as parameter sharing, translational invariance, and sparse connectivity [102].

One of the earliest convolutional architectures was LeNet-5 [102], developed in 1989. Since then, CNNs have become a cornerstone of image-based deep learning tasks. These architectures have grown in both size and depth: AlexNet, introduced in 2012, had 60 million parameters and six layers [101]; VGG, released in 2014, featured 16 layers and between 100 and 138 million parameters [183]; and ResNet, introduced in 2015, scaled from 10 to 1,202 layers (ResNet10–ResNet1202) and had between 1.7 and 19.4 million parameters [79]. VGG emphasized increased depth through additional layers, while ResNet pushed both depth and receptive field size through wider architectures. These advancements were made possible in part by the rapid development of GPU hardware and software,

which enabled the training of such large models within practical timeframes.

B.0.3 Transformer and Attention



(a) Scaled Dot Product **(b)** Multihead Attention

Figure B.4: Left: Scaled dot product attention using query, key, and value vectors. Right: The multi-head attention framework [195].

The Transformer architecture was introduced by Google Brain in [195] and originally designed for machine translation in natural language processing (NLP). It marked a significant departure from recurrent models by using multi-head attention, resulting in state-of-the-art performance and faster training.

At its core, the scaled dot-product attention mechanism operates on three components: query $Q \in \mathbb{R}^{N \times D}$, key $K \in \mathbb{R}^{N \times D}$, and value $V \in \mathbb{R}^{N \times D}$. Here, N is the sequence length and D is the token dimensionality. Conceptually, this setup resembles a key-value lookup system where the query is matched against keys to retrieve the relevant values. The attention weights are computed as:

$$A = \text{softmax}_{\text{row}}\left(\frac{QK^T}{\sqrt{D}}\right) \tag{B.3}$$

Multiplying this attention matrix with the values gives the output:

$$SDP = AV, \tag{B.4}$$

as illustrated in Figure B.4 (a). The matrix A defines how each token relates to others in the se-

quence. This mechanism enables efficient modeling of long-range dependencies. The scaling term $\frac{1}{\sqrt{D}}$ stabilizes the variance of the dot product. Without scaling, the softmax could saturate, resulting in vanishing gradients. For example:

$$\begin{aligned} q_i &\sim \mathcal{N}(0, \sigma), \\ k_i &\sim \mathcal{N}(0, \sigma), \\ \text{Var}\left(\sum_{i=1}^{D} q_i k_i\right) &= \sigma \cdot D, \end{aligned} \tag{B.5}$$

which justifies the normalization.

The transformation from input X to query, key, and value vectors depends on whether self or cross-attention is used. In the original Transformer setup, cross-attention is applied for translation tasks. Given inputs X and X':

$$\begin{aligned} Q &= XW^{Q^T}, \\ K &= X'W^{K^T}, \\ V &= X'W^{V^T}, \end{aligned} \tag{B.6}$$

where W^{Q^T}, W^{K^T}, and W^{V^T} are learned projection matrices. If $X = X'$, the attention is self-attention—i.e., the tokens attend to each other within the same sequence. Since a single attention head may not be sufficient to capture diverse relationships, the model splits the computation into multiple heads, illustrated in Figure B.4 (b). The complete multi-head attention is defined as:

$$\text{Multihead}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h)W^O, \tag{B.7}$$

where:

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{B.8}$$

with $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D \times d_k}$ and $W^O \in \mathbb{R}^{h \cdot d_k \times d_{\text{out}}}$. Typically, $d_k = d_v = D/h$ to distribute dimensions evenly across heads. Each head can focus on different parts of the input or features, akin to filters in convolutional networks. The combined features are then passed to a feedforward network.

Notably, the attention mechanism is equivariant to token reordering, meaning output tokens are affected in the same way as their corresponding inputs under permutation. However, this lack of

