Can Attention-Inspired Artificial Intelligence Provide a Diagnostic Understanding of Colorectal Cancer Imaging Data?

Andrew John Broad

ORCID https://orcid.org/0000-0001-7131-6860

Submitted in accordance with the requirements for the degree of PhD, The University of Leeds, School of Computing, July 2024.

Supervisors:

Dr Marc de Kamps

Dr Alex Wright

Prof Darren Treanor

Centre for Doctoral Training (CDT) in
Artificial Intelligence for Medical Diagnosis and Care,
University of Leeds

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Work carried out for *Chapter 4: Characterising the Whole Slide Image* formed the basis of the article *Attention-guided sampling for colorectal cancer analysis with digital pathology* (Broad et al., 2022), published in the *Journal of Pathology Informatics,* https://doi.org/10.1016/j.jpi.2022.100110.

The work was undertaken by the candidate as corresponding author, while co-authors Dr Marc de Kamps, Dr Alex Wright and Prof Darren Treanor were project supervisors and provided editorial input into the manuscript wording and structure.

Work carried out for *Chapter 5: Feedback Attention*, *Chapter 6: Visualising Feedback Attention* and *Chapter 7: Saccade-like Behaviour with Feedback Attention Models* formed the basis of the article *Object-based Feedback Attention in Convolutional Neural Networks Improves Tumour Detection in Digital Pathology*, which as of 1st Oct 2024 is under review with *Scientific Reports* and available as a pre-print at https://doi.org/10.21203/rs.3.rs-4828783/v1.

The work was undertaken by the candidate as lead author, while co-authors Dr Marc de Kamps, Dr Alex Wright and Prof Darren Treanor were project supervisors and provided editorial input into the manuscript wording and structure. Co-author Dr Clare McGenity reviewed the manuscript's compliance with AI reporting guidelines. Dr Marc de Kamps was declared corresponding author to ensure continuity after the candidate leaves the University.

Acknowledgements

Sincere thanks are due to my supervisors, Marc de Kamps, Alex Wright and Darren Treanor, for sharing valuable domain knowledge and for their feedback and support throughout this project. The NPIC team and CDT sponsors at Roche have also generously given their time and expertise at regular group meetings.

This work uses data provided by patients and collected by the NHS as part of their care and support.

Data was used from the QUASAR trial (QUASAR Collaborative Group, 2007) and I am indebted to the people involved, particularly Dr Gordon Hutchins who carried out the painstaking point classifications that yielded the ground truth data subsequently used in ML model training.

Further model evaluation was carried out using ImageNet-100 data shared on Kaggle by Shekar (2021) as a subset of the ImageNet public challenge dataset (Russakovsky et al., 2015).

Professor Darren Treanor provided valuable insight from a consultant pathologist's viewpoint, performing qualitative analysis of AI outputs and relabelling large volumes of resampled patch images for statistical analysis of algorithms in later chapters.

Solutions to software development issues were found on Stack Overflow, supplied by users *Georgy* and *Vojtech Vozda* (density contours and IoU calculations), *Hooked* (PyPlot memory management), *Joe Kington* and *Flabetvibes* (Gaussian KDE), *Raed Mughaus (*confidence intervals), *Brandon W* (random file selection in shell scripts) and *mbpaulus* (leaf node configuration in CNNs).

This work was undertaken on ARC4, part of the High-Performance Computing facilities at the University of Leeds, UK. This would not have been possible without expert support and training from the ARC team, particularly Martin Callaghan, John Hodrien and Alex Coleman.

This work also made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester and York.

I would also like to thank the management and fellow students in the Centre for Doctoral Training in AI for Medical Diagnosis and Care, for creating a mutually supportive research environment, which persisted online during the height of the Covid-19 pandemic while the group was dispersed.

Finally, I would like to thank my wife Siobhan for her support throughout, and especially for the encouragement to start writing early, for which I am now very grateful.

Dedicated to
Alison Broad (9/5/1967 to 12/6/2024)
and
Linda Broad (23/1/1970 to 24/10/2015)

Abstract

Digital pathology workflows provide high-resolution whole slide images (WSIs) for assessing diseases such as cancer at a cellular level. A worldwide shortage of pathologists limits the adoption of labour-intensive analysis techniques. This is potentially a role for Artificial Intelligence (AI). However, current AI models need images in the order of 200x200 pixels, while WSIs are of gigapixel size. AI-based systems must address this scale discrepancy without overlooking diagnostically important features in the WSI.

Work in this thesis was motivated by human visual attention, where relevant features of an input scene are selected in response to goals in executive brain regions, avoiding processing the whole scene at full resolution.

Two novel WSI processing pipelines incorporated attention-like algorithms. The first used a thumbnail image to map tumour density, controlling the sampling density of full-magnification patches for classification with a convolutional neural network (CNN). A later pipeline introduced weighted regular sampling (WRS) to mitigate sampling biases. The estimated class distributions yielded the tumour outline and tumour stroma ratio (TSR), a predictor of disease severity.

A novel Feedback Attention Ladder CNN (FAL-CNN) used feedback attention, significantly increasing classification accuracy from 79.33% to 82.82% ($p<0.001$) with 9-class colorectal cancer patches. Top-to-bottom and local-group feedback were combined to generate attention masks for the forward path. Increased accuracy with ImageNet-100 showed the approach to be transferrable. In the WRS pipeline, TSR error was substantially reduced at pathologist-selected locations, suggesting application in a TSR measurement tool.

Visualisations of attention masks in the FAL-CNN highlighted informative tissue regions. A novel saccade model resampled the input patch to align the centre-focused FAL-CNN on these regions. The model discovered salient features even when outside the initial patch. Pathologist relabelling of resampled patches confirmed the saccade model's ability to locate nearby regions of tumour, a potentially valuable behaviour in cancer WSI analysis.

Contents

Abbreviations used

2D	Two Dimensional
3D	Three Dimensional
AG	Attention Gate
AHP	Attention Heatmap Pipeline
AI	Artificial Intelligence
ARC	Advanced Research Computing
AUC	Area Under Curve
BA	Bland-Altmann
BB; BBox	Bounding Box
C	Channels
CBAM	Convolutional Block Attention Module
CI	Confidence Interval
CIR	Computationally Intense Research
CM	Confusion Matrix
CNN	Convolutional Neural Network
CoA	Centre of Attention
CV	Cross Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
Dmax	Maximum (tumour) density
DSC	Dice Similarity Coefficient
EPSRC	Engineering and Physical Sciences Research Council
F1	F1 Score, synonymous with DSC
FAL-CNN	Feedback Attention Ladder Convolutional Neural Network
FAM	Feedback Attention Module
FB	Feedback
FC	Fully Connected
FES	Feature Embedding Store
FF	Feedforward
fMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FPC	False Positive Correction
FROC	Free Response Operating Characteristic
GDA	Goal-Directed Attention
GPU	Graphics Processing Unit
GT	Ground Truth
H	Height
H&E	Haematoxylin and Eosin
HPC	High-Powered Computing
HR	Hazard Ratio
IoU	Intersection over Union
IT	Inferior Temporal
KDE	Kernel Density Estimator
LSTM	Long Short-Term Memory
LTHT	Leeds Teaching Hospitals NHS Trust
MAE	Mean Absolute Error
MIL	Multi-Instance Learning
MIT	Massachusetts Institute of Technology

ML	Machine Learning
MSc	Master of Science
MSCN	Multi-Scale Convolutional Network
MSE	Mean Squared Error
N8	Group of 8 research-intensive universities in Northern England
NCRF	Neural Conditional Random Field
NHS	National Health Service
NN	Neural Network
NPIC	National Pathology Imaging Cooperative
PoT	Proportion of Tumour
pp	Percentage Points
QC	Quality Control
QUASAR	QUick And Simple And Reliable
RCNN	Region-based Convolutional Neural Network
RCT	Randomised Controlled Trial
ReLU	Rectifying Linear Unit
ROI	Region of Interest
SA	Self-Attention
SD	Standard Deviation
SE	Standard Error
SGD	Stochastic Gradient Descent
STA	Source Target Attention
SVS	File extension used for WSI image storage in Aperio systems
TCGA	The Cancer Genome Atlas
TIFF	Tagged Image File Format
TN	Thumbnail image
TN	True Negative
TP	True Positive
TSR	Tumour Stroma Ratio
TTP	Tile-by-Tile Pipeline
U-Net	U-Shaped Neural Network
V1, V2, V4	Ventral areas 1, 2 and 4
VIA	VGG Image Annotator
VIG	Visualisation Image Generator
VGG	Visual Geometry Group, University of Oxford
W	Width
WMH	White Matter Hyperintensities
WRS, WRSP	Weighted Regular Sampling [Pipeline]
WSI	Whole Slide Image
XAI	eXplainable AI
XML	eXtended Markup Language

Figures

Tables

1 Introduction

1.1 Motivation

This research explores the use of Artificial Intelligence (AI) in extracting clinically valuable information from medical images, particularly histopathology whole-slide images (WSIs) for colorectal cancer (CRC).

Histopathology is the study of diseases of tissue at cellular level. Traditionally, tissue samples were stained and mounted on glass slides for viewing with an optical microscope. High-magnification scanners are replacing the microscope, and the resulting whole slide images (WSIs) are stored digitally. Digital pathology (DP) supports rapid retrieval and sharing of images, reducing case transfer times to assigned pathologists and accelerating diagnostic workflows. Users can rapidly zoom between a whole-slide thumbnail view of the sample and high magnification views of structures at a cellular scale. Second opinions can be rapidly provided by off-site experts, viewing WSIs remotely (NPIC, 2023).

Despite the many benefits of DP, diagnostic bottlenecks exist at Leeds Teaching Hospitals NHS Trust (LTHT) due to a worldwide shortage of qualified pathologists (Acs et al., 2020). Any measure that reduces their workload will improve image throughput, diagnosis time and subsequently patient care.

The central problem of using AI to analyse histopathology images is the size of a WSI. Images can be up to 100,000×100,000 pixels, whereas many convolutional neural networks (CNNs) used in image recognition have input dimensions in the order of 200×200 pixels. This means that WSIs are often processed as many separate image patches, raising questions of where and how frequently to sample the patches from the WSI for efficient processing without losing important structural context.

This thesis examines solutions to this problem inspired by attention in human vision. Attention is the animal kingdom's solution to the problem of rapidly detecting features of interest in a busy, high-resolution input. Rapid eye movements called *saccades* fixate on objects initially perceived in low-resolution peripheral vision. This is often guided by motion, to which the peripheral vision is more sensitive. Further information is then gathered at higher resolution by the central region of the eye, the fovea. The fovea is of a limited size, so information from multiple saccades must be combined into a larger internal representation. This process uses a lower processing bandwidth than would be possible if analysing the whole scene at full resolution simultaneously (Mnih et al., 2014).

In human attention, *feedback* plays an important role (van der Velde and de Kamps, 2001) in detecting objects of interest in a cluttered scene. For example, if a subject is asked to search for a square object, a representation of this object in higher neural layers is fed back to lower visual layers to increase sensitivity to image features such as sharply defined corners or straight edges. Neurons responding to objects in the visual field will then fire more strongly when these features are detected. This information is passed back up through feedforward connections, causing higher layers to direct a visual saccade towards these areas of the scene. Thus the individual's attention is directed to instances of the square object.

Both feedback attention and saccade behaviours were reproduced in this work.

1.2 Aims and Objectives

The overarching goal of this project was to contribute to knowledge in computational histopathology, by developing novel and diagnostically useful attention-based AI techniques for processing medical images at WSI scale.

1.2.1 Visualising Cancer in the WSI

The research question addressed here was: Can attentional processes allow an efficient WSI patch sampling regime be developed, so that key diagnostic results such as tumour Region of Interest (ROI) and Tumour Stroma Ratio (TSR) can be derived with similar accuracy to a more computationally expensive tile-by-tile approach?

A prototype system existed for visualising distributions of tumour and other cells, by applying a CNN classifier to a WSI divided into many small tiles. It was planned to extend this mechanism into a more efficient processing pipeline, using spatial attention to select a smaller number of salient image patches for analysis. It was expected that this would reduce processing time whilst maintaining the accuracy of diagnostic information derived from the patch distribution. This work (Section 1.3.3 and Chapter 4) would also assess the performance of existing CNN models in the processing pipeline.

Novel CNN models described in Sections 1.2.2, 1.3.4 and Chapter 5, would also be assessed in this pipeline. This work is described in Chapter 8.

1.2.2 Feedback Attention

An original objective of this research was to simulate feedback processes in human cognition using machine learning (ML), investigating potential applications in whole slide imaging for colorectal cancer. It was proposed to add feedback mechanisms to CNNs to evaluate the effect on their performance, in terms of classification accuracy and any consequent improvement in the accuracy of ROI and TSR prediction in the WSI pipeline.

This would address the research question of whether the classification accuracy of a feedforward CNN can be enhanced by adding neural feedback pathways, emulating those in the primate ventral stream.

Additional research questions were raised during this work: Do visual representations of the underlying spatial attention masks reveal informative objects and structures in the input image? Can this information be used to guide a saccade-like movement to align the target with the model's centre of attention (CoA), and would this result in improved diagnostic performance? Would the performance of the WSI pipeline introduced in Section 1.2.1 be enhanced by incorporating the new feedback CNN?

A research methodology was planned, involving a series of iterative enhancements to a feedforward CNN, progressively adding top-down attention elements and evaluating model performance before moving to the next enhancement. To address further emerging research questions, spatial plots were scheduled for development, and it was planned to evaluate these results with a consultant pathologist to gain a biological understanding of any highlighted tissue structures. Object-oriented programming techniques were indicated, to allow software to modularised and interconnected into the required model and pipeline architectures.

1.3 Summary of Work

1.3.1 Background

A literature review (Chapter 2) was carried out, encompassing attention processes in cognitive neuroscience and their application in computational models used in medical imaging.

Neural networks and analytic processes using feedforward (bottom-up) and feedback (top-down) attention were reviewed. The *neural blackboard* was examined as a mechanism for binding features extracted by multiple processes.

Existing applications of attention were explored, in problem domains including general image processing, low-resolution medical imaging, and whole slide imaging and patch analysis in digital pathology.

Tumour Stroma Ratio (TSR) was assessed as a predictor of survival in CRC cases. Several studies were reviewed that showed that higher levels of stroma, or lower proportions of tumour (PoT), were associated with more aggressive disease progression.

1.3.2 Data

Data sources are discussed in Chapter 3. Work for this thesis used data from the QUASAR trial (Gray et al., 2007), where tissue surgically removed from CRC patients was stained and scanned as WSIs to assess the effects of chemotherapy after surgery. Later annotations by Hutchins et al (2018) provided ground truth (GT) data for each WSI, defining ROI outline and point classifications of tissue in a *virtual biopsy* region, manually selected at the area of highest perceived tumour cell density.

Patches of 224×224px size were extracted from WSIs that met quality control (QC) criteria. Each patch was centred on an annotated GT location from the point classification, resulting in a labelled 9-class image set for use in training CNN classifiers. Further sets of patch images were derived from this data as required in later experiments.

To test transferability of new model architectures to other problem domains, the ImageNet-100 dataset (Shekar, 2021) was also used in CNN training and feedback visualisation work.

1.3.3 WSI Processing Pipelines

Chapter 4 presents two novel WSI processing pipelines, each using attention mechanisms based on processes in human visual cognition. In the Attention Heatmap Pipeline (AHP), low-resolution thumbnail tiles were analysed by a convolutional neural network (CNN) to generate a heatmap of likely tumour regions. These determined the sampling density at which full-magnification image patches were to be classified using a further CNN trained on cellular-scale images.

The Weighted Regular Sampling Pipeline (WRSP) used a modified algorithm to reduce sampling biases observed in the AHP, while still directing the action of the CNN classifier towards regions of suspected tumour cells. Multiple well-known CNN classifiers were compared in a benchmarking test, leading to the adoption of VGG19 in this role. Tumour regions of interest (ROI) were predicted with an F1 Score of 83.6% (95% CI 80.5 to 86.7%), representing strong agreement with the ground truth over multiple ROIs of widely varying shapes and sizes.

TSR was estimated from distributions of tumour and stroma patches predicted by the image pipelines. The mean TSR error magnitude, relative to that calculated from pathologist annotations, was less than 20%.

The WRSP and benchmarking results were later published as *Attention-guided sampling for colorectal cancer analysis with digital pathology* (Broad et al., 2022).

Analysis of results concluded that the choice of sampling location and the classification accuracy of the CNN both contributed to the accuracy of TSR prediction. Enhancements to the CNN using feedback attention were therefore explored as a means of boosting model accuracy.

1.3.4 Feedback Attention Models

For Chapter 5, feedback pathways were added to the VGG19 classifier. Inspiration was drawn from literature modelling local feedback within regions in the ventral stream, and from top-to-bottom feedback architectures based on U-Net. A hybrid Feedback Attention Ladder (FAL-CNN) model combined these structures and was trained for optimum feedback activations between multiple combinations of convolutional groups. A feature embedding store (FES), acting as working memory, allowed the model to derive its final classification from the results of the feedforward convolutions over multiple feedback cycles.

The FAL-CNN gave a classification accuracy with 9-class colorectal cancer patch data of 82.99%, a significant increase of 3.50 percentage points (pp) ($p<0.001$) relative to the 79.37% measured with the VGG19.

With ImageNet-100, the FAL-CNN classification accuracy was 83.28%, an of 2.39pp ($p<0.001$) relative to the VGG19 baseline accuracy of 80.89%. While this result does not reach the current state of the art for ImageNet (now approximately 90%), it demonstrates a significant improvement in performance due to our feedback techniques, which we recommend for incorporation in newer architectures such as EfficientNet.

1.3.5 Visualising Feedback Attention

For Chapter 6, spatial distributions of feedback attention activations were plotted at each feedback level in the FAL-CNN and superimposed on the input image. With colorectal cancer patches, averaged over multiple images, the distributions revealed a central focus around the central pixel where the ground truth label was applied, suggesting the model has learned to examine this area preferentially.

Qualitative examination by a consultant pathologist revealed that the regions of high attention in individual patches corresponded to cellular structures that were relevant to the patch class prediction. Plots of patch images with the attention regions superimposed as contours (Figure 1) are therefore useful from an explainable AI (XAI) perspective, highlighting tissue structures relevant to the FAL-CNN model's class prediction.

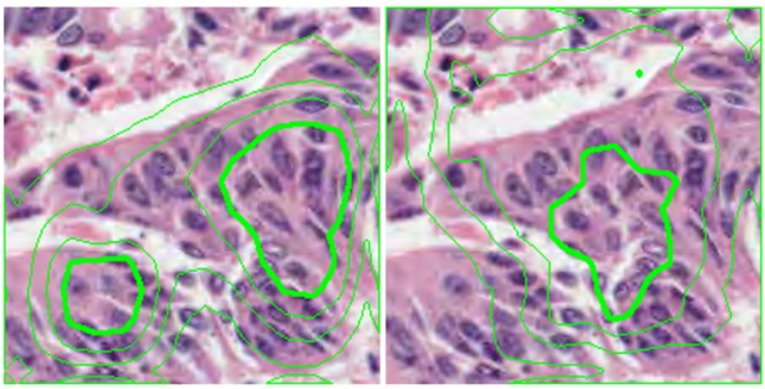


Figure 1: Tumour regions in patch image from CRC resection sample, highlighted by attention distribution in FAL-CNN model

With ImageNet-100, the distributions of feedback activations in higher layers showed strong visual correlation with salient features of the input image, such as a shark's dorsal fin, or the eyes and beak of a bird. Lower-level feedback activations were more fine-grained, highlighting informative textures such as feathers or scales.

1.3.6 Saccade Model

Chapter 7 describes a model implementing a saccade-like behaviour by iteratively moving its sampling region to locate salient features in the surrounding area of the WSI. An embedded FAL-CNN was used to obtain an attention distribution based on the input patch. The input was then resampled from a larger background image, such that the previous centre of attention (CoA) was aligned with the patch centre where the QUASAR-trained FAL-CNN was known to be most sensitive. The new patch was then input to the FAL-CNN to obtain the CoA for the next saccade, and the process repeated for 5 or 10 iterations, generally sufficient for convergence. The final patch class was compared with the GT label for each original input patch.

With ImageNet, agreement rates were significantly higher using the attention-guided saccades (80.71%, 95% CI 80.55 to 80.86%) than with random movements (58.82%, 95% CI 58.55 to 59.09%). *This represents quantitative evidence that attention regions enclose content relevant to the classification result*. The saccade movements tracked towards informative features such as a bird's head or a shark's fin. Notably, this happened even when these features were initially outside the input patch boundary.

With QUASAR-derived patches, the model tracked towards informative tissue, such as regions of tumour adjacent to central lumen or non-informative background. Because of the heterogeneous nature of the images, the resulting classification often differed from the GT class at the original patch centre, resulting in agreement rates of 51.23% (95% CI 49.66 to 52.80%), significantly lower than with ImageNet.

To rectify this, a consultant pathologist was engaged to relabel each patch according to tissue at the centre pixel of the post-saccade location. The rate of agreement with the saccade model over 9 classes was 78.25% (95% CI 74.21 to 82.29%). However, 93.23% (95% CI 90.21 to 96.25%) of expert-relabelled *tumour* patches were correctly identified by the model. This is a strong rate of agreement, given that inter-pathologist agreement rates are typically 85%. Thus, unless the patch was already centred on tumour cells, the saccade model displayed a strong tendency to seek out tumour tissue in regions adjacent to the original patch.

1.3.7 Feedback Attention Model Performance in WSI Pipeline

Chapter 8 revisits the weighted regular sampling pipeline. The original VGG19 classifiers were replaced with FAL-CNN and saccade models, to measure the impact on ROI and TSR predictions of using attention-enhanced CNNs.

This resulted in improved accuracy when calculating TSR at the pipeline's estimated maximum tumour density location. With the FAL-CNN as the 9-class classifier and a further 2-class FAL-CNN dedicated to TSR calculation, a mean error of 18.86% ($\pm 1SE$ 17.68 to 20.04% with 5-fold CV) was calculated. The original VGG19 configuration gave 21.64% ($\pm 1SE$ 20.06 to 23.22%).

At pathologist-selected GT locations, TSR error with a 2-iteration FAL-CNN was 6.84%, compared to 7.57% with the VGG19. This result facilitates application of the model in a pathologist-guided TSR-sampling tool, which would rapidly sample and classify multiple patches around a chosen WSI location, returning an estimated TSR value for use in survival prediction.

2 Background

2.1 AI with Attention

Attention is a process of selecting information from a larger volume of input data (van der Velde and de Kamps, 2001). This can be object-based or feature-based attention. In object-based attention, a whole object is extracted from its background by integrating salient features into a single perceived unit. Feature attention, by contrast, emphasises colours or shapes of objects being sought, such that these features are boosted across the input scene, before being resolved into objects.

Motter (1994) discovered that primates' higher ventral regions (V4) responded most strongly to stimuli when features such as colour or luminance matched those of a prior cue, following a delay of about 200ms, suggesting active feature selection involving a feedback cycle.

Colby and Goldberg (1999) determined that visible space around an observer is represented multiple times in their parietal cortex, which encodes objects and their locations. The authors argue that processing stimuli from sensory inputs to motor outputs does not require a single central representation of the space surrounding the individual. Instead, multiple representations of space are held in the parietal cortex, contributing to an internal assessment of salience which guides spatial attention and spatial memory, and may determine visual saccades or other motor outputs.

De Kamps and van der Velde (2001) built on these studies, showing that object-based attention in primates involves feedback processes in the ventral visual stream. Feedback activations cause disinhibition in lower areas of the visual stream processing features characteristic of a target object. This selectively modulates the level of feedforward activations from representations of the object near the input, thus guiding spatial attention and allowing actions such as saccades to be directed towards the target.

This approach was recreated in the neural networks developed for this thesis, which attends preferentially to objects of the target class via feedback activations, resulting in significant accuracy gains.

2.2 Feedforward Attention

The object-based attention described above contrasts with the attentional processes widely used in existing models. These predominantly use feedforward attention, where regions of an input image are preferentially selected for analysis using a mask, or other biasing signal, derived from the original input. In human vision, this allows objects of prior interest to stand out from their surroundings (Connor et al., 2004), before higher brain regions determine an interest in the object.

This process is purely bottom-up, with no feedback yet from executive brain regions, and can be emulated in a feedforward-only convolutional neural network (CNN). CNNs such as ResNet (Wang et al., 2017) contain feed-forward pathways only. Pixels of the input image are combined by a series of weighted additions and thresholding operations, into low-level features describing colours and textures, then into more abstract, high-level features representing the identity of objects in the original scene. Such models support the addition of convolutional modules to implement feedforward attention.

Woo et al (2018) developed a Convolutional Block Attention Module (CBAM) which can be inserted into a CNN to provide feedforward attention. This was implemented with an initial *channel attention* stage, where a per-channel attention map is derived from the input and used to control the relative levels of the parallel convolutional channels in the model. The next

stage in the CBAM generated a *spatial attention* map to boost certain regions of the image, across all channels. The CNN, with the embedded CBAM, was trained to optimise the goal (e.g. classification) of the outer model. Thus the CBAM learned to emphasise regions and channels of the input feature map that were relevant to the output class prediction. Testing a CBAM within a standard ResNet50 CNN (Wang et al., 2017) in the ImageNet classification challenge (Russakovsky et al., 2015) reduced the ResNet's top-1 error from 24.56 to 22.66%. This is a noteworthy improvement, but confidence intervals are not supplied to confirm significance.

Transformer

The Transformer architecture (Vaswani et al., 2017) was initially used in Natural Language Processing (NLP) for text translation from English to German and French. Previously, recurrent encoder-decoder networks mapped sequences of input symbols into a predicted output sequence. The Transformer implemented encoder and decoder using stacked attention mechanisms to identify relationships between words. Each attention module used scaled dot-product attention, combining key, query and value (QKV) terms to generate an output attention vector. The Transformer uses multiple Self-Attention (SA) modules, in which Q, K and V are all derived from the input vector, to generate an output vector representing important relationships within the input sequence. The Transformer achieved a BiLingual Evaluation Understudy (BLEU) score 7% higher than the previous state of the art (SoA). However, the attention modules have a complexity of $O(n^2)$ for a sequence length n, limiting the length of input and output texts to approximately 25,000 words.

The Vision Transformer (ViT) uses the Transformer principle in image classification (Dosovitskiy et al., 2020) and significantly outperformed SoA models in the ImageNet challenge. To mitigate the quadratic increase in model size with number of pixels, the input image was divided into 16x16 patches, which were processed similarly to words or sentences in a larger text corpus. Multiple attention modules are trained to recognise relationships between regions in the image, such as facial features, that contribute most strongly to the output prediction.

Current leaders in the ImageNet challenge combine models such as ViT or EfficientNetV2 (Tan and Le, 2021) with ensemble training approaches, such as Model Soups (Wortsman et al., 2022) or Meta Pseudo Labels (Pham et al., 2021).

2.3 Feedback and Top-Down Attention

Feedback using blackboard model

The blackboard design pattern was originally proposed for the Hearsay speech recognition system (Erman et al., 1980). Multiple agents independently interact with a central knowledge base, analogous to individuals taking turns to write on a blackboard to solve a shared problem. An overseeing monitor agent schedules the contributing agents and collates the most useful results. In de Kamps and van der Velde's blackboard models (2015; van der Velde, 2018), such agents combine distributed cognitive features into a bound perception, for each object of interest in a visual scene. This model is cited in recent work (Wiggins, 2020) on information dynamics in consciousness and creativity, and was used in Harrison's feedback-based neural models (2012) which have inspired the work in this thesis.

Harrison presented a dynamic architecture for neural networks (NNs), for modelling human attention using feedback pathways. A conventional feedforward NN was trained according to requirements, using traditional backpropagation and gradient descent methods. Each node could then be replaced with a dynamic circuit, whose disinhibition was controlled by feedback from higher layers in the NN. Effectively, the dynamic nodes acted as a common blackboard, where data from forward and reverse streams were combined.

A similar approach is adopted in this current work, where activations from multiple iterations of a feedback model are combined in working memory to generate an optimum class prediction.

Goal-Directed Attention

Luo et al (2021) proposed a classifier exhibiting Goal-Directed Attention (GDA). Their model was based on a conventional VGG16 backbone (Simonyan and Zisserman, 2014), with an additional attention layer which was trained to respond preferentially to predetermined target classes. In training, the loss term was weighted more strongly for output classes in the target vector. Per-channel weights in the attention module were optimised while those in the backbone model were frozen.

Luo hypothesised that the model paid attention to channels in the CNN that were most informative about the target classes, and was able to identify images of the target class amongst other object types or confusing backgrounds.

This increased sensitivity came at the expense of reduced selectivity; the trained model exhibited a static bias toward pre-determined object types, and was prone to false-positive predictions of these types.

Although GDA is driven by executive decisions about which classes are of interest, *it does not have a feedback pathway* and cannot dynamically modify the network's behaviour in response to features of the input image. Also, there was no spatial component to the learned attentional behaviour.

Nonetheless, Luo's architecture demonstrated a reproducible way in which attentional components can be incorporated into an existing feedforward CNN (Figure 2), facilitating the development of feedback models for this thesis.

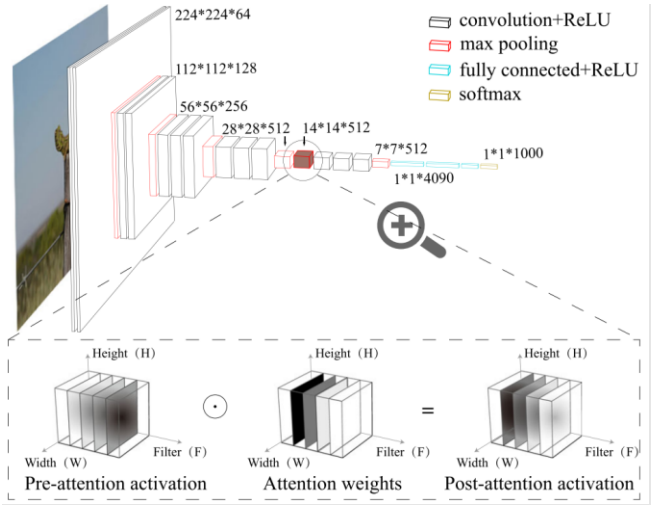


Figure 2: Goal-Directed Attention module in feedforward CNN, showing effect of attention weights on channel activations (Luo et al., 2021)

Network Pruning by Feedback

Cao et al (2019) used feedback in a CNN to implement weakly-supervised visual localisation and segmentation, using only category-level labels. The model was based on a pre-trained VGG16 (Simonyan and Zisserman, 2014), which was used to predict the class label for the input image. A neural pathway pruning algorithm was then used to select neurons associated with the class prediction. This was achieved by optimising neurons to maximise the confidence of the prediction, turning off neurons that contribute the least. Then, backpropagation from the

target neuron to the image space, via the pruned network, was used to generate a spatial energy map to highlight regions of the input that contributed most to the classification.

After several iterations of top-down pruning and optimisation, the energy map consolidated into an effective object localisation output, allowing Cao's model to outperform the state-of-the-art in PASCAL Visual Object Class object localisation challenges (Everingham et al., 2010).

Crowd Counting

Sam and Babu (2018) used top-down feedback in a CNN for counting people in photographs of crowd scenes, via the generation of a crowd density map. In a typical feedforward CNN, crowd-like patterns in the scene caused many false detections. The authors used a feedback (top-down) pathway to generate an attention mask as a correcting signal. This was applied to lower layers of the feedforward (bottom-up) network using multiplicative gating, to favour regions of genuine crowds.

This approach gave lower Mean Absolute Error (MAE) and Mean Square Error (MSE), than previous state-of-the-art models, despite the feedback model having a smaller parameter count.

If parallels are drawn between identifying heads in a crowd scene, and precognising cells or nuclei in a tissue sample, it appears that a CNN architecture with top-down convolutions and multiplicative attention gating has potential value in digital pathology applications.

CORnet and Brain Score

Kubilius, Schrimpf et al (2019) at Massachusetts Institute of Technology (MIT) developed *CORnet*, a family of recurrent CNNs that closely model the primate ventral visual stream, according to the team's own *Brain Score* criteria for neural and behavioural similarity (Schrimpf et al., 2020).

For the Brain Score, neural similarity between brain and CNN was assessed by comparing activations in response to a given input image. Functional magnetic resonance imaging (fMRI) was used to measure transient increases in blood flow in ventral regions. This data was correlated with activations observed in the corresponding layers of the CNN when processing the same image.

Behavioural similarity was judged by comparing image classification accuracy between human subjects and the CNN model, particularly when presented with complex visual scenes. A high combined Brain Score was found to correlate with high top-1 performance in the ImageNet classification challenge.

CORnet used a shallow CNN (Figure 3) with modules corresponding to each main region in the ventral stream (V1, V2, V4 and inferior temporal, or IT). Each module contained a feedback pathway, such that module input activations were gated by a signal derived from the module output. This recurrent behaviour was invoked several times in each execution of the model. Between two and four feedback iterations gave the best trade-off between classification accuracy and inference time.

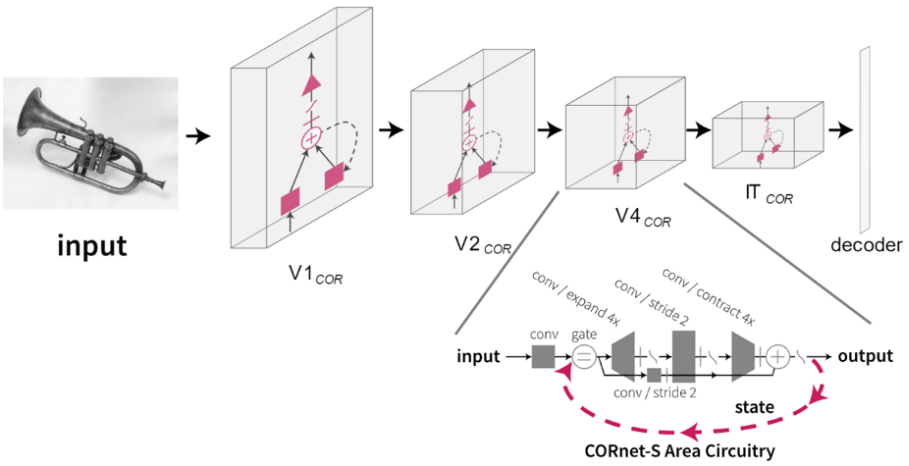


Figure 3: CORnet-S shallow CNN with feedback pathways within convolutional modules (Kubilius, Schrimpf, Kar, Hong, et al., 2019)

The use of feedback led to a higher Brain Score in terms of temporal dynamics, where behavioural accuracy was seen to increase with time from image onset as feedback loops were executed. This correlated with similar behaviour observed in primate brains, particularly when viewing complex images blended from objects at unusual orientations with incongruous backgrounds.

Segmentation Models with Feedback Attention

Tsuda et al (2020) added feedback attention to U-Net models, which are traditionally used for image segmentation (Ronneberger et al., 2015). The enhanced model was used for cell image segmentation in electron microscopy images of larval nerve tissue.

Tsuda used a feedback connection from the final convolution layer in the U-Net's decoder, to the first convolution block at the encoder input (Figure 4). The symmetrical U-Net architecture meant that these layers had the same dimensions ($256 \times 256\text{px} \times 64$ channels), allowing the connection to be made without further transformations.

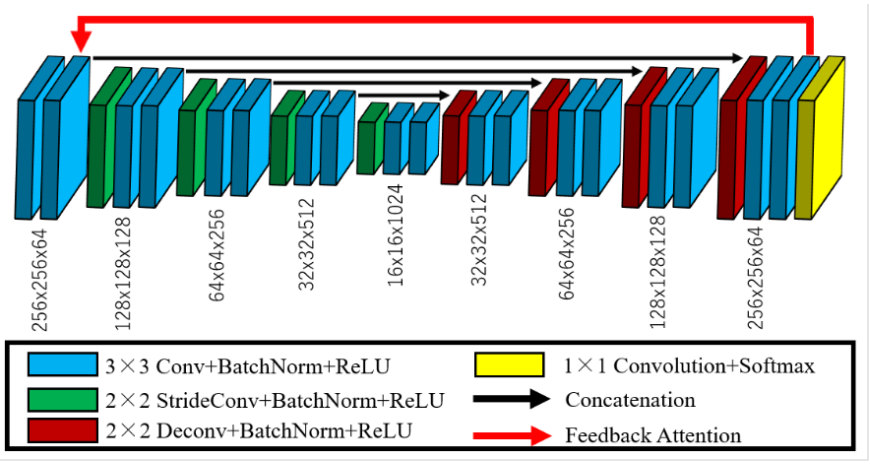


Figure 4: U-Net model with single top-to-bottom feedback attention path (Tsuda et al., 2020)

Two main attentional methods were compared in this model. One used Source-Target Attention (STA), where the input was used as the query term, while the key and value were derived from the output. The second model used self-attention (SA), where query, key and value were all obtained from the output feature maps and the attention map was then combined with the input feature maps by vector addition.

Tsuda's group then compared STA and SA with simpler methods of incorporating the feedback signal, such as vector addition, or a Light Attention Mechanism (Hiramatsu and Hotta, 2020) where spatial attention was controlled by a small feature-extracting module with two convolutional layers.

Self-attention was found to be the most effective feedback type, with a mean IoU of 73.64% relative to the ground-truth segmentations, against 72.72% for STA, 72.56% for Light Attention, 71.81% for additive feedback, and 70.09% for the basic feedforward U-Net. The simpler feedback methods were nonetheless effective, and although marginally less accurate than SA and STA, they do not suffer from the higher memory demands of these types.

These findings informed the design of novel feedback attention models in Chapter 5 of this thesis. It was possible to train more complex feedback models, with multiplicative attention at multiple levels in the encoder, than would have been possible using the computationally expensive STA or SA to process the feedback attention activations.

The FANet (Tomar et al., 2022) represents a similar segmentation model to Tsuda's. An attention mask at the input is iteratively updated, starting from cellular regions determined by Otsu thresholding. Examples were trained with images of colorectal polyps, skin lesions and retinal blood vessels. The performance, measured by F1 (Dice) score and Intersection over Union (IoU), was marginally improved over comparable U-Net derived segmentation models.

Work in this thesis (Section 5.1) combines techniques from feedback-enhanced U-Nets and MIT's CORnet into a novel recurrent Feedback Attention Ladder CNN architecture, FAL-CNN.

2.4 AI and Attention in Medical Imaging

False Positive Detection

False positive detection (Hong et al., 2020) emulates the attentional behaviour of scanning a scene for objects of interest, in which a sequence of high-resolution glimpses is guided by potential matches identified in peripheral vision. Further scrutiny at locations of interest allows incorrect initial matches to be rejected.

Hong used this approach in a two-stage neural network, to identify white matter hyperintensities (WMH) in magnetic resonance imaging (MRI) brain scans of migraine patients. In the first stage, a CNN was used to identify likely WMH locations. The model was trained for high sensitivity, to ensure that as many WMHs as possible were detected, even at the expense of more false positive results.

The output of this CNN was then used as an attention mask for MRI data fed into the second network, which was trained to distinguish true WMH regions from false positives. The use of masks directed the second-stage CNN towards locations of interest, avoiding the need for this CNN to process and eliminate tissue structures from other areas of the brain section. This process improved diagnostic accuracy and computational efficiency.

A similar approach was adopted in WSI processing pipelines in the current work (Sections 4.2 and 4.6), to reject false-positive tumour patches in histopathology image analysis.

Attention-Weighted Segmentation

Oktay et al (2018) proposed an Attention Gate (AG) model, incorporated into a U-Net segmentation model trained to identify the pancreas in abdominal CT scans. The AG model automatically learns to focus on salient target structures. It was trained to suppress irrelevant regions and to highlight salient volumes in a 3D input tensor, eliminating the need for explicit localisation modules around the CNN. The Dice Similarity Coefficient (DSC) for pancreas

segmentation was reported as 0.840±0.087, a significant improvement over state-of-the-art methods for this task.

Schlemper et al (2019) demonstrated the use of AGs in foetal ultrasound screening. Soft attention, using a blurred mask instead of hard cropping of image regions, was chosen for being end-to-end differentiable, allowing the CNN to be trained using standard back-propagation methods. Attention-gated U-Nets informed the development of feedback attention methods in this thesis, which highlight salient areas of the input image for further analysis.

2.5 AI and Attention in Histopathology

Google AI Healthcare (Liu et al., 2019) reported successes in detecting metastatic breast cancer in lymph node biopsies. A 99% Area Under Curve (AUC) at WSI level was published, with 91% at tumour level. This was similar to scores achieved by expert pathologists. The algorithm used randomly selected 128x128 pixel patches, in parallel with larger 299x299 patches for context. The trained Inception 3 networks took "under a minute" to process a WSI on a cloud computer platform, although it was not clear what parallel GPU resources were deployed to achieve this.

The multi-resolution approach was seen again in a study of AI in bladder cancer prediction (Harmon et al., 2020), where "spatially resolved prediction maps" were combined with lymphocyte infiltration features to give each patient a probability score for lymph node metastasis. The model outperformed a clinicopathologic model based on lymphovascular invasion, age and T-stage. The latter gave an AUC of 0.755 (95% CI 0.68 to 0.831), against 0.866 (95% CI 0.812 to 0.92) for the AI-derived prediction score. Data was sourced from the National Cancer Institute and the Cancer Genome Atlas (TCGA).

It should be noted that the above results are on a per-slide level. At a patch scale, accuracy in the GT classification can be limited by pathologist-pathologist agreement rates of approximately 85%, limiting the accuracy possible using models trained on annotated patch data.

Working at the scale of a single 128×128 pixel patch, Jiang et al (2022) proposed a modified Vision Transformer (Dosovitskiy et al., 2020). Jiang's implementation was optimised for tumour cell segmentation in pathology patch images, using an iterative process to reposition 16 mini-patches over cell boundaries, thus focusing on information most relevant to the segmentation result. The team reported only marginal increases in F1 score and IoU relative to SoA models, but nonetheless demonstrated a mechanism for iterative resampling that was transferrable to multiple data sources.

This thesis further explores iterative resampling in a saccade-like sequence, as a tool for locating salient tissue and thereby for validating models that appear to highlight such regions.

Other literature used attention to determine where to inspect for diseased tissue in a WSI:

A Novel Approach to Mitotic Figure Detection in Breast Cancer Histopathology Images using Region Based Convolutional Neural Networks (Rao, 2018)

Rao et al proposed a novel Region-based CNN (RCNN) to grade the rate of cell division (mitosis) in histopathological images, currently a time-consuming job for pathologists.

Performance was expressed using the F1 score, or DSC, chosen for its sensitivity to false positives. Here, true positives were defined as predictions of mitosis that agree with pathologist scores. Rao reported an overall F1 score of 0.955 across several challenge-based data sources, a 6% improvement over previously published models.

Rao's solution used a variant of the Faster-RCNN model, tuned to detect small objects in a 299x299 pixel frame. A region proposal network selects small regions of interest (ROIs), which are classified by a separate detection network based on a pre-trained VGG16 model (Simonyan and Zisserman, 2014). Some convolution layers were omitted, where these represented coarse, abstract-level features that had been shown to limit the minimum detectable object size to 44px at 0.25μm/pixel. This was previously problematic as mitoses were typically only 7-8 microns in size, or 30px at this magnification.

High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection (Cruz-Roa et al., 2018)

Cruz-Roa et al developed an attention-inspired model using adaptive sampling to select patches of interest from breast cancer WSIs, to avoid the need to process all image tiles systematically. This reduced the number of patches required for a reliable classification from 6 million to approximately 2,000, reducing processing time from 24 hours to one minute. The mean DSC was reported as 76%, for the binary classification of cells into invasive/non-invasive types.

The CNN classifier was trained on tiles from whole slide images. Tiles were grouped as inside or outside an expert-annotated region of invasive breast cancer on each slide. The trained classifier was used in an attention-like process, where patches were sampled more densely from the WSI in regions where the model reported high uncertainty of whether the tissue is invasive. This was observed at the boundary between tumour and healthy tissue, where computational resources were therefore focussed in order to estimate the boundary position more precisely.

The optimum sampling distribution was achieved iteratively, starting with patches sampled in a quasi-random pattern. The patches were fed into the CNN classifier to obtain a probability of invasive breast cancer in each location. These were interpolated to provide a pixel-wise probability heatmap. The gradients within this output represented areas of the WSI with a rapid transition between types of tissue. In these areas, a denser sampling distribution would be applied. This whole process was repeated for a fixed number of cycles, or until the calculated Dice coefficient converged.

The CNN used fewer layers than state-of-the-art classifiers optimised on ImageNet. A convolution layer supplied activations to a subsampling pooling layer, before the image features were passed into a fully connected layer feeding two output neurons signalling invasive and non-invasive cells. This suggests that simple image features are sufficient to drive this distinction, and that a more elaborate model risks overfitting.

Cancer Metastasis Detection via Spatially Structured Deep Network (Kong et al., 2017)

Kong et al applied an attention-like approach to breast cancer metastasis detection in lymphatic tissue. Their *Spatio-Net* architecture used a 2D Long Short Term Memory (LSTM) to process spatial sequences of image patches. LSTMs (Hochreiter and Schmidhuber, 1997) are sequential neural networks, often used for sequence prediction in NLP. In Kong's spatial LSTM, the classification of each patch was informed by the class distribution of surrounding patches, expressed as a two-dimensional sequence.

Kong's proposed architecture was based around a 101-layer ResNet CNN (He et al., 2016). This was used to extract a fixed-length vector of image features from each 256×256px tile of the WSI in turn. Kong's novel addition was to use the 2D LSTM to combine the image feature

vectors from a 3x3 grid of neighbouring tiles, allowing larger-scale spatial dependencies to be incorporated into the final calculation of the probability of tumour in each tile.

The average Free-Response Operating Characteristic (FROC) was given as 0.7539 (standard deviation 0.008), compared to a value of 0.7012 for a ResNet-101 without the spatially structured processing. Data for training and testing came from the CAMELYON16 Grand Challenge (Ehteshami Bejnordi et al., 2017).

Multi-Instance Learning

Multi-Instance Learning (MIL) in digital pathology removes the need for multiple expert annotations within each WSI, in pipelines where only a single output per WSI is sought.

The WSI is divided into patches, which are processed individually to generate patch-level representations. These are aggregated and collectively processed into a single representation for the WSI (Gadermayr and Tschuchnig, 2022). Patch processing can be performed by a CNN such as ResNet (He et al., 2016), pre-trained on public data such as ImageNet. This yields embedding vectors that encode salient characteristics of each patch and can be used to train and evaluate the per-WSI patch aggregation elements.

Godson et al (2022) use a variant of this approach for classifying melanoma WSIs into genomic immune subgroups. The aggregation mechanism includes an attention backbone, which calculates patch attention weights that can be plotted in a WSI-scale heatmap showing tissue regions that most strongly indicate the subgroup class.

Many MIL algorithms sample and process all tiled patches in a WSI, which is a computationally expensive process. Breen et al (2023) introduce Discriminative Region Active Sampling for Multiple Instance Learning, which uses attention scores to guide patch sampling towards the most discriminative regions of the WSI. Breen reports a similar AUC to that of standard MIL architectures, but with a 3-fold reduction in processing time.

While this current work explores alternative approaches to sampling patches from WSIs and generating heatmaps of salient tissue, MIL is of interest for future substitution into WSI pipelines. Feedback enhanced CNNs (Chapter 5) are suggested for us in the MIL because they have the potential for generating more accurate patch embeddings than the ResNet.

2.5.1 Tumour Stroma Ratio

The tumour stroma ratio (TSR) is a significant prognostic factor in the treatment of cancer. *Stroma* refers to structural or connective tissue within an organ. In the tumour microenvironment, stroma cells support tumour epithelium and can influence disease progression. Van Pelt et al (2018) report that cancer-associated fibroblasts (CAFs) release growth factors that promote tumour growth. The team have developed guidelines for choosing tissue samples most representative of this infiltration. Slides should generally be taken from the most invasive part of the adenocarcinoma, rejecting samples including necrosis, muscle and large blood vessels. Pathologist performance can vary with ocular quality and magnification, and the quality of H&E staining, factors which are also of concern in AI-augmented image processing.

West et al (2010) examined the role in cancer progression of the relative numbers of epithelial and stromal cells in colorectal cancer. They measured the proportion of tumour (PoT), finding this to have better prognostic performance than TSR. Tissue was sampled from within a $9mm^2$ region near the luminal aspect of the tumour, its boundary with the main lumen or interior of the bowel. Within the square, 300 (+/- 15%) sampling locations were selected using

RandomSpot systematic random sampling (Wright et al., 2015), where sampling points were arranged in a hexagonal grid with a random starting point to minimise sampling bias.

West et al found that tumours with $PoT \leq 47\%$ were associated with significantly lower cancer-specific survival, with a Hazard Ratio (HR) of 2.087, 95% CI = 1.088-4.003.

The "*QUick And Simple And Reliable*" (QUASAR) trial was a randomised study of adjuvant chemotherapy in colorectal cancer (QUASAR Collaborative Group, 2007). The trial evaluated the additional survival benefit from chemotherapy, in patients who were deemed to have a lower risk of disease recurrence following surgical resection.

QUASAR data was further used in a study of stroma density in the tumour ROI (Hutchins et al., 2018). High tumour stroma (>50%) was associated with increased rates of disease recurrence, at 31.3% vs 21.9% for stroma levels below 50%. With stroma levels above 65%, 40% of patients had recurrent disease within 10 years.

Deep Learning for TSR

A retrospective study in Heidelberg, Germany, (Kather et al., 2019) used deep learning to predict colorectal cancer survival from histology slides. Their CNN model was trained on 100,000 224×224px patches from 86 slides, from the *NCT-CRC-HE-100K* set gathered from biobanks at Heidelberg and Mannheim. Nine tissue classes, *adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma*, and *COAD epithelium* were used, as defined in TCGA. These classes differ from those of the QUASAR dataset. Various popular CNN models were tested, of which VGG19 (Simonyan and Zisserman, 2014) was found to be the most accurate at 98.7%.

Kather's team evaluated hazard ratios (HR) for each tissue class, in relation to survival data. Tissue classes *adipose, debris, lymphocytes, mucus* and *stroma* resulted in $HR > 1$. Notably, *tumour* is missing from this list, even though the analysis was carried out in a region that was previously annotated as tumour. As with TSR, it appears that the concentration of *non-tumour* cell types within the tumour region is prognostic of survival rates.

Zhao et al (2020) extended Kather's work, using a CNN-based pipeline to estimate TSRs from colorectal cancer slides. They employed transfer learning, taking a CNN pre-trained on ImageNet then on Kather's NCT-HE-100K dataset. The model was then applied tile-by-tile to generate a rough segmentation of the WSI, from which tissue class ratios were calculated.

WSIs were sorted into stroma-high and stroma-low categories, with a threshold of $TSR = 48.8\%$ chosen using the maximally selected rank statistics method (Hothorn and Lausen, 2003). Overall survival was found to be significantly reduced in stroma-high patients, with a HR of 1.72 (95% CI 1.24-2.37), again showing tissue class ratios to be useful prognostic factors in colorectal cancer care.

2.6 Summary

The work in this thesis aims to fill the following gaps that have been identified in the current literature:

2.6.1 WSI patch sampling algorithm supporting ROI and TSR calculation

The literature reveals the need for an efficient sampling regime for extracting patches from a WSI, in a pattern that supports accurate calculation of ROI and TSR. Tile-by-tile methods are slow; *Quasi Monte Carlo* methods for iterative resampling (Cruz-Roa et al., 2018) are suitable for estimating ROI but do not supply enough sampling points inside the ROI for other

calculations such as TSR. Novel, alternative attention-based sampling methods are therefore proposed in Chapter 4 and further explored in Chapter 8.

2.6.2 Feedback attention CNN architecture

Several studies reviewed in this chapter conclude that top-down attention enhances the performance of CNNs in classification and segmentation tasks, particularly where the background is cluttered or heterogeneous. Kubilius and Schrimpf (Kubilius et al., 2018) propose brain-like local feedback, within convolutional scale-groups corresponding to V1, V2, V4 and IT regions in the ventral stream. Other researchers (Tsuda et al., 2020; Tomar et al., 2022) confirm the benefit of a top-to-bottom feedback path in U-Net segmentation models, although these only use a single feedback connection to the input layer of the encoder.

The work in this thesis builds on these concepts, proposing a novel Feedback Attention Ladder CNN (FAL-CNN) with multiple local and top-to-bottom feedback circuits that all contribute to the model's accuracy and stability (Chapter 5). Spatial visualisations of the model's attention layers contribute to an understanding of its object-detecting abilities (Chapter 6), which are then exploited in a saccade-like resampling mechanism (Chapter 7) with additional object-tracking and tissue-locating behaviours. This novel model can track to salient objects that are initially out of frame, and locates tumour regions in pathology patch images.

3 Data

3.1 Introduction

This chapter describes the derivation of data sources used in experiments in this thesis. Annotated colorectal cancer whole-slide images from the QUASAR trial (Gray et al., 2007) were used to train and evaluate WSI-processing pipelines (Chapters 4, 8), and patch-scale classifier models incorporating feedback attention architectures (Chapter 5) and saccade-like behaviours (Chapter 7).

The ImageNet-100 dataset of general image classes (Russakovsky et al., 2015; Shekar, 2021) was used to test model generalisability (Section 5.4; Chapter 7), and to examine model behaviour in relation to readily identifiable image features (Chapter 6).

3.2 QUASAR Pathology Images

3.2.1 Introduction

In the QUASAR trial, samples of surgically removed colorectal cancer tissue were mounted, stained with Haematoxylin and Eosin (H&E), then scanned with a Leica Biosystems Aperio XT scanner system. Slides were scanned at 0.49μm per pixel with JPEG 2000 compression at 49.09 compression ratio and a quality factor of 30 (Hutchins et al., 2018). Whole slide images (WSIs) were saved in Aperio SVS file format, a pyramidal TIFF (Tagged Image File Format) that contains a hierarchy of images from full magnification down to a 16x down-sampled thumbnail.

The QUASAR slide images were later labelled by pathologists at the University of Leeds, evaluating stromal morphometry as a tool for predicting disease recurrence (Hutchins et al., 2018). The following annotations were of interest in this current work:

1) Region of Interest (ROI) outlines, showing the main area of cancer in the WSI.
2) Pixel coordinates of locations where cells have been classified by type.

3.2.2 Region of Interest Annotations

An example of the expert-annotated ROI is shown in Figure 5. A pathologist has used Leica ImageScope software to draw an outline around the main area of cancer tissue. This region is mainly made up of tumour cells, which show up as a darker blue-purple due to the higher concentration of haematoxylin-stained cell nuclei.

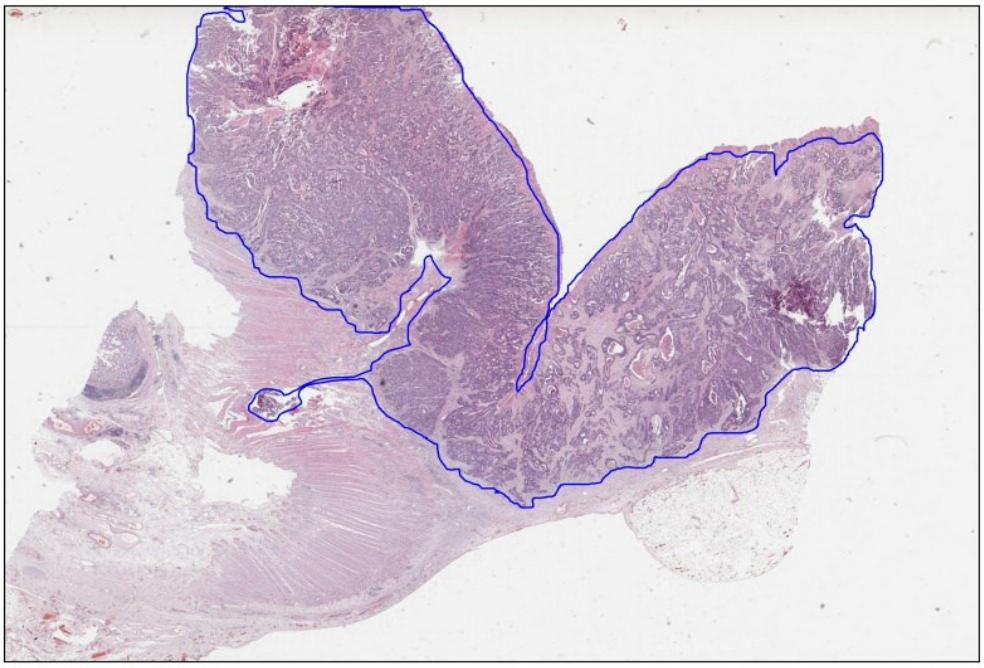


Figure 5: Colorectal cancer section with expert-annotated region of interest (blue outline)

Other diagnostically important tissue types, that contribute to the structure of the cancer, include stroma (connective tissue), lumen (the space enclosed by the epithelial lining of the colon), blood vessels, mucin and necrosis (dead tissue).

The proportions and structure of these tissues are important in assessing the stage and severity of the cancer. For example, the proportion of tumour to stroma has been associated with outcome in some cancers (Hutchins et al., 2018).

The tumour ROI outline was stored as a polygon, with vertices in X, Y pixel coordinates at the maximum WSI resolution, in an eXtensible Mark-up Language (XML) file for each WSI.