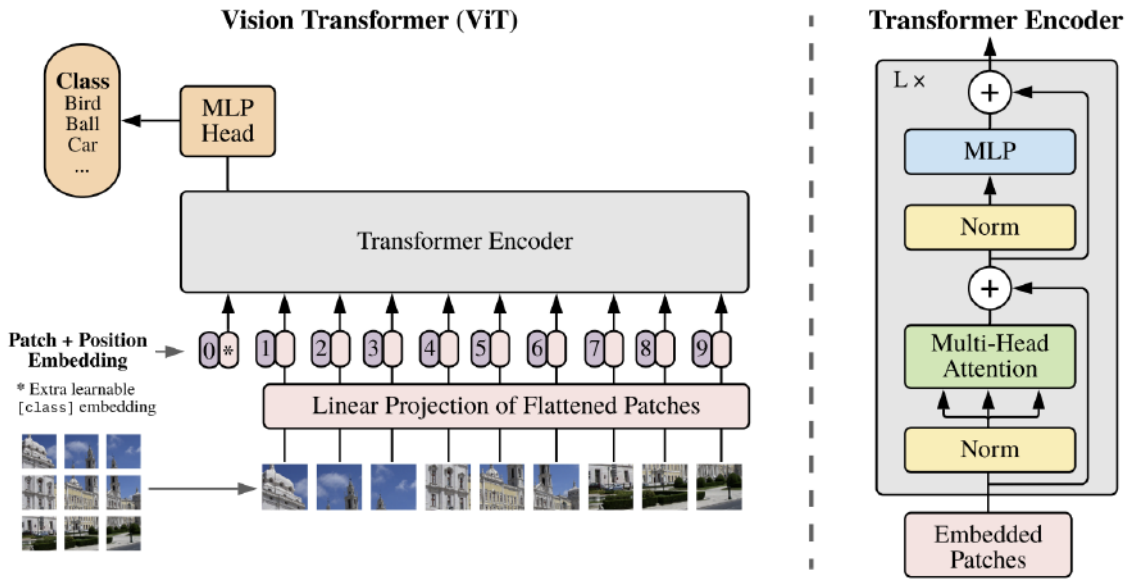


Figure B.5: Vision Transformer architecture [51]

inherent sequence awareness necessitates the addition of positional encodings. [195] introduced sinusoidal positional embeddings:

$$PE_{(pos,i)} = \begin{cases} \sin\left(\frac{pos}{10000^{\frac{i}{D}}}\right) & \text{if } i \bmod 2 = 0 \\ \cos\left(\frac{pos}{10000^{\frac{i-1}{D}}}\right) & \text{otherwise} \end{cases}, \tag{B.9}$$

where pos is the token's position and i is the embedding dimension. These encodings are added to both the input and output token embeddings in the Transformer encoder and decoder.

B.0.4 Vision Transformer

In computer vision tasks, images can be processed as sequences by dividing them into patches, as introduced in [51] and shown in Figure B.5. The image is split into $N = \frac{HW}{P^2}$ patches, where H and W are the image height and width, and P is the patch dimension. Each patch is flattened and linearly projected into a fixed-length token vector. Similar to BERT [48], a special [CLS] token is added to the beginning of the sequence. This token serves as the representation used for classification.

Unlike the original NLP models where the [CLS] token is standard, ViT variants sometimes opt to aggregate all token outputs post-attention, depending on the task. Later, we will argue that relying solely on aggregated patch tokens may not always be the most effective approach, particularly in fine-grained visual classification scenarios.

Chapter C

Transformer Explainability

In the broader machine learning literature, attention mechanisms—despite their intuitive appeal—are not universally accepted as valid explanations of model behavior. Several studies have questioned the reliability of attention scores for interpretability. For instance, [89, 112] argue that attention weights do not necessarily correlate with feature importance or causal influence, and thus should not be treated as direct explanations.

However, others present a more nuanced view. Wiegreffe and Pinter (2019) in "Attention is Not Not Explanation" [211] argue that whether attention can be considered an explanation depends on several factors: the model architecture, the constraints imposed (such as sparsity), and the intended interpretability goal. They emphasize that the question should not be "Is attention an explanation?" but rather "When and how can attention be used as an explanation responsibly?"

Similarly, Serrano and Smith (2019) use input ablation techniques to investigate whether attended tokens truly influence model predictions [177]. Their results show that attention can sometimes—but not always—identify influential inputs, reinforcing the idea that raw attention scores alone are not inherently interpretable.

Despite these caveats, attention mechanisms remain useful as a window into model behavior, especially in the context of Vision Transformers (ViTs). In ViTs, self-attention mechanisms relate image patches through a learned relevance structure, with a special [CLS] token aggregating information across the image. This makes it possible to visualize how the model distributes focus across different regions of the input.

As outlined in Figure B.4 and Equation B.3, the attention mechanism measures the relationships between patches in the case of self-attention. Vision Transformers (ViTs) implement self-attention while also incorporating the [CLS] token from BERT, as discussed previously. In this setting, the attention weights connecting the [CLS] token to the rest of the image tokens can be interpreted as highlighting which patches contribute most significantly to the classification—essentially separating foreground from background [33].

Assume an input image X and a ViT model $f(X)$ composed of L layers, each containing N_h attention heads. The attention matrix for layer l and head n_h is denoted $A^l_{n_h}$. The image is partitioned into $N = \frac{HW}{P^2}$ patches, with H and W representing image height and width, and P denoting the patch side length. These patches are linearly projected into tokens of dimension K, and a [CLS] token is added at the start. After positional encoding is applied, the input has dimensions $1 \times (N+1) \times K$, assuming batch size 1. These tokens are processed through the stack of attention layers, with the [CLS] token ultimately used for classification.