3.2.3 Cell Classification Annotations

Pathologists recorded the predominant cell type at sampling points within the ROI, within a 3mm × 3mm box placed by the pathologist where the density of tumour epithelial cells was perceived to be highest. Within this box, a grid of sampling locations was determined using RandomSpot software by Wright (2015) to place points in an evenly distributed hexagonal grid.

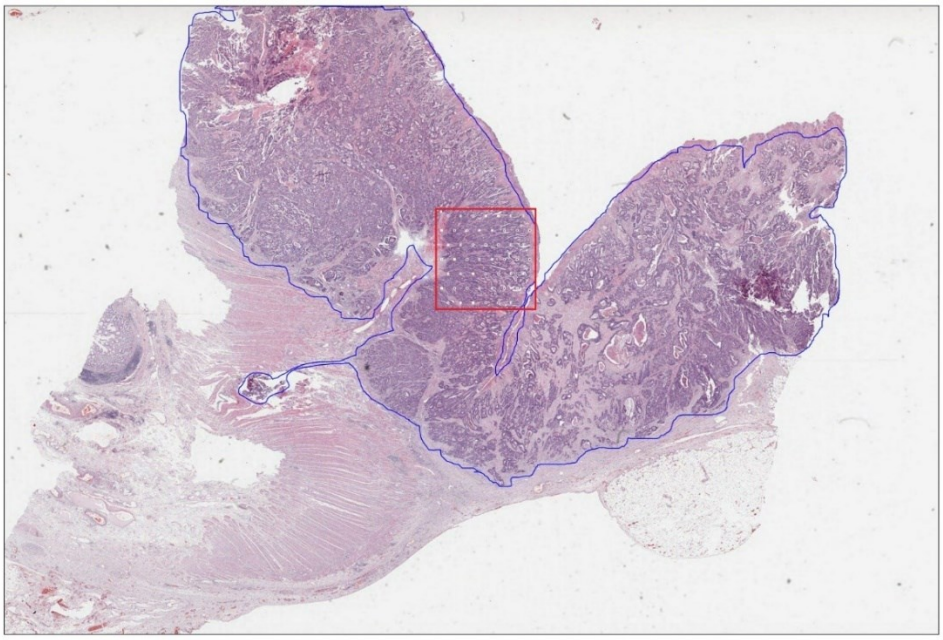


Figure 6: WSI with overlaid ground-truth annotations.

RandomSpot-based sampling points are in red box, over tissue with highest tumour density.

Thus 50 (±15%) sampling points were defined for each WSI. At each location, the surrounding cellular structure was examined to classify the location as *tumour, stroma or fibrosis, necrosis, vessels, inflammation, lumen, mucin, muscle* or *non-informative* (Figure 7). Class identifiers and sampling coordinates were stored in an additional XML file per WSI.

The classification of *tumour* was assigned to *tumour epithelial* cells within the cancer region, representing tumour cells which have mutated from normal epithelial cells in the bowel. The QUASAR dataset does not contain annotated examples of *normal epithelium*, as the ROI was deliberately defined to exclude this tissue. AI-based processing steps were implemented in Section 4.4.2 to distinguish these tissue types.

3.2.4 Patch Extraction

A whole slide image can contain several billion pixels. In contrast, effective neural network models in image classification operate at far lower input image sizes, such as 224 x 224 pixels for VGG19 (Simonyan and Zisserman, 2014) It was therefore necessary to extract multiple small image patches from each QUASAR WSI, to train CNN models at this scale.

Patches were extracted from the WSI using coordinates in the 'box-scores' XML files containing the ground truth cell classifications. Each patch image was cropped from the WSI using the ***OpenSlide*** Python library, then saved into a subdirectory according to its GT class.

Algorithm for patch extraction process

```
For each SVS file (WSI) in input directory:
      Locate and open the corresponding box-scores XML file.

      For each sampling location in XML collection:
            Parse vertices for centre of sampling point
            Parse ground truth (GT) classification code
            Add Box object to list, storing position and class

      For each Box on list:
            Extract patch from WSI, centred on Box coordinates
            If patch does not overlap others of different class:
                  Save patch to subdirectory according to GT class
                  Create and save copies, rotated by 90°, 180° and 270°.
```

The above algorithm was implemented in Python and executed on a subset of 690 Quality Control (QC)-passed WSIs.

Each patch region was compared with regions already extracted, to exclude any that overlapped with patches with a different tissue class.

This was required when extracting from low-magnification thumbnail images, where each patch covers a larger area of the WSI. Here, overlapping training patches had a substantial impact on CNN classification accuracy. However, at full magnification in the WSI, the RandomSpot sampling locations were sufficiently well-spaced to avoid overlap at most patch sizes.

Patch extraction was carried out on ARC4, the University of Leeds's high-powered computing (HPC) resource, to satisfy data governance requirements. The resulting directories of patch images were retained here for subsequent use as CNN training data.

3.2.5 Quality Control (QC)

The 690 whole slide images used for patch extraction were a subset of the 2211 cases in the original QUASAR dataset.

Wright et al (2021) examined image defects that caused inaccuracies in automated calculations of tumour stroma ratio (TSR). The images associated with the largest errors revealed faults issues in staining, mounting and scanning, defining criteria for selecting 'good' WSIs. Most rejections were due to weak or faded staining, for both haematoxylin and eosin dyes. Other QC issues included poorly differentiated tumour, necrotic tissue, and folds, tears, bubbles and debris in the tissue section. The remaining 690 QUASAR WSIs were labelled as acceptable quality.

The QC-passed subset was used in this current work, anticipating that this would yield more consistent results when developing an experimental image classifier. Models for use in clinical settings would require further evaluation with patches from all available WSIs, to assess the robustness of the model with typical image data.

3.2.6 Magnification and the Importance of Context

When examining a tissue sample, pathologists often review the slide image at multiple levels of magnification. For QUASAR, cells were examined at full magnification RandomSpot-derived

locations. However, the human analysis would usually begin with a thumbnail, or fully-zoomed-out, view of the WSI.

Larger regions of suspected tumour can often be identified at this level, by their colour, shape and texture. These features then prompt a systematic examination at higher magnifications, viewing smaller-scale structures and ultimately arrangements of individual cells.

Figure 7 shows typical patches extracted from ground-truth sampling locations, for each of the cellular classes in the QUASAR annotation. At larger patch sizes, more structural features are revealed. These may be associated with a classification that is not evident at smaller patch sizes.

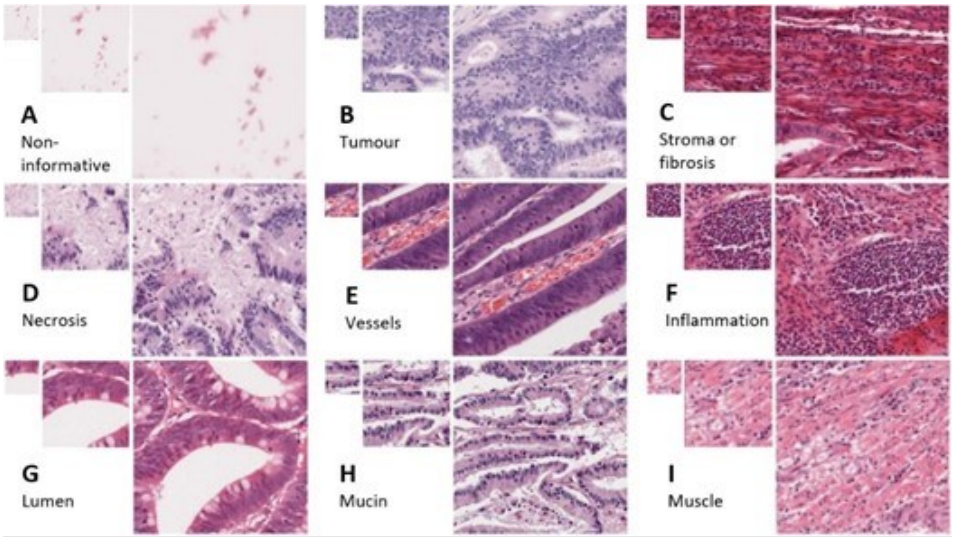


Figure 7: Patches extracted from WSI at ground-truth sampling locations.

Tissue types from L-R, top to bottom: Non-informative, tumour, stroma/fibrosis, necrosis, vessels, inflammation, lumen, mucin and muscle.
Sizes: 100 x 100px, 224 x 224px, 448 x 448px.

For example, a small patch may contain a blank area with the same brightness and colour as the slide background. Such gaps in cellular tissue may be caused by hollow structures, such as lumen or blood vessels, or the boundary of the tissue section. It is only at greater patch sizes that larger-scale structures are revealed and the correct classification can be assigned.

However, with increasing patch size, it becomes increasingly likely that the image will include a mixture of cell types, such as tumour and stroma, confusing attempts to classify the patch as a single type. Patch images were therefore extracted at multiple sizes from full-magnification WSIs and lower-resolution thumbnail images, as detailed in Table 1.

The quoted magnifications follow a common convention in pathology publications, originating in optical microscopy, where for example a 20× objective lens is combined with a 10× eyepiece to give a total magnification of 200×. With the Aperio scanners used in the QUASAR study, this corresponds to a resolution of 0.49μm per pixel.

Patch sizes were chosen as integer fractions or multiples of the input sizes of the CNNs to be evaluated, particularly 224px, 256px and 299px.

Magnification	*Patch height & width (pixels, µm)*		*Total size, GB*	*Notes*
20 ×	56px	27µm	0.47	
20 ×	64px	31µm	0.61	
20 ×	100px	49µm	1.3	Used in early experiments (Broad et al., 2020)
20 ×	112px	55µm	1.5	
20 ×	128px	63µm	2.0	
20 ×	224px	110µm	5.5	Input size of VGG, DenseNet etc
20 ×	256px	125µm	7.3	Input size of 4-layer ConvNet
20 ×	299px	147µm	9.6	Input size of Inception 3
20 ×	448px	220µm	22	
20 ×	512px	251µm	29	
20 ×	894px	110µm	7.6	Scaled to 224px to control image size
1.25 × (Thumbnail)	16px	125µm	0.24	Equivalent to 256px at WSI scale.
1.25 ×	29px	227µm	0.21	
1.25 ×	32px	251µm	0.21	
1.25 ×	58px	455µm	0.37	
1.25 ×	64px	502µm	0.36	Equivalent to 1024px at WSI scale.
1.25 ×	112px	878µm	0.88	
1.25 ×	128px	1,004µm	1.1	
1.25 ×	224px	1,756µm	2.8	
1.25 ×	256px	2,007µm	3.5	
1.25 ×	299px	2,344µm	4.7	For Inception 3. Equivalent to 4784px in WSI.

Table 1: Patch image sizes extracted from QUASAR WSIs.

Patch magnification for WSI sampling = 20x (by convention; effectively 200x), 0.49µm/pixel.
Thumbnail magnification = 1.25x, 7.84µm/pixel.

At the 894px × 894px patch scale, the images were scaled down to 224px × 224px at the point of extraction. This was necessary to reduce disk usage, which otherwise increases with the square of the patch dimension.

3.2.7 Data Governance

Patients in the original QUASAR trial gave written consent for their participation in the randomised controlled trial (RCT). At this time, patients had already undergone resection of colorectal cancer and were then randomly allocated to groups receiving chemotherapy, or observation only.

This current work is covered under NHS ethical approval ***REC 05/Q1205/220*** for analysis of digital pathology images.

Copies of the QUASAR images and XML annotations for this current work were temporarily stored on the ARC4 HPC at Leeds University, with the approval of Dr Treanor's team. The data directories are permission-controlled with access restricted to this researcher and project supervisors.

3.3 Further QUASAR-derived Patch Datasets

3.3.1 *uncertain-class-patches* Dataset

Motivation

To assess new CNN models against more challenging data, further datasets were derived from the QUASAR-based patches. An *uncertain-class-patches* image set was created, containing only patches where a classifier, trained on the 'parent' QUASAR 9-class dataset, reported similarly

high probabilities for more than one tissue class. This dataset was used to evaluate the Feedback Attention Ladder CNN (FAL-CNN) model introduced in Section 5.1.

Methodology

Code used in this section is documented in Appendix Section 1.1.1.

A QUASAR patch trained VGG19 model was used to output a vector of predicted class probabilities for each patch image in its input batch.

Patch image files were copied to a new *uncertain-class-patches* directory, in cases where the largest two predicted class probabilities fell within a given percentage threshold of each other.

This threshold was provisionally set to 25%, such that the patch would be copied if:

$$\frac{highest\ class\ probability - second\ highest\ class\ probability}{highest\ class\ probability} < 0.25 \qquad (1)$$

Algorithm for generating *uncertain-class-patches* patch image set

```
Load pre-trained VGG19 model from file
Initialise data loader with VGG19's original training split

Load Test set via data loader
For each mini batch in Test set:
      Perform model inference on mini batch
      For each patch image in mini batch:
            Select model output (class probability vector)
            Sort class probability vector by magnitude
            If top two class probabilities are within 25%:
                  Copy patch image to target directory
```

This code was executed on 224×224 pixel patches extracted from the QUASAR 9-class patch set described in Section 3.2.4.

Precautions against overfitting

Patches were used from the Test set of the data split whose Training set was used to train the VGG19 model used in the extraction process. Thus, when models were later assessed against the *uncertain-class-patches* set, the model avoided exposure during training to any patches that might also be copied to *uncertain-class-patches* for model evaluation.

Results

2,412 out of the 8,374 patches (28.8%) in the chosen test set were copied to the *uncertain-class-patches* output directory. Table 2 shows the file counts in each class sub-directory.

Table 2: File totals and class directory sizes for uncertain-class-patches output directory

Class	Number of files	Total directory size
0-non-informative	442	38MB
1-tumour	373	38MB
2-stroma-or-fibrosis	529	53MB
3-necrosis	277	28MB
4-vessels	103	11MB
5-inflammation	8	846kB
6-lumen	378	34MB
7-mucin	123	12MB

8-muscle	180	18MB

Examples of patches identified as *uncertain* are shown in Figure 8, captioned with the top two predicted class probabilities from the VGG19 model output.

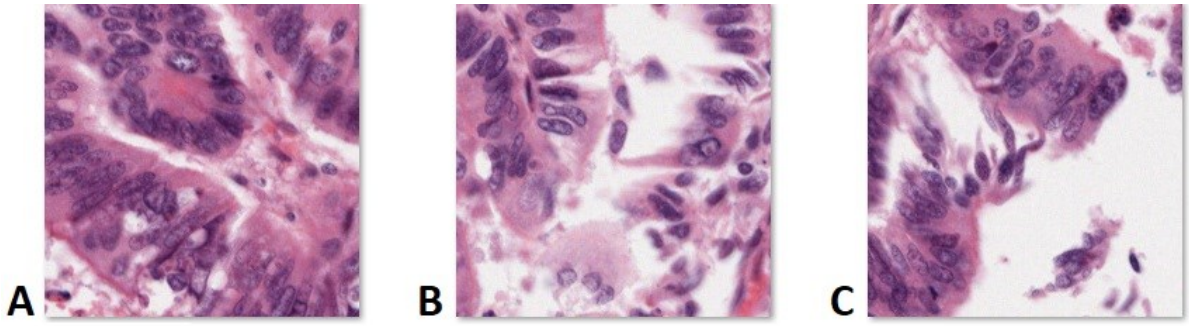


Figure 8: Examples of uncertain-class patch images, labelled as Tumour by pathologist.

A: VGG19 estimated high probability of 1-tumour *then* 0-non-informative.
B: VGG19 estimated high probability of 3-necrosis *then* 6-lumen.
C: VGG19 estimated high probability of 0-non-informative *then* 3-necrosis

Discussion

Figure 8 shows patches that were identified as *uncertain* despite being labelled as *tumour* by human experts. Patch A was correctly predicted as *tumour*, but with a strong secondary probability of *non-informative*. This is thought to be due to the heterogeneous cell types near the centre of the patch.

Patches A through C show increasing fragmentation, with more background becoming visible, perhaps due to tearing of samples prior to mounting. In C, the class has been incorrectly predicted as *non-informative*, consistent with the background being visible at the centre of the patch. It appears that broken and blurred fragments of tissue have led to the secondary predictions of *lumen* and *necrosis* in B and C.

These images are typical of those in the generated *uncertain-class* dataset. These would be used in later experiments as a source of heterogeneous, fragmented or superimposed tissue images, to test whether novel feedback models can discriminate these better than the VGG19 baseline model.

3.3.2 *tumour-stroma-groups* Dataset

Motivation

Wright (2017) established that classification of *tumour* and *stroma* tissue, and subsequent TSR calculations, was most accurate when grouping the eight informative QUASAR tissue classes into parent *tumour* and *stroma* groups. Patches of *tumour, necrosis, lumen* and *mucin* were grouped as *tumour*, while *stroma, muscle, vessels* and *inflammation* were interpreted as *stroma*.

A two-class dataset representing *tumour* and *stroma* groups was therefore created, to explore whether new models could achieve a higher classification accuracy using the simplified image classes.

Methodology

Code used in this section is documented in Appendix Section 1.1.2.

The grouped patch directories were created by copying from the QUASAR 9-class directory built in Section 3.2.4, using patches of size 224×224px. Files were copied from subdirectories listed in a shell script, into output subdirectories according to Wright's parent class groupings.

Where the number of patches in each output group subdirectory was below the number required, images were copied with rotations of 90°, 180° and 270° to make up the image total. Where the total exceeded the required number, patch images corresponding to the excess were randomly selected for deletion.

The shell script was executed twice on ARC4. One job was configured for 20,000 patches per output subdirectory, and one for 12,000. The source directory was copied from the QUASAR 9-class patch directory before rotated patches were deleted.

The 12,000 size was used to generate output directories of exclusively non-rotated images, to train models for comparison with those trained on offset patches (Section 3.3.4). There were sufficient patches of the source classes to yield a total of 12,000 images per grouped class without needing to make rotated copies.

Results

Table 3 and Table 4 show the totals of files copied from the source directories into the new grouped class directories, for 20,000 and 12,000 files per class respectively.

Table 3: Total files copied or deleted to create tumour-stroma-groups patch set

Source directories	Destination directory	Files copied per class	Total files copied	Files deleted	Rotated copies created	Files remaining
1-tumour 3-necrosis 6-lumen 7-mucin	tumour-group	16,101 1,816 2,900 436	21,253	1,253	0	20,000
2-stroma-or-fibrosis 4-vessels 5-inflammation 8-muscle	stroma-group	11,113 539 29 523	12,204	0	7,796	20,000

Table 4: Total files copied or deleted to create tumour-stroma-groups-12000 patch set

Source directories	Destination directory	Files copied per class	Total files copied	Files deleted	Rotated copies created	Files remaining
1-tumour 3-necrosis 6-lumen 7-mucin	tumour-group	16,101 1,816 2,900 436	21,253	9,253	0	12,000
2-stroma-or-fibrosis 4-vessels 5-inflammation 8-muscle	stroma-group	11,113 539 29 523	12,204	204	0	12,000

Discussion

Two balanced datasets have been created. The 20k-per-class set required 7,796 rotated copies to be generated in the *stroma-group* class, while the 12k-per-class set was generated without

rotated copies. It was anticipated that the 20k set would result in a higher model accuracy because of the larger data size, for reduced overfitting, and because of the regularisation effect of training with randomly selected rotated copies.

Meanwhile, the 12k set provided a baseline for comparison with similarly distributed offset patches (Section 3.3.4). Rotated copies were also omitted from the latter set, to avoid moving the ground truth pixel into an unexpected quadrant.

Although balanced in terms of patch totals in *tumour-group* and *stroma-group* outputs, these datasets implicitly contain an unbalanced distribution of tissue classes. This is due to the large variation in occurrence between the widely occurring tumour and stroma, and smaller classes such as inflammation and mucin.

3.3.3 Offset Patches

Motivation

Previous QUASAR-derived patches were extracted from the WSI to place the ground truth pixel at the centre of the patch. Later chapters examine whether distributions of feedback and feedforward activations track the ground truth location when it is moved relative to the patch boundary. To support this, further sets of training patches were extracted from QUASAR WSI with a constant positional offset, such that the GT class would apply at a non-central point. This data is used to evaluate the novel FAL-CNN in Section 5.2.

Methodology

As in Section 3.2.4, patch images were extracted from QUASAR WSIs at locations specified in the ground truth annotation XML, then written to subdirectories according to the annotated class. For the new dataset, the patch extraction code was modified to allow the required X,Y offset to be specified by an additional argument in the associated HPC shell script. The offset was added to the coordinates associated with the ground truth *box-scores* XML data, and the patch was sampled from the WSI using the offset location as the patch centre.

Algorithm for patch extraction process

```
For each SVS file (WSI) in input directory:
      Locate and open the corresponding box-scores XML file.

      For each sampling location in XML collection:
            Parse vertices for sampling point
            Add X,Y offset to sampling point
            Parse ground truth classification code
            Add Box object to list, storing offset position and class

      For each Box on list:
            Extract image patch, centred on offset sampling point
            If patch does not overlap others of different class:
                  Save patch to subdirectory according to GT class
```

The generation of rotated copies was disabled, to preserve the position of the ground truth pixel relative to the patch centre.

The extraction was run per Section 3.2.4, using the 690 QC-passed WSIs as input data.

An offset of (-56px,-56px) was used, corresponding to a translation upwards and leftwards of 27μm, a quarter of the patch width, to place the GT pixel at the centre of the bottom-right

quadrant of the patch. This offset was chosen so that any resulting shift in the distribution of model activations would be clearly identifiable.

Note that the patch coordinates follow the Python image indexing convention, where Y-coordinates are measured from the top downwards.

Results

Table 5 shows the totals of offset patch files extracted into each class subdirectory, alongside the resultant directory size.

Table 5: File totals and class directory sizes following offset patch extraction

Class directory	Files in class directory	Total size, MB
0-non-informative	4620	362
1-tumour	19247	1945
2-stroma-or-fibrosis	11172	1126
3-necrosis	1815	180
4-vessels	539	54
5-inflammation	29	3.3
6-lumen	2899	248
7-mucin	436	40
8-muscle	523	50

Figure 9 shows a sample *tumour* patch, offset by (-56px,-56px), alongside its non-offset counterpart.

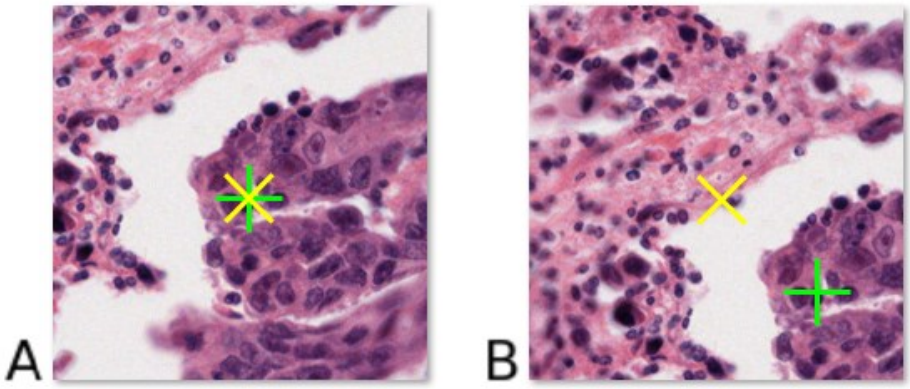


Figure 9: (A) Non-offset tumour patch and (B) same patch offset by (-56px,-56px). "x" denotes patch centre, "+" represents GT label coordinates.

Discussion

In the offset (right-hand) patch in Figure 9, the central mass of tumour nuclei in the LH patch has been moved downwards and rightwards relative to the original version, so that this feature is now located in the bottom right quadrant as required.

Thus, a patch set has been created with consistently offset positions relative to the GT locations. The absence of rotated copies in smaller classes has resulted in a less balanced dataset than in Section 3.2.4. Nonetheless, this dataset facilitates measurements of spatial distributions of model activations against a known, off-centre GT pixel.

3.3.4 *offset-tumour-stroma-groups* Dataset

Motivation

Classification accuracy was observed to increase when training models with the *offset-patches* dataset (Section 5.2.3) and with the 2-class *tumour-stroma-groups* dataset (Section 5.3.3). It was proposed to investigate the performance of models trained on data that embodied both concepts simultaneously. A two-class, offset-sampled *offset-tumour-stroma-groups* dataset was conceived.

Methodology

The grouping method previously used to build the *tumour-stroma-groups* directories (Section 3.3.2) was applied to the 9-class directory generated during offset patch extraction (Section 3.3.3). 12,000 files of each group were specified in script parameters.

Results

Table 6 shows the totals of offset patch files that were collated into *tumour-group* and *stroma-group* subdirectories.

Table 6: Total offset patch files copied into subdirectories to create offset-tumour-stroma-groups dataset

Class directory	Files in class directory	Total size, MB
stroma-group	12,000	1157
tumour-group	12,000	1126

Discussion

This extraction task resulted in a balanced binary *offset-tumour-stroma-groups* dataset with 12,000 offset patches in each of the required *tumour-group* and *stroma-group* categories.

3.4 ImageNet-100

3.4.1 Motivation

A general-purpose image set was sought, to test the generalisability of new model architectures that had previously performed well with QUASAR data.

In datasets of commonplace objects, object boundaries are identifiable without specialist knowledge, allowing immediate comparison with model attention distributions that might reveal intelligent behaviours.

ImageNet (Russakovsky et al., 2015) was of initial interest as a popular benchmark library containing 1000 general image classes. The ImageNet-100 subset shared on the Kaggle competition site (Shekar, 2021) has 100 classes of animal, fish and bird photographs selected from ImageNet, and was chosen for experiments in this work because of its more manageable size. The ImageNet-100 training set is 16GB, against ImageNet's 160GB, facilitating file copying and model training on a single HPC node within a reasonable timescale.

3.4.2 Methodology

Code used in this section is documented in Appendix Section 1.1.3.

ImageNet-100 Training and Test sets were downloaded from Shekar's Kaggle page (Shekar, 2021) then uploaded to the ARC4 HPC.

For readability, and for consistency with the naming convention used in QUASAR class subdirectories, the class subdirectories in ImageNet-100 were renamed using the WordNet identifier (WNID), implicit in the directory name, to look up the English class description for

each category from metadata downloaded with ImageNet-100. The class description was combined with an index number, based on the category's position in the CSV file, to create a human-readable directory name for each image class, with the format {index}_{category}.

3.4.3 Results

Table 7 shows the relative sizes of the Training and Test directories extracted to ARC4.

Table 7: Directory sizes following extraction of ImageNet-100 dataset

Purpose	Class subdirectories	Files per class	Total files	Total size, MB
Training	100	1,300	130,000	16,384
Test	100	50	5,000	712

The 100 class subdirectories were named according to their order in the supplied CSV, from *000-chambered_nautilus* to *099-cock*. Sample images representative of ImageNet-100 are shown in Figure 10 through Figure 13.

Figure 10: ImageNet-100 example, class 056-oystercatcher

Figure 11: ImageNet-100 example, class 053-vine_snake

Figure 12: ImageNet-100 example, class 059-goldfinch (American Goldfinch)

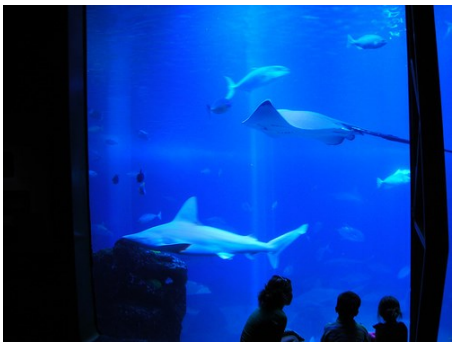


Figure 13: ImageNet-100 example, class 009-tiger_shark

3.4.4 Discussion

A 100-class set of diverse general-purpose images, containing clearly identifiable objects of each class, has been provided for model training and assessment. The results of training activities involving this dataset are detailed in Section 5.4.

4 Characterising the Whole Slide Image

4.1 Summary

Sections 4.3 through 4.6 of this chapter include content that was adapted for publication as *Attention-guided sampling for colorectal cancer analysis with digital pathology* (Broad et al., 2022), in the *Journal of Pathology Informatics*.

This work was designed to satisfy a need, revealed during a review of literature (Section 2.5), for an efficient WSI-sampling algorithm that would support reliable estimation of tumour region of interest (ROI) and tumour stroma ratio (TSR).

To this end, experiments were performed to visualise tissue class distributions in WSIs of colorectal cancer sections. CNNs were used to classify multiple image patches, allowing colour-coded distribution plots to be created. Results obtained using simple tile-by-tile classification revealed the need for a more efficient algorithm for sampling image patches from the WSI.

Novel algorithms were inspired by human attention processes, first using a heatmap-guided sampling approach, then a weighted regular sampling algorithm using a higher sampling density within the predicted tumour region.

From the predicted cell distributions, further diagnostic outputs were derived: An estimate of the ROI outline, and the prognostically useful Tumour Stroma Ratio (TSR) within this area.

4.2 Attention Heatmap WSI Processing Pipeline

4.2.1 Motivation

Prior work (Broad et al., 2020) used a 4-layer CNN to classify every 256×256 pixel tile in a WSI. The predicted tissue class distribution was plotted as a colour-coded image (Figure 14). This gave a detailed representation of cell type across the WSI, but at the expense of computation time, in some cases taking over an hour to process a single image. Furthermore, patches in background and non-tumour regions were sampled at high density, potentially exposing diagnostic calculations to noise from less diagnostically relevant areas of the WSI. A more efficient and selective sampling algorithm was therefore required.

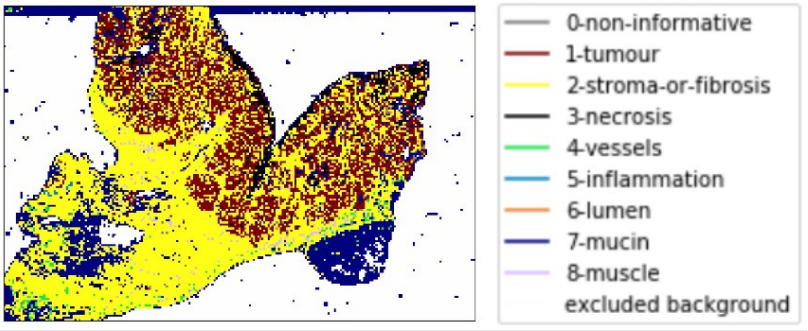


Figure 14: Tile-by-tile classification plot for colorectal cancer WSI (Broad et al., 2020)

Where to look: Using a thumbnail heatmap to determine sampling distribution

A novel *Attention Heatmap Pipeline* (AHP) was proposed. Patches were loaded from the WSI at sampling locations guided by a spatial 'heatmap' of tumour probability derived from low magnification 'thumbnail' patches. This is analogous to human vision, where areas of interest identified in low resolution peripheral vision trigger high-definition foveal glimpses to collect further information from these locations. Similarly, in digital pathology, tissue is examined at low magnification to select ROIs for more detailed inspection at high magnification.

The trouble with epithelium: Distinguishing tumour from healthy tissue

As discussed in Chapter 3, the QUASAR dataset categorises epithelial tissue, when found inside the cancerous region, as *tumour* tissue. These cells are very similar in appearance to normal epithelium found elsewhere in the tissue section.

Annotated examples of normal epithelium were not included in the QUASAR dataset, nor in public challenge data such as Camelyon (Ehteshami Bejnordi et al., 2017). A classifier trained on available data might therefore be expected to mis-identify normal epithelium as tumour, resulting in false positives outside the annotated ROI. These was found to reduce the accuracy and clarity of measurements and visualisations based on these results, particularly the prediction of the ROI in unseen WSIs

Methods were therefore explored for distinguishing normal from tumour epithelium, using a CNN dedicated to this classification task and incorporated as an additional stage in the WSI processing pipeline.

Clustering for ROI estimation

Areas of tumour patches within the visualisation plots showed strong visual correlation with the expert-annotated ROI outlines. Clustering algorithms were therefore assessed for use in estimating the tumour ROI in previously unseen WSIs.

4.2.2 Methodology

Computing Resources

Code used in this section is documented in Appendix Section 1.2.1.

The experiments were carried out on ARC4, part of the HPC facilities at the University of Leeds, to take advantage of powerful NVIDIA GPU nodes and to keep data within the University network as required by the QUASAR data sharing agreements.

Software was developed in Python, using *PyTorch* machine learning packages with *SciKitLearn* and *MatPlotLib* for data visualisation. Linux shell scripts were used to upload code to ARC4, and to instantiate a local Python environment in which to execute the code. Logging functions were developed to record model training data such as loss and accuracy values, and confusion matrices.

Data Split Management

In experiments prior to this work, twenty-four randomly chosen WSIs from the dataset were used to evaluate the Attention Heatmap Pipeline. CNNs used in the pipeline were initially trained on a random 70% split of the overall set of patches from all available WSIs. It was therefore likely that patches derived from WSIs in the pipeline evaluation set, were also encountered during training, causing overfitting and yielding misleadingly high accuracy scores.

Also, when model training exceeded the 48 hours maximum session on ARC4, it was necessary to resume training in a new HPC job, to complete the desired number of training epochs. By default, this created new random split, and resulted in higher classification accuracies than in similar jobs that completed within a single 48 hour session. It was suspected that this was due to overfitting, where that patches from the eventual test set were also likely to have been present in the training set of the initial HPC job.

Before embarking on systematic benchmarking of CNN models, a mechanism was required to manage the data split in a predictable manner, so that predetermined sets of test and training

images would persist across multiple ARC jobs. Further hold-out sets were also required, for evaluating pipelines with previously unseen WSIs.

The new system allowed a new data split to be created for each experiment. The vector of WSI image numbers in each training, test and evaluation set was persisted in an *SQLite* database against an experiment ID. This way a consistent data split was available for training tasks that continued across multiple ARC jobs, provided the same experiment ID was used for each job.

The split was created by allocating whole WSIs, rather than individual patches, to each set. This decision was based on findings by Nir et al (2019), who compared cross-validation approaches in grading prostate cancer images using AI. Here, a 'leave-patients-out' technique proved more accurate than 'leave-patches-out'.

In the QUASAR dataset, there is generally one WSI per colorectal cancer *case*, rather than one per patient. It was assumed that a 'leave-WSIs-out' approach would approximate to Nir's recommended 'leave-patients-out' split, in the absence of a patient-WSI mapping giving an exact per-patient grouping. Image patches were therefore grouped by their originating WSI before the split was made.

A ratio of 481 training WSIs, 100 test and 100 validation images was used for the data split, using image patches extracted from a total of 689 QC-passed WSIs. The 100 validation WSIs allowed statistics, such as percentage of tumour inside the annotated ROI, to be calculated with narrower confidence intervals than were possible using the earlier 24 WSIs.

Thumbnail patch heatmap

Patches at 'thumbnail' magnification were extracted from 690 QC-passed QUASAR WSI files. The patch extraction process in Section 3.2.4 was applied to images from the low-resolution 1.25x magnification tier provided in each SVS file. Patches were sampled at locations scaled from the maximum-resolution *box-scores* coordinates in the ground-truth XML data.

Patches with sizes 8×8px, 16$\times$ 16px, 32×32px, 64×64px, 128×128px and 256×256px were extracted for use in training. 224×244px patches were extracted later for subsequent use with VGG19 and related models.

Because the ground-truth sample locations were clustered together in 3×3mm regions, it was anticipated that the larger thumbnail patches might overlap. This was expected to be problematic for overlapping patches of different class labels, because models would then be expected to produce different class outputs despite the shared input image region. The patch extraction process was therefore modified to reject overlapping patches with differing class labels.

A ***ConvNet*** CNN with 4 convolution layers, as defined in Table 8, was trained to classify thumbnail patches into the 9 QUASAR tissue classes. Data loader transformations were used to scale the thumbnail patches to the 256×256px input size required by the CNN. The model was trained using Stochastic Gradient Descent (SGD) with learning rate (LR) = 0.001 and momentum = 0.9, over 40 epochs. These hyperparameters were found experimentally to provide stable convergence to a low loss level.

Table 8: ConvNet CNN definition

Layer	Input Channels	Output Channels	Type	Kernel Size
1	3	16	Convolution	3x3
2	16	24	Convolution	4x4
3	24	32	Convolution	4x4

4	32	48	Convolution	3x3
FC1	*	512	Fully Connected	
FC2	512	10	Fully Connected	

Heatmap plots were generated by applying this CNN to image tiles extracted sequentially across the WSI's entire thumbnail image. A tumour probability value was obtained by taking the output for the *1-tumour* class before the final SoftMax stage. This value determined the brightness of the tile in the output plot, with the brightest red showing the maximum value of $p(tumour)$.

Generating sampling patches from low-resolution tumour heatmap

The vector of tumour probabilities associated with the low-resolution heatmap was used to control the attention of further processing at high magnification.

A Sample Pattern Generator (SPG) was developed, to determine the sampling locations where image patches would be loaded from the WSI at full magnifications. Here, the WSI space was divided into a grid, with each box representing the area of a low-resolution tile projected onto the coordinate space of the full magnification image. This was typically 1024×1024px in the WSI, for a 64×64px thumbnail tile. The tumour density value used in the heatmap determined the number of sampling patches in the corresponding grid box, such that between one and nine sampling patches were allocated in each box.

Each sampling patch was randomly positioned inside its parent box, to minimise sampling bias due to aliasing effects. If overlap was detected with an existing patch within the box, a new random placement was chosen. This was attempted up to 10 times before the patch was rejected, as a trade-off between computation time and accurate sampling density.

A sample tumour heatmap, and the sampling pattern derived from it, are shown in Figure 15 and Figure 16 respectively.

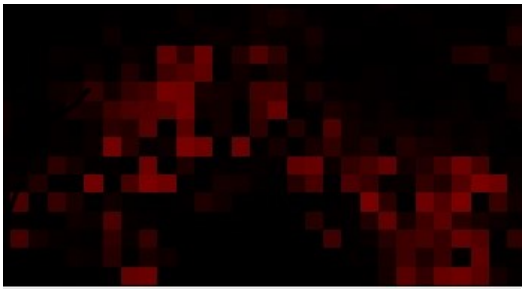


Figure 15: Thumbnail-derived tumour probability heatmap

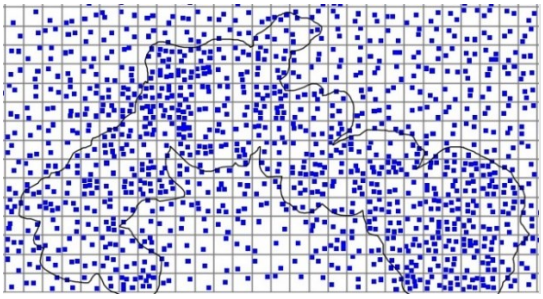


Figure 16: Patch sampling locations based on probabilities from heatmap.

'Parent' tiles are shown by grey gridlines.

Distinguishing tumour from healthy tissue

During the development of the Attention Heatmap Pipeline, it was discovered that epithelial tissue outside the tumour ROI was often mis-classified as tumour by the CNN trained on the 9 QUASAR patch classes. This resulted in many false-positive tumour patches outside the ground-truth ROI, reducing the accuracy of ROI predictions which were based on clusters of predicted tumour patches.

A CNN was therefore trained to distinguish between tumour and normal epithelium. For this purpose, "tumour" patches identified outside the expert-annotated ROI were categorised as normal epithelium.

Data for this task was collected by adding a temporary step to the data analysis pipeline, whereby patches classified as *tumour* were output to either *true positive* or *false positive* data subdirectories, depending on whether they fell within the ground truth ROI (Figure 17). Initially this was done for 24 randomly selected WSIs. After balancing the number of images in each directory, this provided approximately 15,800 training patches of each class.

CNN classifiers were trained using this data, for the binary classification task of distinguishing between tumour and normal epithelium. At first, the four-layer *ConvNet* architecture was used, followed later by VGG16, VGG19 and selected feedback attention models (Chapter 8).

For the later models, a new TP/FP patch directory was created by running the pipeline against multiple data splits, and new TP/FP models were trained against this.

The classifier was then used in the pipeline to reject false-positive tumour patches (see CNN3 in Figure 18). Only patches already classified as *tumour* were processed in this way. Patches that were reclassified as *normal epithelium* were subsequently rejected from the output plot, and excluded from calculations involving tumour patches.

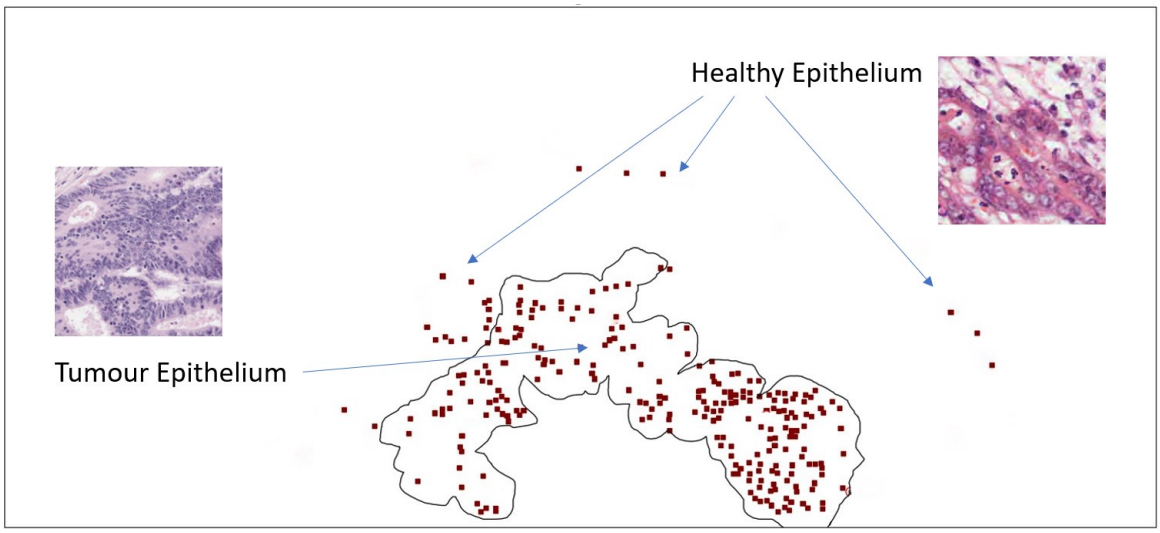


Figure 17: Tumour and normal epithelium patches predicted as tumour, inside and outside the annotated tumour ROI

Clustering for ROI estimation

DBSCAN, or Density-Based Spatial Clustering of Applications with Noise (Ester et al., 1996), was used to estimate the ROI from distributions of tumour patches in the WSI.

DBSCAN examines each point in a set of 2-D coordinates and searches for neighbouring points within a radius ϵ. If more than a chosen number N of neighbours are found, a cluster is declared and the points are allocated to it. The neighbours of these points are then examined, further extending the cluster if N or more points are within distance ϵ. If a point does not have N neighbours within radius ϵ, it is categorised as noise and is not allocated to a cluster.

The *SciKit* Python implementation of DBSCAN (SciKit-Learn, 2020) was used in the pipeline. Once DBSCAN had allocated tumour points to clusters, a further function generated polygons enclosing the points in each cluster. Each point was inflated into a circle of radius ϵ, using the *Shapely Geometry* package (Adair et al., 2020). Thus the cluster yielded a set of overlapping circles, the union of which was used create the output polygon(s). A further step eroded these polygons by 0.5ϵ, to fit the resulting boundary more closely to the outermost tumour points in the cluster.

The resulting polygons provided an estimate of the ROI boundaries, based on the predicted tumour distribution.

Measuring clustering accuracy

The accuracy of the predicted ROI was evaluated by comparing the polygons generated by the clustering algorithm, with the original annotated ROI boundary.

The F1, also known as Dice Similarity Coefficient (DSC), and the Intersection over Union (IoU, or Jaccard Index) are the two most common overlap-based metrics for comparing 2D regions (Reinke et al., 2021). IoU measures the overlap (intersection) between two regions A and B as a proportion of the union area, while F1 compares the intersection to the total area of the two regions:

$$F1 = \frac{2|A \cap B|}{|A| + |B|} \qquad (2)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \qquad (3)$$

F1 and IoU scores were calculated for each WSI in the hold-out evaluation set of 100 WSIs, and were used to derive mean values and confidence intervals for each scoring method.

The two main clustering variables, the minimum neighbour count N and the search radius ϵ, were manually optimised for maximum F1 score, with values of $\epsilon = 2000$ (approximately 1.0mm at 0.49μm/pixel) and $N = 3$ being adopted in the pipeline.

Pipeline

Figure 18 shows the complete Attention Heatmap Pipeline for predicting the distribution of tumour cells in an unseen WSI, using the low-resolution tumour heatmap to control the patch selection in the WSI, and the false-positive detection algorithm as a noise-reducing output stage.

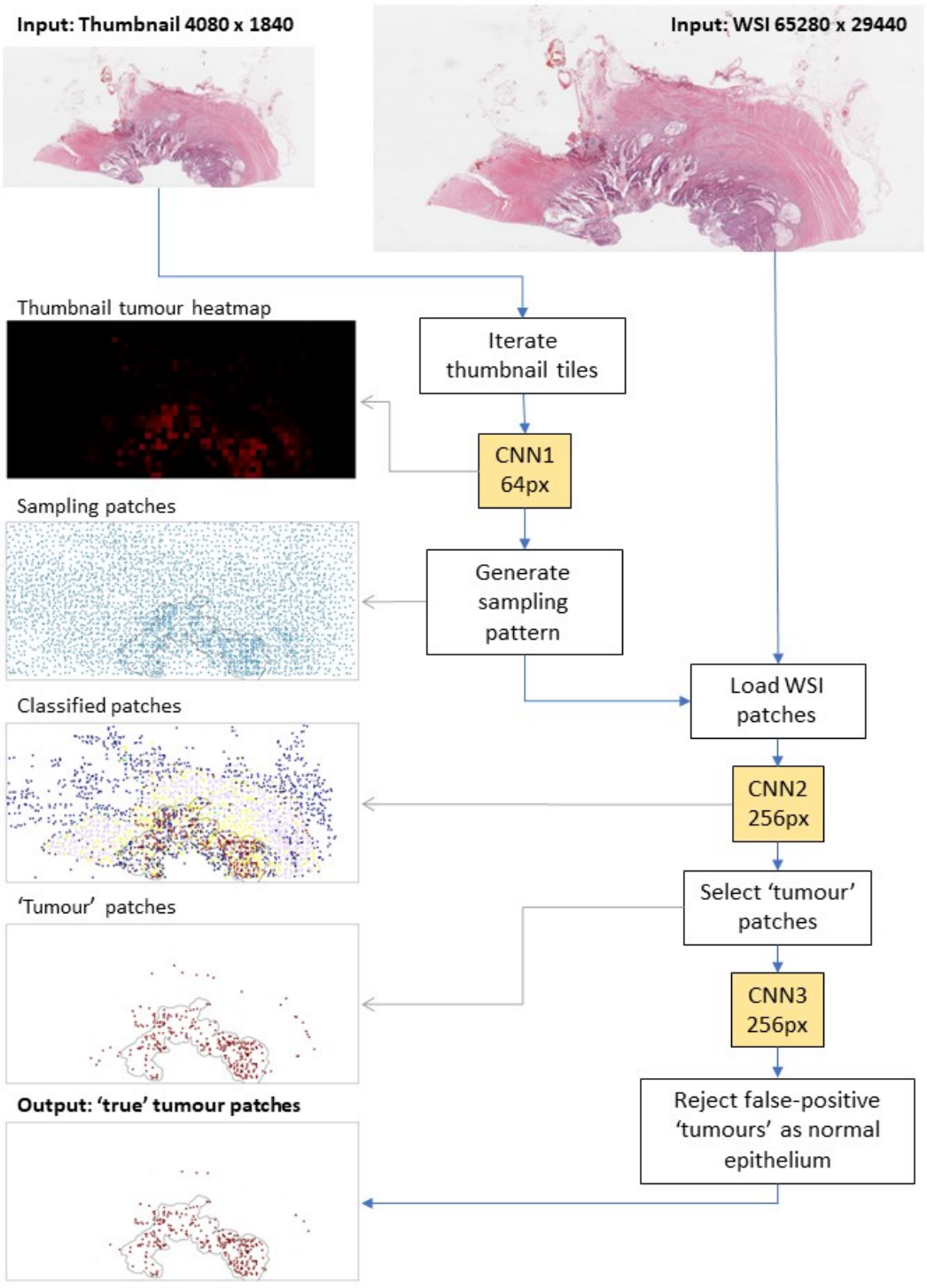


Figure 18: Attention Heatmap Pipeline (AHP), predicting tumour distribution using whole-slide and thumbnail images

The patch size used in sampling from the full magnification WSI was chosen to match the input size of the high-resolution classifier, CNN2. This was initially 256×256px for *ConvNet* classifiers. Later CNNs also used 224×224px and 299×299px patches.

4.2.3 Results

Thumbnail patch heatmap

Table 9 shows classification accuracies for CNNs trained on each size of thumbnail patch. Each result was compared with that obtained with a training set where overlapping patches were retained.

Table 9: Results for training 4-layer ConvNet on thumbnails, excluding overlapping patches with conflicting classifications

Thumbnail patch size (height and width)	Number of patch images extracted and retained	Training time (40 epochs, NVIDIA V100 on ARC4)	Training time per epoch	Accuracy before overlaps rejected	Accuracy after overlaps rejected
16px	59,046	8:40h	16m	61%	51%
32px	52,735	9:40h	14m	64%	59%
64px	30,396	7:20h	10m	66%	68%
128px	24,575	6:28h	9m	65%	70%
256px	22,775	11:25h	17m	62%	72%

The heatmap plot in Figure 20 represents the probability of tumour predicted by the thumbnail patch CNN for image tiles extracted sequentially across the thumbnail image in Figure 19. The patch sampling distribution derived from this heatmap is shown in Figure 21.

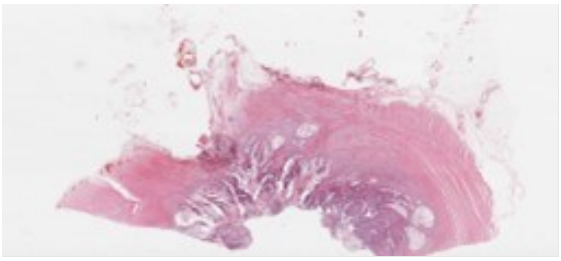


Figure 19: Thumbnail image of colorectal cancer WSI

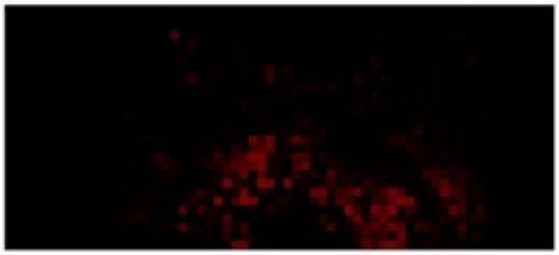


Figure 20: Heatmap of predicted tumour density, from execution of CNN for all thumbnail tiles

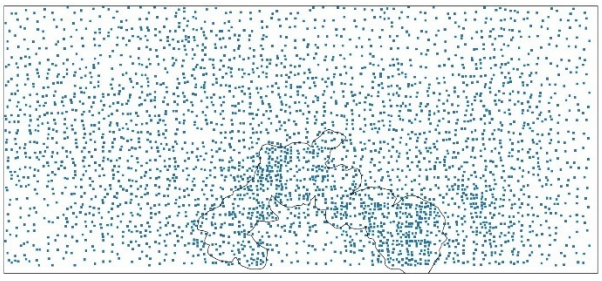


Figure 21: Patch sampling pattern derived from tumour distribution in thumbnail patches

High-Resolution Classification of Selected Patches

Full-resolution image patches were loaded from locations in the WSI determined by the heatmap-derived sampling pattern. These patches were classified using the 9-class CNN (CNN2 in Figure 18). The classified patches were colour-coded and plotted in a 2-D coordinate space to match the WSI bounds, providing a visual map of the distribution of cell types.

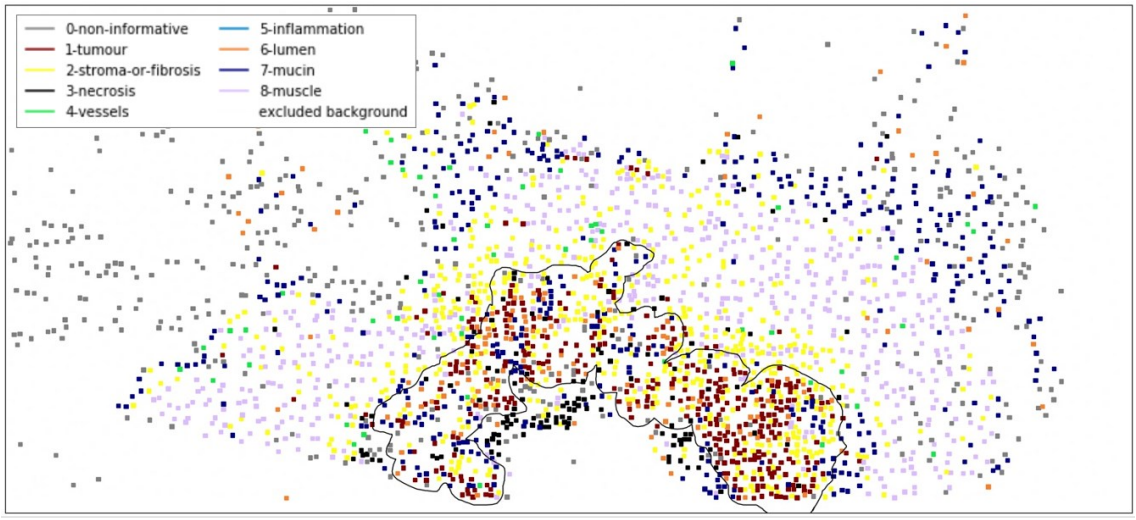


Figure 22: Patch classification plot generated by Attention Heatmap Pipeline

Figure 22 shows a typical plot from this stage of processing. Regions of tissue types are clearly differentiated here, such as muscle, mucin and stroma. The majority of tumour patches (red) are within the ROI originally marked by the pathologist.

Distinguishing Tumour from Healthy Tissue

As shown in Table 10, both networks returned a higher accuracy in this binary classification than was seen in the 9-class classifier.

Table 10: Training results for CNN distinguishing true/false positive tumour patches

CNN	Training epochs	Training duration on ARC4	Accuracy (% correct classifications)
ConvNet	40	16h	88%
VGG16	50	20:42h	91%

Figure 23 shows the result of this process when applied to the clusters of patches classified as *tumour* by CNN2 in the pipeline. In this example, about 80% of the patches outside the tumour region were identified as false positives by CNN3, and excluded from ROI and TSR calculations.

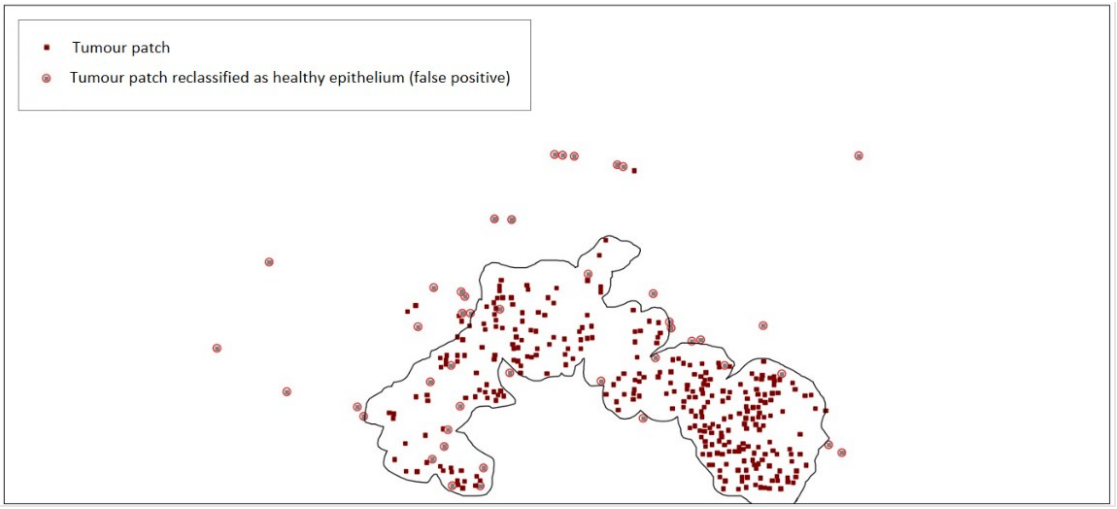


Figure 23: False positive detection: Predicted tumour patches reclassified as normal epithelium, with ground-truth ROI overlaid for reference

Table 11 shows the performance of the pipeline, with and without FPC and using different CNN architectures. This was initially expressed as the *percentage of predicted tumour points falling inside the annotated ROI*. Prior to the development of clustering algorithms, this measurement was used as a proxy for the ability to predict the region of tumour in an unseen WSI.

This data was logged for 100 WSIs in a validation hold-out set. The mean was calculated from these outputs, and 95% confidence intervals based on the mean, variance and sample size were obtained using the *t.interval* function in the SciPy statistics package.