One basic method of visualizing attention, denoted I_{raw}, involves extracting the first row of the final attention matrix A^L, averaged across all heads [195]. This captures how much the [CLS] token attends to each patch. The result is reshaped and upsampled to the original image size. Such maps are used in DINO [26], a self-supervised framework that encourages alignment between local and global views, leading to improved attention to semantically meaningful regions.

Attention rollout and flow [1] aim to model how attention distributes through layers. In the rollout method, attention is recursively multiplied through layers as follows:

$$\hat{A}^{li} = \begin{cases} A^{li}\hat{A}^{l(i-1)} & \text{if } i > j \\ A^{li} & \text{if } i = j \end{cases} \tag{C.1}$$

Here, $j = 0$ and li refers to the current layer and index. The final result $\hat{A}^{li}$ is reshaped and interpolated to match the spatial resolution of the input image. Attention flow, a graph-theoretic variant, is more computationally expensive with complexity $\mathcal{O}(L^2 \cdot n^4)$, compared to rollout's $\mathcal{O}(L \cdot n^2)$, which limits its scalability.

Layer-wise Relevance Propagation (LRP) for transformers [32] merges LRP techniques [11] with the attention rollout mechanism. For each layer l, the modified attention $\bar{A}^l$ is computed as:

$$\bar{A}^l = I + \mathbb{E}_{nh}\left(\nabla_c A^l \odot R^l\right)^+, \tag{C.2}$$

where ∇_c is the gradient with respect to the target class c, $\odot$ represents the element-wise product, R^l is the relevance at layer l, and $\mathbb{E}_{nh}$ averages over heads. The operator $(\cdot)^+$ retains only the positive components, while I ensures skip connections are preserved.

The final attribution map is obtained by cumulatively multiplying the modified attention matrices:

$$A_{LRP} = \prod_{l=0}^{L} \bar{A}^l. \tag{C.3}$$

The resulting matrix A_{LRP} is then reshaped and scaled back to the input image size, yielding the visualization I_{LRP}.

Each of these methods provides a different perspective on what the ViT model focuses on during inference: raw attention maps offer simplicity, rollout captures deep attention paths, and LRP-based approaches offer class-specific explanations. Each comes with trade-offs in terms of interpretability and computational cost.

Chapter D

Bayes' Rule

To begin this section, we provide a concise overview of Bayes' Rule. Consider a dataset $\mathcal{D}$ comprising n observations, where $\mathcal{D} = \{X, y\}$, with X denoting the input features, y the corresponding outputs, and β the parameters of our model. The posterior distribution over the parameters is given by:

$$p(\beta \mid y, X) = \frac{p(y \mid X, \beta)\, p(\beta)}{p(y \mid X)}, \tag{D.1}$$

Alternatively, in plain language:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}, \tag{D.2}$$

This identity is known as Bayes' Rule [155]. For clarity, we briefly define each of the components involved:

- **Marginal likelihood (evidence)**: This term, which serves as a normalization constant, integrates over all possible parameter values to account for uncertainty. The process of computing it—called marginalization—is akin to taking a weighted average over all models, where the weights correspond to the prior plausibility of each model [65].
- **Prior**: Represents our beliefs or assumptions about the distribution of parameters before observing any data.
- **Likelihood**: The probability of observing the data y given inputs X and specific parameter values β.
- **Posterior**: The updated distribution over the parameters after incorporating observed data, reflecting both prior beliefs and new evidence.

In a Bayesian setting, this framework enables us to incorporate prior knowledge into data-driven models, effectively balancing learned information with assumptions about the problem domain.

Chapter E

Point Estimates

It is helpful to unify a few optimisation goals to reveal their underlying similarities. Let's begin with a basic example: linear regression.

Problem Consider a dataset $\mathcal{D}$ comprising N samples, denoted as $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$, where $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ with each $\boldsymbol{x}_i = [x_{11}, ..., x_{nr}]$, and $\boldsymbol{Y} = \{y_1, ..., y_n\}$. Both x_{ij} and y_i are scalar values. This setup represents a straightforward matrix formulation in which the response vector $\boldsymbol{Y}$ is modeled as the dot product of a coefficient matrix $\boldsymbol{\beta} = [b_1, ..., b_r]$ and the feature matrix $\boldsymbol{X}$, plus an additive noise term $\boldsymbol{\epsilon} = \boldsymbol{I}\epsilon$, where $\boldsymbol{I}$ is a $1 \times N$ identity matrix and ϵ denotes the noise. Thus, we write the model as $\boldsymbol{Y} = \beta^T \boldsymbol{X} + \boldsymbol{\epsilon}$. While this matrix-based expression is elegant and insightful, it is only practical when a closed-form solution is available—something not always possible. So, what is the procedure for solving it?

E.1 System of Linear Equations

Let's define the assumptions under which a solution exists:

- No noise term is present, i.e., $\epsilon = 0$.
- The matrix is square: $r = n$.
- The matrix has full row rank—its rows are linearly independent.
- The dataset is internally consistent and free of contradictions.
- All elements of $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{Y}$ belong to $\mathbb{R}$.

Under these idealised assumptions, the problem becomes trivial. Referring back to the matrix formulation:

$$\begin{aligned} \boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{Y}, \\ \boldsymbol{X}^{-1}\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}^{-1}\boldsymbol{Y}, \\ I\boldsymbol{\beta} &= \boldsymbol{X}^{-1}\boldsymbol{Y}, \\ \boldsymbol{\beta} &= \boldsymbol{X}^{-1}\boldsymbol{Y}, \end{aligned} \tag{E.1}$$

Here, I is the $n \times n$ identity matrix. This solution merely requires inverting the input matrix to retrieve the coefficient vector. While this method appears attractively simple, it is rarely applicable in real-world regression. First, this scenario only pertains to linear regression; extending it to classification or non-linear models introduces additional complexity. Second, matrix inversion becomes computationally burdensome for large n. Techniques like Gauss-Jordan elimination used for computing inverses have a time complexity of $\mathcal{O}(n^3)$, and this assumes that $\boldsymbol{X}$ is invertible—which isn't guaranteed. Moreover, the assumption of a zero-error model, although it ensures data consistency, severely limits the realism of the model. In practical scenarios, incorporating and modeling the error term is crucial.

E.2 Ordinary Least Squares

The ordinary least squares (OLS) method is a foundational approach in linear regression. The standard model is expressed as:

$$\boldsymbol{Y} = \boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon}, \tag{E.2}$$

where $\boldsymbol{\beta}$ is a vector of unknown coefficients with dimensions $r \times 1$, and $\boldsymbol{\epsilon}$ represents the noise or error term, shaped $N \times 1$.

We rely on the following key assumptions:

- The rows of $\boldsymbol{X}$ are linearly independent (full row rank).
- No exact linear relationships exist among the explanatory variables.
- The error terms are homoscedastic (equal variance) and uncorrelated.
- All variables $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{Y}$ lie in $\mathbb{R}$.

The aim is to find an estimate of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, that best fits the data. Once $\hat{\boldsymbol{\beta}}$ is known, we can compute the residuals:

$$e = \boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}, \tag{E.3}$$

where e is the residual vector, measuring the discrepancy between actual and predicted values. Note that e is not the same as ϵ—the latter represents inherent observation noise, while the former captures predictive error.

To quantify this error, we define an objective function:

$$G(\boldsymbol{\beta}) = e^T e, \tag{E.4}$$

which is the residual sum of squares (RSS), a common metric for model fit. The goal is to minimise this quantity, leading to the optimisation problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} G(\boldsymbol{\beta}). \tag{E.5}$$

To solve it, we differentiate the objective function:

$$\frac{\partial e^T e}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0, \tag{E.6}$$

resulting in the normal equations:

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}, \tag{E.7}$$

where $\boldsymbol{X}^T\boldsymbol{X}$ is the Gram matrix. The second derivative of $e^T e$ is $2\boldsymbol{X}^T\boldsymbol{X}$, confirming a minimum when $\boldsymbol{X}$ is of full rank. If $\boldsymbol{X}$ is not full rank—possibly due to $N < r$ or linear dependence among features—then perfect multicollinearity exists, and a unique solution may not be achievable.

As long as the feature vectors are linearly independent, the unique solution to the problem is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{E.8}$$

Still, this requires inverting the matrix $\boldsymbol{X}^T\boldsymbol{X}$, which may become computationally intensive for large datasets. Once $\hat{\boldsymbol{\beta}}$ is computed, the predicted values $\hat{\boldsymbol{Y}}$ are obtained via:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{E.9}$$

To assess prediction variance, we use the reduced chi-squared statistic, which normalises the residual error by degrees of freedom. The estimator for this variance is:

$$s^2 = \frac{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}})}{n - r}, \tag{E.10}$$

where s^2 corresponds to the variance parameter σ^2 in the OLS framework. While OLS is well-suited for parameter estimation in linear regression models that do not explicitly assume a probability distribution, it turns out that OLS is also a special case of maximum likelihood estimation under certain conditions.

E.3 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a fundamental statistical technique for estimating the parameters of a probability distribution that most likely explains the observed data [60]. Unlike the OLS framework, MLE explicitly assumes that the residual term $\boldsymbol{\epsilon}$ is drawn from a probability distribution—commonly a Gaussian distribution with zero mean, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

The assumptions underlying this model are:

- The matrix $\boldsymbol{X}$ has full row rank—each row is linearly independent.
- No exact linear dependency exists among the independent variables.
- Error terms are homoscedastic and uncorrelated.
- The residuals $\boldsymbol{\epsilon}$ are random variables drawn from a distribution.
- All vectors $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{Y}$ lie in $\mathbb{R}$.