Table 11: Effects of false positive correction in the Attention Heatmap Pipeline

CNN type	CNN Accuracy (9-class)	% Tumour in ROI, Pre FP Correction	95% Confidence Interval	% Tumour in ROI, Post FP Correction	95% Confidence Interval
ConvNet	72%	**81.9**	79.0 – 84.6%	**87.2**	84.8 – 89.5%
VGG16	74%	**84.7**	80.9 – 88.4%	**94.4**	91.9 – 96.8%
VGG16 Pre-Trained	78%	**91.6**	89.1 – 94.2%	**97.1**	96.2 – 98.0%

Clustering for ROI Estimation

Figure 25 shows the predicted ROI outline generated by the clustering algorithms, enclosing tumour points derived from a typical QUASAR WSI (Figure 24). The original ROI is overlaid for comparison.

For the example WSI shown, the F1 Score was 0.876. The mean value over the 100-WSI validation set was F1 = 0.773.

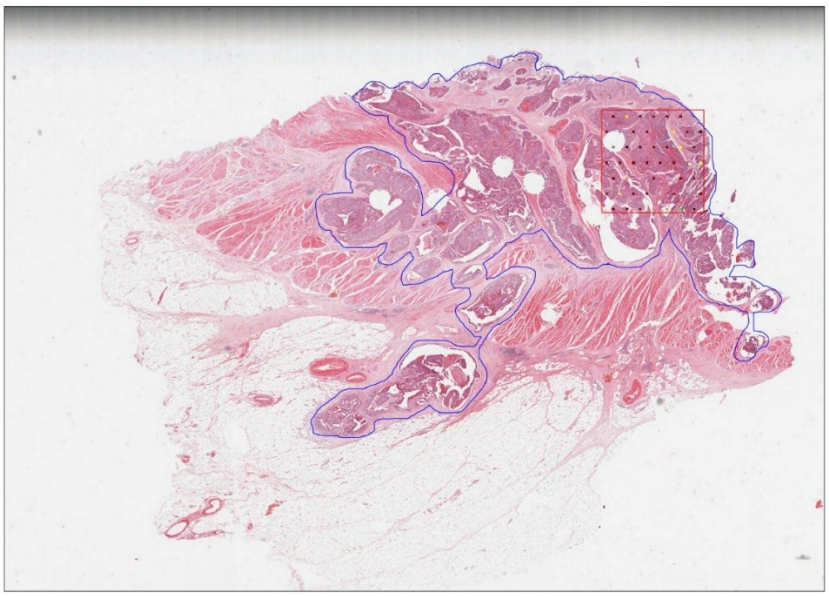


Figure 24: QUASAR WSI 57623.svs with expert-annotated ROI (blue outline)

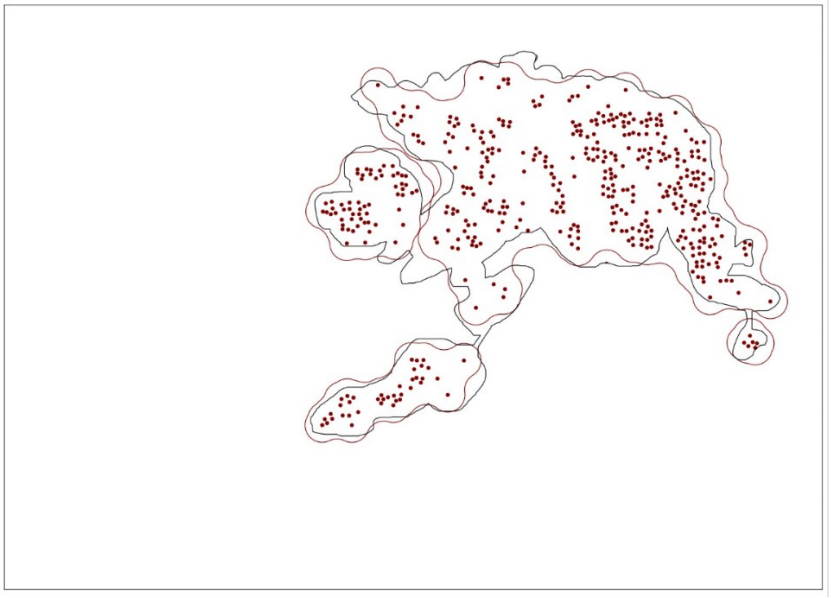


Figure 25: DBSCAN clustering of predicted tumour points (red outline) with expert-annotated ROI annotation (black outline)

Following the implementation of the clustering algorithm, the accuracy of the resulting ROI outline was calculated as IoU and F1 (Dice) score.

Table 12: ROI estimation performance in the Attention Heatmap Pipeline

CNN type	IoU w.r.t GT ROI	IoU 95% CI	F1 score w.r.t GT ROI	F1 score 95% CI
ConvNet	61.4%	57.2 – 65.6%	73.6%	69.5 – 77.8%
VGG16	62.3%	54.7 – 70.1%	73.1%	65.9 – 80.3%

4.2.4 Discussion

Thumbnail patch heatmap

In the thumbnail CNN, the rejection of overlapping patches prior to training was found to improve classification accuracy when using larger patch sizes, which might be expected to overlap more with their neighbours. However, the number of patch images available for CNN training was reduced at larger thumbnail patch sizes, as more patches were rejected. Generally, the classifier was more accurate at larger thumbnail patch sizes, where more structural context is available from surrounding tissues. At thumbnail magnifications, individual cells were not visible. It appears that the classifier was instead responding to higher-level image features such as colour, texture and structure.

A thumbnail patch size of 64×64px, equivalent to 1024×1024px at maximum resolution, was chosen for use in the pipeline. This was chosen to balance thumbnail classification accuracy with heatmap resolution. On visual inspection across the validation set, the brightest regions in each heatmap appeared to correspond to the darker blue region of densely packed tumour nuclei in the WSI, which pathologists had annotated as the ground truth ROI.

Distinguishing tumour from healthy epithelium

The CNN3 classifier was trained to high accuracy in distinguishing normal and malignant epithelium, allowing WSI pipelines to disregard patches previously identified as *tumour* that the CNN now classified as *normal epithelium*. This approach to false-positive correction improved accuracy in 9-way classification tasks for all three CNN architectures evaluated. The non-overlapping 95% CIs in each row of Table 11 imply that these gains are statistically

significant. The FPC technique was therefore reused in further processing pipelines in Sections 4.2 and 4.5.

The percentage tumour in ROI score was consistently higher than the per-patch accuracy of the 9-class CNN2 (see Figure 18), suggesting that an ensemble effect may have averaged out classification errors in individual patches in the cluster. Using a separate CNN for false-positive correction has reduced the rate of epithelial cells being classified as tumour outside the ROI, increasing the percentage of predicted tumour inside this ground-truth region by up to 9.7 percentage points and thus providing valuable noise reduction ahead of further processing of the pipeline output.

Clustering for ROI estimation

DBSCAN clustering, applied to the spatial distribution of tumour patches generated by the pipeline, yielded an ROI estimate that aligned with the ground truth ROI with an F1 score of approximately 73.6%. Figure 25 shows that the predicted ROI outlines deviate furthest from the ground truth ROI in WSI regions where fewer tumour points are available. However, the prediction of multiple cluster outlines is not necessarily an error. The ground truth data only contains one ROI annotation per WSI, sometimes arbitrarily joining multiple tumour regions into one polygon. It is understood that limitations in the software for the stylus pen required the outline to be drawn in a single stroke. Further deviation between the GT and predicted ROI may be due to the low resolution of the touchscreen used with the stylus, meaning that the ROI annotations were not drawn using the full resolution of the slide.

The ability to predict tumour outlines does not in itself bring novel diagnostic value. Pathologists can quickly identify an ROI by viewing the slide – although they may be grateful of a tool to automate this relatively mundane task. Nonetheless, emulating this behaviour in software allows a region to be defined where an AI-based system can direct its processing resources, to extract further prognostic data from the tumour location whilst saving on processing time and therefore operating costs.

4.3 Benchmarking of Popular CNN Architectures for Cell Classification

4.3.1 Motivation

Initial experiments with the Attention Heatmap Pipeline (AHP, Section 4.2) used a four-layer *ConvNet* CNN architecture from earlier work (Broad et al., 2020). This gave 72% accuracy in 9-way classification, with 81.9% of predicted tumour points within the ROI. It was expected that a deeper, published architecture would improve on these results.

The VGG16, a deep convolutional neural network, was also initially tested in the AHP. The closely related VGG19 (Simonyan and Zisserman, 2014) scored highest on *NCT-CRC-HE-100K* colorectal cancer patch data (Kather et al., 2019). VGG16, having 16 instead of 19 weight layers, was expected to be less prone to overfitting than VGG19. Both variants were reviewed.

Other well-known deep CNN architectures, optimised for the ImageNet challenge (Deng et al., 2009), were selected for comparison. AlexNet, DenseNet, GoogLeNet, Inception, MobileNet, ResNet, ResNext, ShuffleNet and SqueezeNet had been reviewed in histopathology applications (Wang et al., 2021) and were available as library classes in the Python *TorchVision* package (PyTorch, 2021). EfficientNet-B0, the 224px version of a recent ImageNet exemplar (Tan and Le, 2020) was also assessed.

4.3.2 Methodology

Existing Python code for configuring and training a CNN, was extended using a Class Factory design pattern (Gamma, 1995) to allow one of a selection of CNN models to be initialised

according to HPC job parameters. The relevant CNN implementations in TorchVision were loaded via subclasses of a classifier base class, the latter controlling common data loading and data transformation activities. Model training was executed on ARC4 for each model architecture in turn.

Pre-Trained vs Random Initialisation

CNNs in histopathology can often be trained to a higher accuracy when pre-trained on generic image sets such as ImageNet (Rao, 2018; Zhao et al., 2020). Downloadable pre-trained weights are available in TorchVision for many CNNs, and were evaluated in addition to the default version with randomly initialised weights.

Testing in pipeline

Selected trained models were substituted for the main classifier in the Attention Heatmap Pipeline (Section 4.2, Figure 18, CNN2) to investigate the relationship between CNN accuracy and the predicted spatial distribution of tumour patches in the pipeline output.

4.3.3 Results

Table 13 lists the CNN models tested, in descending order of classification accuracy. Notable or extreme results are in bold.

The most accurate CNNs were evaluated in the Attention Heatmap Pipeline. The final column in Table 13 percentage of predicted tumour patches falling within the predicted ROI, before the application of false positive correction (FPC). Values are also shown here for the VGG16 and 4-class ConvNet models used in early experiments.

Table 13: Comparative performance of CNN architectures trained on 9-class QUASAR patches. Notable or extreme results are shown in bold.

CNN Type (PT=Pre-Trained on ImageNet)	Image Width and Height (px)	Classification Accuracy	Attention Heatmap Pipeline % Tumour in GT ROI (pre FPC), where evaluated with given CNN
VGG19 (PT)	224	**79%**	**94.8%**
GoogLeNet	224	**79%**	**92.0%**
EfficientNet-B0 (PT)	224	**79%**	**91.6%**
DenseNet (PT)	224	**78%**	**93.7%**
VGG16 (PT)	224	**78%**	**91.6%**
MobileNet (PT)	224	77%	91.0%
AlexNet (PT)	224	76%	-
AlexNet	224	75%	-
GoogLeNet (PT)	224	75%	-
DenseNet	224	74%	-
VGG16	224	74%	84.7%
MobileNet	224	73%	-
ResNext	224	73%	-
VGG19	224	72%	-
Inception 3	**299**	72%	-
EfficientNet-B0	224	71%	-
ResNet 50	224	71%	-
ConvNet (4-layer)	**256**	71%	81.9%
Inception 3 (PT)	**299**	70%	-
ShuffleNet	224	70%	-
SqueezeNet	224	70%	-
ResNet 18	224	68%	-

Confusion matrices (Figure 26 and Figure 27) were generated for the top two CNNs in Table 13, as sorted by descending accuracy and tumour percentage in ROI.

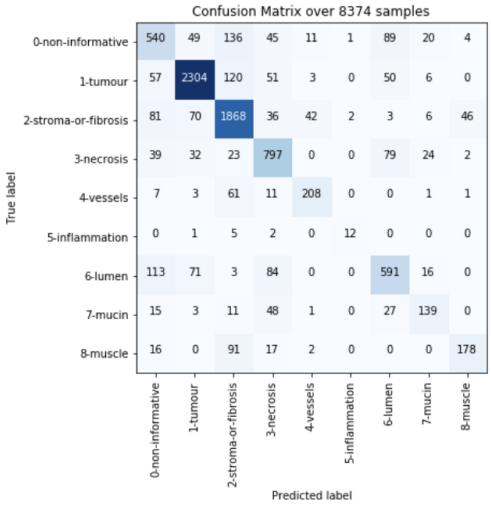


Figure 26: Confusion Matrix for pretrained VGG19

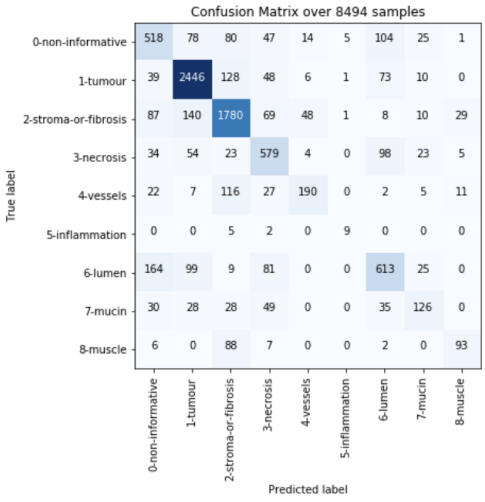


Figure 27: Confusion Matrix for GoogLeNet

Despite similar overall accuracy to VGG19, more non-tumour tissues are falsely identified as tumour.

4.3.4 Discussion

The pretrained VGG19 and non-pretrained GoogLeNet gave the joint-highest classification accuracies at 79%. However, in the Attention Heatmap Pipeline, the VGG19 yielded the highest percentage of tumour points within the ground-truth ROI, suggesting it was better able to distinguish tumour from other cell types. Comparison of confusion matrices (Figure 26, Figure 27) revealed that GoogLeNet incorrectly interpreted many *stroma, necrosis, lumen* and *mucin* patches as *tumour*. Where this occurred outside the ROI, the misclassification of these stroma subclasses would contribute to the observed higher rate of false tumour 'noise' outside the ROI.

With the exception of GoogLeNet, the ImageNet-pretrained CNN models achieved greater accuracy than their randomly initialised counterparts. This suggests that the pretrained models have learned to respond to generic image features, such as shape, colour and texture, that are useful in distinguishing cell types in a WSI.

The pretrained VGG19 was therefore adopted for use in subsequent pipeline experiments.

4.4 Tumour Stroma Ratio

4.4.1 Motivation

Literature (Section 2.5.1) shows that Tumour Stroma Ratio (TSR) is a predictor of disease progression and survival rates in colorectal cancer. The Attention Heatmap Pipeline generated spatial distributions of tumour and stroma cell patches for a WSI, relative to the tumour ROI. It was proposed to use these distributions to derive an estimate of TSR.

4.4.2 Methodology

In the QUASAR ground truth annotations (Section 3.2.3), cell types were sampled at 50 RandomSpot points (Wright et al., 2015) inside a 3mm square box within the ROI of a colorectal cancer WSI. This sampling box represents a 'virtual biopsy', emulating the action of a surgeon taking a tissue core from a live bowel wall. In the WSI, the 3mm box was placed at the region of highest tumour cell density, as perceived by the annotating pathologist, along the luminal aspect (inner wall of the bowel). In a real biopsy, the surgeon would need to sample at this depth to avoid perforating the bowel.

The ground truth TSR was calculated from totals of expert-annotated tumour (T) and stroma (S) patches within the 3mm box, for each WSI in turn. The proportion of tumour (PoT) was similarly calculated:

$$TSR = \frac{T}{T+S} \qquad (4) \qquad PoT = \frac{T}{total\ patches} = \frac{T}{T+S+Others} \qquad (5)$$

The tumour patch count was included in both numerator and denominator for results in the range $0 \leq TSR \leq 1$. Python code was written to evaluate the TSR for collections of classified patches, so that TSR could be calculated from the ground truth class annotations for a WSI, or for classified patches identified by the pipeline.

TSR Sampling Strategies

Multiple algorithms were evaluated for choosing sampling points for TSR calculations, from within the predicted ROI and locations where the estimated tumour density was highest.

For each sampling approach, image patches were loaded from the required locations, and classified using the CNN. TSR and PoT were then calculated from the total patches in each output class. These were compared with the values based on the ground truth annotations in the QUASAR XML files, giving an error score (ϵ_{TSR}) where:

$$\epsilon_{TSR} = \frac{\left|TSR_{predicted} - TSR_{GT}\right|}{TSR_{GT}} \quad (6)$$

Error values were logged for each sampling strategy, against each WSI. The mean error magnitudes and related confidence intervals were then calculated for the 100 WSIs in the evaluation dataset.

A. Sampling at Ground Truth Locations

Image patches were sampled from the WSI at locations defined in the QUASAR ground truth XML data, and the TSR calculated from the CNN's classification of these patches. Figure 28 shows the 3×3mm region of expert-labelled GT patches for a sample WSI.

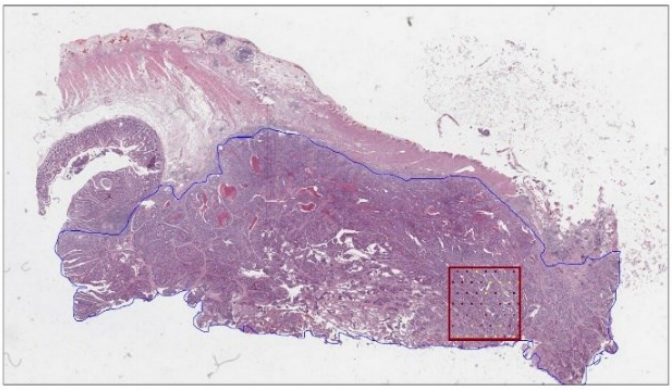


Figure 28: Original WSI with ground truth annotations

Red box = pathologist-selected sampling area for TSR calculations
Blue outline = ground truth ROI

This sampling approach is not possible for previously unseen WSIs, where there is no pathologist-defined 3mm sample box. Nonetheless, TSR results at these locations allowed a baseline error rate to be calculated, for comparison with other sampling strategies.

B. Sampling over Predicted ROI

Initially, a pattern of approximately 100 patches was generated on a regular grid occupying the entire predicted ROI. Patches were extracted from the WSI at these locations, with sizes of 224×224px or 256×256px depending on the CNN model used. The patches were classified using the CNN, after which the totals of patches having class *tumour* or *stroma* were used to calculate the TSR.

C. Sampling above 80% Tumour Density in ROI

A *KernelDensityEstimator* (KDE) class encapsulated the *gaussian_kde* function in the *SciPy.stats* package (SciPy community, 2021). The KDE took a cluster of points with arbitrary X,Y coordinates, interpolating these into a regularised X,Y grid of density values. A further step converted the density grid into a contour enclosing density points with a value over 80% of the maximum tumour density.

The sampling region was defined as the intersection of 80% tumour density contour with the predicted ROI. From this region of the WSI, 100 evenly spaced patches were sampled on a square grid.

D. Sampling in 3mm Box at Maximum Tumour Density within ROI

This strategy was designed to emulate the behaviour of the pathologist in creating the ground truth annotations, by restricting the sample patches to a 3mm square 'virtual biopsy' region.

The tumour density matrix from the KDE was intersected with the predicted ROI. The centroid of the resulting region was taken as the approximate location of the peak tumour density. A 3mm (≈ 6000 pixel) square box was centred on this point. Inside the box, 100 patches were sampled for the TSR and PoT calculations.

This process was carried out twice, using collections of predicted patch classes from before and after the false positive correction (FPC) stage in the pipeline (Figure 18, CNN3). This was expected to reveal the effect on TSR accuracy of re-classifying some tumour patches as non-tumour epithelium.

E. Sampling in 3mm Box at Maximum Tumour Density within WSI

This strategy ignored the predicted ROI boundary and used the centroid of the >80% tumour region wherever that fell in the WSI. This approach was intended to mitigate any positioning errors arising from an inaccurate estimate of the ROI due to clustering errors or a sparse tumour distribution. The TSR calculation here was based on post-FPC classification results only.

F. RandomSpot Sampling in 3mm Box at Maximum Tumour Density within WSI

A 3mm box, located as above, was sampled using an algorithm based on RandomSpot (Wright et al., 2015), where a hexagonal grid was adopted to reduce sampling bias due to edge effects. Here, the algorithm was configured to arrange 120 patches on a hexagonal grid within the 3mm box.

A *RandomSpotSamplePatternGenerator* Python class was developed to apply the algorithm. Sampling points were arranged in a grid of equilateral triangles, with a random starting position. The grid spacing and starting point were then adjusted iteratively until the required number of sampling points fell within the specified boundary (Figure 33).

Analysis of Outliers

Ground-truth and predicted TSR values were plotted for all WSIs in the evaluation set. Bland-Altman (BA) plots (Altman and Bland, 1983) were used to identify WSIs with the largest errors. The difference between pairs of values, $\epsilon_{TSR} = TSR_{GT} - TSR_{predicted}$, was plotted against the mean of the pair, such that $\epsilon_{TSR} > 0$ represents underestimated tumour in the pipeline output.

4.4.3 Results

Sampling distributions

Figure 29 shows the output of the Kernel Density Estimator for tumour patches in the AHP, using the WSI shown in Figure 28. In Figure 30, the 80% density region has been used to generate a distribution of 100 sampling patches.

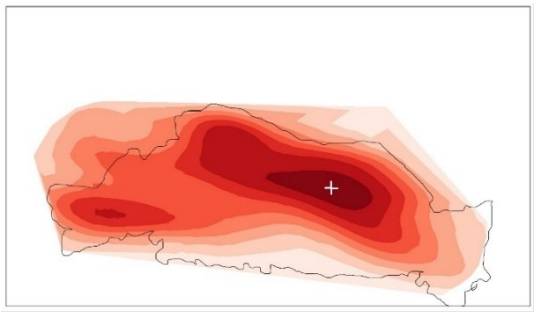


Figure 29: Predicted tumour density with maximum density point

Grey outline = ground truth ROI
White cross = maximum density using KDE

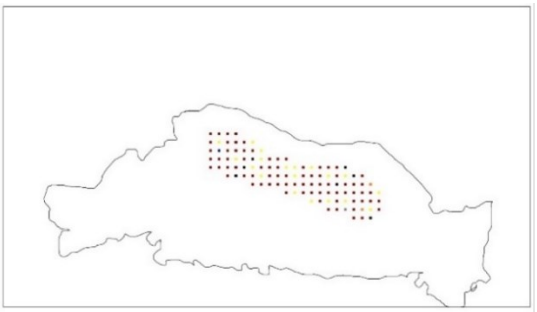


Figure 30: Patches sampled and classified in ≥80% density region

Figure 31 shows the tissue distribution predicted by CNN2, over approximately 100 patches sampled within a 3mm box centred on the maximum tumour density point, emulating the activity of the annotating pathologist. In Figure 32, a RandomSpot-derived hexagonal grid containing approximately 120 patches replaces the previous square grid. Figure 33 shows this distribution in close-up.

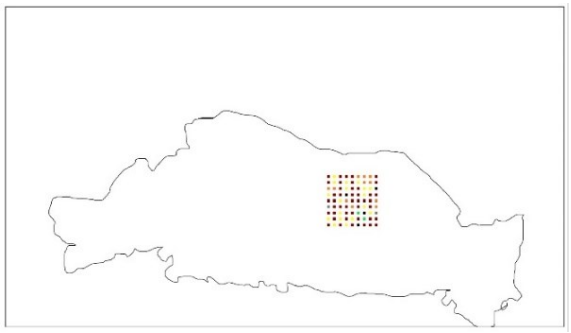


Figure 31: Patches sampled and classified in 3mm box at max predicted tumour density

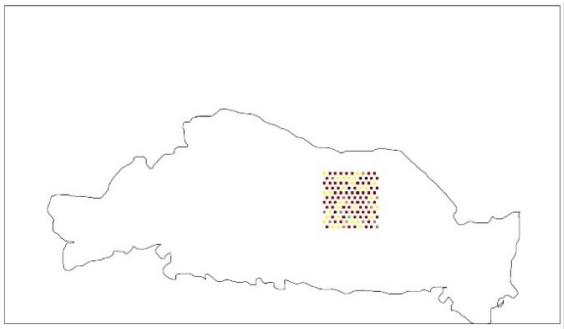


Figure 32: Patches sampled and classified in same 3mm box with RandomSpot layout

Figure 33: Close-up of RandomSpot patches showing hexagonal grid

Red=tumour, Yellow=stroma

TSR Accuracy

Table 14 shows the accuracy of the TSR calculation when sampling and classifying image patches according to the above regimes. These results were obtained using a pre-trained VGG19 as CNN2 in the Attention Heatmap Pipeline, over 100 WSIs in the hold-out evaluation set.

Table 14: Mean TSR error for various sampling strategies, using VGG19 in Attention Heatmap Pipeline

Sampling Region (using post-FPC tumour points unless stated)	TSR mean error as $TSR_{GT} - TSR_{predicted}$ with 95% CI		TSR mean error magnitude with 95% CI	
A) Ground Truth Locations	0.00	-0.02-0.01	8.67%	6.91-10.43%
B) Predicted ROI	0.02	-0.00-0.05	22.32%	14.45-30.18%
C) Maximum Tumour Density (>80%) in ROI	-0.08	-0.11--0.05	28.43%	19.58-37.29%
D) 3mm Box at Max Tumour Density within ROI (pre FPC)	0.00	-0.03-0.0	24.04%	18.36-29.72%
D) 3mm Box at Max Tumour Density within ROI	0.01	-0.02-0.05	23.51%	17.90-29.13%
E) 3mm Box at Max Tumour Density within whole WSI	-0.03	-0.06--0.01	23.14%	15.86-30.42%
F) 3mm Box at Max Tumour Density Point with RandomSpot layout	-0.05	-0.07--0.02	23.32%	16.69-29.94%

Analysis of Outliers

Figure 34 shows the distribution of predicted and actual TSR for the 100 evaluation WSIs, as a Bland-Altman plot. WSIs outside the ± 1 Standard Deviation (SD) band are shown in red.

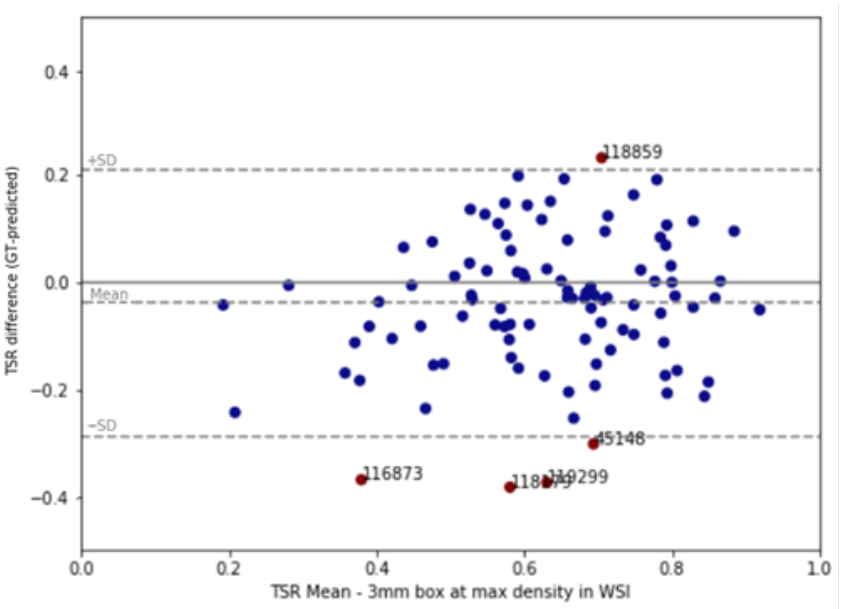


Figure 34: Bland-Altman plot of TSR errors, for 3mm sampling box at max tumour density, using attention-based pathway with VGG19 classifier.

QUASAR image number is given for outliers (red), further than ± 1SD from the mean TSR difference.

Pipeline-generated plots for one such outlier, 119299.svs, are shown in Figure 35 to Figure 38.

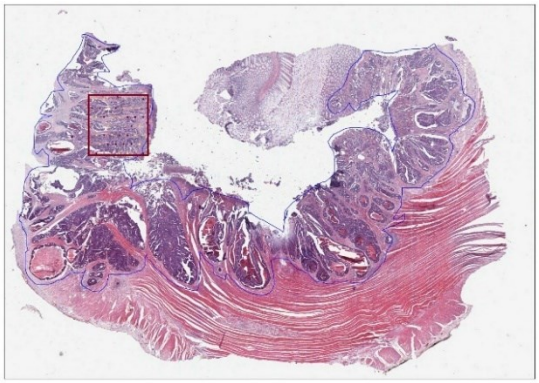


Figure 35: Example outlier: WSI with heterogeneous tumour distribution (QUASAR WSI 119299.svs).

Red outline = 3mm sampling region used for pathologist's GT annotation.

Figure 36: Thumbnail-derived tumour heatmap for 119299.svs

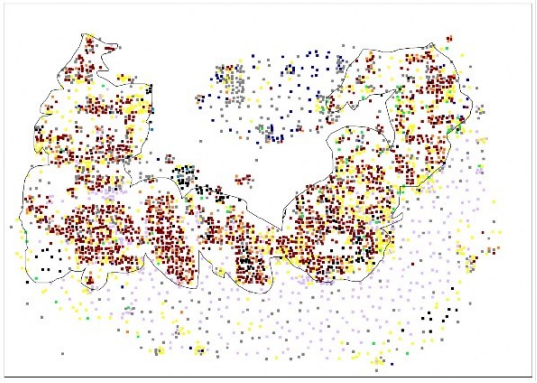


Figure 37: Patch classification results at locations determined by heatmap

Red = tumour
Yellow = stroma

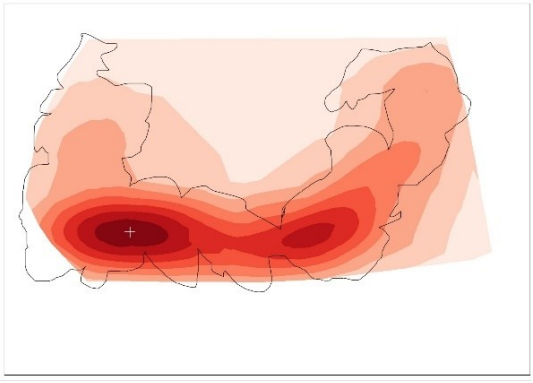


Figure 38: KDE tumour density plot showing max tumour density (white +)

This occurs at a different point to the pathologist's GT sampling location shown in Figure 35.

4.4.4 Discussion

TSR error magnitude was smallest when sampling at the original locations used to record the ground truth classifications. This is as expected, as the classifier was using patches centred on similar cell structures to those viewed by the annotating pathologist. The residual error at this sampling location is likely to be due to classification errors in the CNN, in combination with rounding errors in the TSR calculation due to the limited sample size of 50 patches per WSI.

Sampling within a 3mm box at maximum tumour density is closest to the behaviour of the annotating pathologist, and gave better results than sampling throughout the 80% tumour density contour. Using a RandomSpot-based sampling distribution gave a lower error than a square grid. This is as expected due to the reduced sampling bias associated with the RandomSpot pattern. Some benefit may also be due to the higher number of sample patches fitted into the 3mm box by the RandomSpot algorithm.

Sampling across the whole predicted ROI gave marginally lower error rates than sampling in a 3mm box at maximum tumour density. However, the error percentage was still high and would

result in many patients being mis-classified as stroma-high or stroma-low. Further analysis was undertaken to understand the causes of this deviation.

Analysis of Outliers

Many WSIs gave predicted TSRs that were close to the ground truth values. However, the Bland-Altman plot (Figure 34) identified 4 WSIs below the lower SD line, implying a predicted TSR that is significantly higher than reality. Since high TSR is associated with better patient outcome (Hutchins et al., 2018), the overestimated TSR is potentially dangerous in a clinical setting it implies that disease severity is underestimated for these patients, who may then not receive the treatment they need.

Examining one such outlying WSI, 119299.svs, revealed how an anomalous TSR result can arise using the Attention Heatmap Pipeline. Tumour tissue was unevenly distributed in the ROI, resulting in a fragmented thumbnail heatmap (Figure 36). This meant that fewer patches were sampled in some parts of the ROI, leading to lower densities of predicted tumour points. The estimated max tumour density point (Figure 38), where the TSR was calculated, did not coincide with the ground truth sampling location (Figure 35). With a heterogeneous specimen such as this, it was not surprising to observe a large difference between the proportions of cells sampled at the two locations.

Nonetheless, the AHP generated accurate TSR results in many cases. It has also enabled the detection of tumour regions using far fewer patches than would be used in tile-by-tile sampling. However, the observed variations in predicted tumour density, unhelpfully modulated by the sampling density determined by the low-resolution heatmap, resulted in errors in estimated TSR that could not safely be ignored in a clinical setting.

Pipelines that attempt to avoid this scenario, using more uniform sampling regimes, are explored in the following sections.

4.5 Tile-by-Tile Processing Pipeline

4.5.1 Motivation

A tile-by-tile pipeline (TTP) was now required, to provide a baseline for comparison with the AHP and subsequent attention-based systems in this chapter. The TTP would perform patch classification on every tile in a WSI, to provide baseline ROI and TSR error rates measured against the ground truth annotations. These measurements, and the total processing time per WSI, would be used for assessing the relative performance of more selective sampling algorithms.

4.5.2 Methodology

Figure 39 shows the architecture of the TTP. The pipeline configuration, including false positive correction and estimation of TSR and ROI, was the same as used in the attention-based and weighted regular sampling algorithms, using VGG19 for CNN1 and CNN2. A tile size of

224×224px was used to match the CNNs' default input size.

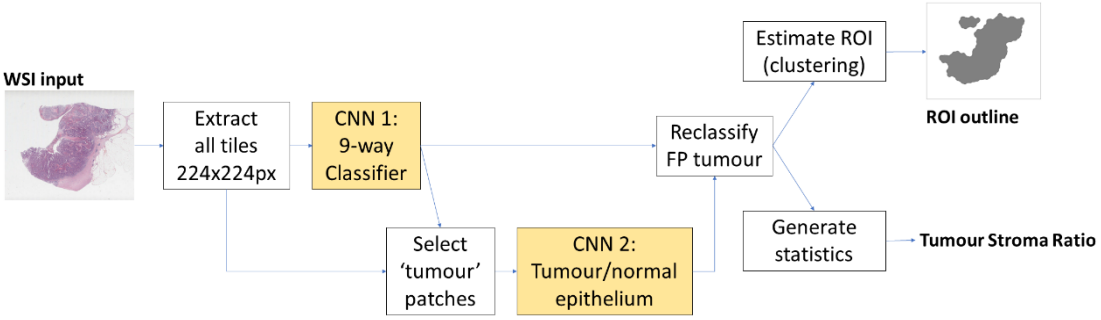


Figure 39: Tile-by-tile WSI processing pipeline

The TTP was executed on ARC4 with 100 hold-out WSIs that were unseen during CNN training and measurement.

The earlier Attention Heatmap Pipeline was also executed using VGG19 classifiers, to facilitate like-for-like comparison with the TTP.

4.5.3 Results

Table 15 shows the comparative performance of Attention Heatmap and Tile-by-Tile pipelines, in terms of patches sampled, and ROI and TSR agreement with the GT annotations. The grid size given for the tile-by-tile pipeline is equivalent to the patch size, while the value for the Attention Heatmap Pipeline represents the 'parent' grid box from which patches are sampled. The 'grid px' size refers to WSI pixels at 20× magnification, corresponding to 0.49μm/pixel.

Table 15: Comparative performance of tile-by-tile and attention-based WSI pipelines

Pipeline	Grid size px	Grid size μm	Time per WSI (mins)	Patches per WSI	Tumour patches	Tumour ROI agreement (F1 score)	Lowest TSR error	Lowest TSR RMSE
AHP	1024	502	10	3029	1357	76.1%	21.6%	12.2%
TTP	224	110	145	33065	7078	89.3%	25.1%	13.5%

4.5.4 Discussion

The AHP was over 14x faster than the Tile-by-Tile pipeline, which took 145 mins per WSI and processed 11x more patches than the AHP.

This was measured on a single GPU node on ARC4 prior to the introduction of automated CPU parallelisation, which would dramatically reduce processing time in later pipeline experiments (Chapter 8). Nonetheless, the introduction of the TTP benchmark confirms that attention-guided sampling yields a substantial improvement in processing.

The TTP predicted the ROI outline more accurately than the AHP, with nearly 90% agreement (by F1 score) with the ground truth. This appears to be due to the greater number of patch coordinates available to the DBSCAN clustering algorithm. However, the AHP yielded marginally more accurate prediction of TSR.

Further experiments would now aim to improve the accuracy of ROI and TSR prediction using a more consistent attention-based sampling regime, using the TTP as a baseline for performance comparisons.

4.6 Weighted Regular Sampling with Attention

4.6.1 Motivation

This novel approach for selecting patches from the WSI builds upon a published *Quasi Monte Carlo* method (Cruz-Roa et al., 2018) and was published in the *Journal of Pathology Informatics* (Broad et al., 2022).

The Cruz-Roa algorithm determined the tumour ROI starting from a sparse grid of sampling points, classifying patches at these locations, then iteratively increasing sampling density in areas of high gradient in tumour probability, corresponding to transitions between tissue regions, to define the boundary of the ROI.

In the current work, the algorithm was further required to generate a uniform sampling distribution within areas of suspected tumour, for example for TSR estimation. A modified sampling regime was therefore proposed, with additional points being sampled in a regular distribution within the estimated ROI.

It was also expected that uniform sampling within the ROI would mitigate localised spatial biases due to the variable sampling density inherent in the Attention Heatmap Pipeline, supporting a more reliable choice of TSR sampling point.

4.6.2 Methodology

Figure 40 shows the Weighted Regular Sampling Pipeline (WRSP) architecture used in this experiment. Code used in this section is documented in Appendix Section 1.2.2.

Initially a sparse but uniform distribution of patches was extracted from the WSI, and classified by a CNN, resulting in data points such as in Figure 41. If a patch was identified as *tumour,* further sampling patches would be added around it.

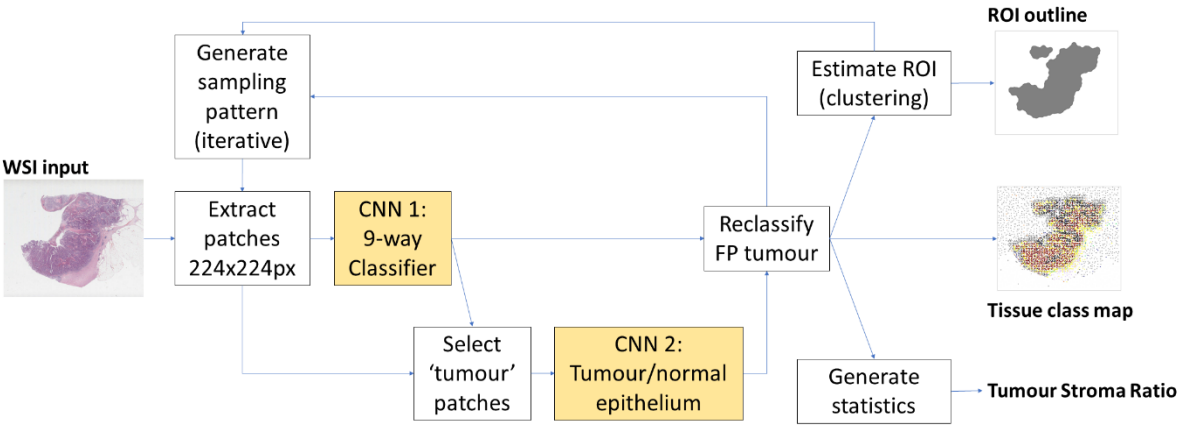


Figure 40: Weighted Regular Sampling Pipeline

Predicts tumour region of interest, tissue class distribution and TSR value for a given WSI.

These new patches were also input to the classifier. This often resulted in additional predicted tumour locations. The resampling process was repeated for one to two further iterations, to capture further detail around the new tumour patches. This resulted in a plot where the tumour region was largely filled by higher-density, regular grids of classified patches (Figure 42).

The clustering algorithm in Section 4.2.2 was then used to estimate the overall ROI around the predicted tumour. A *convex hull* was drawn around the clustered regions, then the whole area within was sampled and classified at the higher sampling density.

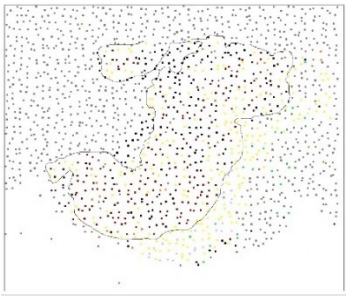


Figure 41: Initial sparse sampling distribution for QUASAR 45258.svs

Red=tumour
Yellow=stroma

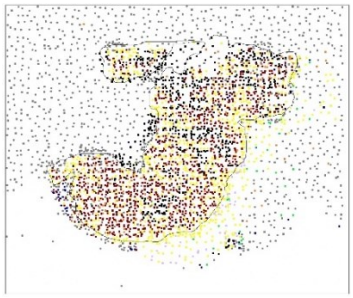


Figure 42: Patch distribution after resampling areas around tumour patches

Figure 43: Patch distribution after resampling within convex hull of predicted ROI

Uniform higher density here improves ROI and TSR prediction

This resulted in a uniform region of sampling (Figure 43) that was expected to enclose all relevant tumour and stroma patches. Where the clustering yielded multiple distinct groupings of predicted tumour, rather than a single continuous ROI, these would be joined together at this stage. A further clustering operation would then be performed on these uniformly higher-density tumour patches, to provide a yet more accurate estimate of the ROI.

False-positive correction was applied during the above process, using CNN2 to exclude normal epithelium patches mis-classified as tumour.

Tumour Stroma Ratio

TSR was calculated from the collections of classified patches, using sampling methods described in Section 4.4.

An additional method exploited the uniformly distributed set of patches now available across the predicted ROI. Python code was developed to apply a 3mm square sliding window to these patches, recording totals of tumour and stroma patches at 3mm intervals within the ROI. This enabled density plots to be generated for tumour and stroma, and therefore TSR, inside the predicted ROI.

4.6.3 Results

The following results were obtained from pipelines using ImageNet-pretrained VGG19 models, at a parent grid size of 1024px or 502μm unless otherwise stated.

ROI Prediction

Table 16 shows the accuracy of predicting the ROI, relative to the GT annotations, using the WRSP with 1 and 2 resampling iterations. Previous results from the Attention Heatmap Pipeline (AHP) are included for comparison. ROI accuracy is expressed as Intersection over Union (IoU) and F1/Dice score.

Table 16: ROI Prediction accuracy for WSI processing pipelines with 1024px grid size

Pipeline	Re-sampling Iterations	Tumour % in ROI (post FPC)	IoU w.r.t GT ROI	IoU 95% CI	F1 w.r.t GT ROI	F1 95% CI
AHP	n/a	96.9	66.28%,	62.04-70.52%	77.29%	73.39-81.18%
WRSP	1	96.4	70.89%	66.60-75.18%	80.76%	76.97-84.54%
WRSP	2	93.3	74.20%	70.56-77.83%	83.58%	80.45-86.72%

Tumour Stroma Ratio

Table 17 summarises the accuracies of TSR predictions obtained using the TSR patch sampling methods employed in the pipeline. TSR mean error is expressed as $TSR_{GT} - TSR_{pred.}$ such that an excessively high TSR prediction results in a negative value. Results in bold are referenced in the Discussion section below.

Table 17: TSR accuracy statistics for Weighted Regular Sampling Pipeline

Sampling Region (using post-FPC tumour points unless stated)	TSR mean error as $TSR_{GT} - TSR_{pred.}$ with 95% CI		TSR mean error magnitude with 95% CI	
Ground Truth Locations	0.00	-0.02-0.01	**8.67%**	6.91-10.43%
Predicted ROI (100 points)	**0.08**	0.05-0.10	**22.87%**	18.09-27.65%
Maximum Tumour Density (>80%) in ROI	-0.05	-0.08--0.02	24.01%	18.00-30.01%
3mm Box at Max Tumour Density within ROI (pre FPC)	0.00	-0.03-0.03	26.77%	20.06-33.47%
3mm Box at Max Tumour Density within ROI	0.01	-0.02-0.05	26.96%	20.28-33.65%
3mm Box at Max Tumour Density over whole WSI	-0.01	-0.04-0.01	23.25%	17.17-29.33%
3mm Box at Max Tumour Density Point	-0.05	-0.07--0.02	20.54%	15.80-25.28%
3mm Box at Max Tumour Density Point with RandomSpot layout (120 points)	-0.03	-0.05--0.00	**19.86%**	14.33-25.38%
Mean sliding window output over predicted ROI	0.03	0.01-0.06	**21.69%**	16.27-27.11%

Effect of varying pipeline parameters

Table 18 shows the effect of varying grid size and number of resampling iterations, on the accuracy of predicting TSR and the ROI outline.

Table 18: WRSP performance metrics with varying grid size and resampling iterations

Pipe-line type	Grid size px	Grid size μm	Re-sampling iterations	WSI proc. time (mins)	Patches per image	Tumour ROI: F1	Lowest TSR error	Lowest TSR RMSE	TSR sampling location
AHP	1024	502	-	10	3029	76.1%	21.6%	12.2%	ROI
WRSP	1024	502	1	23:19	4041	79.3%	19.9%	12.7%	ROI
WRSP	1024	502	2	23:05	7007	83.0%	19.4%	12.9%	dmax
WRSP	768	376	1	25:12	6542	83.2%	21.2%	12.7%	dmax
WRSP	768	376	2	36:27	12016	86.5%	20.3%	11.8%	dmax
WRSP	640	313	1	29:05	7826	83.6%	18.5%	11.3%	dmax
WRSP	640	313	2	44:18	14257	86.6%	19.5%	12.4%	sliding
TTP	-	-	-	145	33065	89.3%	25.11%	13.5%	dmax

TSR sampling locations:
ROI = predicted ROI (100 patches)
dmax = 3mm box at max tumour density point with RandomSpot layout (120 patches)
sliding = sliding 3mm window within predicted ROI

Further analysis of TSR performance

Figure 44 through Figure 46 show distributions of tumour and stroma patch densities, and the resulting TSR distribution, for a sample WSI.

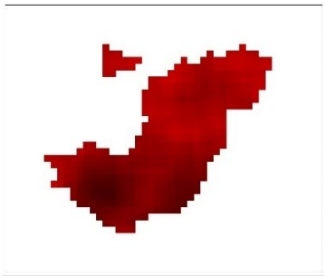


Figure 44: Tumour distribution for 45258.svs using sliding 3mm window in predicted ROI

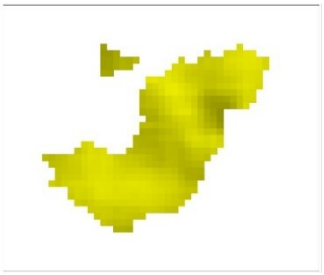


Figure 45: Stroma distribution for 45258.svs using sliding 3mm window in predicted ROI

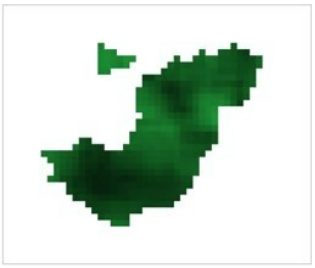


Figure 46: TSR heatmap derived from tumour and stroma distributions

The Bland-Altman plot in Figure 47 shows the distribution of TSR errors from the 100 evaluation WSIs. Outliers in red show the associated SVS image number.

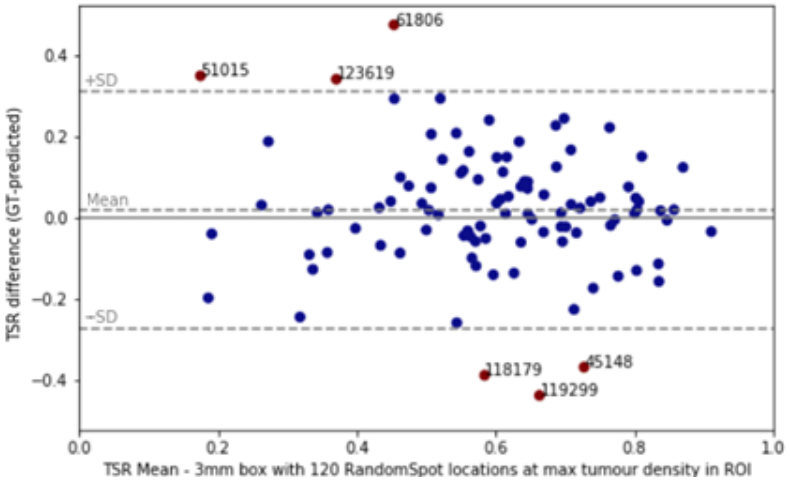


Figure 47: Bland-Altman plot of best-case TSR errors in weighted regular sampling pipeline, using VGG19 classifier.

Outliers in red (>1SD from mean) represent the greatest discrepancy in TSR from the ground truth value, often due to sampling location being different to that used for the GT.

Figure 48 shows the WSI with the GT sampling region overlaid. The pipeline sampling locations are shown in Figure 49, against the GT ROI outline.

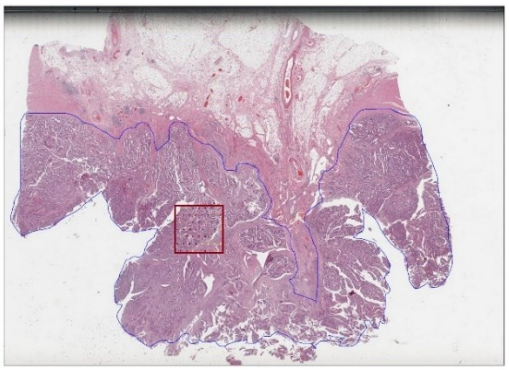


Figure 48: WSI 61806.svs with pathologist's chosen 3mm sampling region (red box)

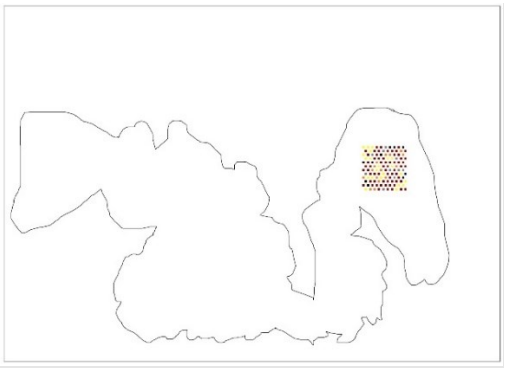


Figure 49: TSR sampling locations for WSI 61806.svs, using 3mm box at predicted maximum tumour density.

Location determined by pipeline differs from pathologist's ground truth sampling region.

4.6.4 Discussion

The Weighted Regular Sampling Pipeline (WRSP) replaced the thumbnail-derived heatmap from the earlier Attention Heatmap Pipeline (AHP) with a more uniform sampling pattern. The new technique remains attention-inspired, starting from 'glimpses' of cells taken a low sampling density, then directing further processing to where features of interest (tumour patches) are detected.

ROI estimation

The ROI predicted by the WRSP is significantly more accurate than that from the AHP, when comparing IoU with 95% confidence intervals. IoU is increased by approximately 4 percentage points (pp) when using one iteration of resampling around tumour patches. This increases by a further 4pp with an additional resampling iteration. F1 score similarly increases by approximately 3pp between pipeline variants.

It appears that the WRSP, particularly when using an additional sampling iteration, supplies a dense distribution of patches in the tumour region, which enhances the spatial definition of the derived ROI outline.

TSR estimation

Error rates in TSR estimation have decreased relative to the AHP. Remaining errors are thought to be due to the accuracy of the CNN classifier, and to the pipeline's choice of sampling location within the ROI.

Averaging TSR over the whole predicted ROI gives an error size of 22.87% (Table 17), but introduces a positive bias. This represents underestimated TSR, or overestimated stroma levels, which would result in an unduly pessimistic prognosis for the patient. Using the sliding window method within the ROI gave a marginally more accurate result, with reduced bias. The larger TSR error here reflects the difference in sampling distributions, between the area of highest tumour cell density chosen by the annotating pathologist, and the larger ROI enclosing varying cell densities.

Sampling in a 3mm box at maximum tumour density gave the most accurate TSR result of the sampling methods that are compatible with unseen WSIs. The approach gave an error rate of 19.86%, but is known to be sensitive to variations in tumour density at the chosen sampling location.

Unsurprisingly, the TSR error was lowest, at 8.67%, when sampling at locations chosen by the pathologist for the ground truth tissue annotations. Having eliminated errors attributable to variations in TSR sampling location, this remaining error is likely to reflect accuracy limitations in the CNN classifier, and uncertainty in the original annotations due to tissue heterogeneity in the GT patches.

The TSR density plots offer further insight into tissue distribution within the ROI. In Figure 44, the highest density of tumour is represented by the dark region at the bottom left of the ROI. The density distribution of stroma was approximately the inverse of this (Figure 45), but the combined TSR varies widely over the ROI (Figure 46).

This highlights the importance of a suitable method for selecting and aggregating spatially varying TSR distributions, when a single score is required for the WSI. The pipeline's sampling location for TSR measurements does not always align with that chosen by the pathologist. For example, the annotating pathologist favoured sampling locations near to the luminal aspect (interior bowel wall), a structure which the pipeline did not attempt to identify. Discrepancies in sampling location were seen to account for many of the larger errors in TSR.

Effect of varying pipeline parameters

The accuracy of ROI prediction, measured by F1 score, has improved in the WRSP (79.3% to 86.6%, Table 18) with respect to AHP (76.1%). The most accurate predictions correspond to the largest number of patches sampled.

TSR errors were marginally reduced relative to the AHP. It appears that the increased number of sampling points in the ROI facilitates more precise estimation of the location of peak tumour density, in cases where this is used to determine TSR sampling location. The greatest reduction in TSR error occurred with the smallest parent grid size of 640px using a single resampling iteration. For most pipeline configurations, the most accurate TSR predictions occurred when sampling over a RandomSpot-based triangular grid of 120 patches at the point of maximum estimated tumour density (dmax).

WSI processing time is greater in the WRSP than in the AHP. WSI processing time increases in proportion with the number of patches being processed. This in turn increases with the number of resampling iterations and with decreasing grid size. Nonetheless the WRSP remains substantially faster than tile-by-tile processing.

Analysis of TSR outliers

The most extreme outlier in the BA plot in Figure 47, WSI 61806, represents a much lower predicted TSR than the ground truth, with a TSR difference of nearly 0.5 out of a possible 1.0. This implies that stroma levels were overestimated in the pipeline, which might lead to a pessimistic survival prediction for the patient, perhaps resulting in unnecessarily aggressive treatment regimes.

The sampling box placed at peak predicted tumour in Figure 49 was found to be in a different part of the ROI from the pathologist's selection, shown in Figure 48. The latter was placed near to the luminal aspect (interior bowel wall), a consideration not made by the pipeline when selecting its sampling region. The extreme difference in TSR prediction in this outlying case may be due the very heterogeneous nature of the tumour ROI, which would account for large variations in tumour density between these sampling locations.

Later work sought to improve classification accuracy of the CNN used in this pipeline, supporting the selection of optimum sampling locations for more accurate TSR and ROI predictions.

5 Feedback Attention

5.1 Feedback Attention Ladder CNN (FAL-CNN)

5.1.1 Motivation

Chapter 4 demonstrated a processing pipeline for extracting diagnostically useful information from colorectal cancer WSIs. The accuracy of predictions of tissue distribution, ROI outline, and TSR, appeared to depend on the accuracy of the CNN used for patch classification. At this stage, the best-performing CNN model was the VGG19.

This chapter investigates the benefit of adding top-down attention pathways to this widely used feedforward classifier. Literature (Kubilius et al., 2018; Tsuda et al., 2020; Tomar et al., 2022) demonstrated variations on this approach that improve model performance, particularly where input samples have mixed or superimposed elements involving different image classes. It was expected that this scenario would occur widely in pathology image patches, due to their heterogeneous nature and variations in mounting, staining and scanning quality.

This chapter introduces a novel hybrid model derived from two prototype architectures, initially inspired by the above literature:

1) CNN with feedback loops local to each emulated brain region, V1, V2 and IT

2) Feedback models based on U-Net networks adapted to generate feedback activations from the output, to control spatial and channel attention in the input layers.

Models were first trained for the task of classifying 9-class QUASAR patches (Section 3.1). In preliminary experiments, where each prototype architecture above was developed separately using a VGG19 backbone model, classification accuracy was enhanced in each case by around 2pp for QUASAR images (Section 3.2), rising to 8pp with the *uncertain-class-patches* subset containing patches with ambiguous or indeterminate content (Section 3.3.1).

The proposed hybrid model combined the two distinct neural behaviours from the prototype architectures, in the expectation that combining local and long-distance feedback pathways would further boost accuracy.

In initial prototypes, the use of multiple feedback iterations unexpectedly impaired classification performance. The predicted output class was determined by a decoder output generated in the last iteration alone – the model's "final answer". This implies a less accurate result in cases where the attention region drifts away from the target object during feedback iterations, or in scenarios where the initial prediction is the most accurate and subsequent feedback does not generate an improved output.

A further enhancement was therefore proposed, in which feature embeddings from the backbone encoder were aggregated during the initial forward pass and after each subsequent feedback cycle. This was designed to emulate localised working memory in a living cognitive system, to derive an output class prediction from the most relevant feedforward activations.

5.1.2 Methodology

Code used in this section is documented in Appendix Section 1.3.

The Feedback Attention Model CNN model, *FAL-CNN* is shown in Figure 50. The upper feedforward encoder path is based on a VGG19 'backbone' with multiplying feedback attention modules (FAM) inserted before the first convolutional layer of each spatial scale level.

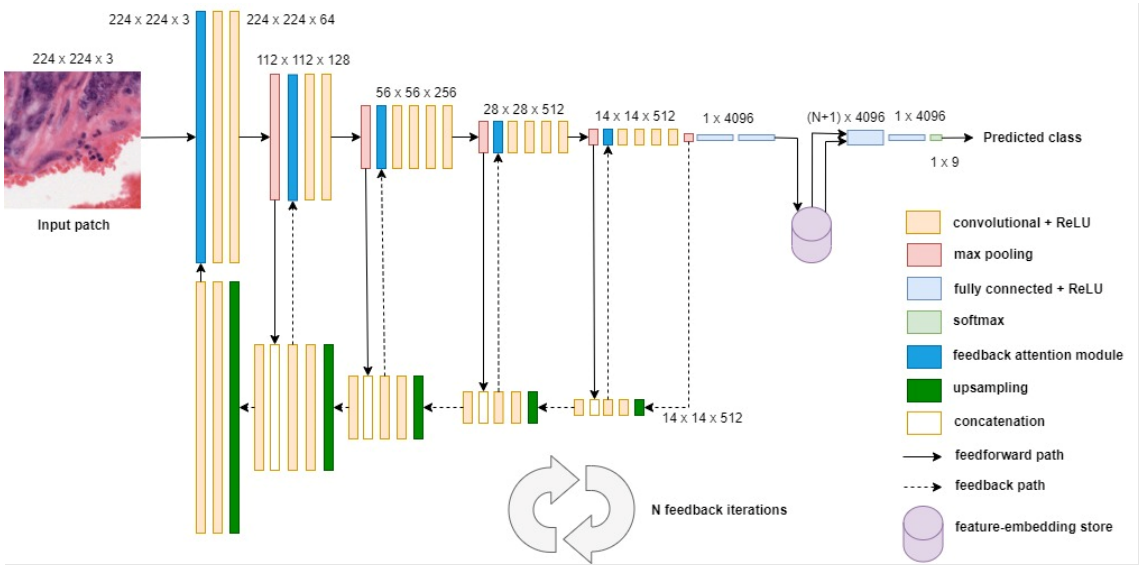


Figure 50: Feedback Attention Ladder CNN (FAL-CNN) model.

Incorporates multiple feedback neural pathways, and feature embedding store to aggregate feature embeddings over multiple processing iterations.

The FAM is shown in more detail in Figure 51. The module's output O, for input I and feedback tensor F with weights W_f and biases B_f, is given by:

$$O = I(W_f F + B_f) \tag{7}$$

This modulates feedforward activations at each pixel and each channel according to feedback activations in F. These are obtained from outputs of the decoder stages in the lower feedback path in Figure 50. This path corresponds to the decoder section of a U-Net model, and includes forward skip connections to preserve spatial resolution in the output.

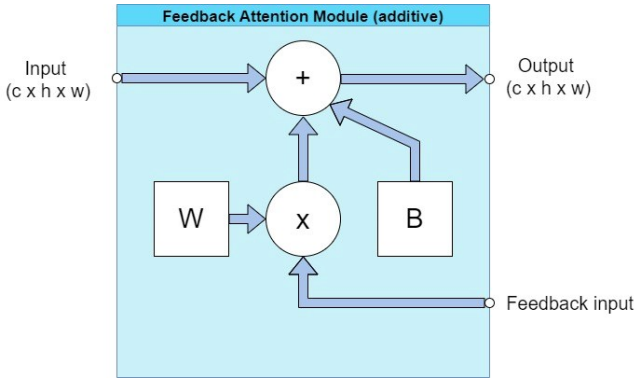


Figure 51: Multiplicative Feedback Attention Module

A feature-embedding store (FES) was used to combine encoder outputs across all feedback iterations. The FES used a tensor with dimension $BC(N+1)$, where N is the number of feedback iterations, C is the number of channels in the fully connected layers (=4096 for VGG19) and B is the image batch size. Each forward-pass result of size BC was stored in the FES, with a memory offset determined by the current iteration number. This resulted in an aggregated tensor containing embeddings from each forward pass.

An additional fully connected layer was then used to reduce the stacked embeddings in the FES back to size BC. This combined feature embedding was then used to generate the output class prediction via further fully connected (FC) layers and a final *softmax* stage, as used in the original VGG19 architecture.

The FAL-CNN was invoked according to the following steps:

Algorithm of FAL-CNN invocation

```
Input: Batch of patch image tensors

Initialise feature-embedding store tensor with size BC(N+1)

Apply null feedback to feedback attention modules (FAM) in feedforward
path

Call feedforward encoder layers on input image batch
Call 2x FC layers
Store FC layer output embeddings at offset 0

For each feedback iteration i:
      Call feedback decoder layers with current encoder output
      Apply decoder group output activations to corresponding FAMs
      Call feedforward encoder layers on input image batch
      Call 2x FC layers
      Store FC layer output embeddings at offset BC(i+1)

Call final FC layers with the N+1 stored embeddings

Return:
Batch of predicted output class probabilities from encoder output
```

Training

Model configurations with 0 to 4 feedback iterations to layers 0,5,10,19,28 were trained against the QUASAR 9-class patch set. The zero-iteration configuration was included to test the effect of the additional FC layer in the feedforward model with no feedback applied.

Training weights were initialised by copying from the PyTorch pre-trained VGG19 model into equivalent layers in the feedforward pathway. Models were trained for 200 epochs, with an initial LR of 0.0003 and momentum of 0.9, using the PyTorch *StepLR* learning rate scheduler to reduce the LR by a factor of 0.7 every 30 epochs. This was found to provide optimum convergence and a marginally higher classification accuracy than initial 100 epoch training with no LR reduction.

Statistics

Five-fold Cross Validation (CV) was used. Five data splits were defined, each with an 80%/20% split between training and test sets, allocated such that each test set contained a wholly different collection of images. Models were trained for each split, so that five accuracy scores were available for calculating a mean value.

Bootstrapping was performed by splitting each hold-out Test set into 6 sub-groups and performing inference on each patch in the sub-group. Thirty mean accuracy points were thus generated. These were used to calculate an overall mean classification accuracy $\bar{x}$ and standard deviation σ, from which 95% confidence intervals CI were calculated using the formula:

$$CI = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad (8)$$

The value of 1.96 above is the Z-score required to enclose 95% of a standard normal distribution within the confidence interval (BMJ, n.d.).

P-values were calculated using the Wilcoxon Rank Sum test, with the SciPy function *scipy.stats.ranksums().*

The trained model was further evaluated against the *uncertain-class-patches* dataset with 30 random samples, with mean and 95% CI being calculated as above.

Confusion Matrix difference plot

The Confusion Matrix (CM) difference plot was designed to show the effect of changes to a classifier model on the rates of correct and incorrect classification for each class. Existing code in the CNN training framework was used to generate and log a CM for the trained model, using the whole Validation set. CMs were subsequently extracted from training log files for models before and after the modification under test, these being the VGG19 backbone and the FAL-CNN feedback model.

A Jupyter notebook was used to parse each CM from text inputs, then subtract the reference CM from that of the FAL-CNN model under test (MUT):

$$CM_{diff} = CM_{MUT} - CM_{ref} \quad (9)$$

The difference CM was plotted with cells colour coded on a gradient from blue to red. On the leading diagonal, a positive difference was coded blue and a negative difference red. Off diagonal, the convention was reversed, with an increase in a given cell total resulting in a red coding. Thus, “good” changes contributing to improved classification accuracy were shown as blue, and “bad” changes highlighted in red.

5.1.3 Results

Results tables to complement the plots in this section, including mean accuracies, error bar ranges and p-values, are in Appendix section 2.1.1.

Results with QUASAR 9-class patches

Figure 52 and Table 21 (Appendix 2.1.1) show the mean classification accuracy and confidence intervals for the FAL-CNN feedback model, with 0 to 4 feedback iterations, after 200 training epochs. For these measurements, bootstrapping was employed in combination with 5-fold CV, allowing 95% CI bands to be calculated.

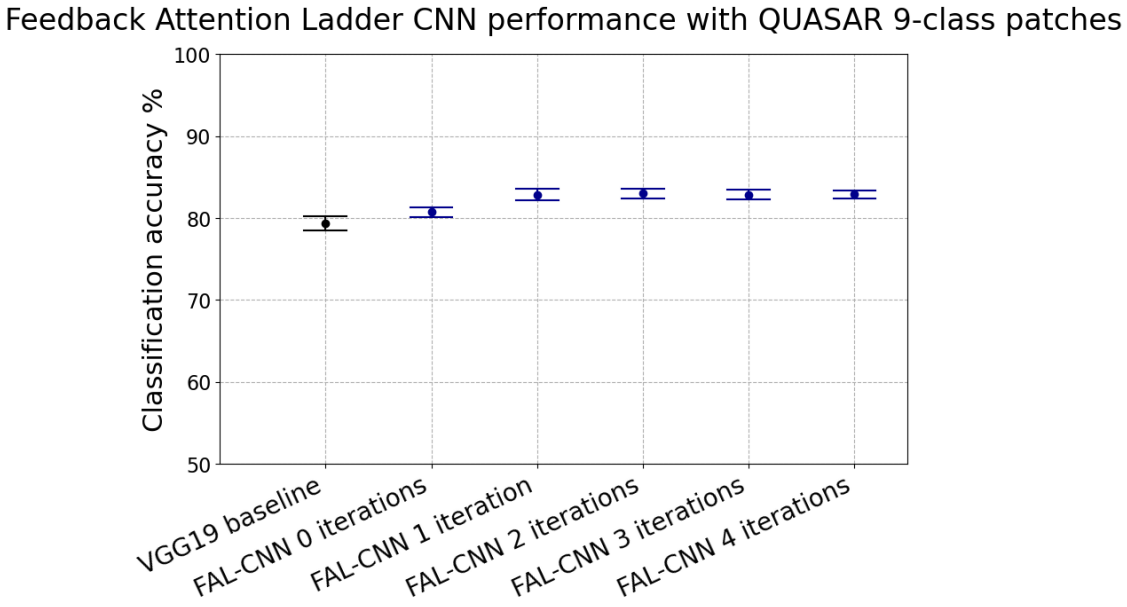


Figure 52: Classification accuracies relative to VGG19 with 95% confidence intervals, for FAL-CNN models with QUASAR 9-class patches

Results show significant benefit of using feedback attention architecture.

Results with uncertain-class-patches

Figure 53 and Table 22 show the classification accuracy for the FAL-CNN feedback model with 0 to 4 feedback iterations, when applied to the *uncertain-class-patches* dataset.

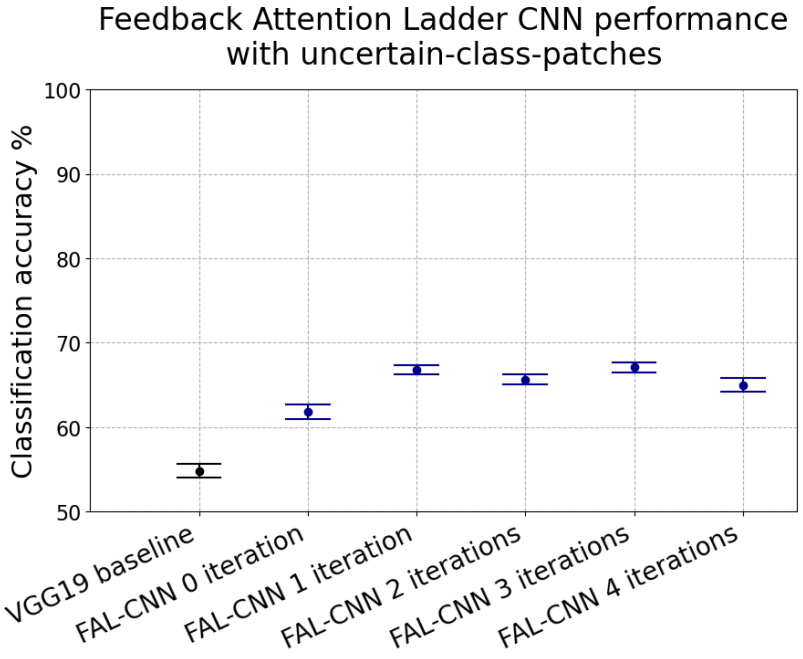


Figure 53: Classification accuracies relative to VGG19 with 95% confidence intervals, for FAL-CNN models with uncertain-class-patches dataset

Results show significant benefit of using feedback attention architecture, particularly in heterogeneous or cluttered input images.

Confusion Matrix difference plot

Figure 54 shows confusion matrices for the VGG19 baseline, and the FAL-CNN model, following training with the QUASAR 9-class dataset. The third panel represents the difference between the two matrices. Decreases in totals on the leading diagonal, and increases in off-diagonal totals, are shown in blue, while increased off-diagonal totals and decreased value on the leading diagonal are shown in red.

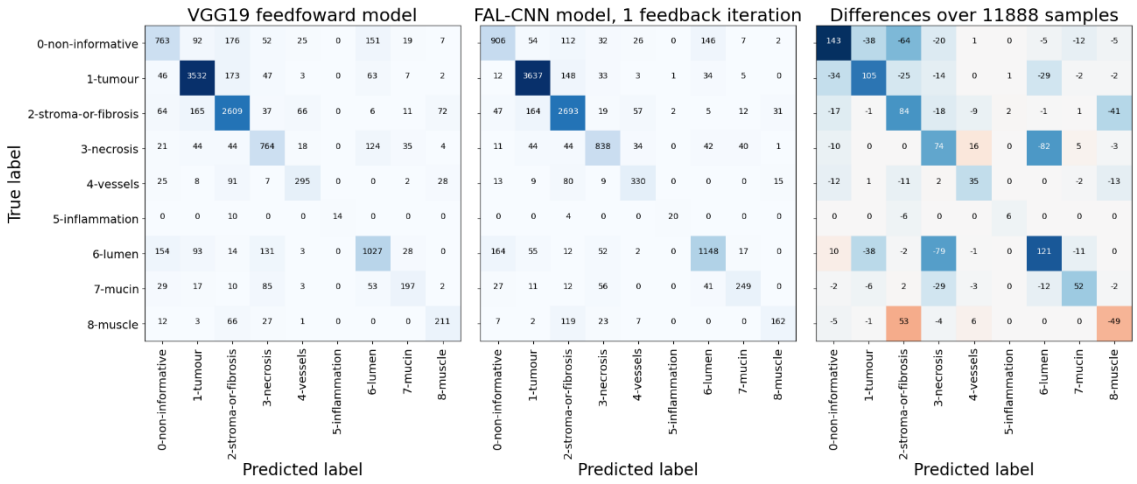


Figure 54: Confusion Matrix difference plot, between feedforward VGG19 and FAL-CNN model

RH panel shows increase in correct identifications (blue cells) due to the use of feedback attention, albeit with increase in incorrect identifications (red cells) for class muscle.

5.1.4 Discussion

In initial experiments, models with feedback enabled to layers 0,5,10,19,28 outperformed the model using only layers 5,10,19,28, confirming the benefit of feedback to the convolutional layers near the model input. These layers are associated with low-level image features such as colour and fine-grained edge detail, implying that the feedback model pays selective attention to such low-level image features.

The FAL-CNN model significantly outperformed the baseline VGG19, giving accuracy gains of approximately 3.5pp with one or more feedback iterations, validated by non-overlapping 95% confidence intervals and a p-value of $p<0.001$.

Interestingly, the zero-iteration feedforward-only configuration gave a 1.60pp increase over the VGG19, the only difference being the additional 4096-channel fully connected layer. It appears that this layer adds extra capability in discriminating the object classes implicit in the 4096-value embeddings at this level in the model, even before feedback is applied.

The use of one or more feedback iterations yielded further accuracy gains, which did not decline over multiple iterations. This suggests that the hybrid feedback system's "ladder" of multiple cross-connections between the feedforward and feedback paths acts to stabilise the feedback activations over multiple iterations.

In addition to the benefit attributed to the extra FC layer, it appears that the model's performance was further boosted by the use of a simulated "working memory" element to accumulate the results of multiple feedback iterations. It appears that output embeddings resulting from multiple feedback iterations can contribute usefully to the final class prediction, but also sometimes diverge from the required result. In the FAL-CNN model, helpful embeddings from earlier iterations – and also from the feedforward pass – are being retained and combined with later iterations' outputs. The model weights at this point have been trained

to provide an optimum mix of these, such that lower iterations' contributions may contribute a "sense check", helping to improve stability during multiple iterations.

When the FAL-CNN model was used with *uncertain-class-patches*, the increased stability over multiple feedback cycles was evident in the highest-yet accuracy gain of 12.26pp being measured after 3 iterations. However, the 1-iteration model is only marginally less accurate than this, and avoids the need for the extra computation required by further feedback iterations. Despite the single feedback cycle, the model still benefits from the "working memory" behaviour.

In the CM difference plot (Figure 54, RH panel), the strong blue leading diagonal shows an increase in correct identifications of all tissue classes, with the exception of *muscle*. Off-diagonal blue cells show a reduction in misclassifications, especially between *lumen* and *tumour* or *necrosis*.

The CM differences were mostly symmetrical around the leading diagonal, the exception being that less *stroma* was misclassified as *muscle*, but more *muscle* was incorrectly identified as *stroma*, than with the baseline VGG19. However, *muscle* is a relatively small class, and is in any case a 'sibling' of *stroma* when grouping into *tumour-group* and *stroma-group* parent classes. The majority of blue cells in the CM represent higher rates of correct classification with most tissue types.

The ladder-like structure of the feedback model represents an efficient implementation of channel and spatial feedback attention. The model includes convolution-based interconnections between all scale-levels, with each connection apparently contributing to the improved classification accuracy of the model. This feedback connectivity offers less independent control than would be possible in a *many-to-many* model having separate convolutional structures for every pathway, but maintains a more compact and efficient model structure.

Later chapters will examine the performance of the FAL-CNN model in the earlier WSI-processing pipeline, and will present visualisations of attention that reveal the simulated cognitive processes that have led to the improvements observed.

5.2 FAL-CNN Performance with Offset Patches

5.2.1 Motivation

It was anticipated that classifiers trained on QUASAR patch images would be most responsive to features near the input patch's centre pixel, at which the ground truth label applies. To explore this behaviour, patches were resampled to place the ground truth point at a different location in the patch, to examine the effect on model accuracy (this section) and attention distributions (Chapter 6).

5.2.2 Methodology

Patch images were sampled from QC-passed QUASAR WSIs, as described in Section 3.3.3, with the ground-truth location centred in the bottom right quadrant of each patch (Figure 9). This *offset-patches* dataset provided training data for the following experiments:

Model evaluated on offset patches

A VGG19 model, and FAL-CNN models using 1, 2 and 3 feedback iterations, previously trained on 9-class centre-labelled patches (Section 5.1), were subsequently executed on *offset-patches* data.

To mitigate overfitting, the offset images used in this evaluation were derived from WSIs listed in the unseen Validation set associated with the data split previously used in model training.

Model inference was performed on the ARC4 HPC, for 30 subsets randomly selected from the Validation set, to provide classification accuracy results for calculation of mean accuracy and 95% confidence intervals.

Model trained and evaluated on offset patches

The VGG19 and FAL-CNN model configurations were subsequently trained with the *offset-patches* dataset. Five-fold CV data splits and training hyperparameters were re-used from earlier model training with standard 9-class QUASAR data (Section 5.1.2). Mean classification accuracy and SE ranges, also obtained using *offset-patches,* were derived from post-training measurements across the five Validation sets.

5.2.3 Results

Results tables to complement the plots in this section, including mean accuracies and error bar ranges, are in Appendix section 2.2.1.

Model evaluated on offset patches

Figure 55 and Table 23 (Appendix 2.2.1) show classification accuracies measured against the *offset-patches* dataset, when processed with FAL-CNN model versions trained on the standard, centre-annotated QUASAR dataset.

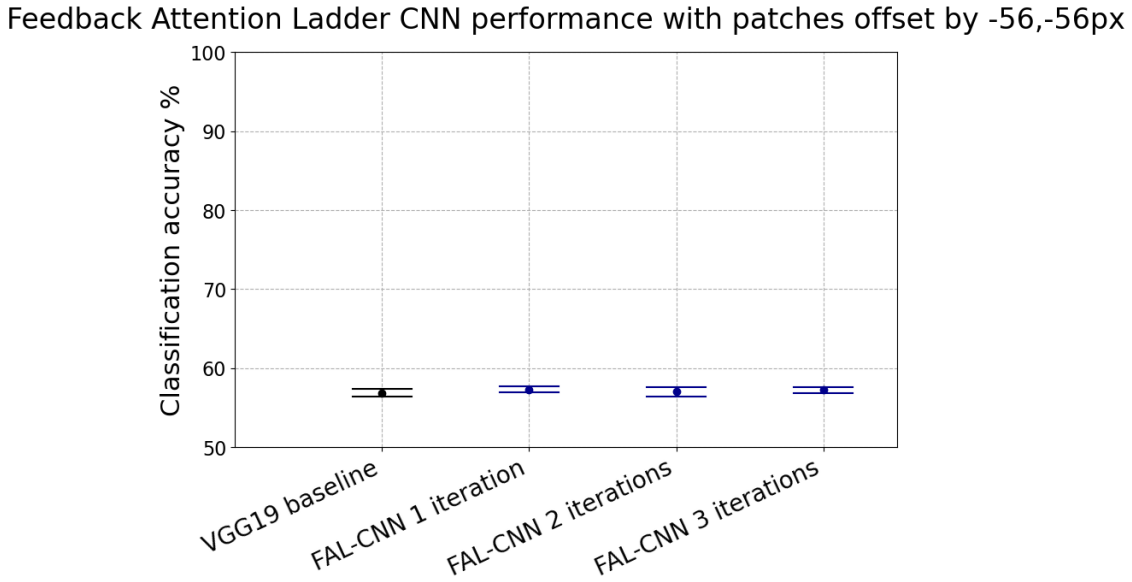


Figure 55: FAL-CNN classification accuracies relative to VGG19 with 95% confidence intervals, trained with standard patches then evaluated with offset-patches dataset

Reduced accuracy across all models shows impact of resampling heterogeneous tissue at a new location, without reassessing GT tissue class label.

Model trained and evaluated on offset patches

Table 24 shows classification accuracies measured after training the FAL-CNN model with the *offset-patches* dataset, alongside the accuracy gain in pp over the equivalent centre-trained model (Section 5.1.3, Table 21). Figure 56 shows classification accuracies with the centre-trained model results included for reference.

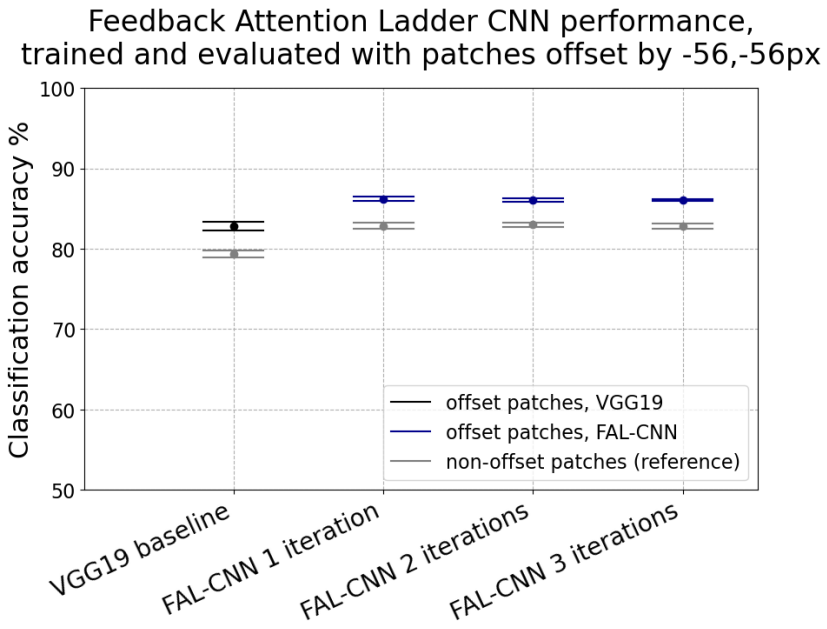


Figure 56: FAL-CNN classification accuracies relative to VGG19 with $\pm 1\ SE$ ranges when trained and evaluated with offset-patches dataset

Earlier non-offset results are shown in grey for comparison. The use of offset patches has boosted classification accuracy across all models.

5.2.4 Discussion

When a model, previously trained on patches centred around the GT pixel associated with the patch label, was then evaluated against *offset-patches*, a dramatic reduction in classification accuracy was observed.

The model was trained to focus on features supporting tissue identification in the centre of the patch, whereas the GT label now applied to another location in the image. Due to the heterogeneous nature of a CRC tissue sample at patch scale, it is likely that the tissue at the GT location is of a different type to that at patch centre, leading to a discrepancy in the predicted class.

Surprisingly, the models trained and evaluated with *offset-patches* showed consistently improved classification performance over models trained with centre-annotated patches. Placing the GT pixel near one corner means that the patch can encompass structural context (albeit only in one direction) from further away from this location than would be possible with centre-labelled patches. This may act as a proxy for using a larger patch size, without exposing the model to the confounding influence of a larger area of heterogeneous tissue types.

Further work examining the spatial distribution of attention within an offset-trained feedback model, to understand how these might influence the model's performance, is reported in Chapter 6, *Visualising Feedback Attention*.

5.3 FAL-CNN Performance with *tumour-stroma-groups* Patches

5.3.1 Motivation

Occam's Razor, also known as the Law of Economy, is a philosophical principle that favours the simpler of any two competing theories (Encyclopaedia Britannica, 2020). In this context, it appeared that the nine tissue classes represented undesirable complexity.

Wright (2017) has shown that TSR can be more accurately calculated when tissue classes are grouped into tumour and stroma 'parent' classes. It was anticipated that models trained with data grouped in this way would yield higher accuracy than observed with 9-class models.

Initial experiments used a dataset with 20,000 images for each class. This dataset was expected to yield the highest classification accuracy, due to the large number of images and the regularisation implicit in the use of rotated copies to balance file totals between classes.

An offset-sampled, grouped dataset was also created, to examine whether the accuracy gain observed with offset-sampled 9-class data (Section 5.2) would be repeatable with the 2-class dataset. A smaller image count of 12,000 per class was chosen to eliminate rotated copies, which would place the GT location in an unexpected quadrant.

A further dataset of 12,000 non-rotated, centre-sampled images per class was created to train models for comparison with offset-trained models.

5.3.2 Methodology

Two 2-class image datasets were derived from previously extracted 9-class QUASAR patches as detailed in Section 3.3.2, with 20,000 and 12,000 images per class respectively.

A VGG19 baseline model and 1, 2, and 3-iteration FAL-CNN models were trained using both datasets, over 50 training epochs for the VGG19 and 200 epochs for the FAL-CNN versions, with an initial LR of 0.0003 and momentum=0.9.

QUASAR offset tumour-stroma-groups

Further VGG19 and FAL-CNN models were trained with the *offset tumour-stroma-groups* dataset described in Section 3.3.4. Training hyperparameters were as used with the non-offset patch sets above.

5.3.3 Results

Results tables to complement the plots in this section, including mean accuracies and error bar ranges, are in Appendix section 2.3.1.

Model trained with tumour-stroma-groups, 20,000 patches per class

Figure 57 and Table 25 show classification accuracies measured after training FAL-CNN with the 20k-per-class *tumour-stroma-groups* dataset version.

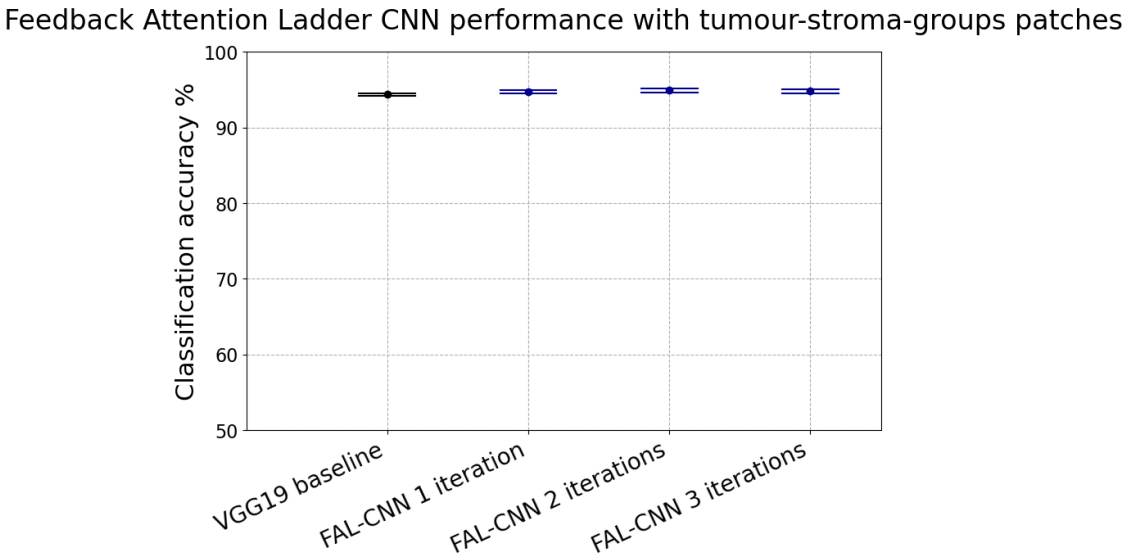


Figure 57: Classification accuracies with $\pm 1\ SE$ ranges for FAL-CNN relative to VGG19, with tumour-stroma-groups dataset

Accuracies are increased by reducing the number of output classes from 9 to 2.

Model trained with tumour-stroma-groups, 12,000 patches per class, un-rotated

Figure 58 and Table 26 show classification accuracies measured after training FAL-CNN with the un-rotated, 12,000 patch-per-class dataset.

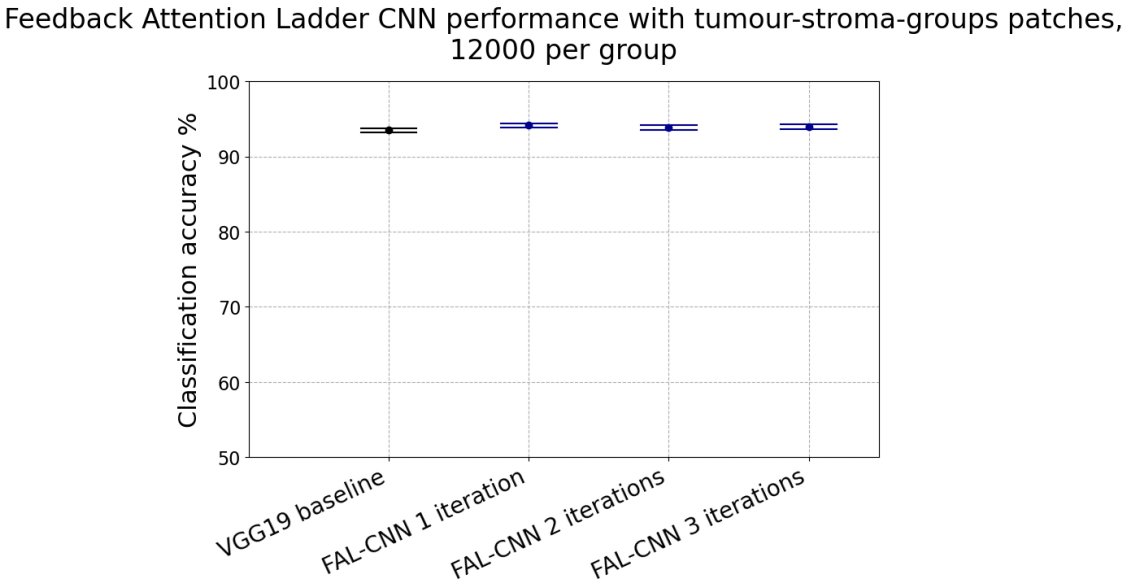


Figure 58: Classification accuracies with $\pm 1\ SE$ ranges for FAL-CNN relative to VGG19, with tumour-stroma-groups-12000 dataset

Model trained and evaluated on offset tumour-stroma-groups *patches*

Figure 59 and Table 27 show classification accuracies measured after training FAL-CNN with the offset-sampled 2-class *tumour-stroma-groups* patch dataset.

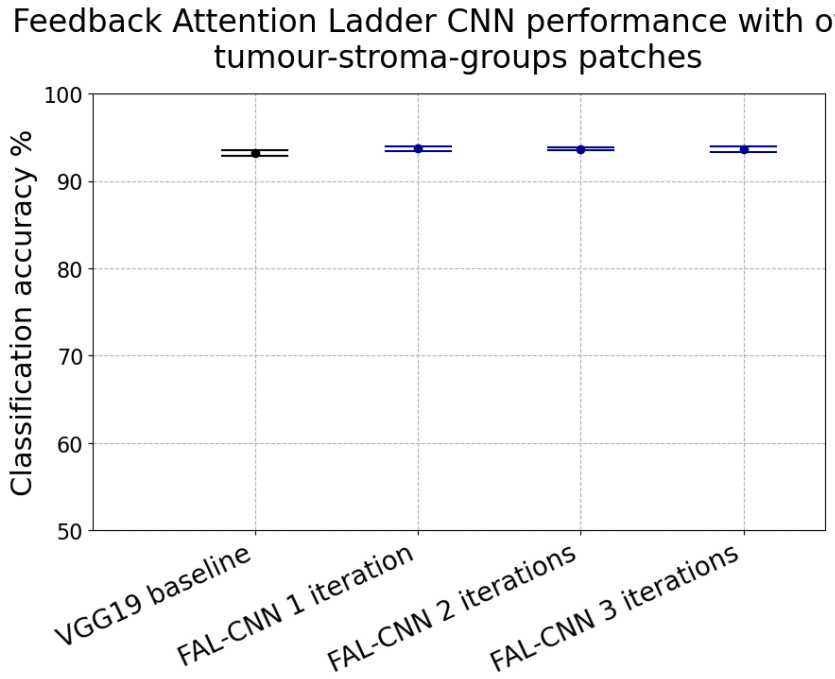


Figure 59: Classification accuracies with ± 1 SE ranges for FAL-CNN relative to VGG19, with offset tumour-stroma-groups dataset

The two-class classification task does not benefit from the use of offset patches.

5.3.4 Discussion

When training with two-class data, grouping the 9-class patch set into parent *tumour-group* and *stroma-group* classes, the resulting classification accuracy exceeded 90% for all baseline and feedback models assessed in this section (Figure 57). At 20k images per class, the VGG19 yielded 94.34%, rising to 94.9% with the 2-iteration FAL-CNN**.** This was marginally higher than the 94.77% of the 1-iteration version, but arguably does not justify the processing overhead of the extra feedback cycle.

The benefit of feedback is less pronounced than in the 9-class case, but is nonetheless substantial, as evidenced by the non-overlapping SE ranges. Also, for accuracies close to 95%, a 0.5pp change in accuracy equates to a 10% reduction in error rate, which can be attributed to the feedback attention mechanism.

With the 12k-per-class dataset, classification scores (Figure 58) were reduced by approximately 1pp for each model, relative to the 20k results (Figure 57). In the latter case, the models benefit from the implicit regularisation in randomly selecting from the extra rotated image copies, and the reduction of overfitting expected with a larger dataset.

The use of 2-class offset-sampled patches, grouped similarly into tumour and stroma parent classes, further reduced accuracy relative to the 12k, in contrast to the significant benefit seen in using the offset technique with 9-class data in Section 5.2.

Nonetheless, the large improvements observed with two-class data suggest that Occam's Razor has been successfully applied to the QUASAR patch set, resulting in 2-class models

capable of distinguishing *tumour-group* from *stroma-group* tissue with significantly higher accuracy than was observed with equivalent models performing 9-way classification.

5.4 FAL-CNN Performance with ImageNet-100

5.4.1 Motivation

ImageNet-100 (Shekar, 2021) contains 100 categories of animal, fish and bird images. VGG19 and FAL-CNN models were re-trained using this further data set because of its easily identifiable subject matter, in the expectation that image regions that were highlighted by the feedback attention mechanism could easily be compared with outlines and identifying features of the target object.

Training with ImageNet-100 would also help to determine the generalisability to other image sets of feedback attention architectures previously optimised for QUASAR data.

5.4.2 Methodology

ImageNet-100 data was downloaded and prepared as described in Section 3.4.2.

Five experiment data splits, with mutually exclusive Test sets, were defined for use in 5-fold cross validation. These were taken from the ImageNet-100 “Train”, within which 5-fold Train/Validation splits were derived for model training and evaluation.

VGG19 feedforward models and FAL-CNN models, with 1, 2 and 3 feedback iterations to layers 0,5,10,19,28, were trained with ImageNet-100 data. Stochastic Gradient Descent optimisation was used, with momentum = 0.9, initial LR = 0.0003 and LR reduction with $\gamma = 0.7$ per 30 training epochs. Weights in the feedforward path of each model were initialised from the pre-trained VGG19 downloaded from the PyTorch ‘Model Zoo’ (PyTorch, 2021). Training was performed over multiples of 50 epochs until convergence was observed.

This training process did not involve the separate Test set provided with ImageNet-100. This was subsequently used as a source of unseen images for further validation of trained models. For this, model accuracy was measured 30 times with random splits, from which the mean classification accuracy and 95% confidence intervals were derived.

5.4.3 Results

Results tables to complement the plots in this section, including mean accuracies, error bar ranges and p-values, are in Appendix section 2.4.1.

Figure 60 and Table 28 show mean classification accuracy with Standard Error (SE) ranges for the VGG19 feedforward model and FAL-CNN with 1, 2 and 3 feedback iterations. The total training epochs required to achieve convergence are also listed in Table 28.

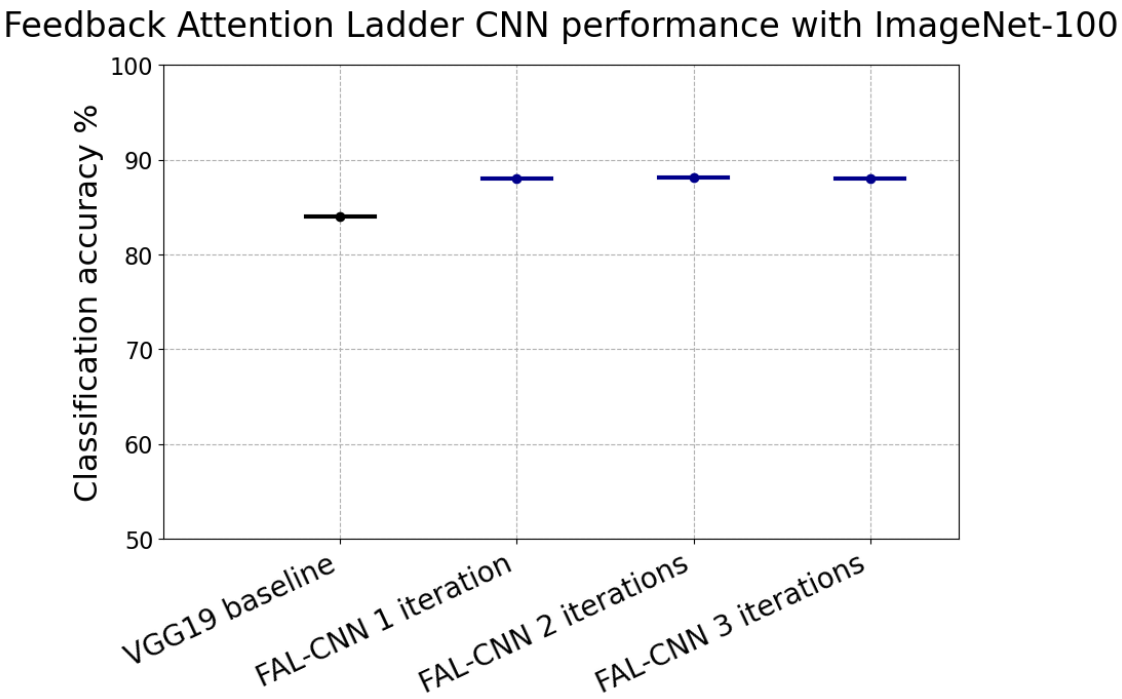


Figure 60: FAL-CNN classification accuracies relative to VGG19 with $\pm 1\ SE$ ranges, trained and evaluated with ImageNet-100

Results show benefit of incorporating feedback architecture but also reflect overfitting, as discussed in Section 5.4.4

Table 29 and Figure 61 show mean classification accuracy with 95% CI for VGG19 and FAL-CNN with 1, 2 and 3 feedback iterations, when evaluated with 30 random splits against the previously unseen ImageNet-100 Test set.

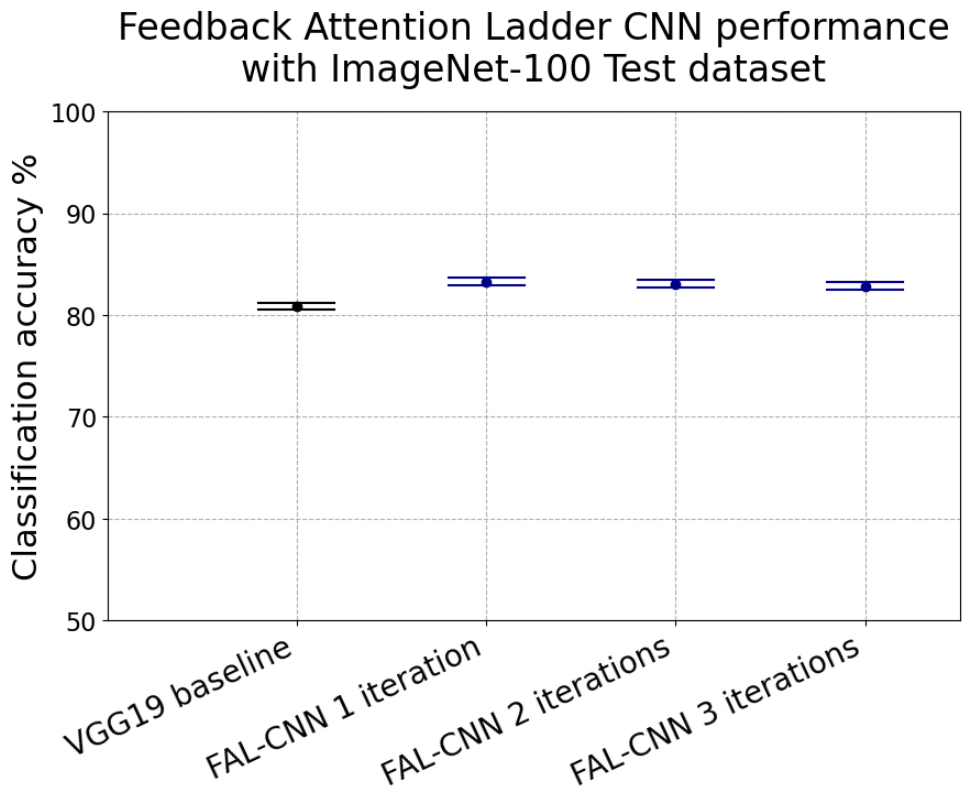


Figure 61: FAL-CNN classification accuracies relative to VGG19 with 95% confidence intervals, evaluated with ImageNet-100 Test dataset

Use of the hold-out Test set mitigates the overfitting associated with Figure 60 and demonstrates a significant increase in accuracy resulting from the use of feedback attention.

5.4.4 Discussion

With ImageNet-100, the feedback attention model yielded a significant improvement in classification accuracy over the feedforward only VGG19. When measuring against the Validation split within each 5-fold CV group selected from the ImageNet-100 Training set, classification accuracy rose from approximately 84% to 88%, a 4pp improvement.

The majority of the benefit occurred with a single feedback iteration (Figure 60). There was a marginal further increase in accuracy when using two iterations, albeit with the computational cost of an extra inference cycle invoking the feedforward and feedback pathways.

When trained models were tested with the unseen hold-out Test dataset, the 1-iteration FAL-CNN showed a significant accuracy gain of 2.39pp over the VGG19, supported by non-overlapping 95% confidence intervals and a p-value of $p<0.001$. However, across all models, accuracies were 3-4pp lower than those seen with Validation sets during training (Figure 61).

This discrepancy appears to be due to overfitting during training. Feedforward module weights were initialised from a downloaded VGG19, pre-trained with ImageNet-1k. As ImageNet-100 contains a subset of these images, the final trained model is likely to have been contaminated by exposure to images subsequently used for post-training accuracy measurements. The higher scores in Table 28 and Figure 60 are therefore likely to involve overfitting, and the results in Table 50 and Figure 105 give a truer representation of the models' performance.

Subsequent evaluation of FAL-CNN in feedback visualisation and saccade-based processing was therefore performed using images from the ImageNet-100 Test set exclusively.

Despite the above issues, these results illustrate the benefit of using feedback attention mechanisms with ImageNet-100 data. The FAL-CNN model architecture has proved to be generalisable across image sources beyond the original digital pathology patches.

Further experiments (Chapter 6) were developed, to visualise and understand the feedback processes involved.

6 Visualising Feedback Attention

6.1 Motivation

Feedback attention has been shown to improve model accuracy when classifying a range of pathology and general purpose image sets. This chapter examines the extent to which feedback activations involved in model evaluation are spatially aligned with objects or regions of interest in the input image. These were expected to be important in understanding the behaviour of the model, en route to a qualitative and, where possible, quantitative understanding of salient image regions.

From an Explainable AI (XAI) perspective, there is a clear value in tools that that can highlight diagnostically relevant tissue in support of an AI model's prediction.

ImageNet and QUASAR-trained FAL-CNN models were analysed. It was expected that attention distributions for ImageNet could be easily compared with images of everyday objects, since object boundaries are apparent without specialist knowledge.

For each data source, feedback activations were captured and plotted at each feedback layer, to explore the model's attention to features at different depths in the model. The large number of channels at the feedback layers presented a further challenge in interpreting multiple feedback distributions. This section therefore explores several methods for combining, visualising and assessing these results.

6.2 Methodology

Code used in this section is documented in Appendix Section 1.4.

6.2.1 Model Enhancements

Visualisation experiments were carried out using the FAL-CNN classifier model, configured for feedback to layers 0,5,10,19,28. The model was modified to return a list of feedback activations for each iteration of feedback (Figure 62) during model execution. Each nested list comprised $B \times C_l \times H_l \times W_l$ tensors for each feedback layer, where B is the batch dimension and C_l, H_l and W_l are the channel, height and width dimensions at layer l in the model. The entire collection was returned as an extra output tuple value alongside the model's inference output tensor.

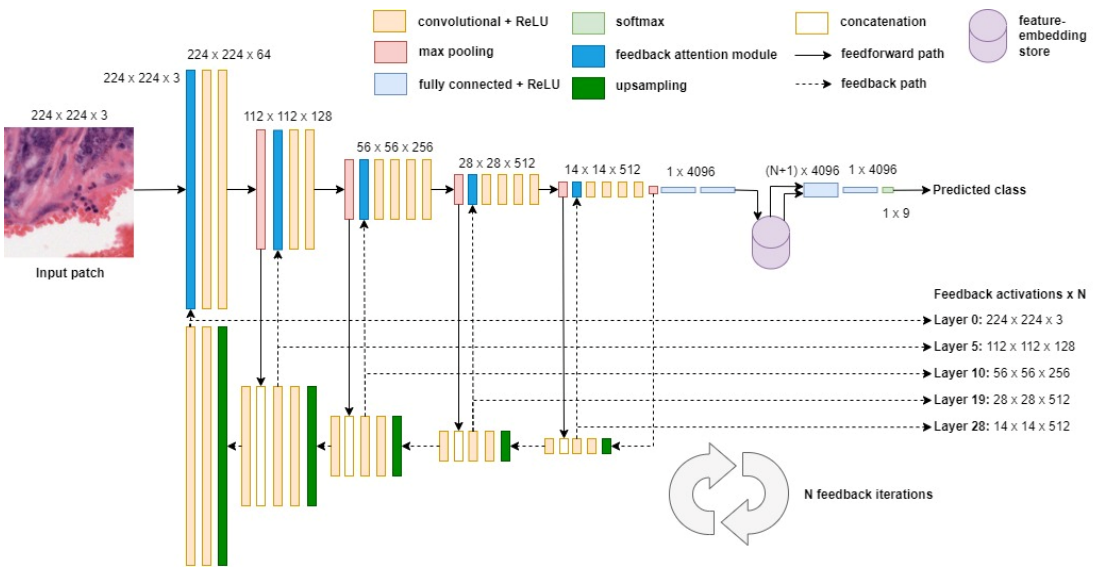


Figure 62: Feedback Attention Ladder CNN with additional outputs of feedback activations

6.2.2 Feedback Activation Visualisation Plots

A Visualisation Image Generator (VIG) was developed to generate output images from the grouped feedback activations returned from the enhanced models, and to superimpose these onto the input image where required.

Images were chosen from a validation set that was unseen during model training, to avoid generating overfitted attention maps.

Strongest N activations

In each feedback layer, the feedback activations for all channels were sorted by descending median value. The first 8 of these were processed and saved to disk. Each activation matrix was resized to the 224×224 input size, using the SciKit-Image *resize* transform (SciKit-Image, 2023) with interpolation to create a smooth heatmap. Each heatmap was saved as an RGB image file with the blue-to-yellow *viridis* colour mapping (Matplotlib, 2012), which is perceptually uniform and offers good visual contrast.

The resulting heatmaps were displayed using an interactive viewer to load and display the plots for a selected attention model and input patch. Images were arranged in a grid with feedback layers as the horizontal axis. Vertical columns represented each of the strongest channel activations, ordered by decreasing median value.

The patch corresponding to QUASAR WSI 116206, box 29 in the GT annotations, was chosen as representative of tumour features, for which the FAL-CNN correctly predicted the *tumour* class.

A further output panel displayed feedback activations for the strongest channel activation in each feedback layer. Two distinct methods were used to create these plots:

- Alpha channel

  Each resized attention heatmap was used as the 'A' (Alpha, or opacity) channel in an RGBA image based on the input patch. This image was then combined with a white background, to show regions of high attention in full contrast, while regions of low attention were made paler to mirror the model's disinterest in these areas.

- Contour plot

  Each attention heatmap was also expressed as a contour plot over the input patch image. For each of a series of thresholds (0.2, 0.4, 0.6, 0.8), a binary mask was created with a level of 1 where activations exceeded the threshold and 0 elsewhere. The boundaries between these regions were extracted as contour polygons, which were superimposed onto the input image.

  For later statistical analysis, centroid coordinates and area values were derived from the 80% contour polygons and saved to CSV against each input patch filename. These values were scaled to align with the $224 \times 224px$ input patch size.

Channel mean contour plots

The feedback activations for each layer were combined by taking the mean across all channels, resulting in a single $H \times W$ matrix per feedback layer and iteration, where H and W are the image height and width at that layer.

The mean feedback layer activations were rendered as a $224 \times 224px$ contour plot, which was written to disk for each input patch, feedback iteration and layer. One patch of each class, with structures and cellular textures characteristic of that class, was selected for plotting.

The following pseudocode summarises the processing steps to generate the mean activations and resulting plot images and statistics:

Algorithm for plotting mean feedback activations from Feedback Attention model

```
Load pre-trained Feedback Attention model from file
Load pre-trained VGG19 model from file

For each patch image in input directory:
      Load image file
      Apply to VGG19 model, for FF-only class prediction
      Apply to Feedback Attention model, for class and FB activations

      Derive output filename from FF and FB predictions

      If bounding boxes available:
            Load bounding box from XML
            Combine with input image
            Save to disk

      For each feedback iteration performed by model:

            For each feedback layer in 0,5,10,19,28:
                  Select H x W x C feedback tensor for layer

                  Sort tensor by total level per channel (descending)
                  For each channel in top N:
                        Select H x W tensor
                        Normalise and resize to match input patch
                        Save to disk

                  Take normalised mean over all channel activations
                  Resize mean tensor to match input patch dimensions

                  Convert mean activations to contours/heatmap
                  Combine with input image
                  Save to disk

                  If stats requested:
                        Get centroid and area of 80% contour
                        Write to CSV row against patch name and layer

                        If bounding boxes loaded:
                              Get bounding box centroid and area
                              Write to current CSV row
```

Grouped spatial plots

Feedback activations were combined by taking the mean of activation matrices acquired from 900 images, representing 100 patches of each QUASAR class. These were grouped into grids of patch-sized images using the following combinations of parameters:

- Layer and class:
  A grid of images was created, with each column corresponding to a different feedback layer in the model (layers 0,5,10,19,28) and each row corresponding to a different QUASAR image class (0-non-informative through 8-muscle). This would allow feedback behaviour to be differentiated between tissue types at different levels in the model. This was performed for the 1-iteration FAL-CNN.

- Layer and feedback iteration:
  In this visualisation, feedback activations from a 3-iteration FAL-CNN were combined from patches of all image classes, then presented in a grid with one column per feedback layer and one row per feedback iteration.

Offset patches

Each grouped plot was generated for two further scenarios, mirroring the experimental configurations involving offset patches in Section 5.2:

- Model trained on centre-annotated patches and evaluated with *offset-patches* dataset
- Model trained and evaluated with *offset-patches* dataset

6.2.3 Pathologist Review

A directory of patch images with overlaid attention contours was generated using the VIG, configured to evaluate a QUASAR-trained 2-iteration FAL-CNN model. Separate images were generated for each feedback iteration, with contours representing the mean feedback activations in that iteration.

The contour-enhanced images were reviewed by a consultant pathologist. Examples of each QUASAR tissue class were selected at random. The pathologist examined each image and recorded qualitative observations on the tissue structures and cell types, in regions highlighted by the attention contours.

Contours generated for layer 28 were preferentially examined, as these were found to enclose larger regions of cells, allowing structural context as well as cell types to be assessed.

6.2.4 Statistical Analysis

Scatter plots and histograms were generated from the following contour measurements per patch, as previously written to CSV file (Section 6.2.2):

- Effective area, the total activation value in the $H \times W$ attention matrix, equivalent to the mean pixel value multiplied the total number of pixels
- Centre of mass of $H \times W$ attention matrix
- Centroid coordinates of 80% attention contour
- Area of 80% attention contour

Plots were grouped in grids by layer and class, and layer and feedback iteration, similar to the grouped spatial plots in Section 6.2.2. The results of these measurements are listed in Appendix Section 2.

6.2.5 Visualisation with ImageNet-100

For evaluation with ImageNet-100, one image was randomly selected from each image class subdirectory. Images were taken from the previously unseen ImageNet-100 Test set, to eliminate the risk of overfitted output distributions being generated as a result of encountering images previously used in training.

1-iteration and 3-iteration FAL-CNN versions were executed with each of the 100 test images. Contours were generated using the mean attention activations over all channels, for each feedback layer in the model, then superimposed on the corresponding input image and saved to disk.

Bounding boxes

Bounding box (BB) annotations for objects in the ImageNet-100 dataset were downloaded as XML data from ImageNet (Stanford University, 2020). These were loaded and combined with the corresponding input images, with the results saved to disk for comparison with the contour images.

BB coordinates were saved to a CSV file, per the algorithm in Section 6.2.2. BBs were subsequently compared with the 80% contours of mean feedback activations at each feedback layer in the model, using F1 (Dice) score (Section 4.2.2, equation 2). Distances between BB and contour centroids were also recorded for each sample image.

VIA outlines

F1 scores were similarly derived to compare the 80% attention contours with object boundaries.

Object outlines were manually drawn for the 100 ImageNet sample images, using the online VGG Image Annotator (VIA) tool (University of Oxford, 2023). Annotations were downloaded as a CSV file. F1 (Dice) scores were generated for VIA-sourced annotation polygons, in comparison with the 80% attention contours derived from the FAL-CNN feedback activation outputs. Distances between centroids of the VIA outline and the 80% contour were recorded for each sample image.

6.3 Results

6.3.1 Feedback Attention Visualisation Plots

Strongest N activations

Figure 63 shows the heatmap grid generated for the patch representing WSI 116206, box 29, labelled *tumour*. Columns from left to right represent the 8 strongest channel activations at each layer of the 1-iteration FAL-CNN model, in descending order of median value with the highest activation values shown in yellow.