Though the change from OLS may appear minor, assuming a normal distribution for noise implies that the predicted outputs $\hat{\boldsymbol{Y}}$ also follow a Gaussian distribution—since linear combinations of normally distributed variables remain normally distributed. To illustrate, consider a single observation: $y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i$. Then,

$$\begin{aligned} \mathbb{E}(y_i) &= \mathbb{E}(\boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i) \\ &= \boldsymbol{\beta}^T \boldsymbol{x}_i + \mathbb{E}(\epsilon_i) \\ &= \boldsymbol{\beta}^T \boldsymbol{x}_i, \end{aligned} \tag{E.11}$$

since $\mathbb{E}(\epsilon_i) = 0$ due to the zero-mean Gaussian assumption. Similarly, the variance is:

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i) \\ &= 0 + \text{Var}(\epsilon_i) \\ &= \sigma^2. \end{aligned} \tag{E.12}$$

This follows from the properties of expectations and variances for independent variables. Thus, y_i can be described as a normally distributed variable conditioned on $\boldsymbol{x}_i$:

$$y_i \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{x}_i, \sigma^2)$$

The corresponding probability density function (PDF) for a single observation is:

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2}{2\sigma^2}\right). \tag{E.13}$$

To generalise this across all N data points, we consider the joint likelihood function, defined as the product of the individual PDFs:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma|\boldsymbol{Y}, \boldsymbol{X}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2\right). \end{aligned} \tag{E.14}$$

Here, our objective is to identify the parameters $\boldsymbol{\beta}$ that maximise this likelihood. Formally, we solve:

$$\boldsymbol{\beta}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \sigma). \tag{E.15}$$

Because working directly with the likelihood is numerically inconvenient (especially due to the product of exponentials), we usually take the logarithm, which simplifies the optimisation without changing the optimum (since the log is a monotonic function). The log-likelihood becomes:

$$\boldsymbol{\beta}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma|\boldsymbol{Y}, \boldsymbol{X}). \tag{E.16}$$

Expanding this expression gives:

$$\boldsymbol{\beta}_{\text{MLE}} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2, \tag{E.17}$$

and in matrix-vector notation:

$$\boldsymbol{\beta}_{\text{MLE}} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{E.18}$$

This formulation reveals that MLE, under the assumption of i.i.d. Gaussian noise, leads to the same solution as OLS. In essence, OLS can be interpreted as the maximum likelihood estimator for a linear model with normally distributed errors.

Moreover, MLE can be related to broader optimisation frameworks. For instance, MLE is equivalent to minimising cross-entropy in probabilistic terms. Recall the cross-entropy formulation:

$$\begin{aligned} H(p,q) &= H(p) + D_{\mathrm{KL}}(p\|q) \\ &= \mathbb{E}_{x\sim p(x)}\left[\log\left(\frac{1}{q(x)}\right)\right], \end{aligned} \tag{E.19}$$

Here, $H(p,q)$ is the cross-entropy between the true distribution p and approximation q, and D_{KL} is the Kullback-Leibler divergence [40]. Comparing Equations E.16 and E.19, we see that maximising the likelihood corresponds to minimising cross-entropy when treating q as the model distribution.

This is just one instance where MLE links directly to familiar loss functions. For example, log-loss is the MLE under a Bernoulli (sigmoid) likelihood, and L1 regression corresponds to MLE under a Laplace noise assumption.

MLE is a purely data-driven approach to parameter estimation—it relies entirely on observed data without incorporating prior beliefs. While this is effective in many settings, it can be limiting in small-data regimes where priors (as in Bayesian methods) could provide valuable regularisation.

E.4 Maximum a Posteriori Estimation

Bayes' theorem provides a framework for updating beliefs about model parameters given observed data. If we drop the marginal likelihood term (which is constant with respect to the parameters), we can write the posterior distribution as being proportional to the product of the likelihood and the prior:

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) \propto p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma)p(\boldsymbol{\beta}), \tag{E.20}$$

The goal of maximum a posteriori (MAP) estimation is to find the parameter value that maximizes this unnormalized posterior distribution:

$$\boldsymbol{\beta}_{MAP} = \arg\max_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma) + \log p(\boldsymbol{\beta}). \tag{E.21}$$

MAP estimation introduces prior knowledge about the parameters, which can be beneficial in settings where data is limited. While MAP and MLE may yield similar results when the dataset is large, the inclusion of a prior can improve generalization in small-sample regimes.

Assuming a Gaussian prior on the coefficients, where the prior distribution of $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}$, we write:

$$p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = (2\pi)^{-\frac{r}{2}} \det(\boldsymbol{\Lambda})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})\right) \tag{E.22}$$

Here, r denotes the number of parameters in $\boldsymbol{\beta}$. The likelihood, assuming Gaussian noise in the outputs, can be written in vector form as:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}(\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right) \tag{E.23}$$

Combining the log-likelihood and the log-prior, and assuming a zero-mean prior (i.e., $\boldsymbol{\mu} = \boldsymbol{0}$), the log-posterior becomes:

$$\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \propto -\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta} \tag{E.24}$$

Thus, MAP estimation corresponds to minimizing the following objective function:

$$\boldsymbol{\beta}_{MAP} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\beta} \right] \tag{E.25}$$

This formulation is equivalent to ridge regression, where the regularization term arises naturally from the Gaussian prior. Hence, MAP estimation provides a Bayesian justification for L2 regularization and serves as another example of empirical risk minimization corresponding to a probabilistic inference strategy.

E.5 Point Estimates

Both maximum likelihood and maximum a posteriori estimation methods yield point estimates, meaning they return a single best guess for the model parameters based on the data. These approaches do not characterize uncertainty in the parameter estimates. In contrast, fully Bayesian inference seeks to infer the entire posterior distribution over the parameters, allowing for more robust predictions and uncertainty quantification, particularly valuable when data is scarce or noisy.

Chapter F

Gaussian Processes

A Gaussian Process (GP) defines a distribution over functions $f(\boldsymbol{x})$, such that any finite collection of function evaluations follows a multivariate Gaussian distribution:

$$\begin{aligned} f(\boldsymbol{x}) &\sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x'})) \\ m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ K(\boldsymbol{x}, \boldsymbol{x'}) &= \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))] \end{aligned} \tag{F.1}$$

Here, $m(\boldsymbol{x})$ is the mean function, and $K(\boldsymbol{x}, \boldsymbol{x'})$ is the covariance function (or kernel), which must be positive definite [155]. Though GPs are defined over an infinite input space, they are typically characterized through finite-dimensional marginal distributions. For any finite set of inputs $\boldsymbol{X}$, the corresponding function values $f(\boldsymbol{X})$ follow a multivariate normal distribution:

$$f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X})), \tag{F.2}$$

where $K(\boldsymbol{X}, \boldsymbol{X})$ is the covariance matrix computed between all input pairs in $\boldsymbol{X}$, with $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Intuitively, inputs that are close in feature space should produce outputs that are more correlated; thus, the GP encodes similarity between inputs through the kernel.

We can connect GPs to Bayesian linear regression. Suppose we define a linear model $f(\boldsymbol{x}) = \phi(\boldsymbol{x})^T \boldsymbol{w}$ with a Gaussian prior on the weights, $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p)$. Then the induced distribution over functions is:

$$\mathbb{E}[f(\boldsymbol{x})] = \phi(\boldsymbol{x})^T \mathbb{E}[\boldsymbol{w}] = 0, \tag{F.3}$$

$$\mathbb{E}[f(\boldsymbol{x}) f(\boldsymbol{x'})] = \phi(\boldsymbol{x})^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] \phi(\boldsymbol{x'}) = \phi(\boldsymbol{x})^T \Sigma_p \phi(\boldsymbol{x'}), \tag{F.4}$$

demonstrating that the function values are jointly Gaussian with zero mean and kernel $k(\boldsymbol{x}, \boldsymbol{x'}) = \phi(\boldsymbol{x})^T \Sigma_p \phi(\boldsymbol{x'})$. Given a kernel and a test input matrix $\boldsymbol{X}_*$, we define the GP marginal distribution as:

$$f(\boldsymbol{X}_*) \sim \mathcal{N}(0, K(\boldsymbol{X}_*, \boldsymbol{X}_*)), \tag{F.5}$$

which allows us to draw function samples from the GP prior at the specified input locations.

F.1 Posterior Prediction with GPs

Let us now consider noise-free observations with $\sigma_n^2 = 0$. The training dataset $\mathcal{D} = (\boldsymbol{X}, f(\boldsymbol{X}))$ consists of n input-output pairs. Similarly, the test dataset is $\mathcal{D}_* = (\boldsymbol{X}_*, f(\boldsymbol{X}_*))$ with n_* new input points. The joint distribution over training and test outputs is:

$$\begin{bmatrix} f(\boldsymbol{X}) \\ f(\boldsymbol{X}_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ K(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix}\right). \tag{F.6}$$

Conditioning on the observed training data yields the posterior predictive distribution:

$$\begin{aligned} f(\boldsymbol{X}_*)|\boldsymbol{X}_*, \boldsymbol{X}, f(\boldsymbol{X}) \sim \mathcal{N}(&K(\boldsymbol{X}_*, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1} f(\boldsymbol{X}), \\ &K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1}K(\boldsymbol{X}, \boldsymbol{X}_*)). \end{aligned} \tag{F.7}$$