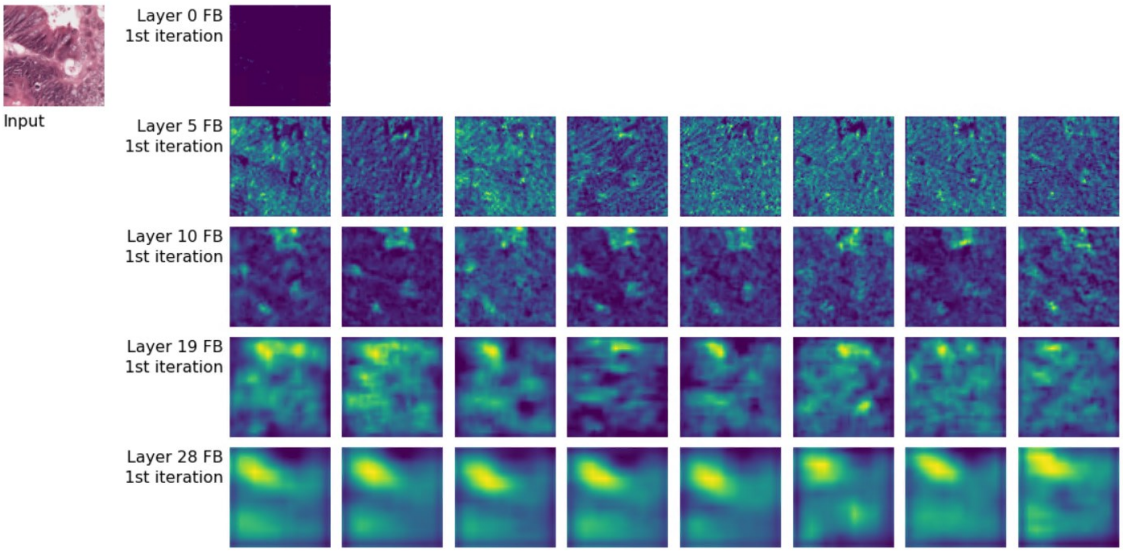


Figure 63: Strongest 8 feedback activations per layer, in FAL-CNN model when processing sample tumour patch.

Figure 64 and Figure 65 show alpha and contour plots, where the activations from the strongest channel, corresponding to the LH column in Figure 63, are superimposed on the input patch image.

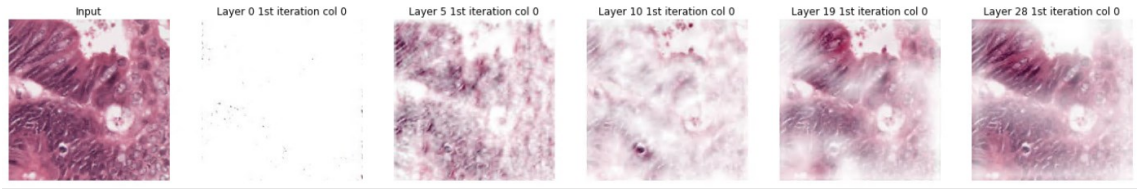


Figure 64: Alpha-channel plot of feedback attention distribution, using strongest feedback activations in layers 0..28

Least-attended tissue regions are 'whited out'.

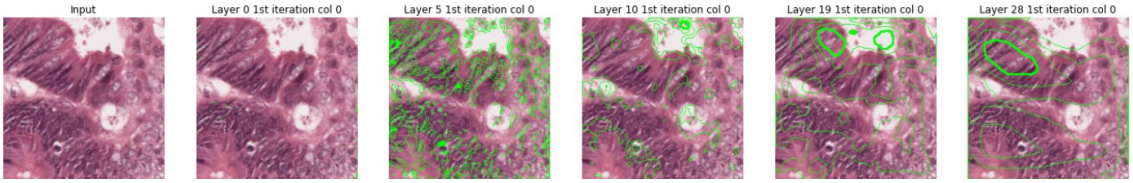


Figure 65: Contour plot of feedback attention distribution, using strongest feedback activations in layers 0..28

All parts of patch remain clearly visible beneath the contour representation of the spatial attention distribution.

Channel mean contour plots

In Figure 67, contours representing the mean activations in each feedback layer of the 1-iteration FAL-CNN model have been superimposed on a sample patch image from each QUASAR tissue class.

Figure 66 shows the effect of multiple feedback iterations on the attention distribution at each feedback layer, for the 3-iteration FAL-CNN model applied to an example *tumour* patch.

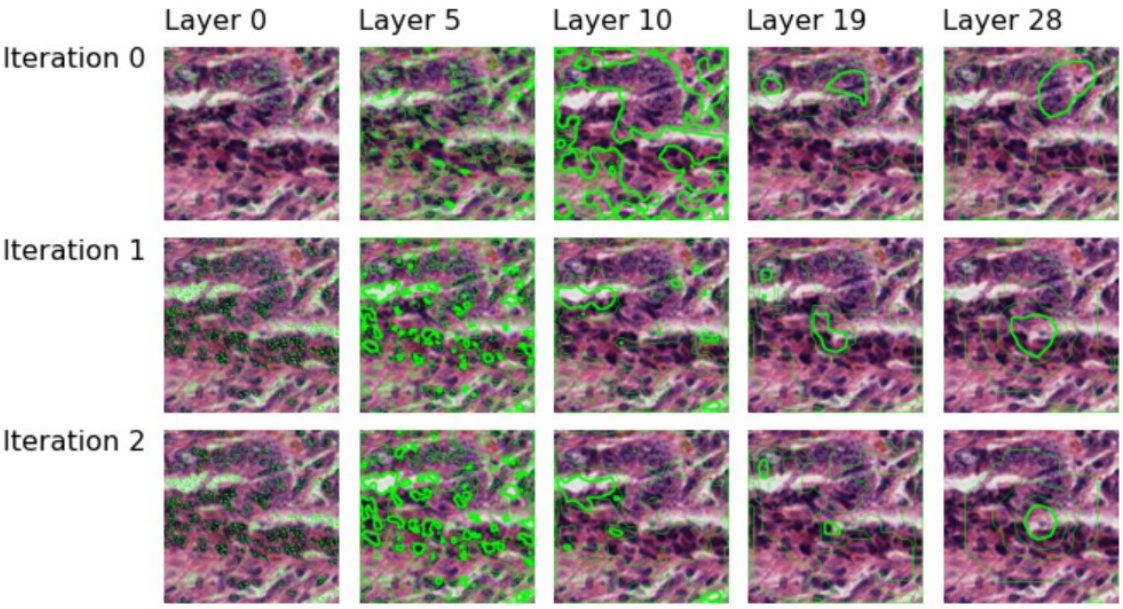


Figure 66: Feedback attention contours by layer and iteration for tumour patch from WSI 52918 box 23

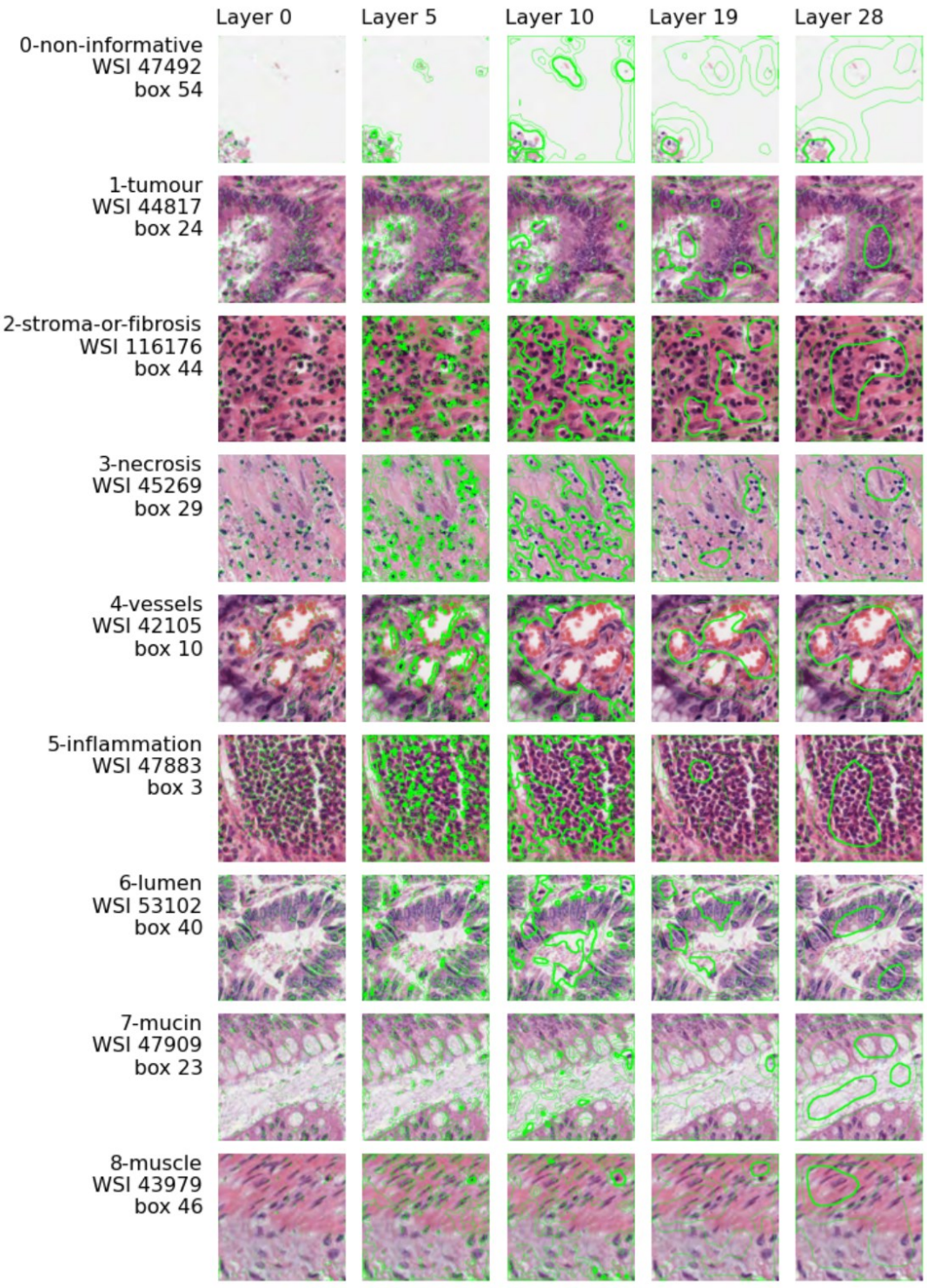


Figure 67: Feedback attention contours by layer, for one patch of each tissue class

Lower layers attend to more granular features, while contours for higher layers highlight larger cellular structures relevant to the predicted classification.

Grouped spatial plots

Figure 68 shows the mean spatial distributions of feedback attention activations grouped by feedback layer and combined across all 9 image classes, for each of 3 feedback iterations.

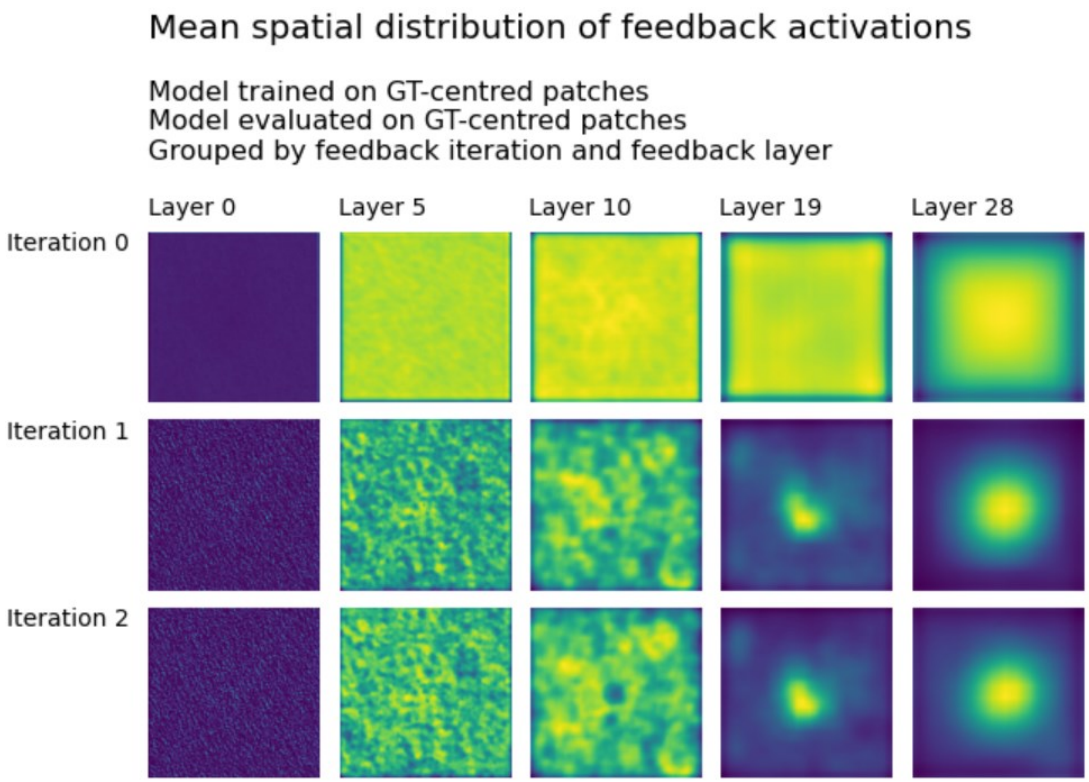


Figure 68: Mean spatial feedback activations over multiple patches, grouped by layer and feedback iteration

Layer 28 attention highlights patch centre, mimicking behaviour of annotating pathologist initially examining cells close to GT location.

Figure 69 shows the spatial distributions of feedback activations in FAL-CNN, averaged from inferences performed on multiple patches, and grouped by layer and image class.

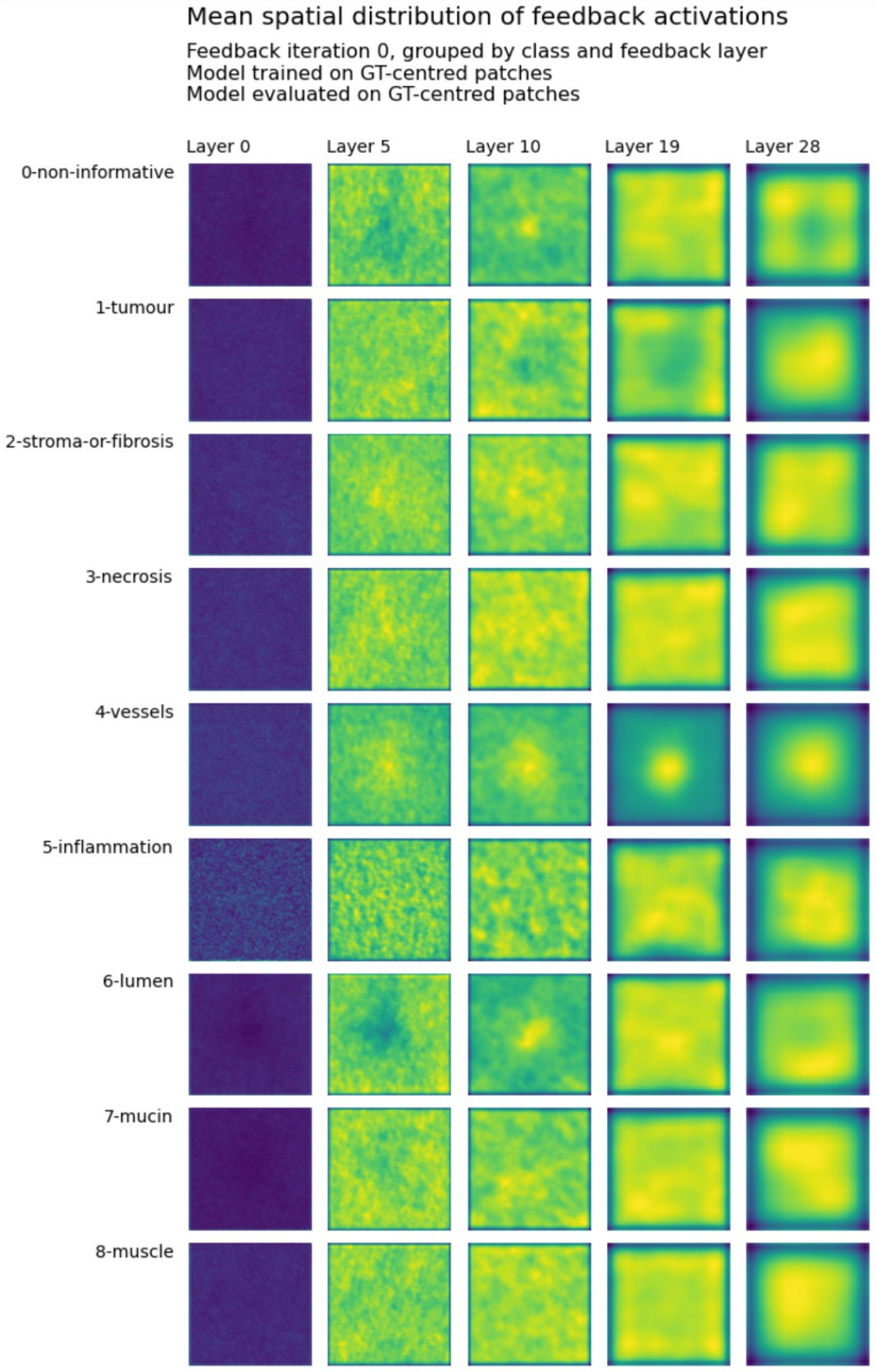


Figure 69: Mean spatial feedback activations over multiple patches, grouped by layer and class

Offset patches

In Figure 70, the previously used model has been executed with patches offset up and left by 56px. Figure 71 shows corresponding results obtained when the offset input data was applied to a FAL-CNN model that was trained with the 56px-offset dataset.

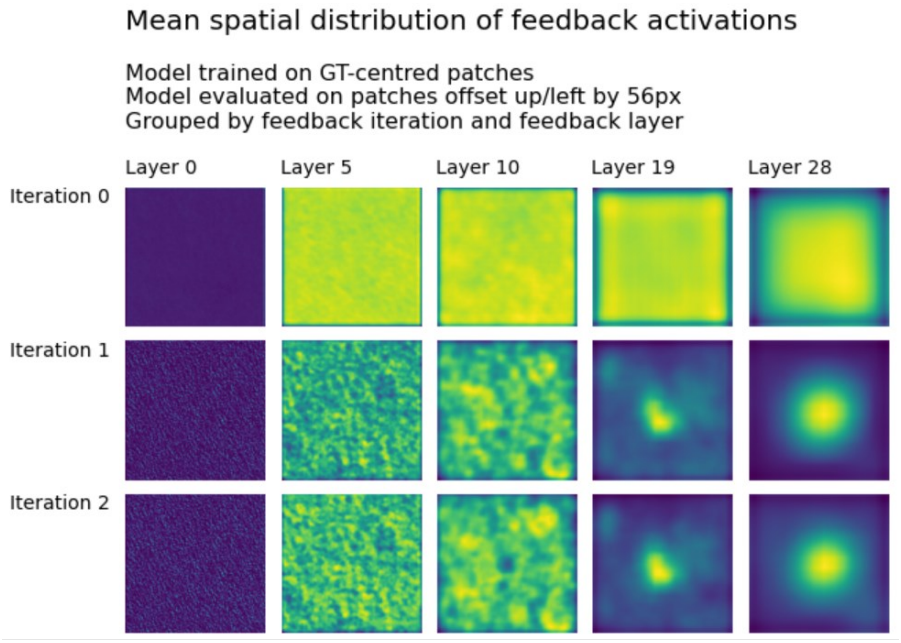


Figure 70: Mean spatial feedback activations over multiple offset patches, grouped by layer and feedback iteration

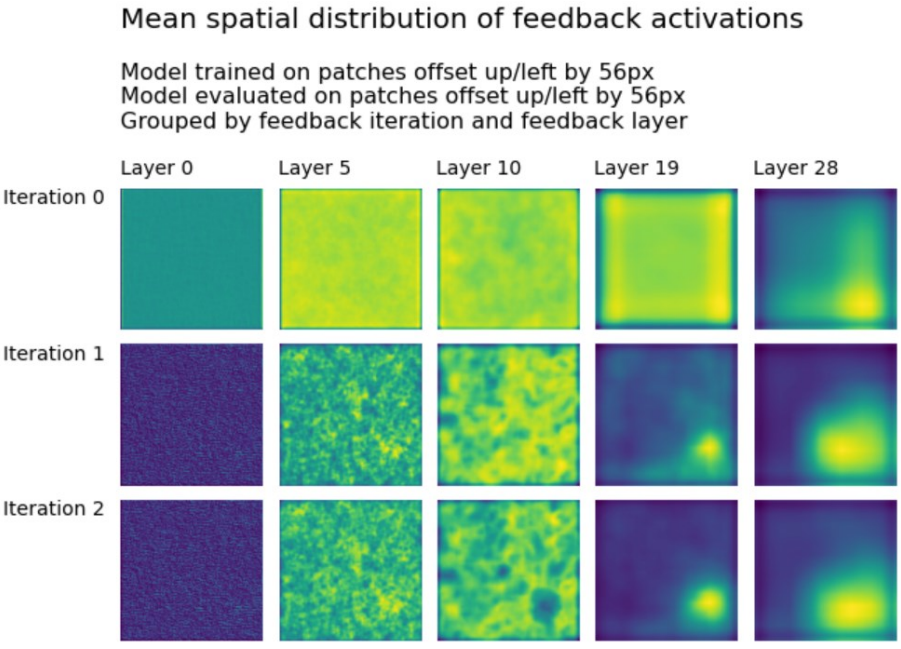


Figure 71: Mean spatial feedback activations in offset-trained model, over multiple offset patches grouped by layer and feedback iteration

Model has learned to attend to region around new GT location in bottom-right quadrant.

For Figure 72, the non-offset model was executed with the 56px offset patch set, while Figure 73 shows the distributions resulting when a model trained with offset patches was executed on similarly offset patches.

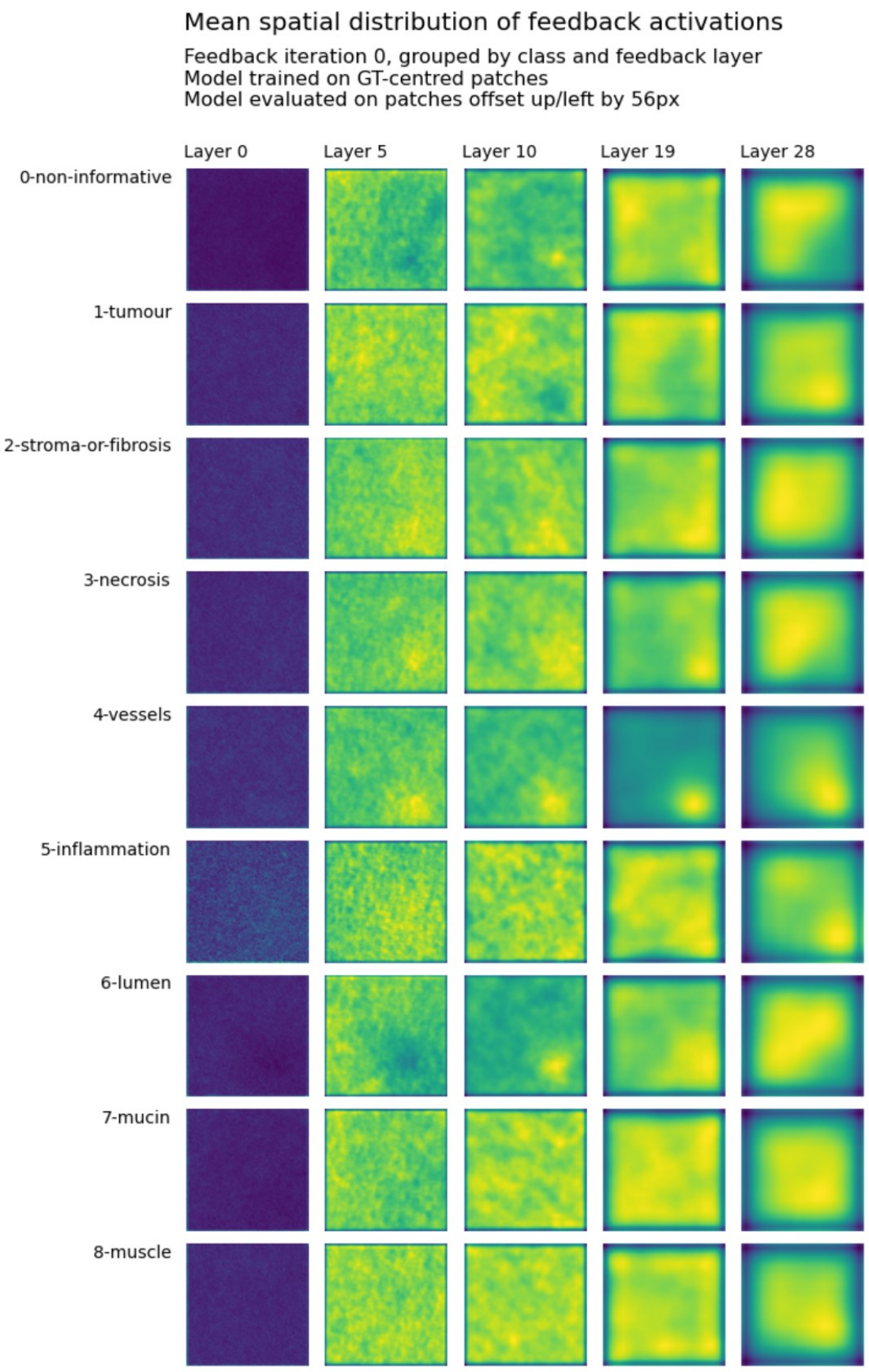


Figure 72: Mean spatial feedback activations over multiple offset patches, grouped by layer and class

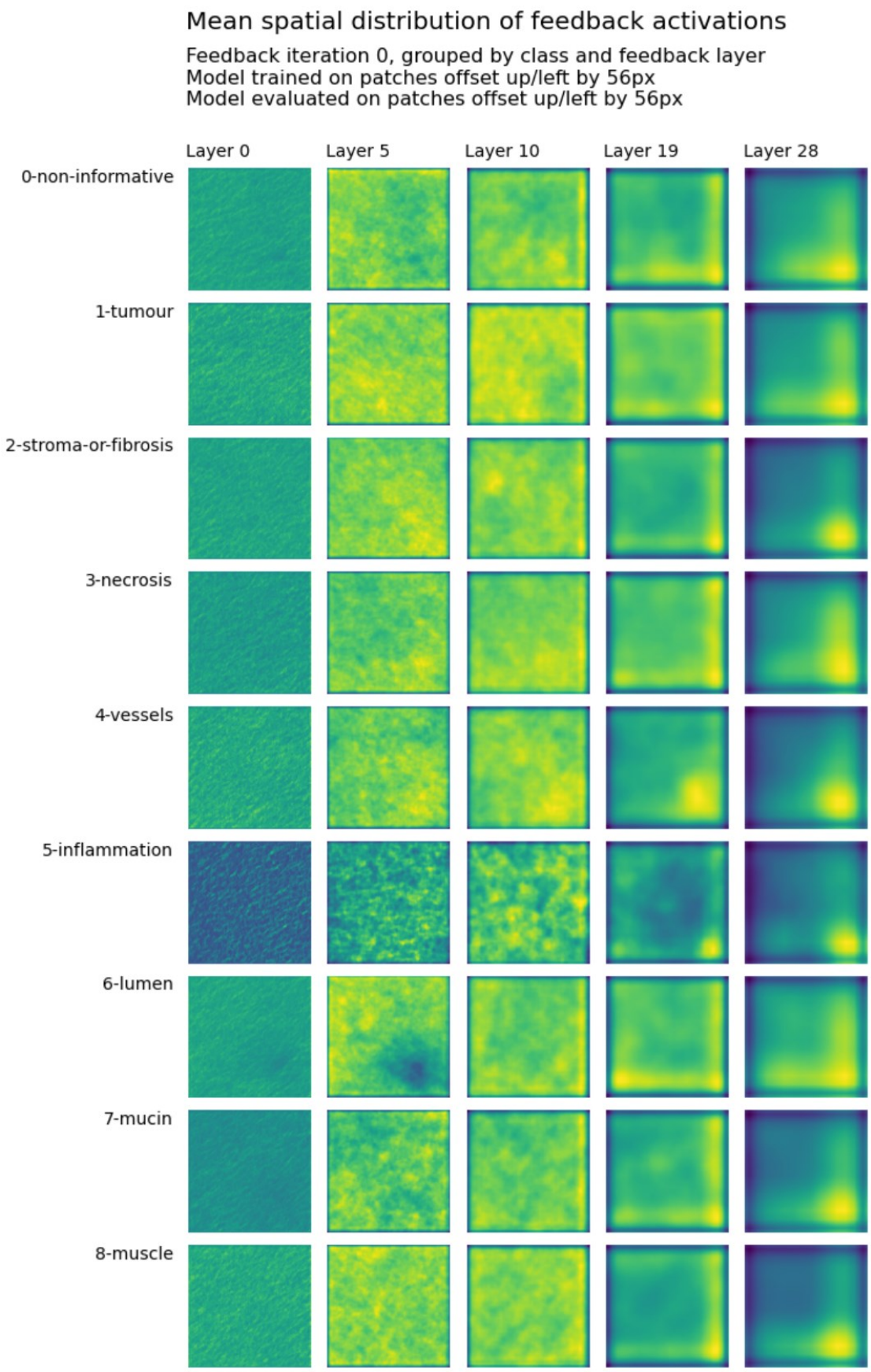


Figure 73: Mean spatial feedback activations for offset-trained feedback model, with multiple offset patches, grouped by layer and class

6.3.2 Pathologist Review

Figure 74 to Figure 88 show the attention contour images of patch images selected for review, with the pathologist's observations quoted. These comments refer to tissue enclosed by the 80% attention contours (thicker green line).

Results are grouped by the original expert-annotated class of the input patch. Most are based on the two-iteration FAL-CNN, where the left-hand image includes attention contours for feedback iteration 1, while the right-hand image represents iteration 2. The exception is WSI 116769 box 22 (Figure 77), which represents feedback iterations in a 3-iteration model.

The WSI number listed below corresponds to the filename of the WSI from which the patch was extracted; the box number is the index of the ground truth annotation that defined the centre of the extracted patch.

Tumour

WSI 42123 box 21: *"Correctly finding cancer gland nuclei; lumen also highlighted"*

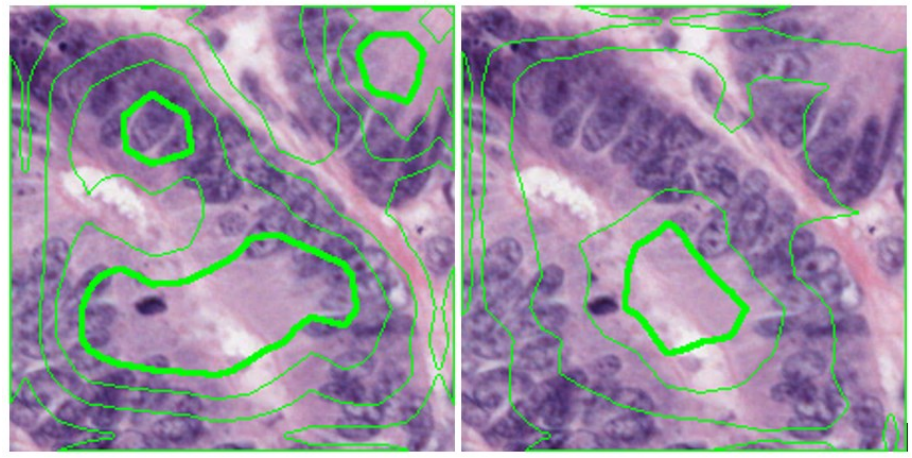


Figure 74: WSI 42123 box 21, attention contours for feedback iteration 1 (left) and 2 (right)

WSI 44888 box 29: "*Displaced nuclei bottom right + necrosis"*

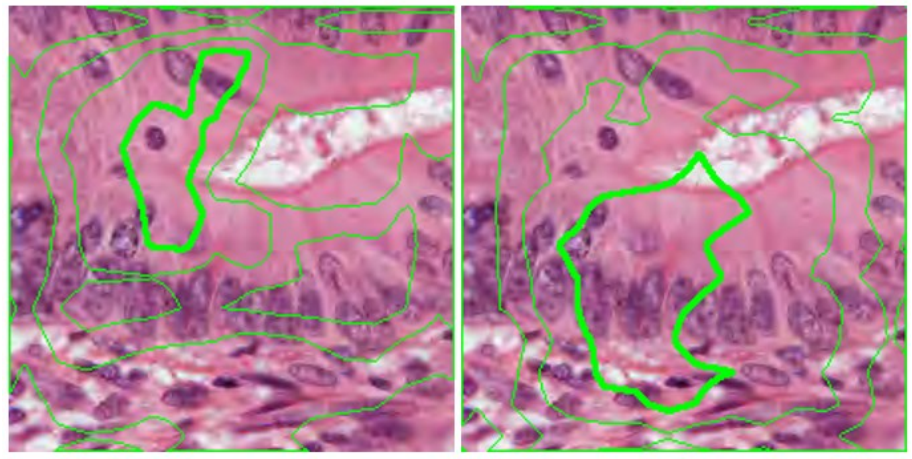


Figure 75: WSI 44888 box 29, attention contours for feedback iteration 1 (left) and 2 (right)

WSI 53434 box 49: "*Good, relevant cancer cells*"

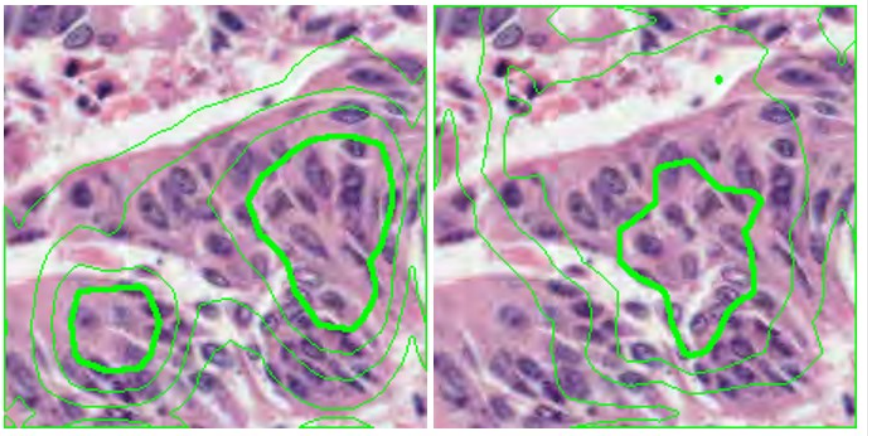


Figure 76: WSI 53434 box 49, attention contours for feedback iteration 1 (left) and 2 (right)

WSI 116769 box 22: " *'Saccades' over multiple feedback cycles in three-iteration model appear to be detecting further relevant parts of patch. Tissue around the central pixel, the location to which the classification should apply, is thought to be misleading in this case.*"

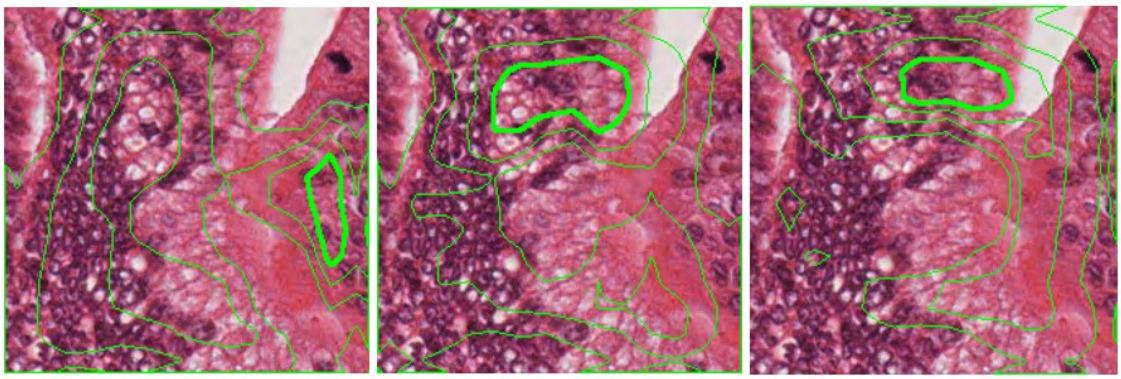


Figure 77: WSI 116769 box 22, attention contours for feedback iterations 1-3 (left to right)

WSI 116840 box 45: "*Centre pixel is borderline (tumour/stroma) … Picking up some cancer but also an area of stroma.*"

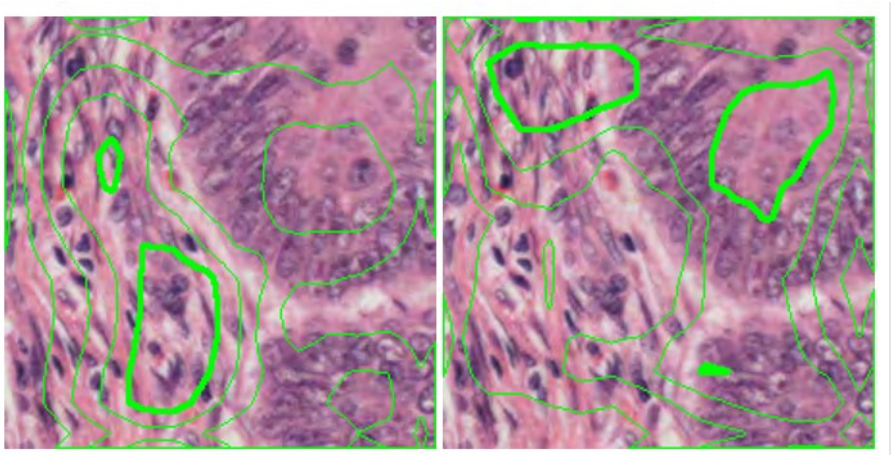


Figure 78: WSI 116840 box 45, attention contours for feedback iterations 1 (left) and 2 (right)

Stroma

WSI 42189 box 6: "E*xample of inflamed stroma detected correctly.*"

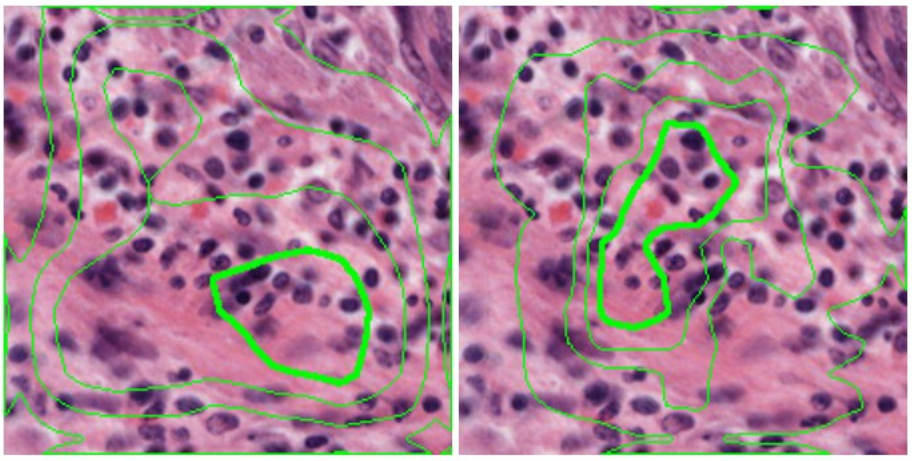


Figure 79: WSI 42189 box 6, attention contours for feedback iterations 1 (left) and (right)

WSI 45269 box 31: "*Impressive results, ignoring cancer to decide patch is stroma.*"

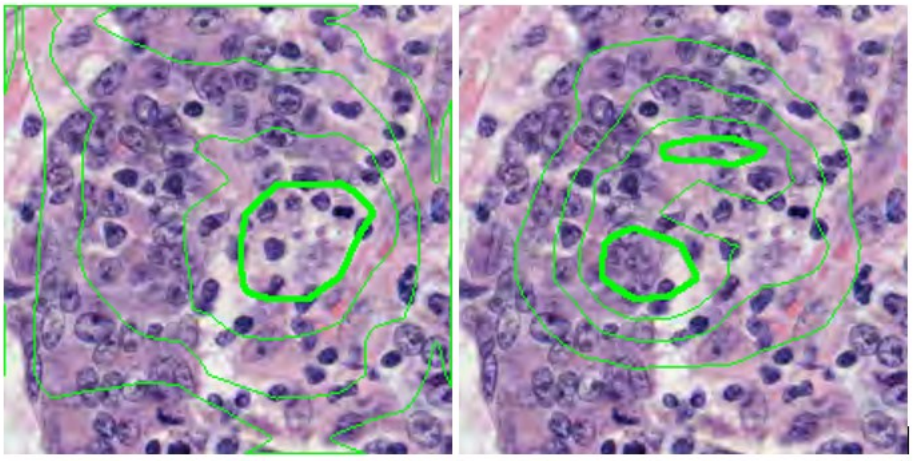


Figure 80: WSI 45269 box 31, attention contours for feedback iterations 1 (left) and 2 (right)

WSI 61084 box 49: "*Good example, stroma highlighted, ignores cancer top left and focuses on stroma in rest of box.*"

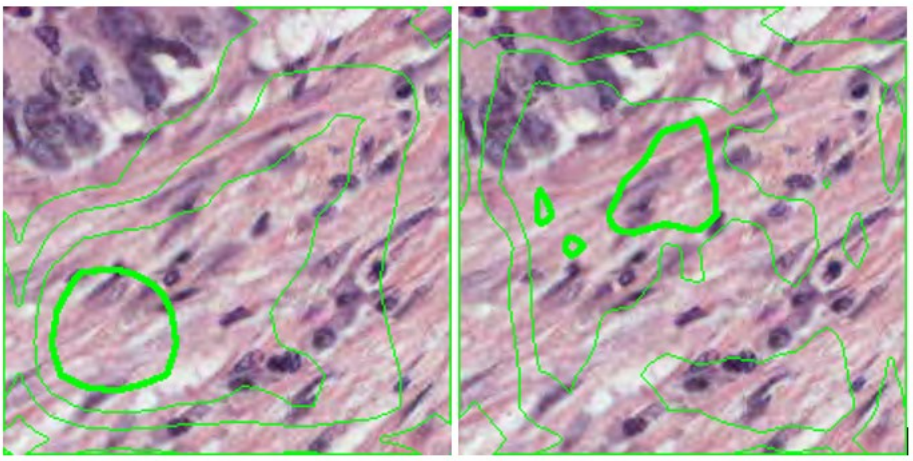


Figure 81: WSI 61084 box 49, attention contours for feedback iterations 1 (left) and 2 (right)

WSI 57404 box 38: "*Ignores cancer, finds stroma as required.*"

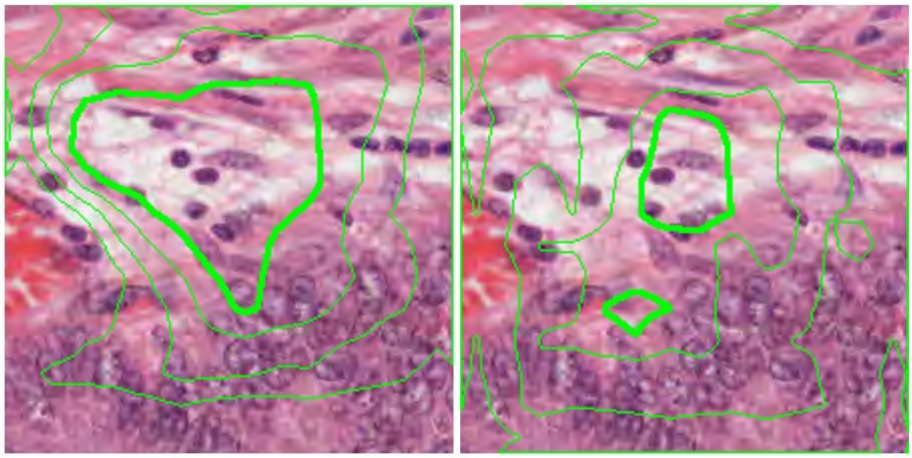


Figure 82: WSI 57404 box 38, attention contours for feedback iterations 1 (left) and 2 (right)

WSI 116873 box 41*: "Some cancer attention, some tumour, despite stroma class."*

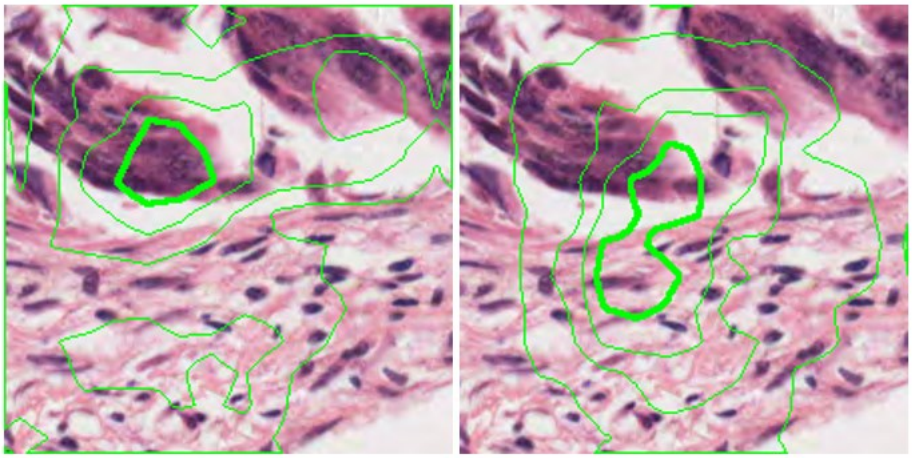


Figure 83: WSI 116873 box 41, attention contours for feedback iterations 1 (left) and 2 (right)

Necrosis

WSI 61413 box 6: "*Targets necrosis in centre of gland.*"

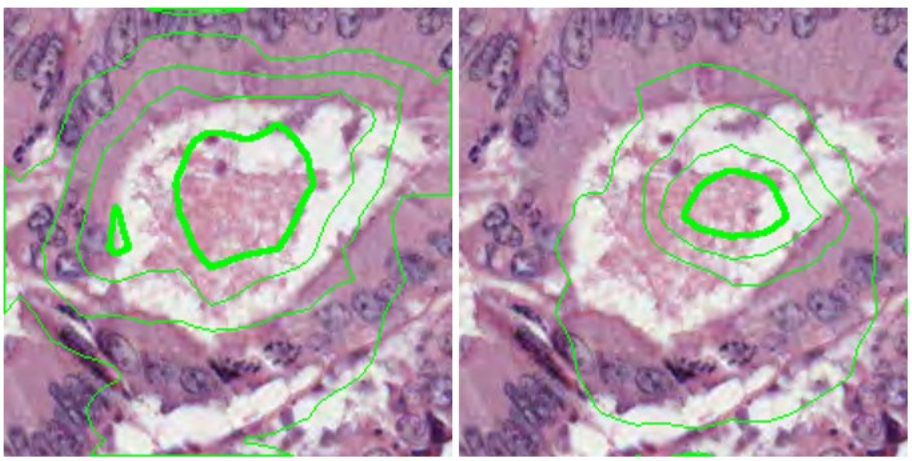


Figure 84: WSI 61413 box 6, attention contours for feedback iterations 1 (left) and 2 (right)

Lumen

WSI 52910 box 15, WSI 53446 box 29: "*Both show lumen prediction based on finding cancer (but 2nd feedback iteration more centred on actual lumen gap).*"

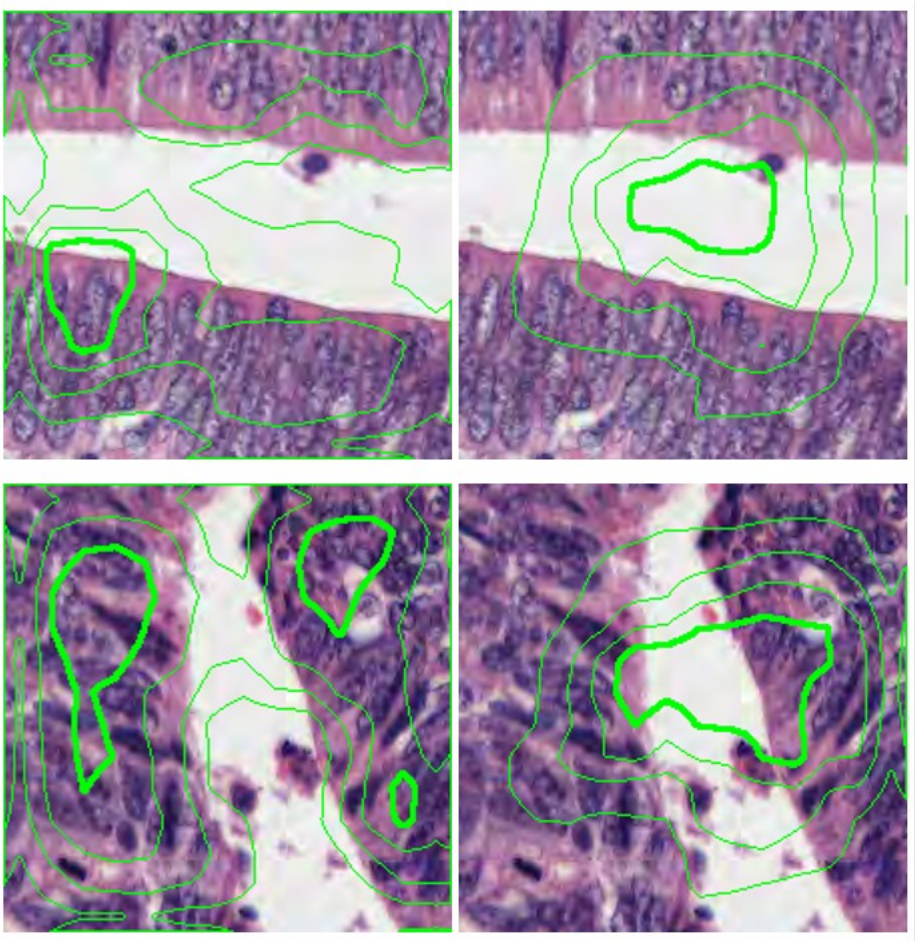


Figure 85: WSI 52910 box 15 (top) and WS 53466 box 29 (bottom), attention contours for feedback iterations 1 (left) and 2 (right)

Vessels

WSI 46684 box 1 (layer 19 feedback)**: "*Vessels – generally responding to red cells! Very impressive detection of tiny vessels.*"**

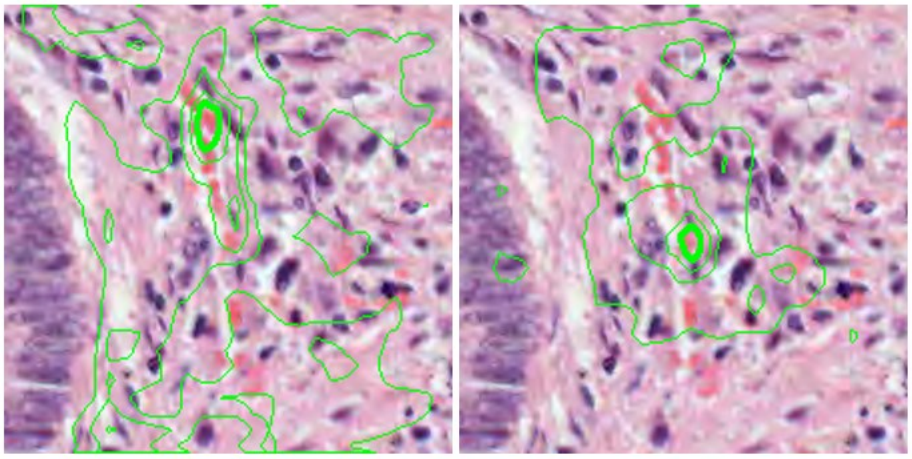


Figure 86: WSI 46684 box 1, attention contours for feedback iterations 1 (left) and 2(right)

Non-informative

WSI 63334 box 39: "*Many types present, results split accordingly? VGG baseline predicts stroma then muscle; feedback corrects this to non-informative.*"

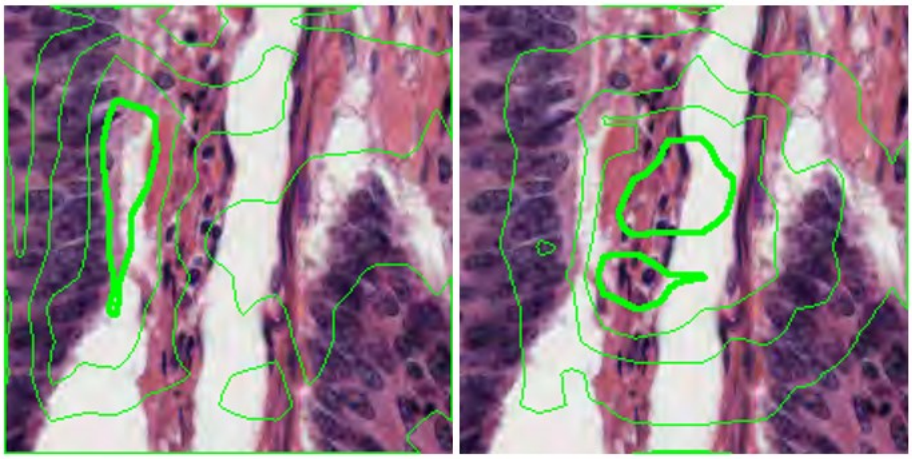


Figure 87: WSI 63334 box 39, attention contours for feedback iterations 1 (left) and 2 (right)

"Some patches are rotated copies of others. Attention region broadly follows rotation."

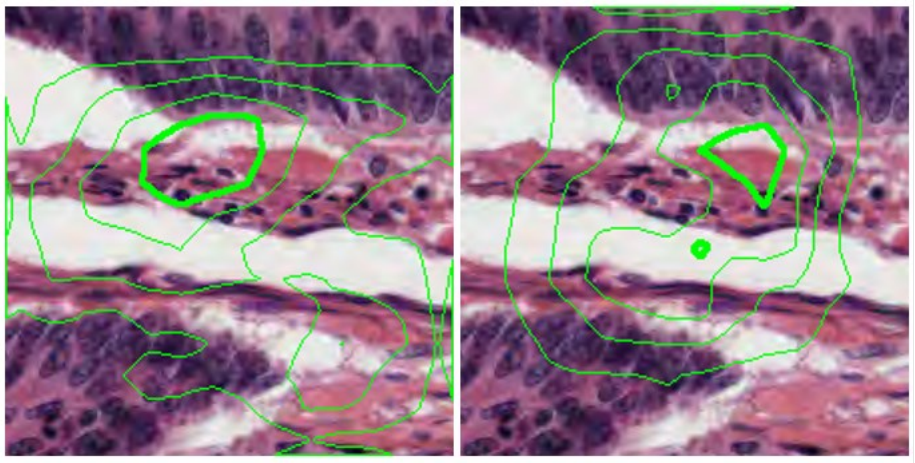


Figure 88: WSI 63334 box 39, rotated 90 degrees clockwise. Attention contours for feedback iterations 1 (left) and 2 (right)

6.3.3 Visualisation with ImageNet-100

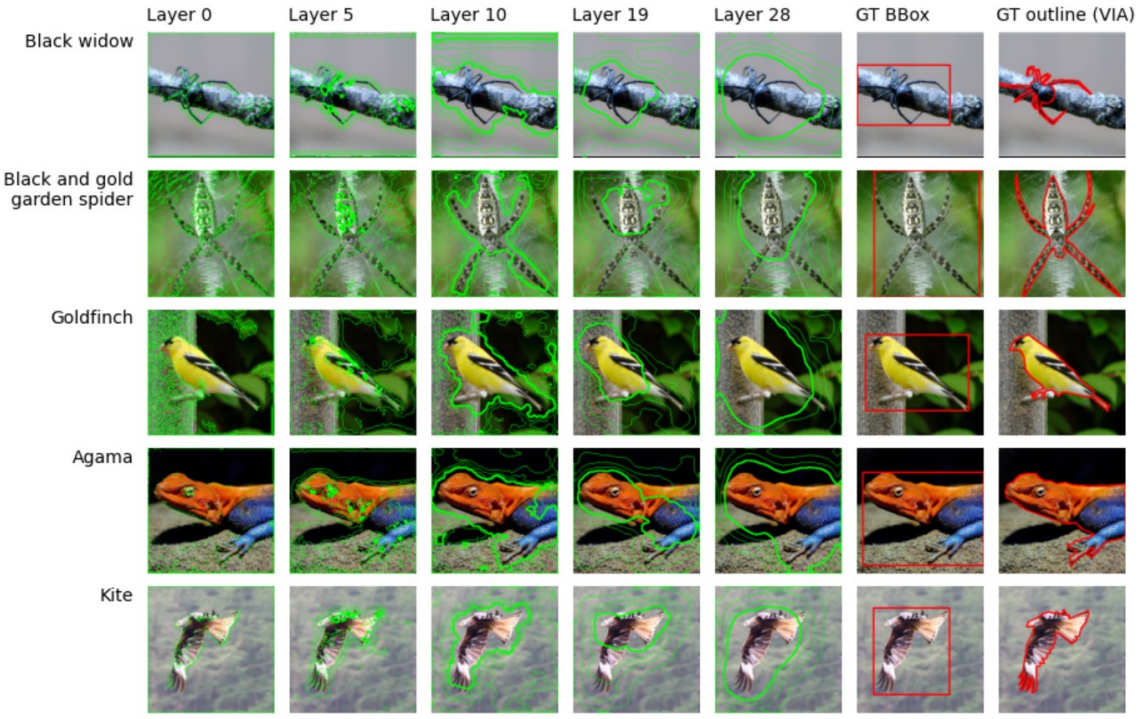


Figure 89: Feedback attention contours and ground-truth annotations for ImageNet-100 sample images, arranged by class and layer

The two rightmost "GT" columns show human-generated bounding boxes and object outlines for comparison with attention regions.

Figure 89 shows attention contours at each feedback layer, generated for selected images from ImageNet-100 with a single-iteration FAL-CNN model. In Figure 90 the agreement between 80% feedback contour for each layer, and the GT bounding boxes, is expressed as an F1 (Dice) score.

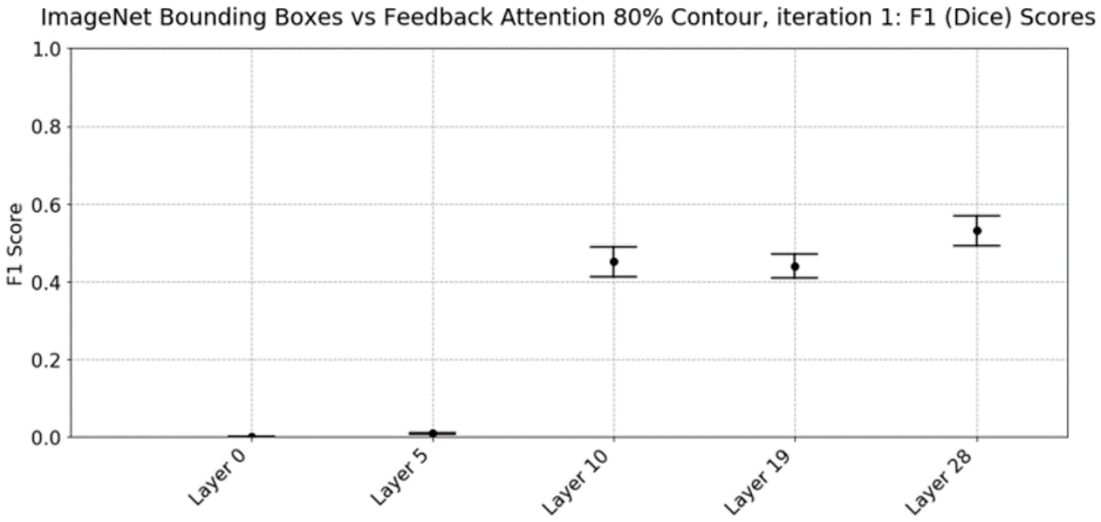


Figure 90: Agreement as F1 score, between layer 28 80% attention contours and ground-truth bounding boxes for ImageNet-100 test images

Higher F1 scores in layers 10, 19 and 28 show tendency for agreement between attention regions in higher layers, and human-annotated object boundaries.

Figure 91 shows the corresponding F1 scores for agreement between the 80% feedback contours and object boundaries manually annotated using VIA.

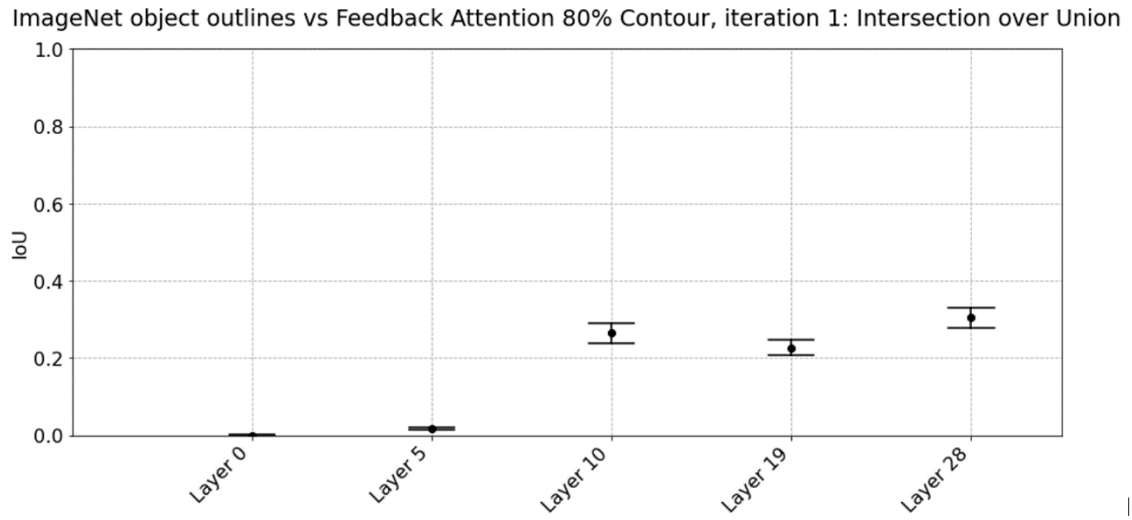


Figure 91: Agreement as F1 score, between layer 28 80% attention contours and GT object outlines for ImageNet-100 test images

For bounding boxes and VIA outlines, distances between the centroids of the GT annotations and the 80% attention contours were analysed. Spatial and frequency distributions of these values are shown in Appendix 3.2. Table 19 shows the resulting mean distance for each annotation method.

Table 19: Mean distances between annotation and 80% attention contour centroids with 100 ImageNet-100 sample images, for Bounding Box and VIA Outline

Manual annotation method	Mean distance between annotation centre and 80% attention contour centroid, pixels	95% Confidence Interval, pixels
Bounding Box	18.13	14.83 to 21.43
VIA Outline	19.43	15.93 to 22.93

6.4 Discussion

6.4.1 Feedback Attention Visualisation Plots

Strongest N activations

In the initial spatial 'heatmap' plot (Figure 63), regions of high feedback activation were larger and less granular in higher layers (19, 28), due to the lower spatial resolution (e.g. 14×14px at layer 28) compared to activations at lower layers. At lower layers, the feedback activations aligned with smaller textural divisions and cellular-level structures.

Feedback activations at layer 0, at the model's input, revealed no visible content in this plot, despite this pathway's contribution to model accuracy. The RGBA plot in Figure 64 revealed sporadic pixel-sized activations. Due to their apparently beneficial role in layer 0 feedback, these are assumed to represent salient features at this low perceptual level.

For a given feedback layer, plots of the strongest 8 feedback activations showed little variation between channels. Subsequent visualisation methods would therefore combine these activations into a single mean activation plot per feedback layer.

RGBA plots

The RGBA plot, with opacity in the 'A' channel controlled by feedback activation strength, showed low contrast "whited out" regions where attention was weakest. Areas showing the greatest visual contrast represented the most strongly attended regions. This plot type was most effective for displaying the larger attention regions in higher feedback layers. However, the very translucent less-attended regions could be confused with non-informative background in some plots, so alternative visualisation methods were sought.

Contour plots and spatial distributions

When attention maps were shown as contour plots, tissue in the underlying patch remained fully visible. This approach provided the clearest representation of the patch image and the attention distribution simultaneously.

Regions of highest attention, inside the 80% contours (shown as bold green outlines in Figure 65 to Figure 66), were found to align with dense nuclei characteristic of the sample *tumour* patch. At lower feedback layers (leftmost columns in Figure 67), the contours highlighted finer details such as clusters of cell nuclei.

The spatial distribution of feedback activations was found to vary with tissue class (Figure 67). For *non-informative* and *lumen* patches, often characterised by a blank central region, stronger attention was paid to surrounding tissue structures, such as clusters of tumour nuclei, than to the patch centre.

Classes such as *vessels*, characterised by smaller, well-defined structures, showed strong visual correlation between the 80% contour and the boundary of the structure, suggesting that the feedback model performed accurate object localisation in these cases.

More uniformly distributed tissue types, such as *inflammation*, *necrosis* or *stroma*, yielded broader attention regions in higher feedback layers, with the 80% contour aligning with the densest or most heterogeneous regions of cells.

For the 3-iteration FAL-CNN model (Figure 66), the first feedback iteration showed attention in higher layers to informative tissue distal to the centre pixel. With further iterations, the attention region became more focused around the image centre. There was a visible 'saccade' in attention from the first to the second feedback iteration. However, this behaviour was not associated with a significant further increase in classification accuracy when using two or more iterations (Section 5.1.3 Table 21).

The observed trends become more apparent when aggregating the spatial distributions of feedback attention across multiple input patches. In Figure 68, the model's attention at higher layers becomes much more focused around the central pixel, while lower-level features are attended more uniformly across the patch. In Figure 69, different distributions are apparent for different classes. *Non-informative* and *lumen* at layer 28 showed strongest attention in a ring around the centre. This is again consistent with the model seeking structural context around a blank central point. Contrastingly, the model showed much stronger central focus when classifying smaller, self-contained structures such as *vessels.*

Distributions for centre-trained model with offset-patches *dataset*

Spatial distributions were plotted for feedback activations generated when processing patches sampled at an offset from the ground truth label point (Section 3.3.3), to examine whether the attention "hotspots" followed the GT point or retained a central focus. The results (Figure 70) were superficially similar to those for the regular centre-annotated patches (Figure 68), in that

overall focus of attention in higher layers was concentrated around patch centre, especially after multiple feedback iterations.

However, evidence of attention moving to the new GT position was embodied in the visible bulge towards the bottom right of the patch area, in the layer 28 activations for the first feedback iteration (Figure 70, top right image).

Further evidence of this behaviour arose when plotting mean activations grouped by class (Figure 72). The attention focus in layer 28 moved away from the bottom right quadrant for blank centred classes *tumour* and *lumen*, but towards this region for localised objects such as *vessels*, and to a lesser degree, *tumour.*

Thus, the attentional response learned by the FAL-CNN appears to be a combination of two behaviours: a static bias towards the patch centre, where the GT class was defined before the model was trained, and a dynamic process which tracks informative structures when these are moved within the patch boundary.

Distributions for offset-trained model with offset-patches *dataset*

When an offset-trained model was executed upon similarly offset patches, the static and dynamic attentional behaviours converged on the same location, at the bottom right of the patch. Figure 71 shows a dramatic shift of attention to this area, especially in the layer 28 feedback activations.

Figure 73 shows the change in attention for each input class. In the higher layers, the focus was consistently in the bottom right quadrant, in a tighter pattern than seen with non-offset usage of the standard FAL-CNN (Figure 69). Attention heatmaps at layer 28 show long vertical and horizontal “tails”, encompassing contextual features further from the GT pixel than would be possible with models trained and evaluated on 224x224px patches labelled at the centre pixel.

Section 7.2.3 will show that offset training enhances model accuracy by a further 3.08-3.34pp relative to equivalent centre-trained models. This now appears to be associated with the offset model’s combination of tighter attentional focus with contextual information sampled further from the focus in given directions.

6.4.2 Pathologist Review

The expert qualitative analysis of colorectal cancer patch images, with overlaid feedback attention contours, gave further insights into the tissue structures being attended by the model. Encouragingly, the pathologist’s detailed assessment showed agreement with salient tissue regions picked out by the model.

In several patches originally labelled as *tumour*, the 80% attention contours enclosed features such as cancer gland nuclei, densely packed nuclei, tumour lumen (Figure 74), displaced nuclei and necrosis (Figure 75). These are all characteristic of a region of tumour, with necrosed tissue a possible result of radiotherapy.

The 80% contour is an arbitrary boundary in a smoothly varying attention profile, and should not be regarded as the only salient region in the image. Nonetheless, the contour is indicative of attended tissue and its centroid can be regarded as a proxy for the peak attention location. The presence of relevant cells here implies that the feedback model amplifies regions of the patch that contribute the most usefully to its final classification output.

Using a 3-iteration FAL-CNN model, the 80% contours for each iteration follow saccade-like movements, outlining additional salient tissue regions (Figure 77). This has potential value in an XAI application, highlighting cellular features that support the model's class prediction.

The model's attention behaviour varied dramatically according to tissue type. In the spatial distribution plots (Section 6.3.1), compact structures such as *vessels* resulted in highly localised attention regions that tightly enclosed the input structure. This is evident in Figure 86 where the layer 19 attention contours accurately demarcate blood vessels and the red cells within.

By contrast, patches labelled as *stroma* (Figure 79) were found to be relatively uniform, resulting in broader attention contours that highlighted apparently random areas of cells. This suggests that prediction of this tissue class does not rely on particular objects or structures, and is instead informed by cellular textures and colour distributions.

The overall centre focus of the model is apparent in Figure 80 to Figure 82, where surrounding cancer regions are ignored and the patch is classified according to the central *stroma* tissue. In Figure 84, the central *necrosis* informs the class prediction, despite the tissue being surrounded by a cancer gland structure.

Structural context further from the centre is attended when classifying *lumen* (Figure 85). Here, the diagnosis is informed by surrounding cancer tissue, which implies that the blank centre represents a tumour lumen.

Non-informative predictions also arise from a blank centre with context from adjacent structures. In Figure 87, the patch's multiple surrounding tissue types appeared to confuse the VGG19 feedforward model, which initially reported high probabilities of *stroma* and *muscle*. The feedback model adjusted this to *non-informative* after attending to various heterogeneous features around the central gap.

The attention patterns approximately followed the rotation of patches through 90 or 270 degrees (Figure 88), suggesting that the attention mechanism is not biased to horizontal or vertical directions. Inspection of the patches examined in this section suggests that the attention mechanism works over a range of staining conditions and image sharpness.

6.4.3 Visualisation with ImageNet-100

When the FAL-CNN model was evaluated with images of everyday objects, attention contours from the higher feedback layers showed strong spatial agreement with observed object boundaries (Figure 89). Layer 28 attention selected a region broadly surrounding the object, while layer 19 more closely followed the object outline.

At lower feedback layers, smaller features characteristic of the image class were highlighted. Layer 10 contours often followed a more detailed outline around distinctive features such as the legs on a spider or a bird's protruding feathers. The lowest feedback layers, 0 and 5, showed the model attending to smaller salient objects such as eyes, and textures such as feathers.

These layers also responded strongly to some background textures, such as a spider's web or the seed in a bird feeder. This information can provide useful context, as seen in pathology examples, but would be detrimental in classifying an object that was photographed against an unexpected background. This highlights a well-known limitation of CNN classifiers, observed by Tulio and Ribeiro (2016) in a classifier trained to distinguish wolves and husky dogs, where the model learned to respond only to the presence or absence of a snowy background. In the FAL-

CNN model, it is likely that the higher layers' attention to the foreground object helps to control this behaviour by reducing activations arising from unhelpful background regions.

As noted by Geirhos (2019), CNNs generally respond more to surface textures than to larger structures which may also be important for accurate object identification. The increased object-level focus of the attention model appears to mitigate this, contributing to the enhanced accuracy over the feedforward backbone CNN.

Attention contours at the 80% level were compared with bounding boxes and annotations drawn using VIA, to assess the model's value as an object location tool. F1 scores were higher for bounding boxes despite these being of a different shape to the object outline. Attention contours alone cannot be relied upon for accurately identifying object boundaries.

However, stronger correlation was observed between the centre points of the object annotations and the 80% attention contours. The mean distance was 18-19 pixels, meaning that the model can locate object centres to within 9% of the 224px image width.

7 Saccade-like Behaviour with Feedback Attention Models

7.1 Motivation

Visualisation results showed that feedback activations can direct a classifier model's attention to salient off-centre image content, contributing to increased classification activity.

I observed that a model trained on centre-annotated patches will exhibit an overall attentional focus towards the centre of the image.

Together, these findings prompted a new research question: What would happen if the patch were re-sampled from its WSI, so that the region highlighted by feedback is moved into the centre of the patch, to coincide with the model's inherent focus?

Such an algorithm would emulate the *saccade* behaviour in animal vision, in which executive brain regions direct rapid eye movements to align the central fovea with features of interest in the input scene.

7.2 Methodology

Code used in this section is documented in Appendix Section 1.5.

A *saccade model* was developed to perform iterative repositioning of the $224 \times 224px$ patch sampling region, for input to the CNN, within a larger input patch of $448 \times 448px$ size. The smaller patch was passed to a 1-iteration FAL-CNN classifier (Section 5.1), returning a spatial distribution of feedback activations which were then used to determine a new sampling region for the next saccade cycle. The model behaviour is summarised by the following algorithm:

Algorithm for execution of Saccade model

```
Input:
448x448px image

Sample central 224x224px patch from input image

For each of N saccades:
      Apply sampled patch to FAL-CNN feedback attention model
      Derive centre point of mean feedback activation (layer 28)
      Calculate offset from centre of patch
      Sample new 224x224px patch from input image with this offset

Return:
Arrays of predicted class and feedback activations per saccade
```

Figure 92 shows the behaviour of the new Saccade model, in which the embedded feedback attention model generates attention distributions to determine the next crop region, over *N* saccade cycles.

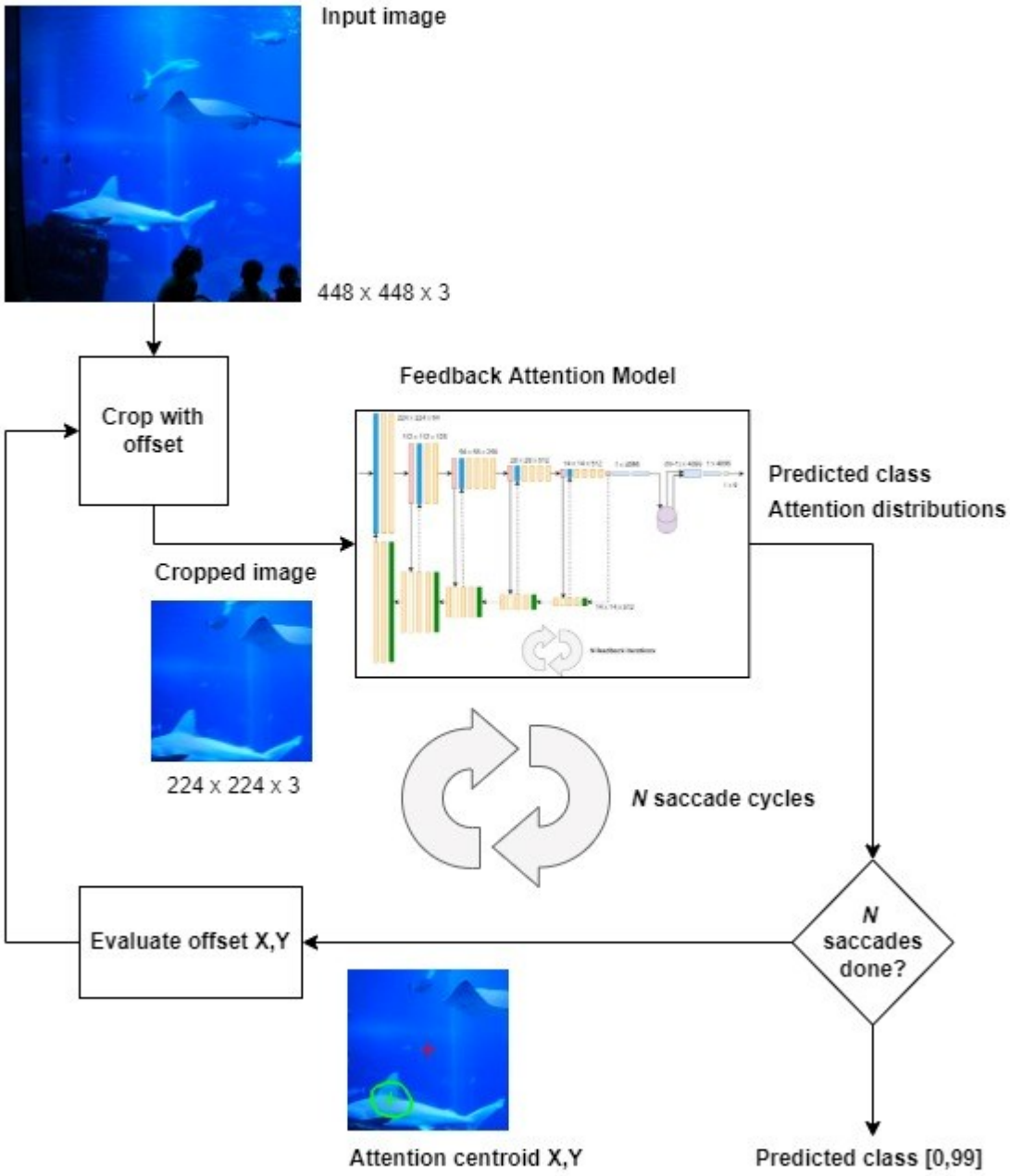


Figure 92: Saccade model system diagram, with sample image from ImageNet-100

Input crop region is progressively repositioned to align with objects at centre of attention region generated by FAL-CNN model.

7.2.1 Model Variants

Several algorithms were tested for calculating the centre of attention (CoA) and hence the new sampling region offset:

- Centre of Mass (CoM) of mean layer 28 feedback activations.
- Centroid of largest 80% attention contour of mean layer 28 feedback activations, as a proxy for peak spatial attention.
- Random sampling location, offset by random X and Y amounts between 0 and 224px upon each saccade cycle. This approach was used to obtain baseline measurements, to establish whether attention-controlled saccades outperformed random sampling.

The saccade model was initially evaluated with two datasets:

- 9-class patches extracted from QUASAR colorectal cancer resection WSIs, as described in Section 3.2.4, here using a patch size of $448 \times 448px$.
- ImageNet-100 images (Section 3.4.2), cropped and scaled to $448 \times 448px$ by the run-time data-loader code.

Later analysis was performed with a 2-class $448 \times 448px$ *tumour-stroma-groups* dataset, compiled by grouping patches from the above 9-class dataset according to parent class group, as defined in Section 3.3.2.

7.2.2 Evaluation of Saccade Models

For each dataset, the saccade model was evaluated against images randomly sampled from the validation set, as previously undertaken when evaluating models against *uncertain-class-patches* data in Chapter 5. Classification accuracy was measured over 30 random sample sets, to obtain mean and 95% CI values.

Pre-saccade $(N = 0)$ classification accuracies were recorded as a baseline against which to compare model performance over multiple saccades. Model accuracy was further measured over 1, 2, 5 and 10 saccade cycles. These values were chosen to explore model convergence and the trade-off between execution time and eventual accuracy.

7.2.3 Visualisation of Saccade Sequences

A 5-saccade model was processed by the VIG, which was extended to capture the $224 \times 224px$ patch region sampled in each saccade, superimposed with feedback attention contours and a cross marking the centroid of the 80% contour.

The $448 \times 448px$ input region was also plotted, with the sequence of patch locations superimposed as colour-coded $224 \times 224px$ boxes.

These plots were combined to create a saccade sequence plot showing the 5 sampling locations and associated attention regions.

7.2.4 Pathologist Reclassification of Post-Saccade QUASAR Patches

The class predicted by the saccade model is determined by the distribution of tissue in a newly sampled region, at an offset location from the original input patch. The new patch may then represent a different tissue class from the original GT label. In order to assess the post-saccade accuracy, updated class labels were therefore required.

A consultant pathologist reviewed patch images from the final sampling locations of an 8-saccade model, for multiple QUASAR-derived input patches. They were invited to classify the new patch at the centre pixel into one of the 9 QUASAR tissue classes.

This review took place over two sessions:

Experiment A: 4 input classes

407 input patches were randomly selected from 4 classes, in a distribution matching the proportions of the classes in the 9-class QUASAR $448 \times 448px$ patch set used in this chapter: 144 of *1-tumour*, 104 of *2-stroma-or-fibrosis*, 114 of *3-necrosis* and 45 of *6-lumen.*

This reduced class set was chosen following discussion with a consultant pathologist, to focus their relabelling effort on tissues of greatest importance in tumour evaluation. *Lumen* was of interest as an example of hollow tissue structures, with near-white background visible at the centre, where the attention-guided saccade was expected to move towards neighbouring tissue types.

The VIG was configured to apply the saccade model to this image set over 8 saccade cycles, using an embedded 1-iteration 9-class FAL-CNN to generate class predictions and attention distributions to guide each saccade. 407 post-saccade patches were saved with a red cross overlaid at the centre. Identifiers for the original and predicted classes were encoded in the filename for later comparison.

The pathologist examined each resulting image and assigned a new class label, according to cells present at the red cross, and surrounding contextual tissue. Labels were manually logged in a CSV file against the original filename. For each image, the class prediction was compared with the new label to generate Boolean agreement values, which were counted and expressed as a percentage of the total file count.

Similarly, per-class agreement scores were generated for the above results when grouped by the pathologist's new label.

A further agreement score was derived by grouping model predictions and new labels into *tumour-group* and *stroma-group* using the parent class definitions in Section 3.3.2. Agreement rates were calculated as the proportion of image files where the model output and expert label were assigned to the same parent group.

Experiment B: 9 input classes

Two further sets of saccade output patches were later compiled for expert re-labelling as follows:

- 400 patches of all 9 tissue classes were randomly selected from the QUASAR $448 \times 448px$ patch set, then processed using an 8-saccade model with 1-iteration 9-class FAL-CNN**.**
- 400 patches randomly selected from a 2-class $448 \times 448px$ *tumour-stroma-groups* dataset, derived from the 9-class input data according to Section 3.3.2. Patch images were processed using an 8-saccade model with an embedded 1-iteration, 2-class FAL-CNN trained on *tumour-stroma-groups* data**.**

In each case, the pathologist examined and labelled the final saccade output image and agreement rates were calculated between saccade model predictions and the new expert labels. For *tumour-stroma-groups,* the 2-class model output was directly compared with the pathologist's group label for each image.

Binomial Proportion Confidence Interval

Binomial Proportion Confidence intervals (BCI) (Brown et al., 2001) were estimated for each agreement score as:

$$BCI = p \pm z\sqrt{\frac{p(1-p)}{n}} \tag{10}$$

Where p is the proportion of Boolean *true* values in a set of size n, and z is the value in the normal distribution corresponding to the required confidence level. Here, $z = 1.96$ was used for a 95% CI.

7.3 Results

7.3.1 Evaluation of Saccade Models

Results tables to complement the plots in this section, including mean accuracies and error bar ranges, are in Appendix section 4.1.1.

QUASAR 9-class patches

Figure 93 and Table 30 (Appendix 4.1.1) show classification accuracies, with 95% confidence intervals, measured for saccade models with varying numbers of saccade cycles, using CoM, 80% contour centroid and random methods to offset the sampling regions.

Baseline results are also included for zero saccades, equivalent to executing the embedded attention model against the central $224 \times 224px$ region of the input patch.

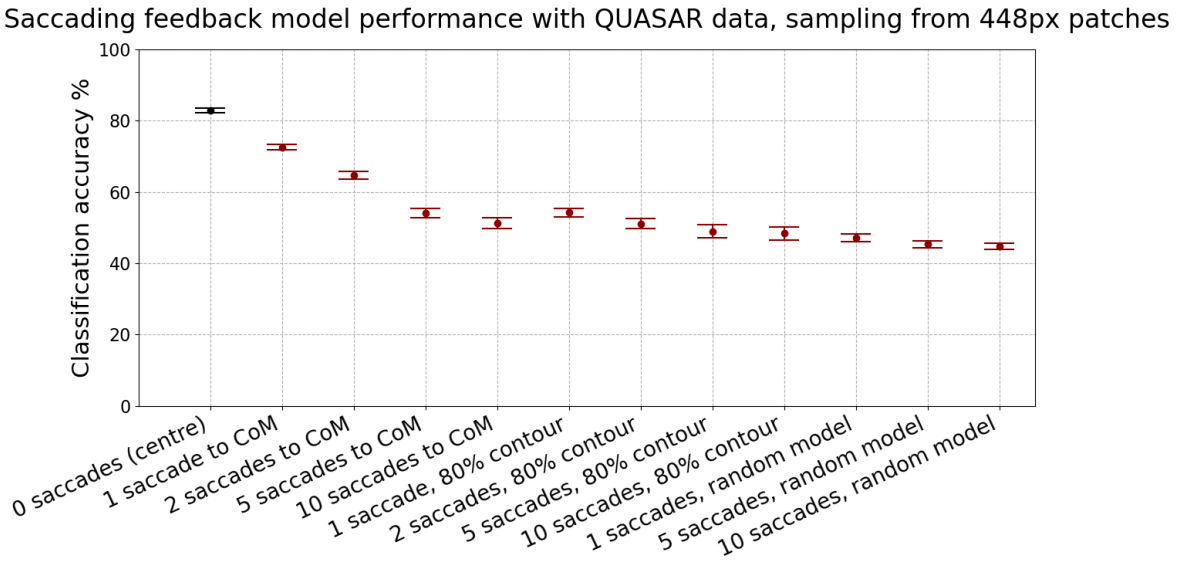


Figure 93: Classification accuracies with 95% confidence intervals for saccade models, relative to zero-saccade FAL-CNN model, with QUASAR 9-class patches

Red points show decreased accuracy with saccade approach, relative to non-saccading model

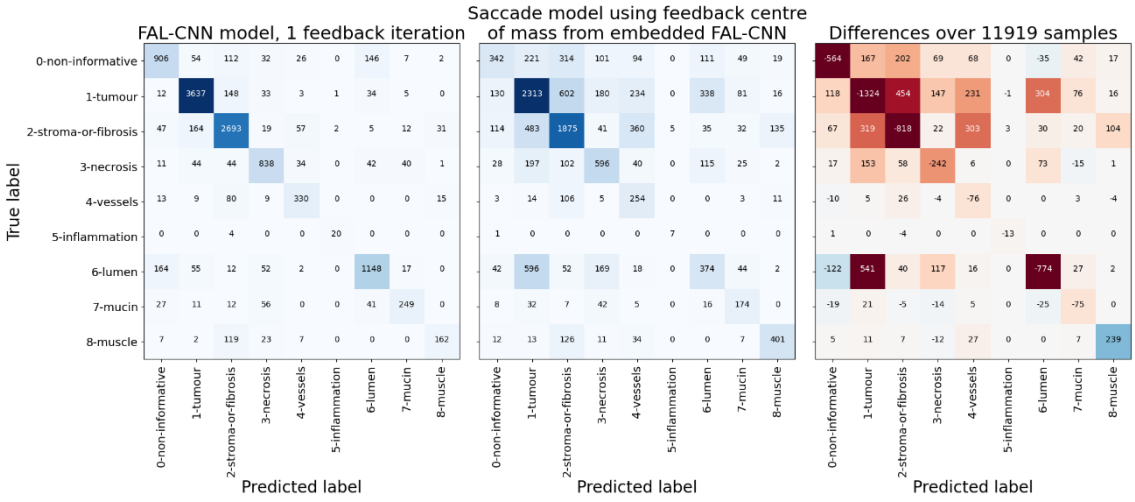


Figure 94: Confusion matrix difference plot, for FAL-CNN and 10-saccade CoM model

Red panels on right-hand difference matrix show dramatic reduction in accuracy relative to original GT class labels.

The confusion matrix (CM) difference plot in Figure 94 shows the per-class differences in classification accuracy between a non-saccading 1-iteration FAL-CNN and the 10-saccade, CoM-based saccade model.

Off-diagonal values highlighted in red in the right-hand grid represent the largest increases in misclassifications, while the darkest red squares on the leading diagonal represent the largest reductions in true classifications for each class.

ImageNet-100

Figure 95 and Table 31 (Appendix 4.1.1) show classification accuracies for each saccade model variant, relative to the zero-saccade FAL-CNN baseline, for ImageNet-100 input images at $448 \times 448px$.

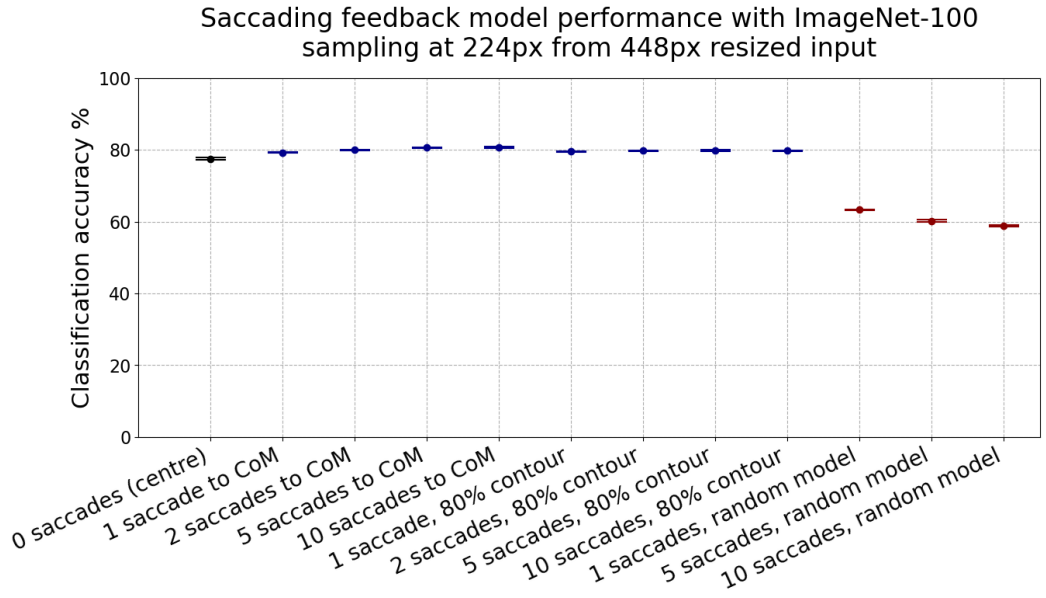


Figure 95: Classification accuracies with 95% confidence intervals for saccade models, relative to zero-saccade FAL-CNN model, with ImageNet-100

Here, saccade action improves on model accuracy measured after initial iteration, sampled at centre of input image. Red points show decreased accuracy with random saccades, ignoring attention region, confirming that saccade behaviour attends to image regions relevant to the expected class prediction.

7.3.2 Visualisation of Saccade Models

QUASAR 9-class patches

Figure 96 shows sequences of crop regions used by the saccade model over 5 cycles, for examples where the final class prediction matches the original GT label.

In each row, the "Saccade 0" plot represents the central $224 \times 224px$ region sampled before the first execution of the embedded classifier. The bold green outlines represent the 80% attention contours at layer 28 after the classifier is executed, before the next saccade.

The sequence of crop regions' locations relative to the larger $448 \times 448px$ input is shown in the larger left-hand panel.

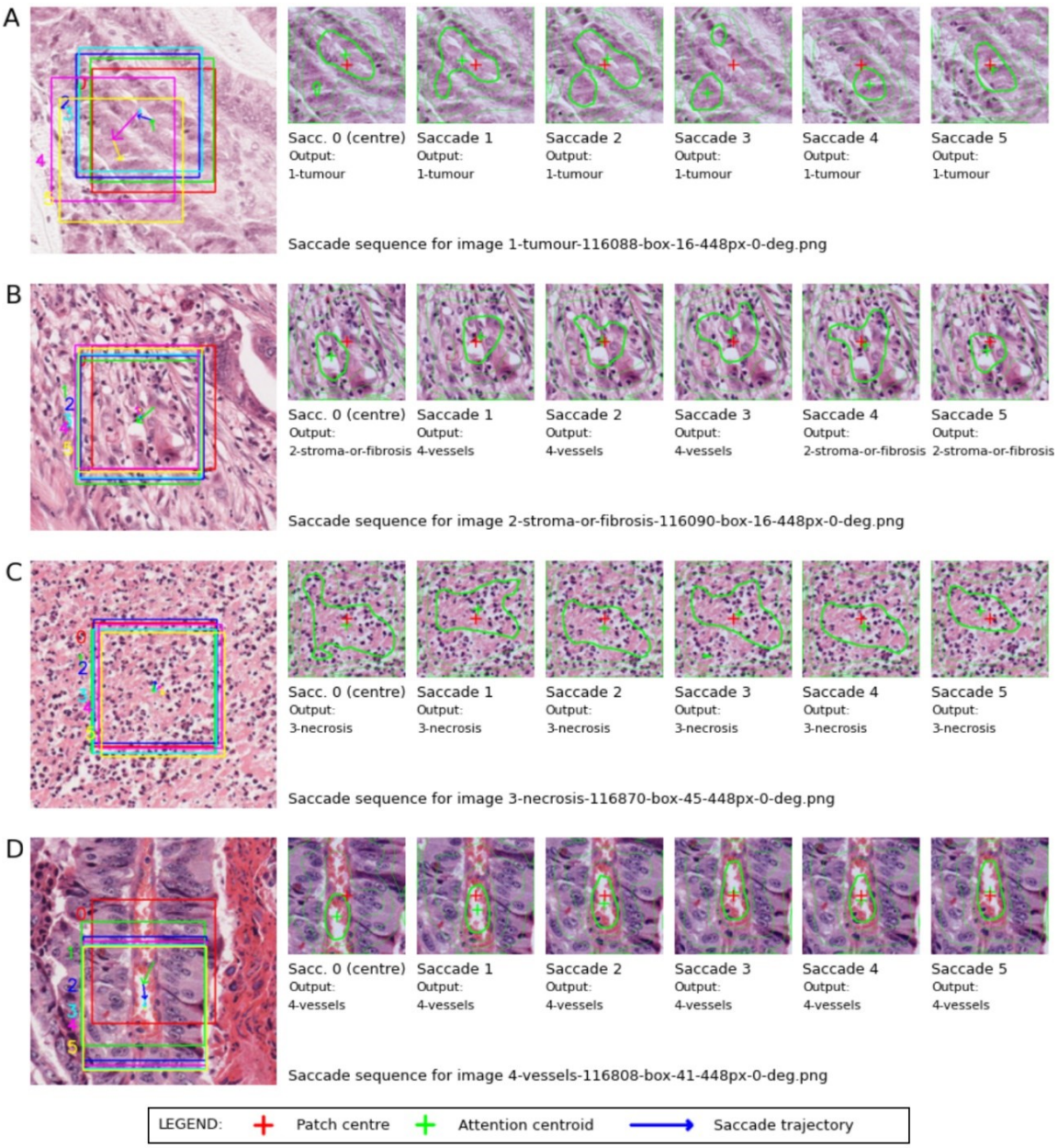


Figure 96: Example saccade sequences for (A) tumour, (B) stroma, (C) necrosis and (D) vessels, where model's final class prediction agrees with GT class

Figure 97 shows selected examples of patches where the saccade process resulted in a different class prediction to the ground truth label.

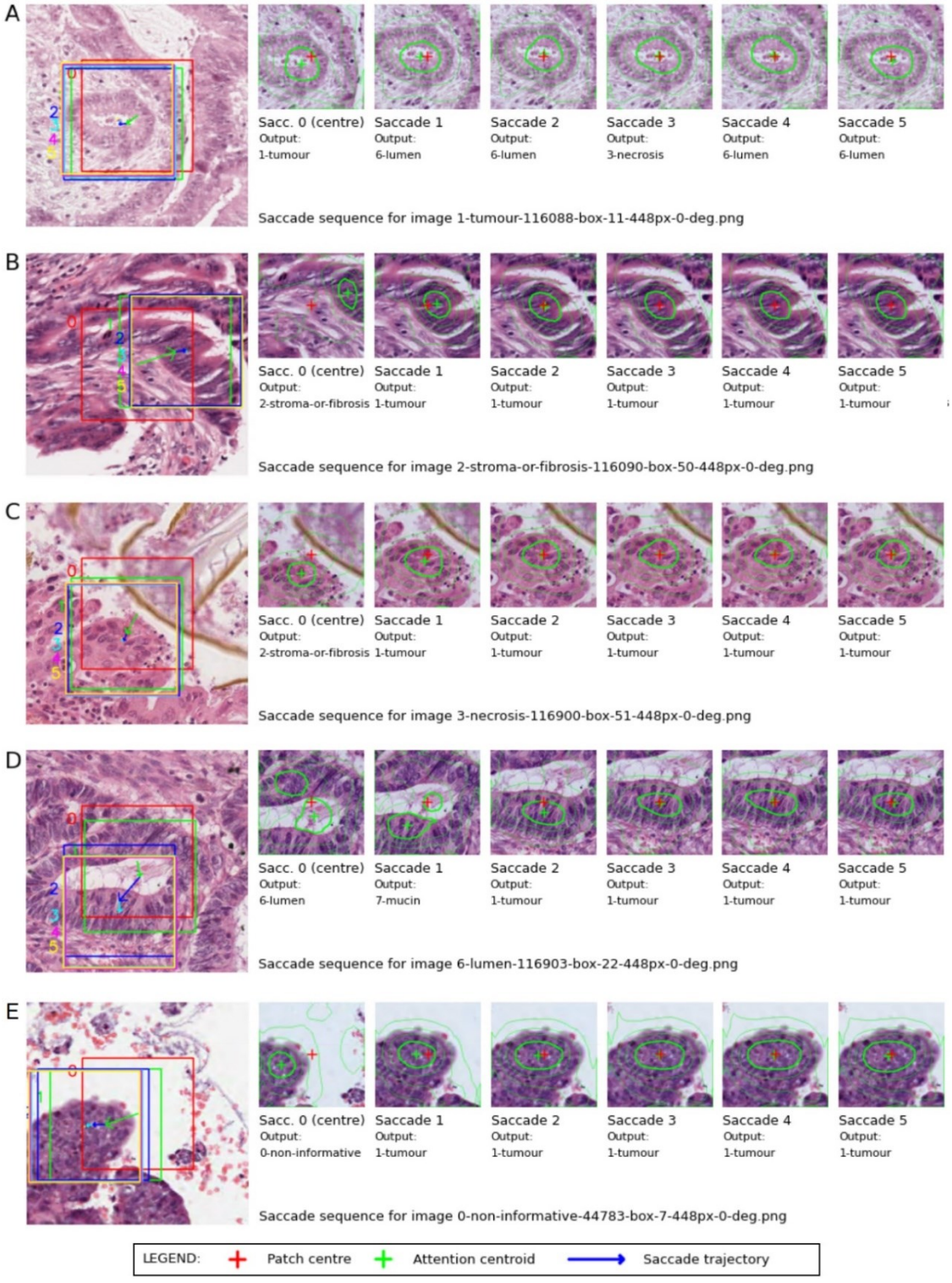


Figure 97: Example saccade sequences for (A) tumour, (B) stroma, (C) necrosis, (D) lumen and (E) non-informative tissue, where model's final class prediction disagrees with GT class

Saccade model centres sampling area on nearby tissue of different type to GT – often finding tumour *cells instead.*

ImageNet-100

Figure 98 shows saccade sequence plots for ImageNet-100 input images, obtained using an ImageNet-trained FAL-CNN within the saccade model. Examples were chosen to illustrate attentional and tracking behaviours for various image classes.

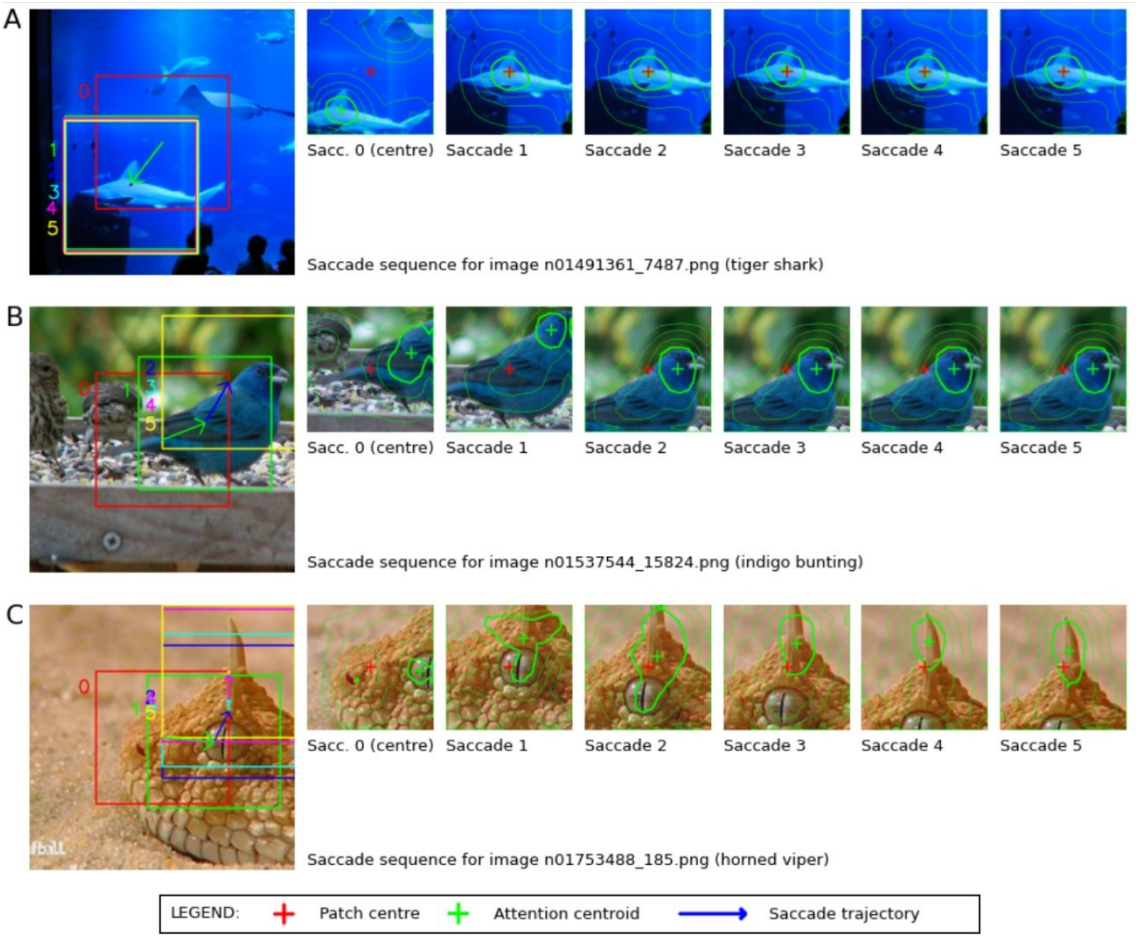


Figure 98: Example saccade sequences for ImageNet-100 classes (A) tiger shark, (B) indigo bunting and (C) horned viper

Sampling region tracks towards salient features of target.

7.3.3 Pathologist Reclassification of Post-Saccade Patches

Results tables to complement the plots in this section, including mean accuracies and error bar ranges, are in Appendix section 4.1.2.

Figure 99 shows the mean agreement rate between the pathologist-relabelled saccade model output patches, and the saccade model's own predictions at these locations, for Experiments A and B.

Earlier classification accuracy results for the VGG19 and FAL-CNN**,** and the saccade model agreement with input patch labels, are included for reference.

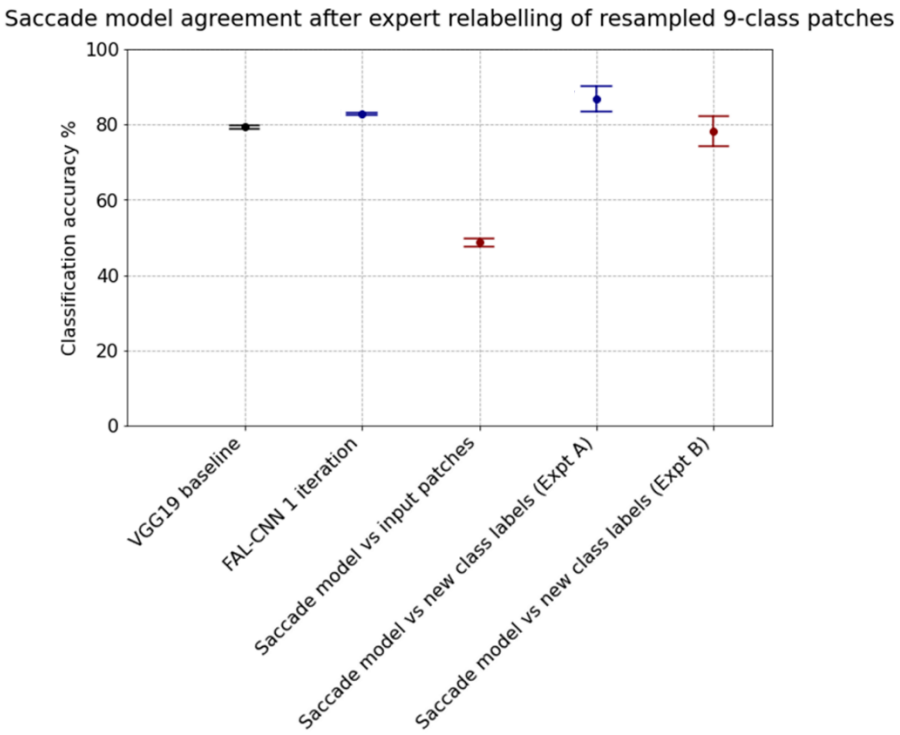


Figure 99: Summary of 9-class classification accuracies of FAL-CNN relative to VGG19 with and without saccade process, with 95% confidence intervals, showing rates of agreement with expert-relabelled post-saccade patches

Red points show decreased accuracy relative to non-saccading baseline model

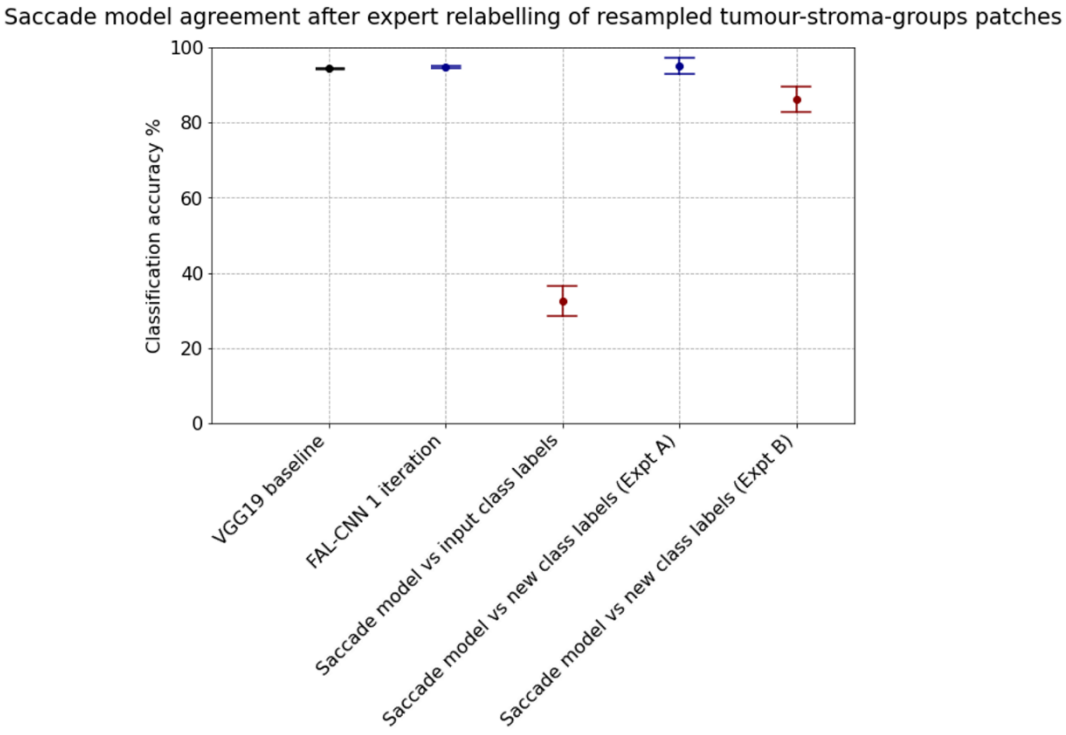


Figure 100: Summary of tumour-stroma-groups classification accuracies of FAL-CNN relative to VGG19 with and without saccade process, with 95% confidence intervals, showing rates of agreement with expert-relabelled post-saccade patches

Red points show decreased accuracy relative to non-saccading baseline model

In Figure 100 (above), saccade model agreement rates with the expert-relabelled post-saccade patches from Experiment A and B (Section 7.2.4) are shown alongside previous classification accuracy results for VGG19, FAL-CNN and saccade model, for patches collated as *tumour-stroma-groups* per section 3.3.2.

Confusion matrices in Figure 101 and Figure 102 show the class-by-class agreement rates between pathologist labels and saccade model predictions, at the re-sampled locations after 8 saccades.

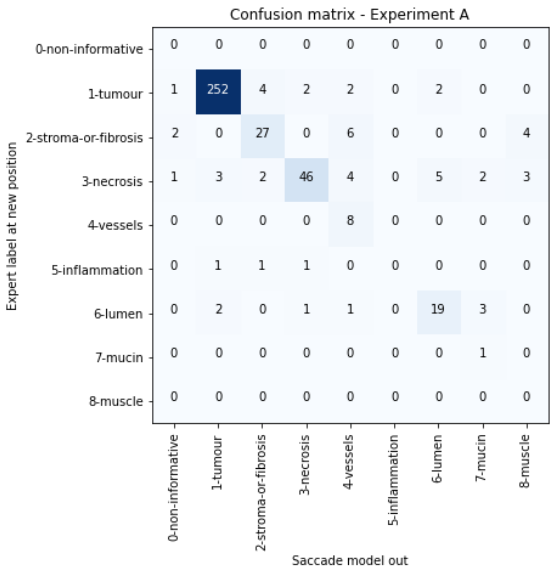


Figure 101: Confusion matrix for expert-assigned label vs saccade model prediction, Experiment A

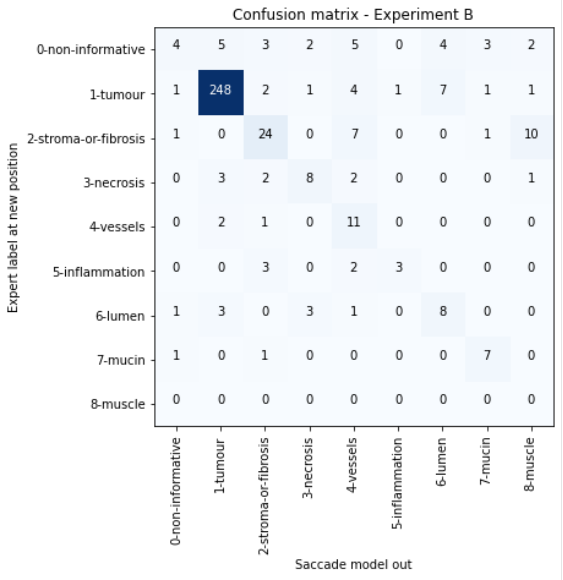


Figure 102: Confusion matrix for expert-assigned label vs saccade model prediction, Experiment B

Figure 103 and Figure 104 show a breakdown of agreement rates per class, obtained by grouping the previous results by the pathologist's label. BCI ranges and total samples per class are listed in Table 34 and Table 35 in Appendix 4.1.

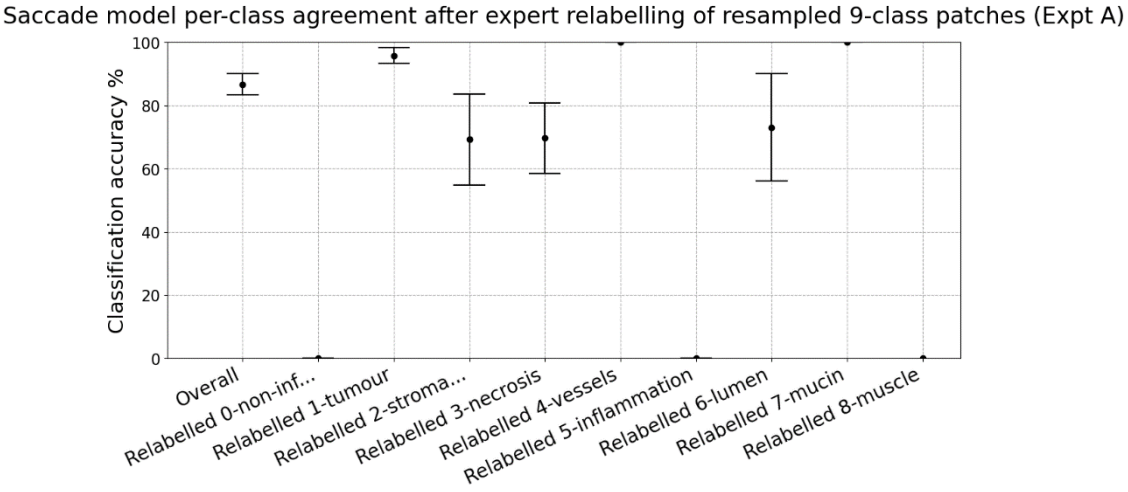


Figure 103: Per-class breakdown of agreement rates between saccade model output class and relabelled final sample location, with 95% binomial confidence intervals – Experiment A (4 input classes)

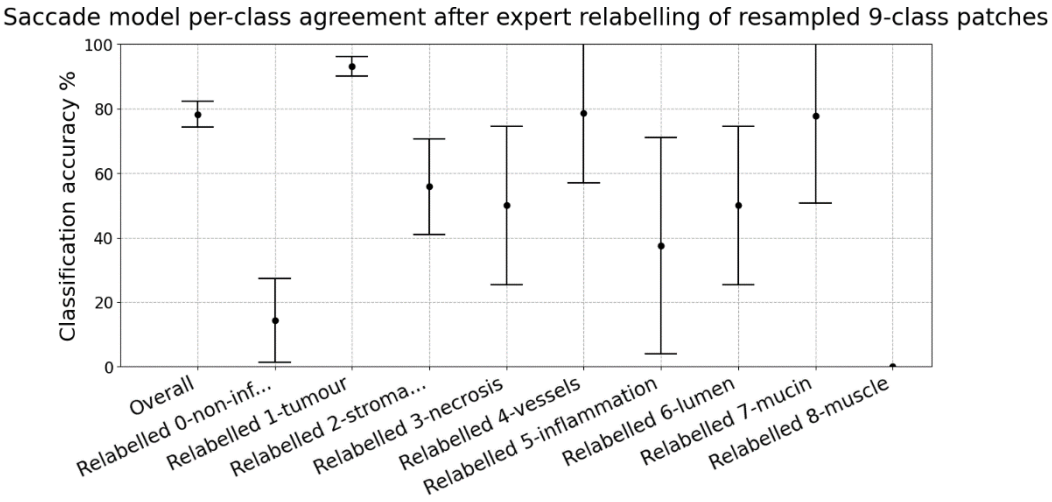


Figure 104: Per-class breakdown of agreement rates between saccade model output class and relabelled final sample location, with 95% binomial confidence intervals – Experiment B (9 input classes)

7.4 Discussion

High rates of agreement, between saccade model predictions and expert labels at post-saccade patch locations, confirm that the FAL-CNN model attends to diagnostically useful regions of tissue within CRC pathology images.