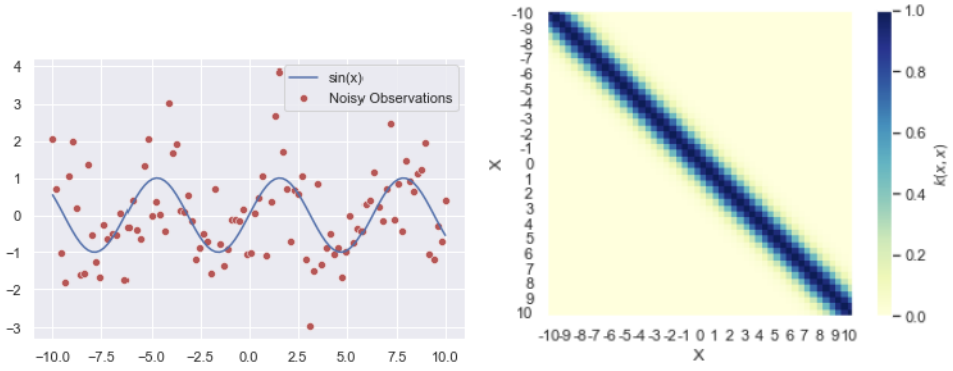
Now consider the more realistic case where observations are corrupted by Gaussian noise, $\boldsymbol{y} = f(\boldsymbol{X}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma_n^2 \boldsymbol{I})$. The joint distribution becomes:

$$\begin{bmatrix} \boldsymbol{y} \\ f(\boldsymbol{X}_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I} & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ K(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix}\right), \tag{F.8}$$

leading to the posterior distribution:

$$\begin{aligned} f(\boldsymbol{X}_*)|\boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(&K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}]^{-1} \boldsymbol{y}, \\ &K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}]^{-1} K(\boldsymbol{X}, \boldsymbol{X}_*)). \end{aligned} \tag{F.9}$$

F.2 Illustrative Example



(a) Noisy observations of $\sin(x)$ **(b)** RBF kernel illustration

Figure F.1: Setup for Gaussian Process regression

To build intuition, consider the true function $f(x) = \sin(x)$ over $x \in [-10, 10]$ with additive Gaussian noise, $\epsilon \sim \mathcal{N}(0, 1)$. We use the radial basis function (RBF) kernel:

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right), \tag{F.10}$$

which defines similarity based on distance. Figure F.1b shows this kernel for a set of input pairs.

Given this setup, we define a finite set of $N = 50$ input locations $\boldsymbol{X}$ and evaluate the GP prior by drawing samples from:

$$f(\boldsymbol{X}) \sim \mathcal{N}(0, K(\boldsymbol{X}, \boldsymbol{X})). \tag{F.11}$$

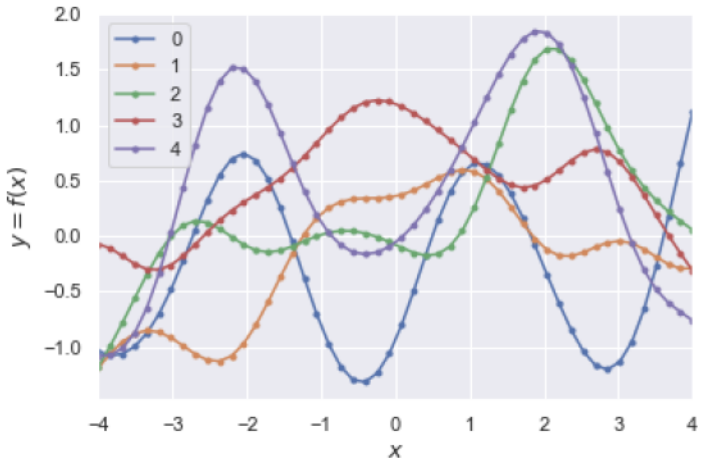


Figure F.2: Five function samples from the GP prior using the RBF kernel

Figure F.2 displays five realizations of functions from the prior. Each function corresponds to a draw from a 50-dimensional multivariate Gaussian.

We now incorporate observations and use the posterior (Equation F.9) to make predictions over 100 test points in $[-6, 6]$, based on 20 noisy training observations.