Initially, however, the saccade process was found to be detrimental to agreement rates when using QUASAR data and measuring class predictions against input patch labels. Larger numbers of saccades resulted in further apparently degraded performance, representing disagreement with ground truth classes labelled at the original sampling location(Figure 93).

Nonetheless, the accuracy of the saccade model when resampling around the centre of attention was at least 3.65pp higher than when using random positional offsets, confirmed to be significant by non-overlapping 95% confidence intervals. This demonstrates that that the attention-guided model version attended preferentially to tissue regions that contributed to the correct class prediction.

The CM difference plot (Figure 94) shows that saccade-like behaviour led to frequent incorrect identification of tissues such as *lumen* or *non-informative*. These were instead reported as *tumour* or *necrosis*.

Visual inspection of patch sampling sequences in the saccade process (Figure 96) showed the crop region converging on informative tissue such as *tumour*, or towards small, tightly defined object such as vessels. However, *lumen* and *non-informative* tissue are often characterised by an unpopulated central area surrounded by other tissue types. In these scenarios, the saccade model was found to reinforce the behaviour of its embedded FAL-CNN model, whose feedback distributions focused on the more informative surrounding tissue. This led to a prediction based on the newly attended region, which was often of a different class to the previously centred tissue (Figure 97). This accounts for the observed high rate of classification errors relative to the original GT classes.

After expert re-labelling of the resampled patches, higher rates of agreement were measured between the saccade model's predictions and the new class labels at the new locations (Figure 99). In Experiment A, the agreement rate of 86.73% exceeded the classification accuracy of the

non-saccading FAL-CNN by 3.87pp. The 4-class dataset of *tumour, stroma, necrosis* and *lumen* patches was chosen here to allow the pathologist to focus on diagnostically important or challenging tissue type.

Experiment B used images randomly selected from the QUASAR 9-class dataset and yielded an agreement rate of 78.25%. Although now lower than the accuracy of non-saccading models, this nonetheless represents a significant increase of 32pp over models using random saccades.

Results obtained using a 2-class *tumour-stroma-groups* trained FAL-CNN followed a similar pattern to the 9-class results. Subsequent analysis therefore focused on anomalies in the 9-class model predictions.

Confusion matrices for Experiments A and B (Figure 101, Figure 102), representing the class-by-class relationship between saccade model predictions and new pathologist labels, show that the dominant output class was *tumour*, despite the more balanced distribution of tissue types at the input. High rates of pathologist agreement with this prediction suggest that this apparent bias towards *tumour* does not involve frequent false positive identifications of this class. Rather, the model appears to seek and converge upon regions of genuine tumour tissue within the larger input patch, selecting crop regions where *tumour* is the central, predominant class. We believe that the model's strong response to this class arises from the tumour tissue's combination of distinctive textures and structures, at a scale compatible with the patch size.

In Experiment B, agreement rates were lowest between *non-informative* model outputs and subsequent expert labels (Figure 102). Post-saccade patches associated with this class prediction were instead labelled by the pathologist as *tumour, stroma, vessels* or *lumen*. The *non-informative* class appears to be particularly challenging for a ML model, as the original label can be applied in response to poor image quality, or to many permutations of heterogeneous tissue types, particularly around an unpopulated central region. Other tissue classes have more distinctive, consistent structures which the model is able to associate more strongly with given output predictions, suggesting the *non-informative* class could be excluded altogether from future model training.

Limitations

The per-class grouping of agreement rates (Figure 103 and Figure 104) show wide BCIs for classes other than tumour, due to the smaller total of samples of each type (Appendix 4.1, Table 34 and Table 35). Agreement rates of 0% and 100% occurred where the class totals were very small, and should be therefore disregarded as statistically weak.

The saccade model's observed tendency to converge on *tumour* regions shows potential value in tumour detection applications, but may provide a biased result when detection rates are compared with other classes such as *stroma*. Further experiments (Chapter 8) were therefore performed to assess the model's efficacy in measuring TSR in the WSI pipeline.

ImageNet

With ImageNet-100, the saccade model outperformed a non-saccade FAL-CNN loading just the central $224 \times 224px$ region of a $448 \times 448px$ input image (Figure 95). Increasing the number of attention-guided saccades resulted in increasing classification accuracy. Contrastingly, 'random walk' saccades resulted in significantly reduced performance, often focusing on background or uninformative parts of the target. This further supports the conclusion that image regions selected during attention-guided saccades contained features that contributed to a correct class prediction.

Plots of saccade sequences for multiple input images (Figure 98) show the resampled image regions becoming centred on identifying features, such as a shark's dorsal fin (A) or the horn of a horned viper (C).

Additionally, the saccade process often discovered salient features *outside* the initial crop region. With the indigo bunting (B), the bird's head was not visible in the initial input to the FAL-CNN. Nonetheless, the saccade process brought the head into frame over several saccades, converging on an emerging attention 'hotspot' around the eye and beak.

When comparing model output with input classes at pre-saccade locations, the saccade model performed better with ImageNet than QUASAR. This is attributed to the relatively sparse distribution of objects in ImageNet images, such that a saccade movement is less likely to encounter an object or region of a different class. By contrast, QUASAR patch images are highly heterogeneous. Saccade behaviour often converges on different tissue to that originally at the patch centre, which necessitated expert relabelling at the new location.

The proportion of the WSI width encompassed by these saccade movements is very small compared to eye movements in nature. Nonetheless, the saccade process here is effective because of the rapid spatial variation of tissue types and structures across the patch area, allowing the new model to locate nearby *tumour* amongst less informative tissue.

8 Feedback Attention Model Performance in WSI Pipeline

8.1 Motivation

Section 4.5 describes a published ***Weighted Regular Sampling* Pipeline** (WSRP) for WSI analysis, performing segmentation of the tumour region of interest (ROI) and estimation of tumour stroma ratio (TSR) for assessment of disease prognosis. The accuracy of this pipeline appeared to be limited by the accuracy of the VGG19 model used for patch-level classification.

This assumption was tested here by replacing the VGG19 with the most accurate feedback attention CNNs, both as the main 9-class classifier and for false positive correction (FPC). A further experiment used the Saccade model in the pipeline to examine the effect of this model's tendency to seek out tumour tissue.

In each case, the pipeline's ability to predict cancer ROI and TSR was measured against ground-truth data derived from QUASAR WSI annotations (Sections 3.2.2 and 3.2.3).

This work represents further exploration of the aims and objectives introduced in Section 1.2.1 and explored in Chapter 4, for visualising cancer in the WSI.

8.2 Methodology

The shell script used to execute the pipeline on ARC4 was parameterised to allow feedback-enhanced CNN models to be substituted for the main 9-class classifier (CNN1 in Figure 105), and for the 2-class classifier (CNN2) used for FPC.

The 9-class FAL-CNN models from experiments in Section 5.1 were reused in the pipeline as CNN1.

Further feedback attention models were trained to perform FPC using the 2-class *tumour epithelium/normal epithelium* dataset used in the earlier training of VGG19 FPC models, in Section 4.2.2. FAL-CNN models using 1 and 2 feedback iterations were trained for 200 epochs, using SGD optimisation with LR of 0.0003 and momentum of 0.9.

An optional third model path was provided, allowing a 2-class *tumour-stroma-groups* model (CNN3) to be specified for TSR calculation. Where this was not supplied, the TSR calculation defaulted to using CNN1 and FPC classification outputs per the original pipeline. The extended 3-CNN architecture is shown in Figure 105.

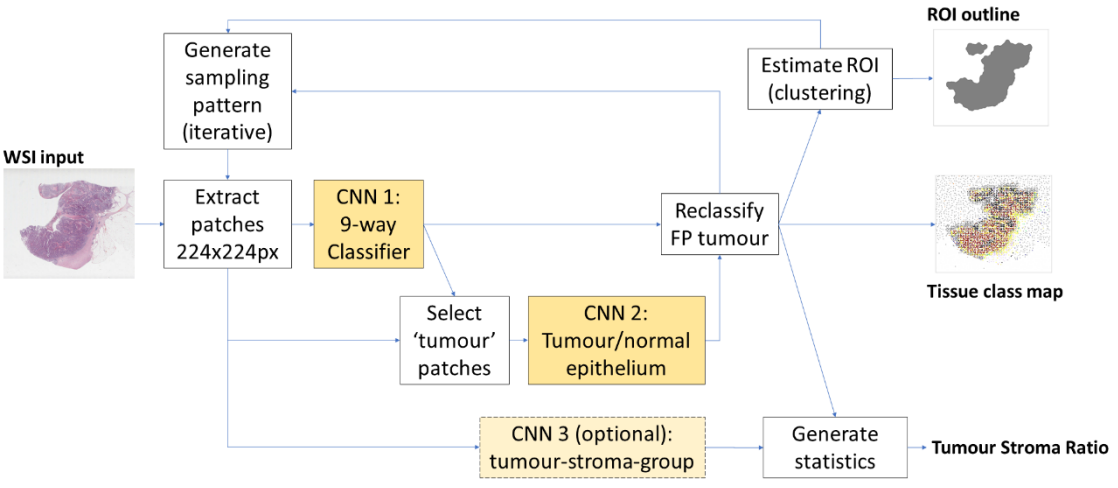


Figure 105: Enhanced Weighted Regular Sampling Pipeline (WRSP) with optional two-class CNN for TSR calculations

The pipeline code was further modified to allow Saccade models (Chapter 7) to be used as CNN1. In this scenario, the post-saccade patch location was used to reload patches from the WSI for use in CNN2. This ensured that the FPC classifier acted on the same tissue region as that loaded by the saccade model in its final saccade. It was anticipated that using the Saccade model would shift each patch in the pipeline's sampling pattern to centre on nearby informative tissue, rather than simply reporting the class for each original patch, thus enhancing the pipeline's prediction of tumour ROI outline and TSR.

The pipeline used a parent tile size of 1024×1024 pixels, from which multiple 224×224 patches were sampled over two iterations of ROI optimisation. This was consistent with earlier experiments in Chapter 4.5.

8.2.1 CNN Model Combinations

Table 20 shows the combinations of CNN models chosen as CNN1, 2 and 3 in the pipeline, with the abbreviated names used in subsequent plots and tables.

Table 20: Combinations of CNN models used in WSI processing pipeline

Abbreviated label	CNN1 (9-way classifier)	CNN2 (FP correction)	CNN3 (dedicated classification for TSR calculation)
VGG + VGG	VGG19	VGG19	None
VGG + FAL1	VGG19	FAL-CNN 1-iteration	None
VGG + FAL2	VGG19	FAL-CNN 2-iteration	None
FAL1 + VGG	FAL-CNN 1-iteration	VGG19	None
FAL1 + FAL1	FAL-CNN 1-iteration	FAL-CNN 1-iteration	None
FAL2 + VGG	FAL-CNN 2-iteration	VGG19	None
FAL2 + FAL2	FAL-CNN 2-iteration	FAL-CNN 2-iteration	None
FAL3 + VGG	FAL-CNN 3-iteration	VGG19	None
FAL2 + VGG + FAL2 t-s-g	FAL-CNN 2-iteration	VGG19	FAL-CNN 2-iteration *tumour-stroma-groups*
5-saccade FAL1 + VGG	5-Saccade model using 1-iteration FAL-CNN	VGG19	None

8.2.2 Cross Validation

Pipeline measurements in this section were taken five times for each model combination, once for each split in a 5-fold cross validation (CV) set. WSIs were chosen from the test set of each data split, and processed using CNNs that were trained against patches from WSIs in the corresponding training set. Thus, all patches analysed during pipeline execution were unseen during training, to mitigate the risk of overfitting.

8.2.3 TSR Distribution: Bland-Altman and Scatter Plots

Tumour and stroma patch counts were captured during pipeline execution, for each TSR sampling method, for each WSI in the test set of each 5-fold CV split. These statistics were

used to evaluate TSRs for each WSI and sampling method. These values were plotted alongside TSRs derived from the ground truth class annotations.

TSR values, for model combinations of interest, were displayed using scatter plots and Bland-Altman plots. The latter were used to show the bias and spread of values relative to the mean. Points were plotted for 686 WSIs, combined from results from all 5 CV splits, The values were grouped into colour-coded ranges corresponding to 'high' and 'low' TSR. This was intended to reveal cases where pipeline inaccuracies would risk placing a patient into an incorrect risk band in any clinical implementation of the pipeline. For this, a TSR threshold of 0.488 was used to separate high and low, as proposed by Zhao et al (2020) to categorise patients by survival prognosis.

8.3 Results

8.3.1 TSR Estimation

Figure 106 shows the TSR error evaluated for each model combination. The error was calculated per Equation 6, Section 4.4.2, comparing the ground truth TSR at the pathologist-selected area of highest tumour cell density with pipeline estimates of TSR, evaluated at this location (green) and at the pipeline's own estimated location of maximum tumour density (blue).

Mean error rates and error bar values are listed in Table 36, Appendix 5.1.

TSR distributions for results labelled A, B, C and D below are shown in more detail in Scatter and Bland-Altman plots in Figure 107 to Figure 110.

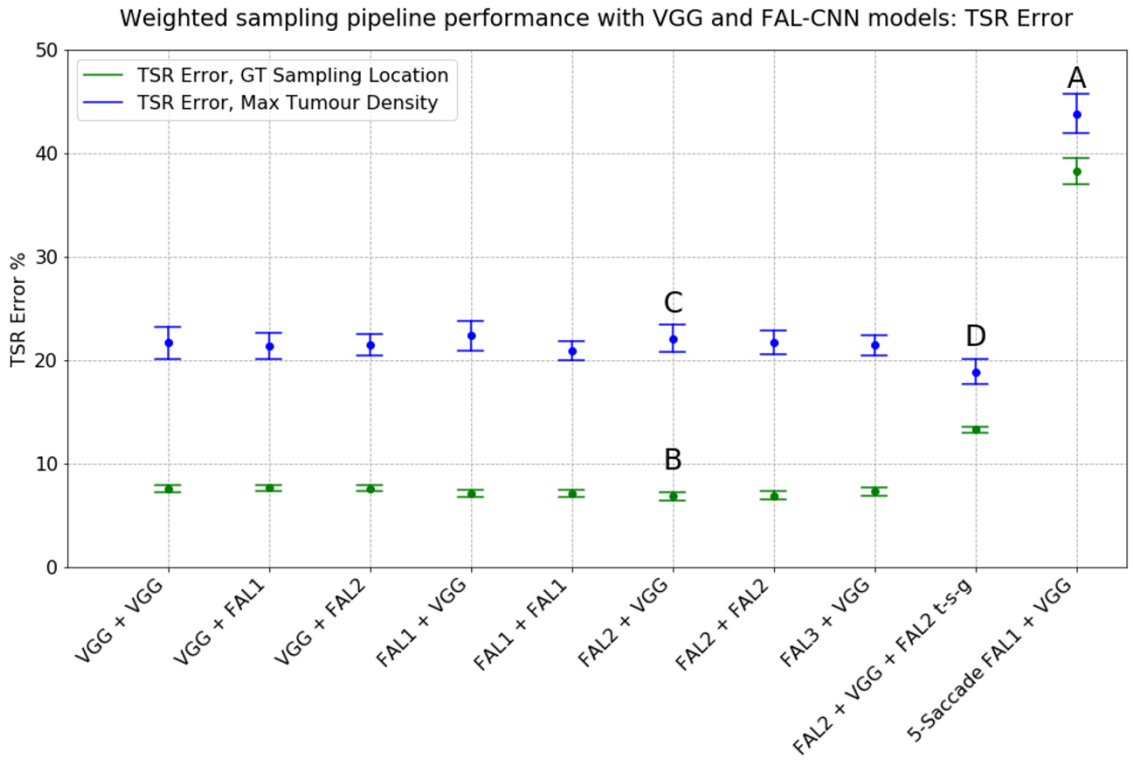


Figure 106: TSR error rates in Weighted Regular Sampling Pipeline, using various combinations of feedforward and feedback CNN classifiers

TSR calculation is most accurate when using additional "tumour-stroma-group" classifier (D). Accuracy is unsurprisingly higher at pathologist-selected GT locations (green plots). Saccade model shows greater error (A) due to its tendency to seek tumour tissue over stroma.

TSR distribution: Bland-Altman and scatter plots

Figure 107 to Figure 110 show scatter and Bland-Altman plots for selected model combinations and sampling locations, as labelled A, B, C and D in Figure 106.

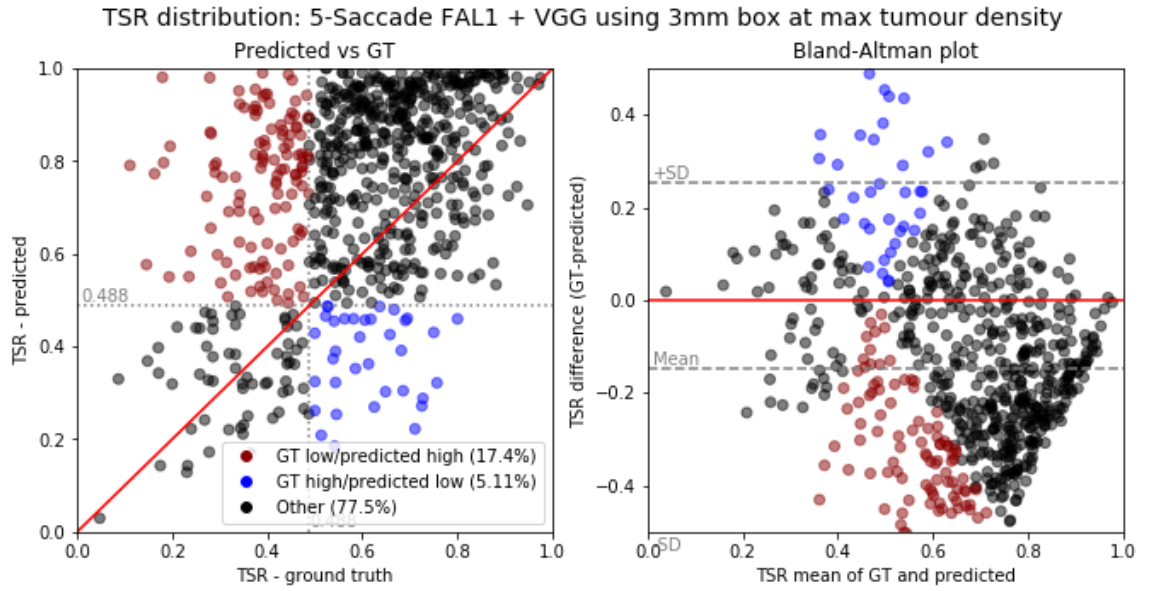


Figure 107: (A) Scatter and Bland-Altman plots of pipeline-predicted vs actual TSR, using 5-Saccade model for TSR calculation in 3mm box at maximum tumour density location determined by pipeline

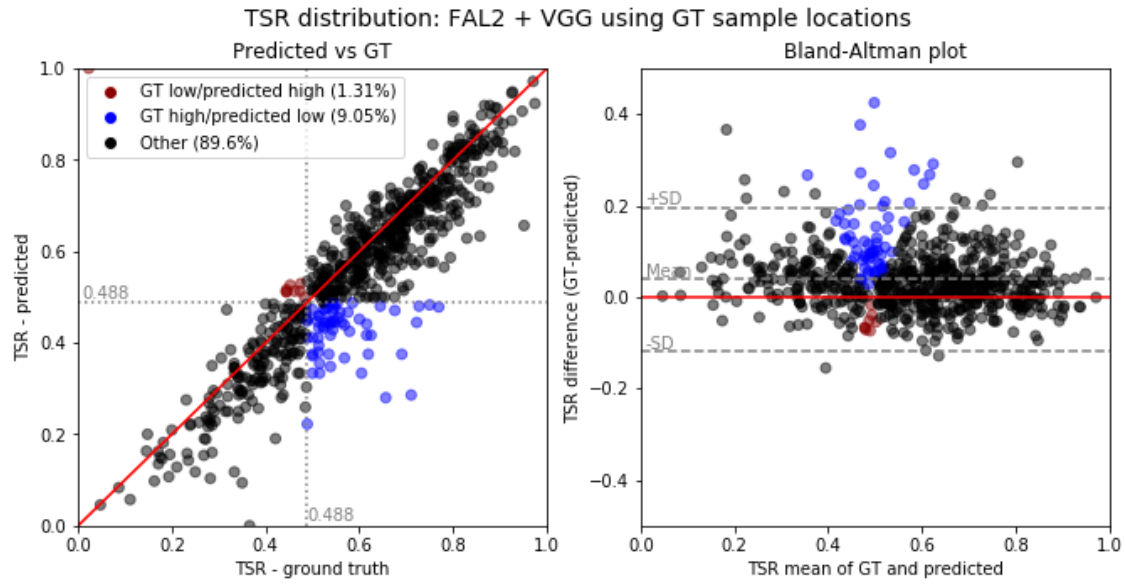


Figure 108: (B) Scatter and Bland-Altman plots of pipeline-predicted vs actual TSR, for FAL2+VGG models at GT sample locations

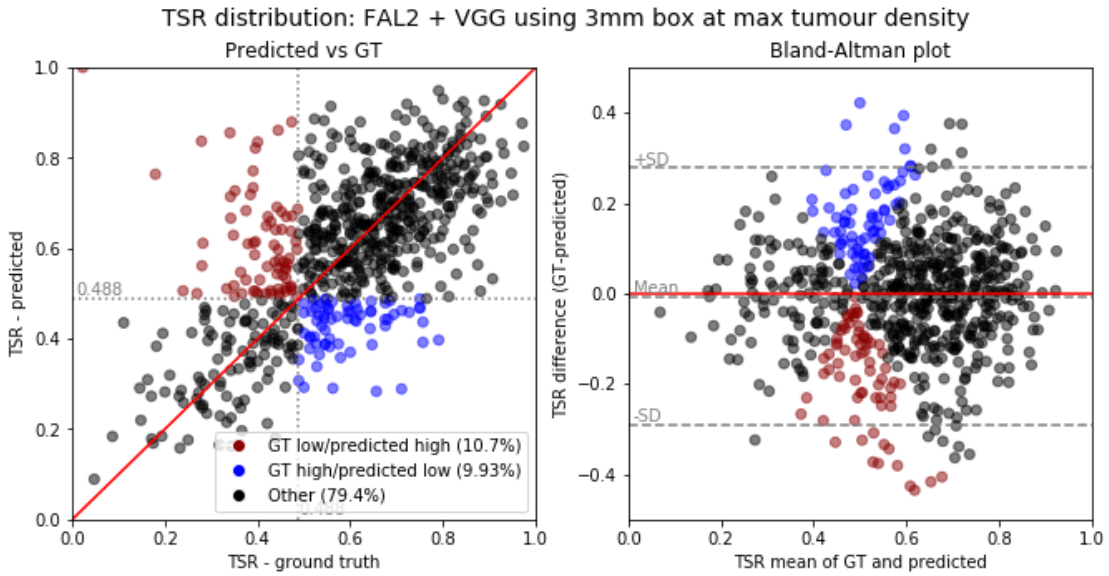


Figure 109: (C) Scatter and Bland-Altman plots of pipeline-predicted vs actual TSR, for FAL2+VGG models at maximum tumour density location determined by pipeline

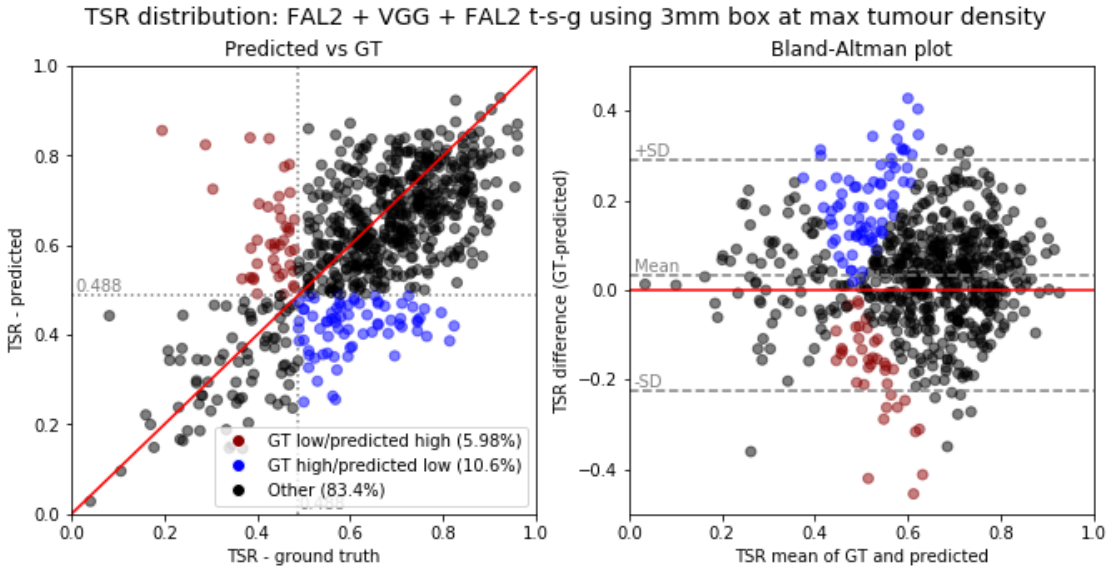


Figure 110: (D) Scatter and Bland-Altman plots of pipeline-predicted vs actual TSR, using tumour-stroma-groups model for TSR calculation in 3mm box at maximum tumour density location determined by pipeline

8.3.2 Tumour ROI Estimation

Figure 111 shows rates of agreement between the GT ROI, derived from expert annotations, and the pipeline ROI estimate. Agreement is expressed as F1 (Dice) score and Intersection over Union (IoU). Error bars represent a $\pm 1SE$ range either side of the mean of the CV results. These values are listed in Table 37, Appendix 5.1.

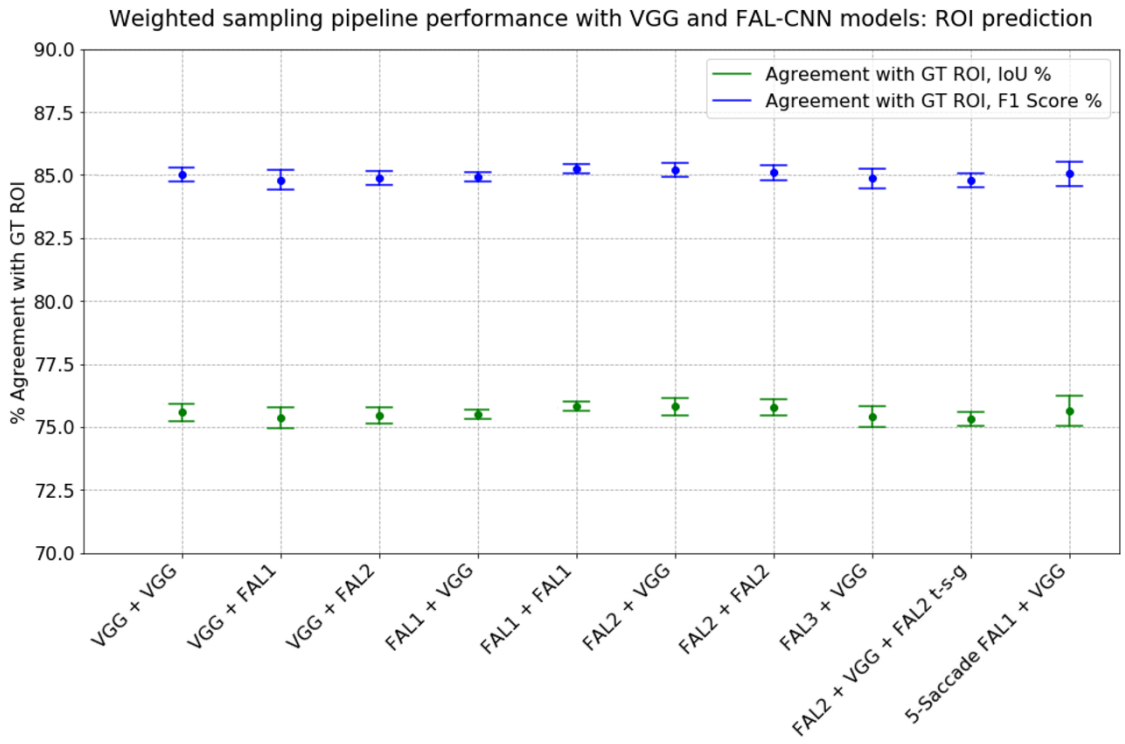


Figure 111: Rates of agreement with $\pm 1SE$ range between WSI pipeline and GT ROI annotations, for combinations of feedforward and feedback CNN classifiers

Only marginal benefit is gained using FAL-CNN. Errors in VGG19 are mitigated by FP correction.

8.3.3 WSI Processing Time

The mean time to process a WSI is plotted in Figure 112, for each model combination. The value was calculated by dividing the total duration of the HPC job, for each CV split, by the total number of WSIs in the split.

Mean durations with error bar values are listed in Table 38, Appendix 5.1.

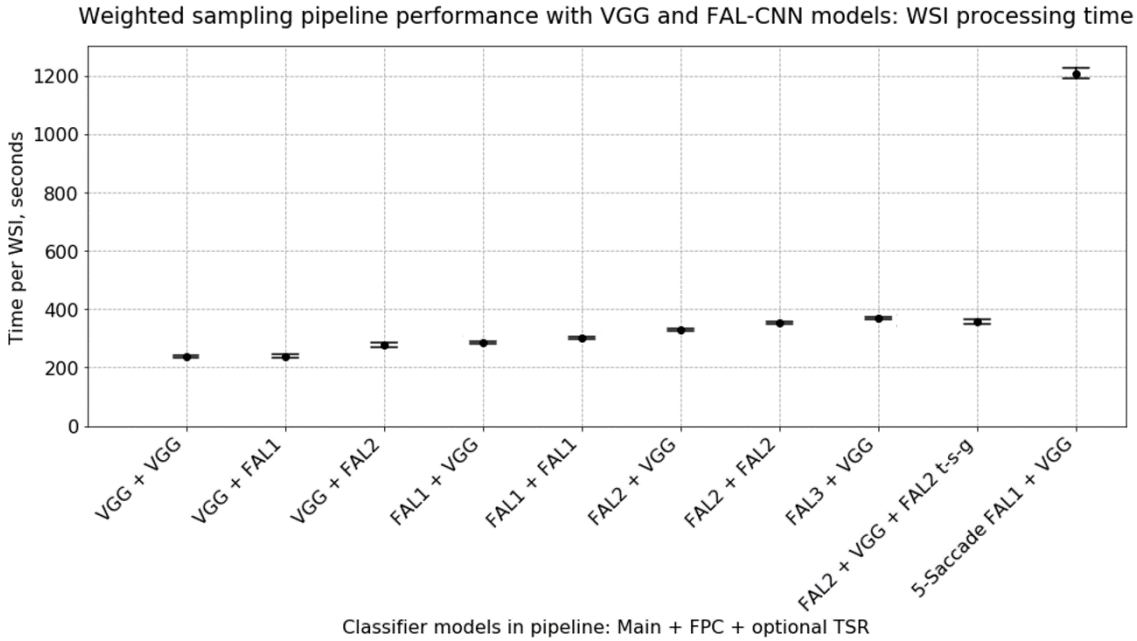


Figure 112: Mean pipeline processing time per WSI with $\pm 1SE$ range, for combinations of feedforward and feedback CNN classifiers

Saccade model is 4-5x slower due to multiple iterations of model execution.

8.4 Discussion

8.4.1 TSR Estimation

The use of feedback attention models in the WSI pipeline resulted in more accurate estimation of tumour stroma ratio than with the VGG19 (Figure 106). The benefit was most apparent when sampling at the pathologist-selected GT locations. Here, the 2-iteration FAL-CNN (FAL2) reduced the TSR error rate to 6.84% from the 7.57% observed with VGG19. This is an improvement of 0.73pp or approximately 10%.

When using sampling locations chosen by the pipeline itself, the greatest reduction in TSR error was obtained with a *tumour-stroma-groups*-trained FAL2 model dedicated to the TSR calculation. TSR error was reduced to 18.86%, an improvement of 2.78pp or approximately 13% over the VGG19-based pipeline.

Using a saccade model for patch classification was found to be highly detrimental to TSR accuracy, with a worst-case error rate of 43.81%. Analysis of the Scatter and Bland-Altman plots (Figure 107) revealed a high level of bias, with many WSIs having excessively high predicted TSR. This is likely to be due to the known tendency of the saccade model to seek out tumour regions in preference to stroma and other tissue classes, resulting in higher reported proportions of tumour.

This behaviour is potentially harmful, risking a falsely optimistic prognosis for the patient due to the lower predicted rates of *tumour stroma*. The red-coded points, in the top-left quadrant of the scatter plot in Figure 107, represent 17.4% of cases that would be incorrectly placed in a high TSR group, by exceeding Zhao's threshold of 48.8%. The ground-truth TSR values, derived from pathologist annotations, are lower, indicating higher stroma levels associated with more aggressive cancers. These could be incorrectly diagnosed as less severe in red-coded cases.

By contrast, when using FAL2 to evaluate TSR at GT sampling locations, scatter plot points (Figure 108) were clustered close to the diagonal, reflecting the strong agreement measured between GT and predicted TSR in this scenario. Only 1.31% of points fell in the dangerous top-left ‘red’ quadrant, and all of these were close to the 48.8% threshold. The offset mean level in the BA plot reveals an overall tendency towards predicting lower TSR than the ground truth. 9.05% of cases were in the bottom-right ‘blue’ quadrant, representing a risk that this group of patients may be given a pessimistic prognosis leading to potentially unnecessary treatment.

When TSR was calculated at a peak tumour location selected by the pipeline, a wider spread of points was produced (Figure 109). Approximately 20% of cases were placed in an incorrect high-stroma or low-stroma group. The distribution in the BA plot shows a mean value near to zero, indicating negligible overall bias towards tumour or stroma.

When using an additional *tumour-stroma-groups* model as CNN3 in the pipeline, the improved mean TSR accuracy is evident in the reduction in the proportion of ‘red’ cases to 5.98% (Figure 110). The three-model combination yields the lowest overall TSR error rate at sampling locations chosen automatically by the pipeline. The 10.6% rate of ‘blue’ cases, at risk of a misdiagnosis of aggressive tumour, would nonetheless necessitate further triage stages in a practical workflow.

Nonetheless, grouping patches into the tissue categories of interest has enabled more accurate tumour and stroma counts to be generated. This represents a useful application of *Occam’s Razor,* reducing complexity by simplifying the classification problem.

8.4.2 Tumour ROI Estimation

In Tumour ROI estimation (Figure 111), the pipeline's performance with feedback attention models was similar to that observed with the basic VGG19. F1 scores, relative to ground truth annotations, were in the range 84.78% to 85.24%. Marginal increases were measured when using feedback models for CNN1 and CNN2.

It appears that any benefit arising from the improved patch classification accuracy is diluted by the averaging behaviour of the pipeline, which uses the spatial distribution of multiple tumour patches to estimate the ROI. Errors in classifying individual patches are often corrected by this approach, so increased model accuracy does not translate into greatly improved ROI prediction.

The use of feedback models for FPC had only a marginal effect on ROI estimation. This suggests that FPC, when combined with the elimination of many stray tumour patches by the clustering algorithm, was already compensating effectively for misclassified patches before feedback methods were introduced.

8.4.3 Processing Time

WSI processing time (Figure 112) increased with the introduction of feedback attention, by a factor of approximately 1.2 to 1.5 for the most accurate model combinations. This was expected, given the extra processing required in the feedback pathways of the FAL-CNN. Unsurprisingly, the longest processing times were seen with the saccade model, which invoked its embedded CNN five times for every patch sampled, and was approximately 5x slower than the VGG19 baseline as a result.

The smallest performance impact from the use of FAL-CNN models was in the FPC role (CNN2). This model is only invoked when tumour is detected by the main classifier (CNN1), so does not affect the analysis time for every patch.

The performance of ARC4 HPC infrastructure has greatly improved since earlier pipeline experiments. A batch script option was introduced, to enable parallel CPU processing and thereby reduce the processing time in between calls to the GPU. A 5.8x reduction in pipeline execution time was observed with VGG19 models, from 23 minutes per WSI in earlier work (Broad et al., 2022) to approximately 4 minutes here.

The inference time of a CNN model remains an important consideration when designing a CNN-based processing pipeline. However, overall performance has been shown to benefit from parallelisation and optimisation, in model and pipeline code and in the supporting HPC architecture, and these factors must also be considered when implementing a WSI pipeline.

8.4.4 Proposed TSR Sampling Tool

It is acknowledged that the pipeline is most effective in calculating TSR when sampling at ground truth locations, which were selected by a pathologist when the QUASAR annotations were created. This is unsurprising, as the algorithm is then using data from the same sampling area as used by the human expert and the residual error is only due to differences in the patch classes evaluated at this location by machine and human.

A fully automated pipeline does not have access to a human-chosen location when processing an unseen, un-annotated WSI. It must therefore select its own optimum sampling location within unseen, un-annotated WSIs. Without performing further survival analysis, it is not possible to determine whether results from the human-selected or machine-selected sampling location correlates best with overall survival.

Meanwhile, it may nonetheless be possible to exploit model combinations that performed well at pathologist-selected locations, to evaluate TSR at WSI locations chosen manually by an expert operator.

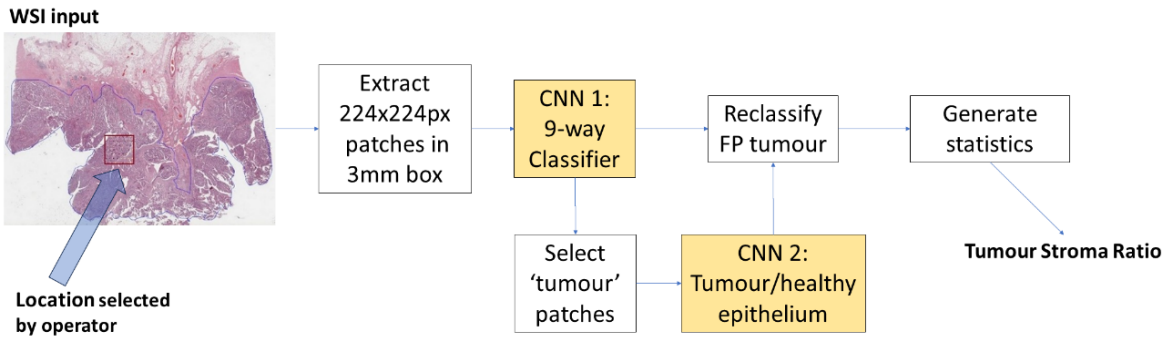


Figure 113: Proposed TSR sampling tool

Pathologist picks location in WSI, then pipeline classifies multiple patches and calculates TSR.

Figure 113 shows such a system, based on components of the current WSI pipeline. The operating scenario is as follows:

1. **User views WSI** at thumbnail and zoomed-in scales, to determine regions of high tumour density. This can be guided by tissue distribution maps generated by the fully automated pipeline.
2. **User selects region of interes**t, e.g. 3mm box at peak tumour near luminal aspect, as used in original studies with QUASAR data.
3. **Tool samples patches** using RandomSpot algorithm to create sample patch distribution.
4. **Tool loads patch images** from these locations, invoking the 9-way classifier (2-iteration FAL-CNN as CNN 1) and FPC stage (VGG19 as CNN 2) for each patch.
5. **TSR is calculated** as $TSR = \frac{T}{T+S}$ from totals of tumour (T) and stroma (S) patches.

The use of an automated tool allows a far greater number of patches to be sampled, than the 50 per WSI that were labelled by the annotating pathologist. Stereology calculations by Wright (2017, p.16) suggest that using in the order of hundreds of patches would increase the accuracy of the TSR calculation. Given that the pipeline takes 4-5 minutes to process an entire WSI, with 1000s of patches, the TSR calculation might be expected to complete in a few tens of seconds on an HPC GPU node.

Results would be grouped into TSR high, low, and borderline. The latter would apply to cases where the result is within a given percentage of the high/low decision threshold, where further expert analysis would be required to mitigate risks of misdiagnosis. In the high and low bands, where the TSR grouping is more certain, a tool such as this would enable diagnostically useful metrics to be generated far more rapidly than would be possible with manual cell counting.

9 Discussion

9.1 Thesis Overview

This work aimed to explore applications of AI in extracting diagnostically useful information from high-magnification whole slide images (WSIs) in digital pathology. Chapter 1 discusses the central problem: The gulf in scale between multi-gigapixel WSIs, and the far smaller image patches that many current AI models can handle.

The work in this thesis addressed this problem using techniques inspired by human vision, in which attention processes guide the eye towards targets that relate to our executive goals, allowing us rapidly to extract relevant information from a complex, cluttered input scene. Simultaneously, neural feedback processes adjust sensitivity in our visual processing layers, so that salient features 'pop out' more vividly.

The application of methods inspired by these biological processes led to a published method for characterising cancer tissue at WSI level, and to the development of feedback attention and saccade-based models that respond to tumour at a patch scale.

9.2 Achievements

Work performed for this thesis resulted in the following key outcomes:

1. Novel WSI processing pipelines (Chapter 4) that select and classify patches from colorectal cancer WSIs, using attention-inspired sampling algorithms to support efficient estimation of tissue distribution, tumour stroma ratio (TSR) and tumour region of interest (ROI).

2. The novel Feedback Attention Ladder CNN (FAL-CNN, Chapter 5) the culmination of an evolving series of CNN classifiers exploring feedback attention architectures, resulting in significantly improved classification accuracy over the equivalent feedforward only model, over diverse data sets.

3. Visualisation plots for attention activations in the FAL-CNN (Chapter 6), revealing object location behaviours that support Explainable AI (XAI).

4. A Saccade Model (Chapter 7) which emulates human eye movements, resampling its input image to align the model's central sensitive region with its previous centre of attention. In this way the model tracks to informative image features, notably areas of *tumour* tissue.

5. Enhanced WSI pipelines (Chapter 8) using FAL-CNN and Saccade models, providing further insight into attention-guided behaviours and boosting TSR accuracy at pathologist-selected locations, with potential application in a diagnostic tool for rapid TSR measurement (Section 8.4.4).

9.2.1 WSI Pipelines

Three approaches to WSI sampling were compared in Chapter 4, in an exploration of methods to reduce the computational effort in extracting diagnostically useful information.

The Attention Heatmap Pipeline (AHP) used a CNN trained on thumbnail patches to obtain a heatmap of tumour probability, based on low-magnification image tiles. The heatmap was used to determine the local density at which high-magnification patches were sampled, thus guiding processing resources towards suspected tumour regions.

Tissue distribution plots of the WSI were obtained by using a CNN to classify the sampled patches. A further CNN distinguished tumour from normal epithelium, correcting false positives. The tumour ROI was estimated by applying clustering methods to the remaining tumour patches. TSR was calculated by counting tumour and stroma patches in a 'virtual biopsy' region of the ROI.

The Weighted Regular Sampling Pipeline (WRSP) refined the AHP method by sampling at regular intervals, more densely where tumour was detected. This reduced sampling biases which had affected TSR accuracy in regions where the low-magnification classifier predicted low tumour density. Figure 114 shows a typical sequence of sampling iterations, where the patch density is highest within the predicted ROI.

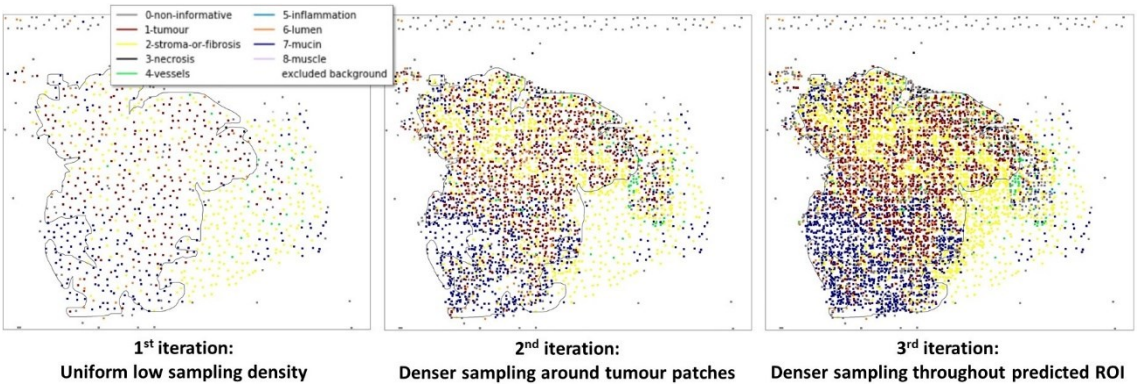


Figure 114: Typical patch sampling sequence in Weighted Regular Sampling Pipeline (Broad et al., 2022)

Finally, a 'brute force' Tile-by-Tile Pipeline, where contiguous patches were classified across the entire WSI, was evaluated for comparison with more efficient sampling methods.

The WRSP was published as *Attention-guided sampling for colorectal cancer analysis with digital pathology* (Broad et al., 2022), reporting processing speeds 3.3x to 6.3x faster than tile-by-tile processing. ROI was predicted with a mean F1 score of 86.6%, and TSR was estimated with an RMS error of 11.3% relative to expert annotations. CNN benchmarking results recommended the use of VGG19, with a classification accuracy of 79%.

The WRSP represents an efficient sampling algorithm for evaluating TSR and tumour outline in a cancer WSI, maintaining accuracy while reducing the total number of patch samples required, and hence the computational cost of extracting these outputs. Previous state-of-the art methods at the time of publication focused either on ROI estimation (Cruz-Roa et al., 2018) or overall TSR on a tile-by-tile basis (Zhao et al., 2020), while the WRSP algorithm is optimised to calculate both of these outputs simultaneously.

Subsequent experiments would aim to improve on the accuracy of the embedded CNNs at patch scale, through the addition of feedback attention mechanisms.

9.2.2 CNNs with Feedback Attention Mechanisms

The FAL-CNN (Chapter 5) is a novel, biologically inspired neural network for patch-scale image classification. The model uses a folded U-Net structure, where decoder outputs provide feedback activations, controlling spatial attention at multiple convolutional levels in the encoder path. Top-to-bottom feedback is combined with local feedback loops around convolutional groups for each spatial scale level. A Feature Embedding Store (FES) was used to aggregate feature embeddings from each forward pass over multiple feedback iterations, so that each cycle contributes to the output class prediction.

The FAL-CNN configuration yielded significant increases in classification accuracy, relative to the unadorned VGG19 architecture of the model's encoder backbone, for multiple datasets. With ImageNet-100 the accuracy increase was 2.39pp ($p<0.001$), and with 9-class CRC pathology patches, a 3.50pp ($p<0.001$) increase was measured. With the adversarial *uncertain-class-patches* subset, the feedback architecture increased classification accuracy relative to VGG19 by up to 12.26pp ($p<0.001$).

These results confirm that feedback attention improves discrimination in CNN classifiers, especially for images with mixed or ambiguous content. This was shown to be due to the model's ability to emphasise objects and structures relevant to the output class predictions, whilst inhibiting regions associated with unhelpful background content. While other attention-based CNNs exist, the FAL-CNN is particularly "brain-like", emulating feedforward and feedback neural pathways in the ventral stream in a manner that benefits object identification.

9.2.3 Feedback Attention Visualisations

In Chapter 6, attention activations at each feedback level in the FAL-CNN showed spatial correlation with salient image features. With ImageNet-100, the higher feedback layers highlighted larger features of the target object, such as a bird's head or a shark's dorsal fin. In lower layers, attention distributions mimicked finer details such as feathers.

Similarly, when using 9-class CRC pathology patches, feedback activations showed that the FAL-CNN attended to informative tissue features at multiple scale levels. In lower layers, the feedback contours were aligned with nuclei and other textural features. In higher layers, regions of tissue such as tumour and stroma were highlighted.

Given the increased accuracy of models using feedback attention, it is inferred that the most-attended regions in the feedback activations represent content that is most relevant to the class prediction. Visualisations of the feedback attention in our model are therefore potentially useful in bringing XAI to Digital Pathology, as a tool for object location and for highlighting salient tissue in patch-level images.

9.2.4 Saccade Model

When averaged over multiple patches, the attention distributions revealed a central focus, consistent with the annotating pathologist's behaviour in applying a class label to a single nominal pixel whilst examining surrounding tissue for context.

The saccade model (Chapter 7) exploited this tendency by resampling the input patch, to align the most strongly attended image features with the centre region where the model is most sensitive. This behaviour is analogous to foveal vision in humans and enabled the model to discover informative structures, even where these were outside the initial sampling region.

Comparison of accuracies when tracking to peak attention, against accuracies seen using random offsets, provided statistical proof that the attention regions represent image features salient to the classification result.

Expert relabelling of resampled input patches yielded class labels that agreed with the model's final class prediction in 76.9% of cases. For *tumour*, the agreement rate was 93.23%. This suggests that the FAL-CNN responds strongly to tumour, such that the saccade model actively locates tumour in the neighbourhood of the initial patch (Figure 115). This behaviour notably unites several goals of this research, combining brain-like feedback attention and saccade-based attention in a model capable of discerning and focusing on cancer tissue in a DP image.

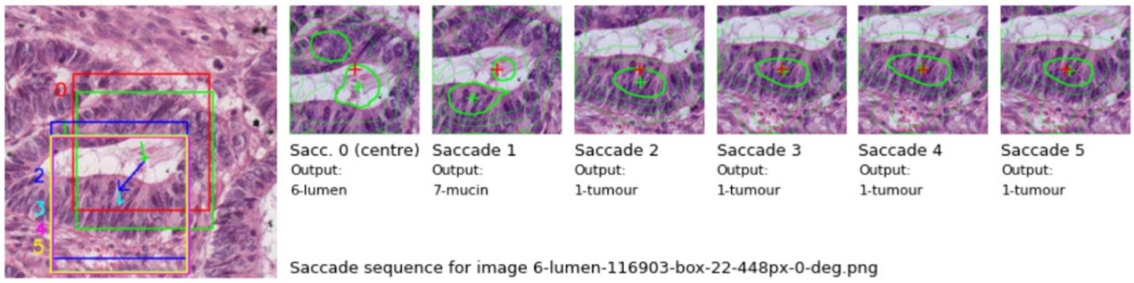


Figure 115: Example saccade sequence with lumen-labelled patch, converging on tumour cells

9.2.5 Attention-Inspired Models in WSI Pipeline

In Chapter 8, the new FAL-CNN and Saccade model were substituted into the WRSP. Use of the FAL-CNN was beneficial in calculations of TSR at pathologist-selected locations, reducing the error rate by approximately 10% relative to the VGG19 baseline and thereby confirming that the feedback-enhanced classifier can improve the performance of diagnostic image processing at a WSI scale.

The saccade model was known to fixate on tumour regions and therefore gave misleadingly low stroma rates and high TSR, potentially resulting in falsely optimistic predictions of disease progression. For this reason it is not indicated for use in the WRSP.

ROI estimation was mostly unaffected by the use of the FAL-CNN and Saccade models, suggesting that errors in identifying individual tumour patches are already being mitigated by the false positive correction, averaging and noise-reducing clustering behaviours built into the pipeline.

9.2.6 Generalisability

WSI pipelines

WSI processing pipelines in Chapter 4 were designed to predict ROI and TSR for images of tissue from colorectal resection surgery obtained in the QUASAR study. To generalise to other data sources – such as resection images from other parts of the body – would require new classifier CNNs trained to distinguish tumour features from normal cells of the new tissue type.

The pipelines in their present form would be less useful for analysing slides of *biopsy* samples. Commonly used needle biopsies differ in form from resection samples, having been taken from long, narrow cylindrical tissue cores. As such they are unlikely to encompass the complete cancer volume. Calculations of ROI would be less meaningful as a result, except for use cases where the expected cancer regions are much smaller than the core diameter.

FAL-CNN

The FAL-CNN classifier was tested on QUASAR 9-class patch data (Section 5.1) and generic images from ImageNet-100 (Section 5.4). Similar accuracy gains were observed despite the differences in data type and the relatively unbalanced class distribution of the pathology patches. This suggests that the FAL-CNN would generalise well to other image types of similar pixel size.

Experiments with the deliberately adversarial *uncertain-class-patches* set (Section 5.1.3) show that the FAL-CNN is particularly effective in distinguishing very heterogeneous or ambiguous patches, suggesting that the model would generalise well to images with a cluttered background, and to images of mixed object types.

Saccade model

The saccade model (Chapter 7) was initially developed using 9-class patch images. Expert relabelling was necessary before model accuracy could be assessed, because of the heterogeneous nature of the patches. The GT class label was only applicable at the very centre of the patch, and new labels were therefore required at the new centres of the post-saccade resampled patches.

Using ImageNet-100, this problem was not encountered as the GT class was generally applicable to the whole image, or the largest object therein. Then, the saccade action was found to improve accuracy relative to the 'pre-saccade' sample at the centre of the input image, showing that the model is tracking beneficially towards salient image features.

The successful evaluation of this model with two such diverse data sets suggests it would generalise well to other image types, with the qualification that relabelling may be required for more heterogeneous images.

9.3 Conclusions and Future Work

The improved TSR results at pathologist-selected locations suggest a practical application of FAL-CNN in a tool for rapidly measuring TSR at pathologist-selected locations. Such an AI-assisted application would classify large enough numbers of patches to provide an accurate TSR estimate, in workflows where the time required for fully manual sampling might otherwise make this measurement impractical. The TSR sampling tool could be integrated into WSI viewing workflows, where the user is first guided to peak tumour regions using tissue distribution plots generated by the WRSP. The resulting TSR values could be used to annotate the WSI, and from there be stored against the patient record. For clinical adoption, a trial involving multiple pathologists would be required to compare their TSR estimates with each other and with those generated by our model. These results would ideally then be correlated with survival data to confirm prognostic accuracy.

The WRSP currently uses hard-coded parameters for sampling intervals and clustering radii, and uses fixed algorithms for establishing TSR sampling points. It is anticipated that using machine learning to determine pipeline parameters will allow performance to be further optimised, using a loss term chosen for the required balance between ROI or TSR accuracy and processing speed.

The time-per-WSI of such a pipeline can potentially be reduced further through parallel processing. Patch classification tasks can be marshalled across multiple GPU nodes, while CPU multi-threading enables parallel patch extraction and pre-processing.

The CNN benchmarking results (Broad et al., 2022), obtained during the development of the pipelines, led to the adoption of VGG19 as the backbone for subsequent feedback models. Recent leaders in the ImageNet challenge use later models such as EfficientNetV2 (Tan and Le, 2021) with novel training regimes, such as Model Soups (Wortsman et al., 2022) or Meta Pseudo Labels (Pham et al., 2021). Proposed future projects would evaluate these model and training combinations in the WRSP, and develop an enhanced FAL-CNN with a feedforward backbone based on EfficientNetV2. Training and evaluation with the full 1,000-class ImageNet dataset would allow direct comparison with ImageNet challenge leaders.

Further testing of the generalisability of the current Feedback Attention Ladder would involve evaluating FAL-CNN with further datasets, from colorectal cancer and other diseases, in additional to other imaging modalities such as CT or ultrasound.

It is also recommended that the FAL-CNN is assessed as a feature extractor in other WSI processing methods, particularly in MIL algorithms that currently use the less accurate ResNet18 model.

The Saccade model currently interrogates a 448×448px input region. A further experiment is proposed, using larger sections of the WSI to evaluate the tumour-seeking behaviour when locating cancer cells further from the starting position. This suggests a further diagnostic application, where sections of WSI are automatically swept for candidate tumour locations for a pathologist to examine further. In a practical workflow, this would reduce the time taken to review a new WSI for disease, guiding the pathologist's attention to suspicious tissue.

In summary, the work in this thesis has demonstrated that biologically inspired attention mechanisms can contribute to image analysis in digital pathology. These methods were successfully applied in WSI processing pipelines and in patch-level analysis with the FAL-CNN and Saccade models. Answering the question posed in the title of this work, the author strongly believes that attention-inspired artificial intelligence *can* provide a diagnostic understanding of cancer imaging data, in a way that can benefit current clinical pathways.

References

Acs, B., Rantalainen, M. and Hartman, J. 2020. Artificial intelligence as the next step towards precision pathology. *Journal of Internal Medicine*. **288**(1).

Adair, A., Blakey, A. and Freeland, A. 2020. The Shapely User Manual — Shapely 1.7.1 documentation. *Shapely*. [Online]. [Accessed 27 May 2021]. Available from: https://shapely.readthedocs.io/en/stable/manual.html.

Altman, D.G. and Bland, J.M. 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*. **32**(3), pp.307–317.

BMJ n.d. 4. Statements of probability and confidence intervals | The BMJ. *The BMJ | The BMJ: leading general medical journal. Research. Education. Comment*. [Online]. [Accessed 27 June 2024]. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/4-statements-probability-and-confiden.

Breen, J., Allen, K., Zucker, K., Hall, G., Orsi, N.M. and Ravikumar, N. 2023. Efficient subtyping of ovarian cancer histopathology whole slide images using active sampling in multiple instance learning *In*: *Medical Imaging 2023: Digital and Computational Pathology* [Online]. SPIE, pp.248–258. [Accessed 26 September 2023]. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12471/1247110/Efficient-subtyping-of-ovarian-cancer-histopathology-whole-slide-images-using/10.1117/12.2653869.full.

Broad, A., de Kamps, M. and Wright, A. 2020. Predictive Visualisation of Colorectal Cancer Distribution using Artificial Intelligence in Digital Histopathology Imaging *In*: Leeds, UK: University of Leeds, p.8.

Broad, A., Wright, A.I., de Kamps, M. and Treanor, D. 2022. Attention-guided sampling for colorectal cancer analysis with digital pathology. *Journal of Pathology Informatics*. **13**, p.100110.

Brown, L.D., Cai, T.T. and DasGupta, A. 2001. Interval Estimation for a Binomial Proportion. *Statistical Science*. **16**(2), pp.101–133.

Cao, C., Huang, Y., Yang, Y., Wang, L., Wang, Z. and Tan, T. 2019. Feedback Convolutional Neural Network for Visual Localization and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **41**(7), pp.1627–1640.

Colby, C. and Goldberg, M. 1999. Space and Attention in Parietal Cortex. *Annual review of neuroscience*. **22**, pp.319–49.

Connor, C.E., Egeth, H.E. and Yantis, S. 2004. Visual Attention: Bottom-Up Versus Top-Down. *Current Biology*. **14**(19), pp.R850–R852.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A. and González, F. 2018. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PLOS ONE*. **13**(5), p.e0196828.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li and Li Fei-Fei 2009. ImageNet: A large-scale hierarchical image database *In*: *2009 IEEE Conference on Computer Vision and Pattern Recognition*., pp.248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [Accessed 6 May 2021]. Available from: https://openreview.net/forum?id=YicbFdNTTy.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M. and Consortium, and the C. 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. **318**(22), pp.2199–2210.

Encyclopaedia Britannica 2020. Occam's razor | Origin, Examples, & Facts. [Accessed 30 March 2021]. Available from: https://www.britannica.com/topic/Occams-razor.

Erman, L.D., Hayes-Roth, F., Lesser, V.R. and Reddy, D.R. 1980. The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Computing Surveys*. **12**(2), pp.213–253.

Ester, M., Kriegel, H.-P. and Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise *In*: AAAI Press, p.6.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*. **88**(2), pp.303–338.

Gadermayr, M. and Tschuchnig, M. 2022. Multiple Instance Learning for Digital Pathology: A Review on the State-of-the-Art, Limitations & Future Potential.

Gamma, E. 1995. *Design patterns : elements of reusable object-oriented software* [Online]. Reading, Mass. : Addison-Wesley. [Accessed 26 May 2021]. Available from: http://archive.org/details/designpatternsel00gamm.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A. and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231 [cs, q-bio, stat]*.

Godson, L., Alemi, N., Nsengimana, J., Cook, G.P., Clarke, E.L., Treanor, D., Bishop, D.T., Newton-Bishop, J. and Gooya, A. 2022. Weakly-supervised learning for image-based classification of primary melanomas into genomic immune subgroups.

Harmon, S., Sanford, T., Daneshvar, M., Brown, G.T., Yang, C., Mehralivand, S., Shih, J., Jacob, J., Valera, V., Agarwal, P., Choyke, P. and Turkbey, B. 2020. Multiresolution Application of Artificial Intelligence in Digital Pathology for Prediction of Positive Lymph Nodes From Primary Tumors in Bladder Cancer. *Journal of Urology*. **203**(Supplement 4), pp.e929–e929.

Harrison, D.G. 2012. *A Computational Dynamical Model of Human Visual Cortex for Visual Search and Feature-Based Attention*.[Online] PhD, University of Leeds. [Accessed 14 June 2021]. Available from: https://etheses.whiterose.ac.uk/4878/.

He, K., Zhang, X., Ren, S. and Sun, J. 2016. Deep Residual Learning for Image Recognition *In*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*., pp.770–778.

Hiramatsu, Y. and Hotta, K. 2020. Semantic Segmentation using Light Attention Mechanism: *In*: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging*

and Computer Graphics Theory and Applications [Online]. Valletta, Malta: SCITEPRESS - Science and Technology Publications, pp.622–625. [Accessed 11 August 2022]. Available from: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009347206220625.

Hochreiter, S. and Schmidhuber, J. 1997. Long Short-term Memory. *Neural computation*. **9**, pp.1735–80.

Hong, J., Park, B., Lee, M.J., Chung, C.-S., Cha, J. and Park, H. 2020. Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs. *Computer Methods and Programs in Biomedicine*. **183**, p.105065.

Hothorn, T. and Lausen, B. 2003. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*. **43**(2), pp.121–137.

Hutchins, G.G.A., Treanor, D., Wright, A., Handley, K., Magill, L., Tinkler-Hundal, E., Southward, K., Seymour, M., Kerr, D., Gray, R., Quirke, P. and QUASAR trial collaborators 2018. Intra-tumoural stromal morphometry predicts disease recurrence but not response to 5-fluorouracil – results from the QUASAR trial of colorectal cancer. *Histopathology*. **72**(3), pp.391–404.

Jiang, S., Li, J. and Hua, Z. 2022. Transformer with progressive sampling for medical cellular image segmentation. *Mathematical biosciences and engineering: MBE*. **19**(12), pp.12104–12126.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., Jansen, L., Reyes-Aldasoro, C.C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M. and Halama, N. 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*. **16**(1), p.e1002730.

Kong, B., Wang, X., Li, Z., Song, Q. and Zhang, S. 2017. Cancer Metastasis Detection via Spatially Structured Deep Network *In*: M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap and D. Shen, eds. *Information Processing in Medical Imaging*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp.236–248.

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D.L.K. and DiCarlo, J.J. 2019. Brain-like object recognition with high-performing shallow recurrent ANNs. *GitHub*. [Online]. [Accessed 5 October 2023]. Available from: https://github.com/dicarlolab/neurips2019/blob/master/slides.pdf.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D.L. and DiCarlo, J.J. 2019. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs *In*: *Advances in Neural Information Processing Systems* [Online]. Curran Associates, Inc. [Accessed 14 October 2021]. Available from: https://proceedings.neurips.cc/paper/2019/hash/7813d1590d28a7dd372ad54b5d29d033-Abstract.html.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D.L.K. and DiCarlo, J.J. 2018. *CORnet: Modeling the Neural Mechanisms of Core Object Recognition* [Online]. [Accessed 20 September 2021]. Available from: https://www.biorxiv.org/content/10.1101/408385v1.

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D. and Stumpe, M.C. 2019. Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. *Archives of Pathology & Laboratory Medicine*. **143**(7), pp.859–868.

Luo, X., Roads, B.D. and Love, B.C. 2021. The Costs and Benefits of Goal-Directed Attention in Deep Convolutional Neural Networks. *Computational Brain & Behavior*. **4**(2), pp.213–230.

Matplotlib 2012. Choosing Colormaps in Matplotlib. *Choosing Colormaps in Matplotlib*. [Online]. [Accessed 7 July 2023]. Available from: https://matplotlib.org/stable/tutorials/colors/colormaps.html.

Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K. 2014. Recurrent Models of Visual Attention. *Advances in neural information processing systems 3*., pp.2204--2212.

Motter, B.C. 1994. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*. **14**(4), pp.2178–2189.

Nir, G., Karimi, D., Goldenberg, S.L., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Thompson, D.J.S., Black, P.C. and Salcudean, S.E. 2019. Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. *JAMA Network Open*. **2**(3), pp.e190442–e190442.

NPIC 2023. Guide-to-Digital-Pathology-Vol.1.pdf. *The Leeds Guide to Digital Pathology*. [Online]. [Accessed 12 October 2023]. Available from: https://npic.ac.uk/wp-content/uploads/sites/71/2023/05/Guide-to-Digital-Pathology-Vol.1.pdf.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B. and Rueckert, D. 2018. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv:1804.03999 [cs]*.

OpenCV 2023. OpenCV: Finding contours in your image. [Accessed 10 July 2023]. Available from: https://docs.opencv.org/4.x/df/d0d/tutorial_find_contours.html.

van Pelt, G.W., Kjær-Frifeldt, S., van Krieken, J.H.J.M., Al Dieri, R., Morreau, H., Tollenaar, R.A.E.M., Sørensen, F.B. and Mesker, W.E. 2018. Scoring the tumor-stroma ratio in colon cancer: procedure and recommendations. *Virchows Archiv*. **473**(4), pp.405–412.

Pham, H., Dai, Z., Xie, Q. and Le, Q.V. 2021. Meta Pseudo Labels *In*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* [Online]., pp.11557–11568. [Accessed 15 November 2023]. Available from: https://openaccess.thecvf.com/content/CVPR2021/html/Pham_Meta_Pseudo_Labels_CVPR_2021_paper.html.

PyTorch 2021. torchvision.models — Torchvision master documentation. [Accessed 26 May 2021]. Available from: https://pytorch.org/vision/stable/models.html.

QUASAR Collaborative Group 2007. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *The Lancet*. **370**(9604), pp.2020–2029.

Rao, S. 2018. MITOS-RCNN: A Novel Approach to Mitotic Figure Detection in Breast Cancer Histopathology Images using Region Based Convolutional Neural Networks. *arXiv:1807.01788 [cs]*.

Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Rädsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., Farahani, K., Glocker, B., Heckmann-Nötzel, D., Isensee, F., Jannin, P., Kahn, C.E., Kleesiek, J., Kurc, T., Kozubek, M., Landman, B.A., Litjens, G., Maier-Hein, K., Menze, B., Müller, H., Petersen, J., Reyes, M., Rieke, N., Stieltjes, B., Summers, R.M., Tsaftaris, S.A., van Ginneken, B., Kopp-Schneider, A., Jäger, P. and Maier-Hein, L. 2021. Common Limitations of Image Processing Metrics: A Picture Story. *arXiv:2104.05642 [cs, eess]*.

Ribeiro, M.T., Singh, S. and Guestrin, C. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier *In*: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Online]. KDD '16. New York, NY, USA: Association for Computing Machinery, pp.1135–1144. [Accessed 4 November 2020]. Available from: https://doi.org/10.1145/2939672.2939778.

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. **115**(3), pp.211–252.

Sam, D.B. and Babu, R.V. 2018. Top-down feedback for crowd counting convolutional neural network *In*: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, pp.7323–7330.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B. and Rueckert, D. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*. **53**, pp.197–207.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D.L.K. and DiCarlo, J.J. 2020. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*., p.407007.

SciKit-Image 2023. SciKit Image. [Accessed 7 July 2023]. Available from: https://scikit-image.org/docs/stable/api/skimage.transform.html#skimage.transform.resize.

SciKit-Learn 2020. sklearn.cluster.DBSCAN — scikit-learn 0.24.2 documentation. *SciKit Learn*. [Online]. [Accessed 26 May 2021]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

SciPy community 2021. Statistical functions (scipy.stats) — SciPy v1.6.3 Reference Guide. *SciPy.org*. [Online]. [Accessed 3 June 2021]. Available from: https://docs.scipy.org/doc/scipy/reference/stats.html.

Shekar, A. 2021. ImageNet100. [Accessed 22 May 2023]. Available from: https://www.kaggle.com/datasets/ambityga/imagenet100.

Simonyan, K. and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*.

Stanford University 2020. Download the Object Bounding Boxes. *ImageNet*. [Online]. [Accessed 13 June 2023]. Available from: https://image-net.org/download-bboxes.php.

Tan, M. and Le, Q.V. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*.

Tan, M. and Le, Q.V. 2021. EfficientNetV2: Smaller Models and Faster Training. *arXiv:2104.00298 [cs]*.

Tomar, N.K., Jha, D., Riegler, M.A., Johansen, H.D., Johansen, D., Rittscher, J., Halvorsen, P. and Ali, S. 2022. FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation. *arXiv:2103.17235 [cs, eess]*.

Tsuda, H., Shibuya, E. and Hotta, K. 2020. Feedback Attention for Cell Image Segmentation. *arXiv:2008.06474 [cs]*.

University of Oxford 2023. VGG Image Annotator. *Information Engineering*. [Online]. [Accessed 18 August 2023]. Available from: https://www.robots.ox.ac.uk/~vgg/software/via/via_demo.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.

van der Velde, F. 2018. In Situ Representations and Access Consciousness in Neural Blackboard or Workspace Architectures. *Frontiers in Robotics and AI*. **5**(32).

van der Velde, F. and de Kamps, M. 2015. Combinatorial Structures and Processing in Neural Blackboard Architectures [Accessed 1 October 2020]. Available from: https://openreview.net/forum?id=Sk4jDwb_ZH.

van der Velde, F. and de Kamps, M. 2001. From Knowing What to Knowing Where: Modeling Object-Based Attention with Feedback Disinhibition of Activation. *Journal of Cognitive Neuroscience*. **13**(4), pp.479–491.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X. 2017. Residual Attention Network for Image Classification. *arXiv:1704.06904 [cs]*.

Wang, X., Chen, Y., Gao, Y., Zhang, H., Guan, Z., Dong, Z., Zheng, Y., Jiang, J., Yang, H., Wang, L., Huang, X., Ai, L., Yu, W., Li, H., Dong, C., Zhou, Z., Liu, X. and Yu, G. 2021. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. *Nature Communications*. **12**(1), p.1637.

West, N.P., Dattani, M., McShane, P., Hutchins, G., Grabsch, J., Mueller, W., Treanor, D., Quirke, P. and Grabsch, H. 2010. The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. *British Journal of Cancer*. **102**(10), pp.1519–1523.

Wiggins, G.A. 2020. Creativity, information, and consciousness: The information dynamics of thinking. *Physics of Life Reviews*. **34–35**, pp.1–39.

Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. 2018. CBAM: Convolutional Block Attention Module *In*: V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss, eds. *Computer Vision – ECCV 2018* [Online]. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp.3–19. [Accessed 16 September 2021]. Available from: http://link.springer.com/10.1007/978-3-030-01234-2_1.

Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S. and Schmidt, L. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time *In*: *Proceedings of the 39th International Conference on Machine Learning* [Online]. PMLR, pp.23965–23998. [Accessed 15 November 2023]. Available from: https://proceedings.mlr.press/v162/wortsman22a.html.

Wright, A. 2017. *Automated Analysis of Colorectal Cancer*. University of Leeds.

Wright, A.I., Dunn, C.M., Hale, M., Hutchins, G.G.A. and Treanor, D.E. 2021. The Effect of Quality Control on Accuracy of Digital Pathology Image Analysis. *IEEE Journal of Biomedical and Health Informatics*. **25**(2), pp.307–314.

Wright, A.I., Grabsch, H.I. and Treanor, D.E. 2015. RandomSpot: A web-based tool for systematic random sampling of virtual slides. *Journal of Pathology Informatics*. **6**.

Zhao, K., Li, Z., Yao, S., Wang, Y., Wu, X., Xu, Z., Wu, L., Huang, Y., Liang, C. and Liu, Z. 2020. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. *EBioMedicine*. **61**.

Appendix

1 Source Code

Source code and further documentation for the experiments described in this thesis are available at: https://github.com/scajb/histopathology-ai-msc-experiments. This is a private repository but access can be granted on request.

Code for the FAL-CNN and Saccade models, as presented in the author's article *Object-based Feedback Attention in Convolutional Neural Networks Improves Tumour Detection in Digital Pathology,* is also publicly available at https://github.com/scajb/feedback-attention-cnn.

PTH files containing model weights for 1, 2 and 3-iteration versions of the FAL-CNN, trained on ImageNet-100, are available at https://zenodo.org/doi/10.5281/zenodo.10361266.

The following sections describe the code used in correspondingly titled thesis chapters:

1.1 Data Extraction

1.1.1 Creation of *uncertain-class-patches* Dataset

The ***ClassProbabilityLogger*** class, developed during earlier work (Broad et al., 2020) to write the class probability vector to a CSV log file, was extended to copy patch images so a new *uncertain-class-patches* directory, for cases where the largest predicted class probabilities fell within a given percentage threshold of each other.

1.1.2 Creation of *tumour-stroma-groups* Dataset

Grouped parent patch *tumour-groups* and *stroma-groups* directories were created by copying patches of size 224×224px from the QUASAR 9-class directory. A Python function ***BuildBalancedDataDirectories.py*** was written to copy all image files from lists of subdirectories specified in a shell script, into output subdirectories according to Wright's parent class groupings (Wright, 2017).

The same grouping method was applied to the 9-class directory generated during offset patch extraction (Section 3.3.3). ***BuildBalancedDataDirectories.py*** was executed using *patches-offset\offset-x-56-y-56* as the source directory and *tumour-stroma-groups-offset-x-56-y-56* as the output directory.

1.1.3 Renaming ImageNet-100 Class Directories

For readability and for consistency with the naming convention used in QUASAR class subdirectories, the class subdirectories in ImageNet-100 were renamed using the ***ImageNetRenameTrainingDirectories.py*** script. Image class directories were named for the WordNet identifier (WNID) of each image category. The WNID was used to look up the English class description in the ***wnids_and_categories.csv*** file supplied with ImageNet-100. The class description was combined with an index number, based on the category's position in the CSV file, creating a human-readable directory name for each image class.

1.2 Characterising the WSI

1.2.1 Generation of Sampling Distributions

The ***SamplePatternGenerator*** Python class generates collections of ***Box*** objects, each describing the location and size of a single patch location within the WSI. These are determined by a low-resolution input heatmap distribution, such that higher probabilities in the heatmap result in a greater number of patches within a given parent tile.

1.2.2 Weighted Regular Sampling Pipeline

This pipeline is implemented in the ***__main__*** function in ***WeightedRegularPatchPathway.py***, which is called from the shell script ***weighted-regular-patches.sh.*** Pipeline parameters are passed to nested functions using the ***AttentionHeatmapRequest*** class.

The function ***generate_regular_classification_plots()*** iterates through all WSI files in the evaluation set associated with the specified experiment ID, calling ***generate_classification_plots_for_file()*** to output patch distributions and to estimate ROI polygons and TSR values for each WSI.

Aggregated accuracy statistics are generated by the ***AggregatedPatchStatistics*** class, which calculates mean and 95% confidence values for ROI and TSR accuracy calculations and writes these to a CSV file.

1.3 Feedback Attention

The FAL-CNN model was implemented in the ***UNetRecurrentConcatenatingHybridFeedbackCNN*** class, a subclass of ***UNetHybridFeedbackCNN***. The ***UNetBuilder*** was extended to supply encoder and decoder blocks as ***nn.Sequential***-based collections of convolutional and ReLU modules. The ***forward()*** function calls the encoder first with null feedback activations. The decoder is then used to derive feedback activations for use in a second encoder call.

1.4 Visualising Feedback Attention

Feedback activations were superimposed on input images in the Python class ***UNetOutputImageGenerator***. An interactive Jupyter notebook, ***FeedbackVisualisations.ipynb***, was developed to load and display the heatmap plots for a selected attention model and input patch.

Contour plots were generated using the OpenCV ***findContours()*** function (OpenCV, 2023). The contours were superimposed onto the input image using OpenCV ***drawContours()***. These operations were encapsulated in the ***ContourGenerator*** class.

For statistical analysis, the centre of mass of the $H \times W$ attention matrix was calculated using SciPy's ***ndimage.measurements.center_of_mass()*** function (SciPy community, 2021).

1.5 Saccade-like Behaviour with Feedback Attention Models

The saccade model algorithm was implemented in in the ***forward()*** method of the ***SaccadingFeedbackCNNContainer*** class.

In this class, the centre of attention (CoA) determining the sampling region for the next saccade cycle was calculated as the centre of mass (CoM) of the mean layer 28 feedback activations.

Subclass ***ContourSaccadingFeedbackCNNContainer*** overrides the ***get_image_centroid()*** method to calculate the CoA using the centroid of the largest 80% attention contour, again derived from mean layer 28 feedback activations, as a proxy for peak spatial attention.

A further model subclass, ***RandomSaccadingFeedbackCNNContainer,*** implemented a random walk-like behaviour, where the sampling location was offset by random X and Y amounts between 0 and 224px between saccade cycles.

To visualise the saccade sequences, a 5-saccade model was processed by the ***UNetOutputImageGenerator*** class. A new ***plot_saccades()*** method was developed to capture

the $224 \times 224px$ patch region sampled in each saccade, superimposed with feedback attention contours and a cross marking the centroid of the 80% contour.

Further code was added to plot the $448 \times 448px$ input region, with the sequence of patch locations superimposed as colour-coded $224 \times 224px$ boxes.

These output plots were loaded in Jupyter notebook ***SaccadeViewer.ipynb*** to create a combined saccade sequence plot showing the 5 sampling locations and associated attention regions.

2 Feedback Attention

2.1 Feedback Attention Ladder CNN (FAL-CNN)

2.1.1 Results

Table 21: Classification accuracies for FAL-CNN model with QUASAR 9-class patches

Feedback iterations	Classification accuracy %	95% Confidence Interval, % (N=30)	Difference from VGG19, pp	p-value
None: Baseline VGG19	79.37	78.52 to 80.21	-	-
0	80.74	80.16 to 81.32	1.60	< 0.001
1	82.86	82.18 to 83.55	3.47	< 0.001
2	82.99	82.42 to 83.55	3.38	< 0.001
3	82.85	82.22 to 83.49	3.50	< 0.001
4	82.87	82.37 to 83.37	3.43	< 0.001

Table 22: Classification accuracies for FAL-CNN models with uncertain-class-patches dataset

Feedback iterations	Mean accuracy %	95% Confidence Interval, % (N=30)	Difference from VGG19, pp	p-value
None: Baseline VGG19	54.82	54.00 to 55.63	-	-
0	61.82	60.98 to 62.66	7.00	< 0.001
1	66.78	66.23 to 67.32	11.96	< 0.001
2	65.65	65.02 to 66.28	10.84	< 0.001
3	67.08	66.51 to 67.65	12.26	< 0.001
4	64.99	64.16 to 65.82	10.17	< 0.001

2.2 FAL-CNN Performance with Offset Patches

2.2.1 Results

Table 23: Feedback model accuracies with offset input patches

Feedback iterations	Classification accuracy %	95% Confidence Interval	Difference from VGG19, pp
None: Baseline VGG19	56.88	56.43 to 57.34	-
1	57.32	56.94 to 57.70	0.44
2	57.04	56.44 to 57.64	0.16
3	57.23	56.88 to 57.59	0.35

Table 24: Feedback model performance when trained and evaluated with offset-patches dataset

Feedback iterations	Classification accuracy %	$\pm 1\ SE$ range	Difference from VGG19, pp	Difference from centre-trained model, pp
None: Baseline VGG19	82.80	82.27 to 83.33	-	3.43
1	86.20	85.94 to 86.46	3.40	3.34
2	86.07	85.86 to 86.28	3.27	3.08
3	86.07	85.94 to 86.20	3.26	3.22

2.3 FAL-CNN Performance with *tumour-stroma-groups* Patches

2.3.1 Results

Table 25: Classification accuracies for FAL-CNN model with tumour-stroma-groups dataset

Feedback layers	Feedback iterations	Training epochs	Classification accuracy %	$\pm 1\ SE$ range	Difference from offset trained VGG19, pp
None: Baseline VGG19		50	94.34	94.18 to 94.50	-
0,5,10,19,28	1	200	94.77	94.55 to 94.98	0.42
0,5,10,19,28	2	200	94.90	94.65 to 95.15	0.55
0,5,10,19,28	3	200	94.81	94.55 to 95.08	0.47

Table 26: Classification accuracies for FAL-CNN model with tumour-stroma-groups-12000 dataset

Feedback layers	Feedback iterations	Training epochs	Classification accuracy %	$\pm 1\ SE$ range	Difference from VGG19, pp
None: Baseline VGG19		50	93.48	93.22 to 93.75	93.48
0,5,10,19,28	1	200	94.16	93.89 to 94.42	0.67
0,5,10,19,28	2	200	93.85	93.57 to 94.13	0.37
0,5,10,19,28	3	200	93.93	93.62 to 94.23	0.44

Table 27: Classification accuracies for FAL-CNN model with offset tumour-stroma-groups dataset

Feedback layers	Feedback iterations	Training epochs	Classification accuracy %	$\pm 1\ SE$ range	Difference from VGG19, pp
None: Baseline VGG19		50	93.18	92.87 to 93.49	-
0,5,10,19,28	1	200	93.70	93.47 to 93.93	0.52
0,5,10,19,28	2	200	93.67	93.50 to 93.83	0.48
0,5,10,19,28	3	200	93.62	93.31 to 93.92	0.43

2.4 FAL-CNN Performance with ImageNet-100

2.4.1 Results

Table 28: Classification accuracies for FAL-CNN model with ImageNet-100

Feedback layers	Feedback iterations	Training epochs	Classification accuracy %	$\pm 1\ SE$ range	Difference from VGG19, pp
None: Baseline VGG19		100	83.97	83.87 to 84.06	-
0,5,10,19,28	1	250	88.04	87.94 to 88.13	4.07
0,5,10,19,28	2	250	88.09	88.00 to 88.19	4.13
0,5,10,19,28	3	250	88.01	87.93 to 88.09	4.04

Table 29: Classification accuracies for FAL-CNN model with ImageNet-100 Test dataset

Feedback iterations	Classification accuracy %	95% Confidence Interval, % (N=30)	Difference from VGG19, pp	p-value
None: Baseline VGG19	80.89	80.55 to 81.23	-	-
1	83.28	82.88 to 83.69	2.39	< 0.001
2	83.08	82.65 to 83.51	2.19	< 0.001
3	82.86	82.45 to 83.27	1.97	< 0.001

3 Statistical Analysis of Attention regions

3.1 Attention Distributions for FAL-CNN with QUASAR Patches

Figure 116 shows the locations of the centre of mass of the mean attention activations in the FAL-CNN model for each of 100 patches x 9 classes, superimposed on the $224 \times 224px$ input patch area. These results are grouped by feedback layer and feedback iteration.

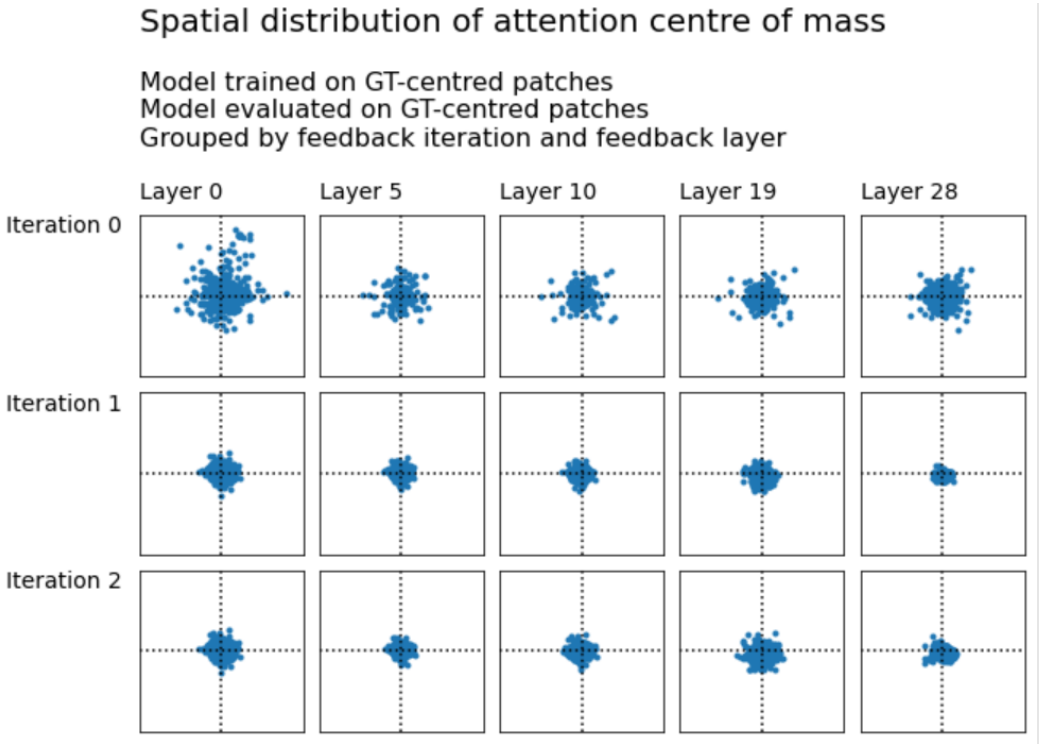


Figure 116: Spatial distributions of attention centre of mass, grouped by layer and feedback iteration

Figure 117 shows the similarly grouped locations of the centroids of the 80% attention contours derived from the same feedback attention activations.

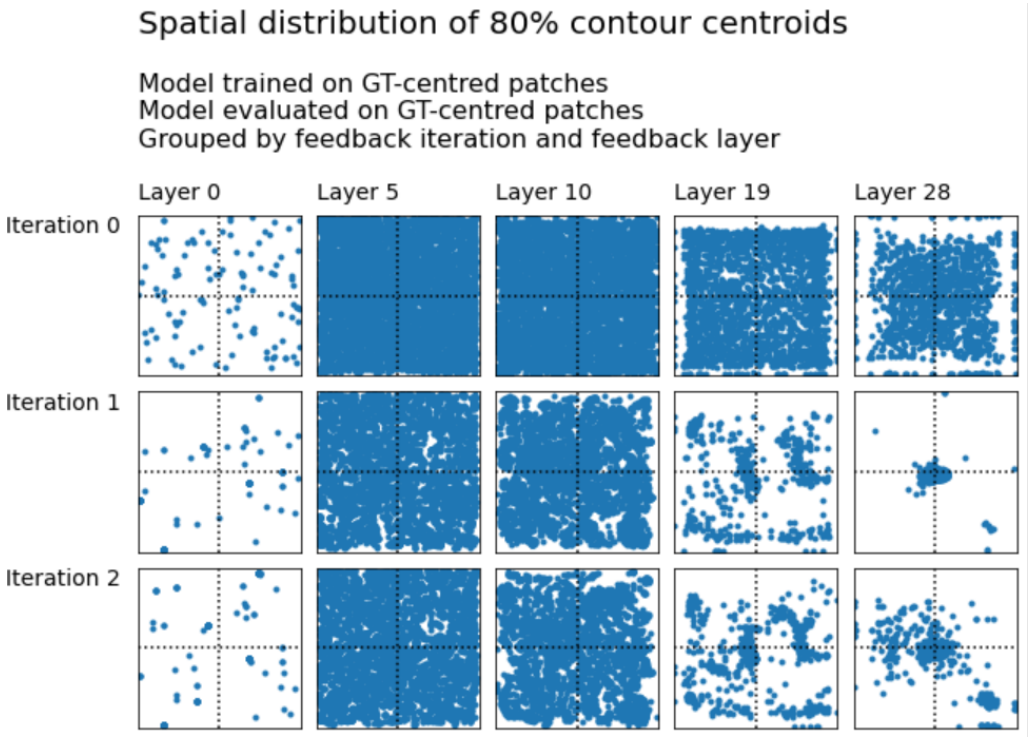


Figure 117: Spatial distributions of 80% attention contour centroids, grouped by layer and feedback iteration

Figure 118 shows the frequency distribution of the effective area, or total value, of the mean feedback activations for each layer and feedback iteration. In each sub-plot, the distribution is shown relative to a maximum of 50176, the theoretical maximum possible sum of a 224×224 array where each pixel has the maximum value of 1.

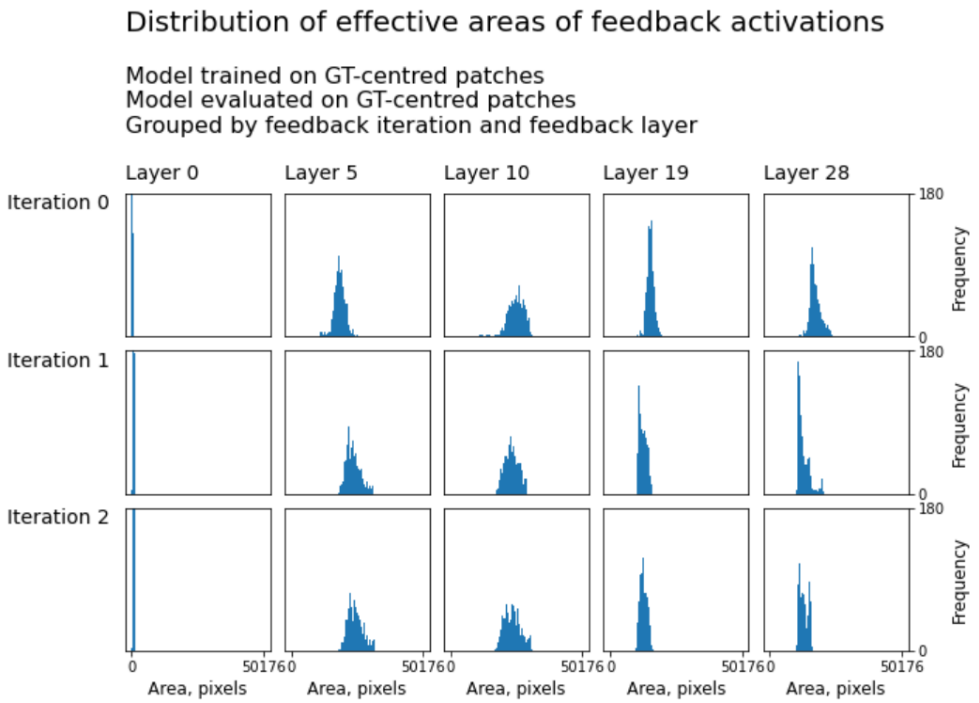


Figure 118: Frequency distributions of effective areas of feedback activations, grouped by layer and feedback iteration

A similar plot in Figure 119 represents the distributions of the area within the 80% attention contours, for each feedback layer and iteration.

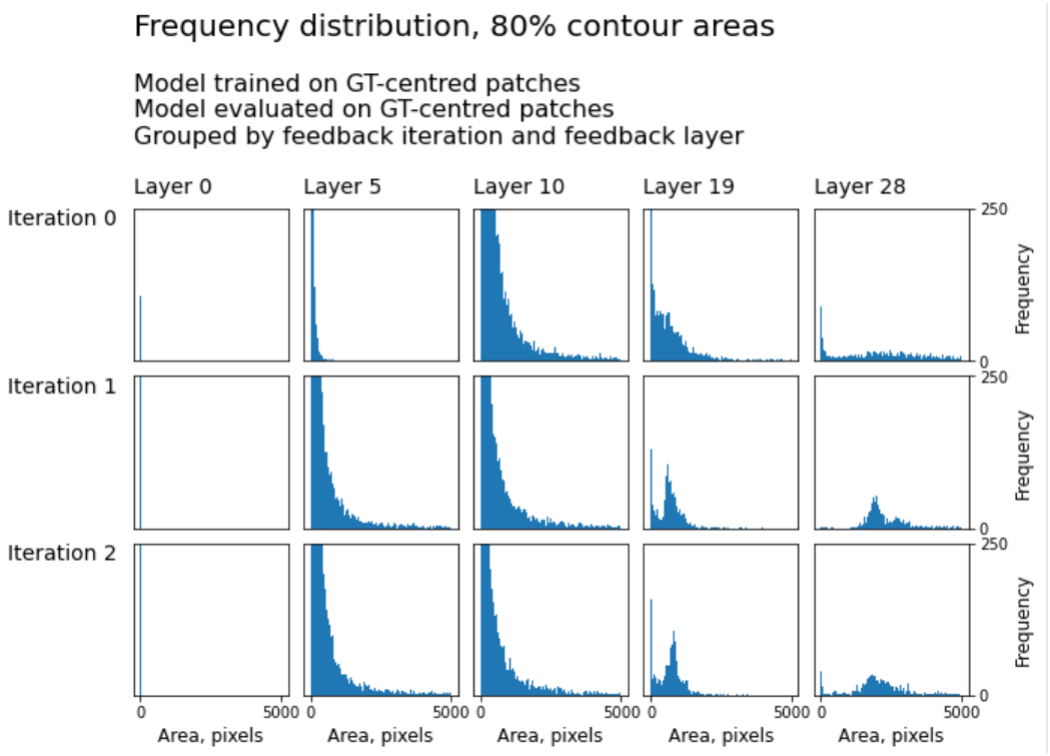


Figure 119: Frequency distributions of areas of 80% attention contours, grouped by layer and feedback iteration

Figure 120 shows the spatial distributions, relative to the input patch area, of the attention centre of mass in the feedback activations of the 1-iteration model variant, grouped by layer and image class.

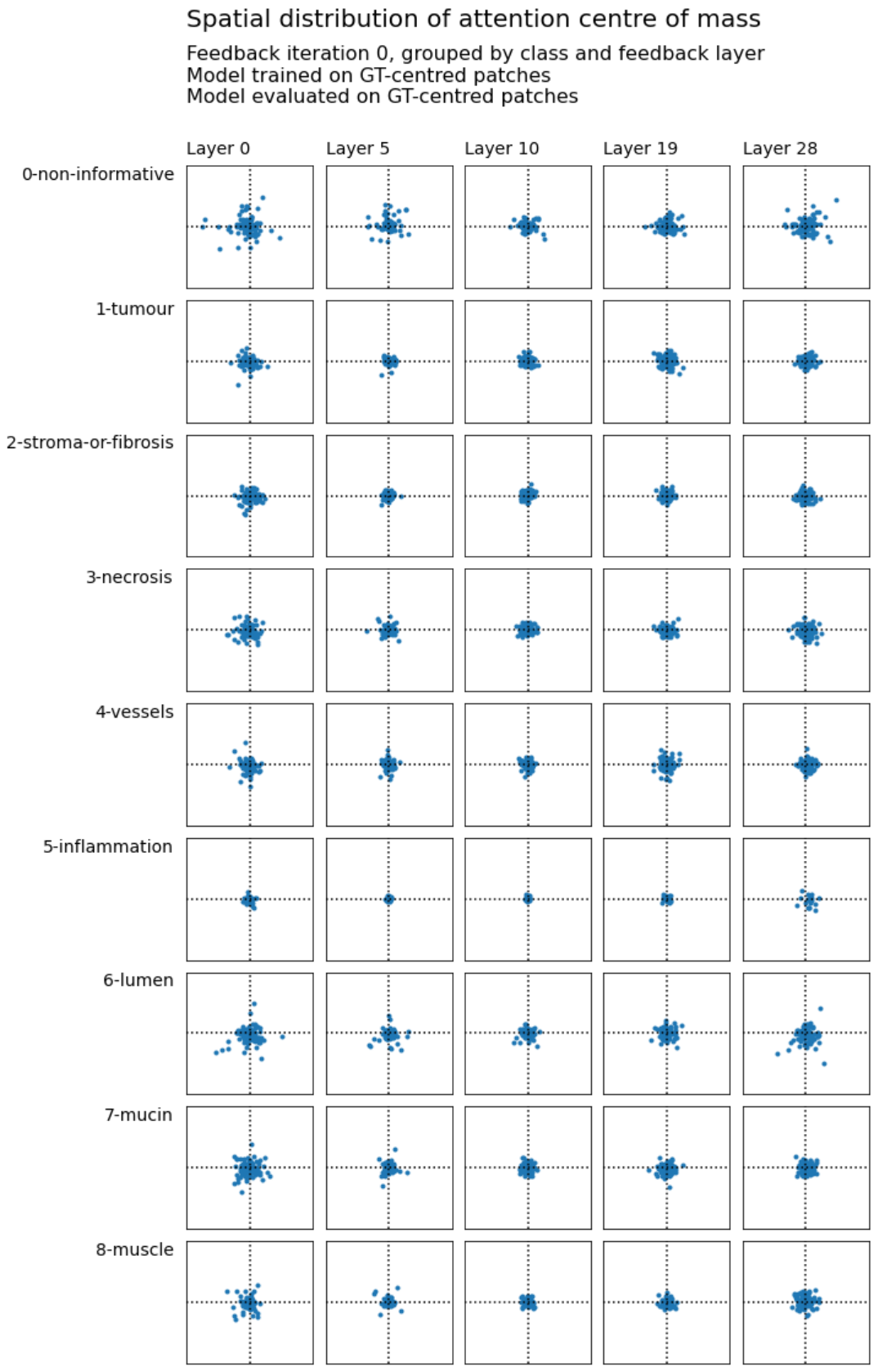


Figure 120: Spatial distributions of attention centre of mass, grouped by layer and class

Figure 121 shows the similarly grouped spatial distribution of the centroids of the 80% attention contours, from the same model's feedback activations.

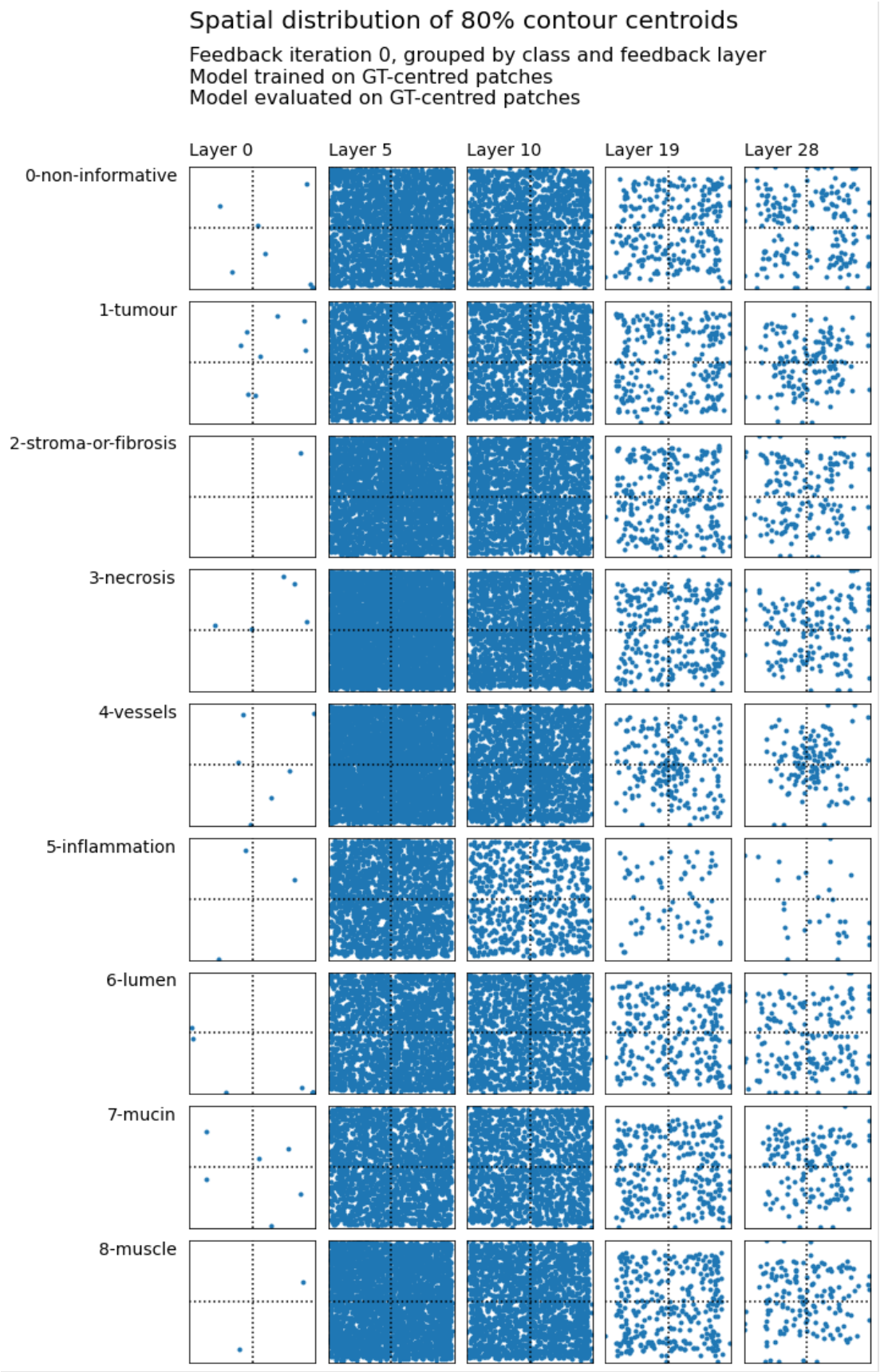
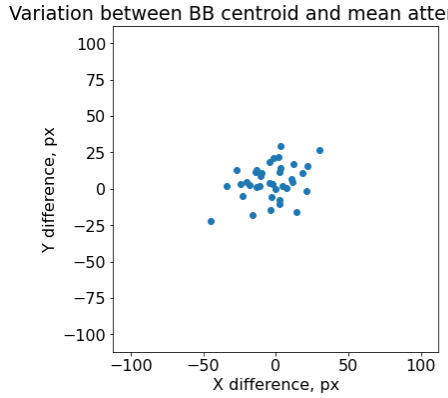
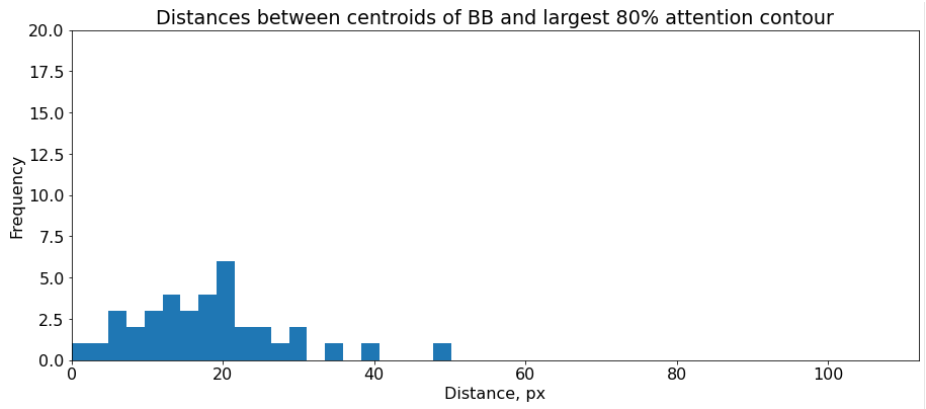
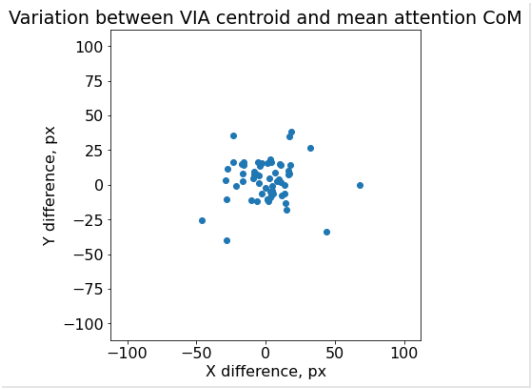


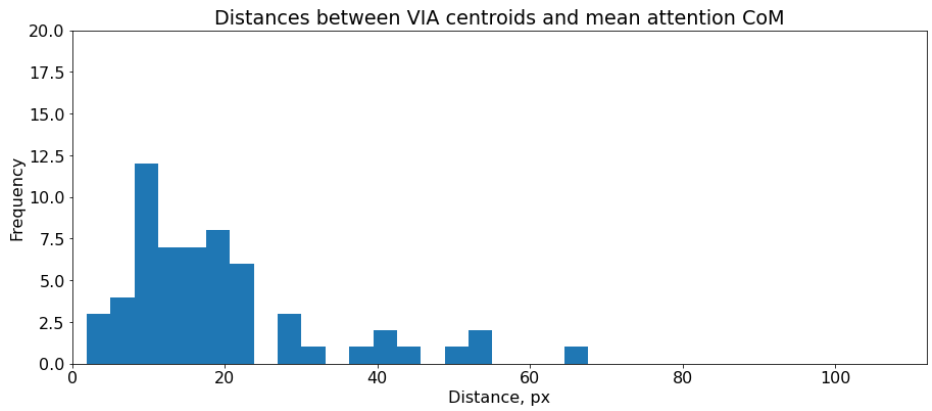
Figure 121: Spatial distributions of 80% attention contour centroids, grouped by layer and class

3.2 Attention Distributions for FAL-CNN vs ImageNet Annotated Regions









4 Saccade-like Behaviour with Feedback Attention Model

4.1 Results

4.1.1 Evaluation of Saccade Models

Table 30: Classification accuracies for saccade models with QUASAR 9-class patches

Number of saccades	Centre of attention method	Classification accuracy %	95% Confidence Interval	Difference from non-saccading model, pp
0	-	82.86	82.16 to 83.56	-
1	CoM	72.55	71.79 to 73.31	-10.32
2	CoM	64.65	63.48 to 65.83	-18.21
5	CoM	54.00	52.69 to 55.32	-28.86
10	CoM	51.23	49.66 to 52.80	-31.63
1	80% contour	54.22	52.98 to 55.45	-28.65
2	80% contour	51.07	49.62 to 52.52	-31.79
5	80% contour	48.95	47.09 to 50.82	-33.91
10	80% contour	48.41	46.58 to 50.25	-34.45
1	Random	47.12	46.12 to 48.12	-35.74
5	Random	45.34	44.40 to 46.28	-37.52
10	Random	44.77	43.93 to 45.60	-38.10

Table 31: Classification accuracies for saccade models with ImageNet-100

Number of saccades	Centre of attention method	Classification accuracy %	95% Confidence Interval	Difference from non-saccading model, pp
0	-	77.53	77.26 to 77.80	-
1	CoM	79.20	79.07 to 79.33	1.67
2	CoM	79.95	79.82 to 80.07	2.42
5	CoM	80.56	80.40 to 80.73	3.03
10	CoM	80.71	80.55 to 80.86	3.17
1	80% contour	79.55	79.47 to 79.63	2.02
2	80% contour	79.70	79.56 to 79.85	2.17
5	80% contour	79.80	79.68 to 79.92	2.27
10	80% contour	79.77	79.64 to 79.91	2.24
1	Random	63.27	63.16 to 63.38	-14.26
5	Random	60.15	59.82 to 60.47	-17.38
10	Random	58.82	58.55 to 59.09	-18.72

4.1.2 Pathologist Reclassification of Post-Saccade QUASAR Patches

Table 32: 9-class classification accuracies with/without saccade process, including agreement with expert-relabelled post-saccade patches

Model	Input data	Reference label source	Mean agreement rate %	±1 SE range / binomial probability CI	Difference from VGG19, pp
VGG19 baseline	QUASAR 9-class patches	QUASAR 9-class patches	79.37	78.94 to 79.79 (SE)	-
FAL-CNN 1 iteration	QUASAR 9-class patches	QUASAR 9-class patches	82.86	82.51 to 83.21 (SE)	3.50
Saccade model	QUASAR 9-class patches	QUASAR 9-class patches	48.63	47.70 to 49.57 (BCI)	-30.73
Saccade model (Expt A)	QUASAR classes 1,2,3,6	Expert-relabelled, resampled patches	86.73	83.44 to 90.03 (BCI)	7.37
Saccade model (Expt B)	QUASAR 9-class patches	Expert-relabelled, resampled patches	78.25	74.21 to 82.29 (BCI)	-1.12

Table 33: Tumour-stroma-groups classification accuracies with/without saccade process, including agreement with expert-relabelled post-saccade patches

Model	Input data	Reference label source	Mean agreement rate %	±1 SE range / binomial probability CI	Difference from VGG19, pp
VGG19 baseline	QUASAR 9-class patches	QUASAR 9-class patches	94.34	94.18 to 94.50 (SE)	-
FAL-CNN 1 iteration	QUASAR 9-class patches	QUASAR 9-class patches	94.77	94.55 to 94.98 (SE)	0.42
Saccade model	QUASAR 9-class patches	QUASAR 9-class patches	32.57	28.54 to 36.61 (BCI)	-61.77
Saccade model	QUASAR classes 1,2,3,6	Expert-relabelled, resampled patches	95.09	92.99 to 97.19 (BCI)	0.74
Saccade model	QUASAR 9-class patches	Expert-relabelled, resampled patches	86.25	82.88 to 89.62 (BCI)	-8.09

Table 34: Per-class breakdown of agreement rates between saccade model output and relabelled final sample location – Experiment A (4 input classes)

Expert-assigned label for post-saccade patch image	Number of patch images	Total in agreement with saccade model output class	Mean agreement rate	Binomial probability confidence interval
All	407	353	86.73	83.44 to 90.03
0-non-informative	0	0	0.00	-
1-tumour	263	252	95.82	93.40 to 98.24
2-stroma-or-fibrosis	39	27	69.23	54.75 to 83.72
3-necrosis	66	46	69.70	58.61 to 80.78
4-vessels	8	8	100.00	100.00 to 100.00
5-inflammation	3	0	0.00	0.00 to 0.00
6-lumen	26	19	73.08	56.03 to 90.13
7-mucin	1	1	100.00	100.00 to 100.00
8-muscle	0	0	0.00	-
Unable to score	1	0	0.00	-

Table 35: Per-class breakdown of agreement rates between saccade model output and relabelled final sample location – Experiment B (9 input classes)

Expert-assigned label for post-saccade patch image	Number of patch images	Total in agreement with saccade model output class	Mean agreement rate	Binomial probability confidence interval
All	400	313	76.90	72.81 to 81.00
0-non-informative	28	4	14.29	1.32 to 27.25
1-tumour	266	248	93.23	90.21 to 96.25
2-stroma-or-fibrosis	43	24	55.81	40.97 to 70.66
3-necrosis	16	8	50.00	25.50 to 74.50
4-vessels	14	11	78.57	57.08 to 100.00
5-inflammation	8	3	37.50	3.95 to 71.05
6-lumen	16	8	50.00	25.50 to 74.50
7-mucin	9	7	77.78	50.62 to 100.00
8-muscle	0	0	0.00	-

5 Feedback Attention Model Performance in WSI Pipeline

5.1 Results

5.1.1 TSR Estimation

Table 36: TSR error rates in WSI pipeline, for combinations of feedforward and feedback CNN classifiers

Main CNN	*TP / FP CNN*	*TSR CNN*	*TSR % err at GT loc*	*TSR at GT loc confidence interval*	*TSR % err at max tumour density*	*TSR % err at max tumour density CI*
VGG	VGG	None	7.57	7.23 to 7.92	21.64	20.06 to 23.22
VGG	FAL1	None	7.62	7.31 to 7.94	21.34	20.08 to 22.61
VGG	FAL2	None	7.60	7.32 to 7.87	21.42	20.40 to 22.44
FAL1	VGG	None	7.12	6.76 to 7.47	22.36	20.93 to 23.79
FAL1	FAL1	None	7.12	6.76 to 7.47	20.87	19.93 to 21.81
FAL2	VGG	None	6.84	6.43 to 7.24	22.05	20.73 to 23.37
FAL2	FAL2	None	6.87	6.46 to 7.27	21.72	20.56 to 22.89
FAL3	VGG	None	7.26	6.87 to 7.64	21.44	20.45 to 22.42
FAL2	VGG	FAL2 tumour-stroma-groups	13.24	12.91 to 13.57	18.86	17.68 to 20.04
Saccade FAL1	VGG	None	38.26	37.00 to 39.53	43.81	41.86 to 45.75

5.1.2 Tumour ROI Estimation

Table 37: Rates of agreement between WSI pipeline and GT ROI annotations, for combinations of feedforward and feedback CNN classifiers

Main CNN	*TP / FP CNN*	*TSR CNN if not main*	*IoU %*	*IoU CI*	*F1 %*	*F1 CI*
VGG	VGG	None	75.57	75.24 to 75.90	85.04	84.76 to 85.32
VGG	FAL1	None	75.36	74.94 to 75.79	84.81	84.40 to 85.21
VGG	FAL2	None	75.44	75.11 to 75.77	84.87	84.58 to 85.16
FAL1	VGG	None	75.50	75.31 to 75.68	84.93	84.75 to 85.11
FAL1	FAL1	None	75.83	75.65 to 76.01	85.24	85.07 to 85.41
FAL2	VGG	None	75.81	75.45 to 76.16	85.20	84.92 to 85.48
FAL2	FAL2	None	75.78	75.46 to 76.10	85.11	84.81 to 85.40
FAL3	VGG	None	75.42	75.00 to 75.83	84.86	84.48 to 85.25
FAL2	VGG	FAL2 tumour-stroma-groups	75.30	75.03 to 75.58	84.78	84.50 to 85.06
Saccade FAL1	VGG	None	75.64	75.03 to 76.24	85.05	84.56 to 85.54

5.1.3 WSI Processing Time

Table 38: Mean pipeline processing time per WSI, for combinations of feedforward and feedback CNN classifiers

Main CNN	TP / FP CNN	TSR CNN	Time per WSI, sec	Time per WSI, confidence interval
VGG	VGG	None	237.24	233.48 to 241.00
VGG	FAL1	None	239.46	232.99 to 245.94
VGG	FAL2	None	276.67	269.32 to 284.02
FAL1	VGG	None	284.37	280.02 to 288.72
FAL1	FAL1	None	303.15	299.12 to 307.18
FAL2	VGG	None	329.55	325.02 to 334.09
FAL2	FAL2	None	353.66	349.26 to 358.07
FAL3	VGG	None	368.56	366.02 to 371.11
FAL2	VGG	FAL2 tumour-stroma-groups	358.48	350.41 to 366.55
Saccade FAL1	VGG	None	1206.60	1188.97 to 1224.22

-