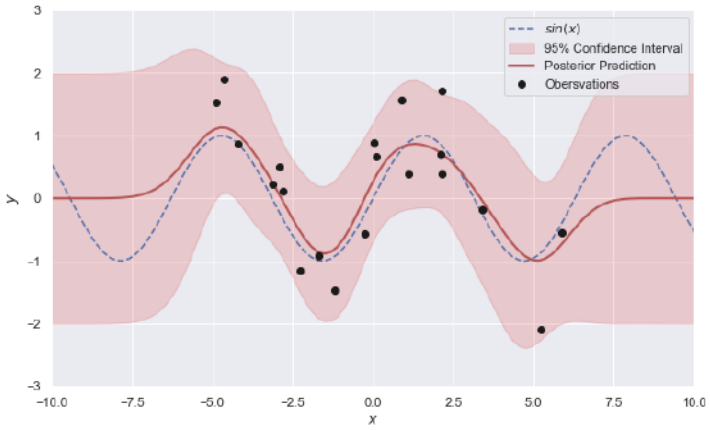


Figure F.3: Posterior GP predictions with 95% confidence intervals

As shown in Figure F.3, the posterior mean aligns closely with observed regions, and uncertainty grows in regions far from data. This is one of the key advantages of GPs: the ability to quantify predictive uncertainty in a principled way.

Chapter G

Extended Chapter 2 Results

Table G.1: Performance evaluation study on pCLE data based on the ADCC and time metrics. Coh is Coherence, Comp is Complexity, AD is average drop and they are reported for completeness. Time(s) is the average time to compute one PA map using a batch size of one.

Architecture	PA method	Coh ↑	Comp ↓	AD ↓	ADCC ↑	Latency(ms) ↓
ResNet18	Original					
	Grad-CAM	90.1	32.7	10.1	76.6	5.6
	Grad-CAM++	90.6	33.1	10.6	76.2	5.5
	SmoothGradCAM++	88.3	27.6	14.3	74.8	70.7
	Score-CAM	90.0	32.3	5.9	80.5	121.4
	Recipro-CAM	91.0	41.2	10.0	72.8	3.5
	Proposed method					
	Grad-CAM	92.5	34.2	11.9	77.7	9.7
	Grad-CAM++	93.2	32.5	12.6	78.2	10.5
	SmoothGradCAM++	92.2	30.6	17.2	75.6	103.9
	Score-CAM	90.0	32.3	5.9	80.5	1267.2
	Recipro-CAM	92.1	37.8	11.8	75.9	36.4
MoblieNetV2	Original					
	Grad-CAM	89.5	21.3	73.8	29.3	8.8
	Grad-CAM++	86.2	30.0	66.9	37.8	8.7
	SmoothGradCAM++	77.7	18.1	76.2	24.5	71.3
	Score-CAM	62.5	33.9	56.3	43.9	315.1
	Recipro-CAM	85.8	32.3	67.1	35.8	5.8
	Proposed method					
	Grad-CAM	89.5	27.1	59.3	48.0	12.5
	Grad-CAM++	90.7	35.9	41.7	59.5	13.4
	SmoothGradCAM++	88.8	22.0	71.3	37.1	88.8
	Score-CAM	62.5	33.9	56.3	43.9	3154.3
	Recipro-CAM	90.2	33.8	48.6	55.5	65.6

Table G.2: Performance evaluation study on ImageNet based on the ADCC and time metrics. Coh is Coherence, Comp is Complexity, AD is average drop and they are reported for completeness. Time(s) is the average time to compute one PA map using a batch size of one.

Architecture	PA method	Coh ↑	Comp ↓	AD ↓	ADCC ↑	Latency(ms) ↓
ResNet50	Original					
	Grad-CAM	98.1	34.0	30.0	67.9	11.3
	Grad-CAM++	98.3	34.8	30.0	67.6	11.28
	SmoothGradCAM++	97.3	34.8	34.8	64.6	84.6
	Score-CAM	98.2	34.8	21.6	74.3	134.0
	Recipro-CAM	97.5	27.8	39.9	63.0	7.4
	Proposed method					
	Grad-CAM	97.7	35.2	22.3	72.4	25.4
	Grad-CAM++	97.7	35.8	22.6	72.1	26.3
	SmoothGradCAM++	97.9	36.5	23.1	71.8	134.0
	Score-CAM	98.2	34.8	21.6	74.3	14147.9
	Recipro-CAM	97.3	29.5	28.8	70.9	86.6

Table G.3: T vs ADCC Study for Resnet-18 and MobileNetV2 on pCLE data.

	PA Method's ADCC			
	ResNet18		MobileNetV2	
T	Grad-CAM	ReciproCAM	Grad-CAM	ReciproCAM
1	76.9	75.1	45	53.2
2	77.1	75.3	47.2	53.9
3	77.1	75.6	46.6	55.1
4	77.2	75.9	46.9	54.9
5	77.5	75.6	47.9	55.4
6	77.6	75.9	47.3	55.2
7	77.6	75.9	47.9	55.3
8	77.6	75.8	48	55.3
9	77.5	75.8	47.8	55.5
10	77.7	75.9	48	55.5
20	77.5	75.9	47.9	55.8
40	77.6	75.9	48.2	56.1
80	77.7	76.1	48	56.1
90	77.7	76.1	48.1	55.8
100	77.7	76.1	48.4	55.8