BDIAP report: Deep learning algorithm for colorectal cancer detection

Kevin Wang, 2019

I undertook this project with the support from BDIAP summer studentship. It was originally thought to be 6 weeks long but instead extended to an 8-week project.

Background

Colorectal cancer (CRC) is the second largest cause of cancer mortality in the UK. With early diagnosis and appropriate treatment, 57% of patients can survive for 10+ years. Both diagnoses and treatment decisions are made manually, typically on glass slides. Manual analysis of tissues requires visual inspection of highly complex cellular structures, which is time consuming, subjective and prone to error. Deep Learning (DL) has the potential to automate this task, improving on speed, objectivity and accuracy. Current research at the University of Leeds uses DL algorithms to classify CRC tissue in order to automatically predict response to therapy. However, it requires cancer tissue to be annotated by a pathologist prior to analysis which has the same disadvantages. This project aims to develop a DL algorithm to detect CRC on digital slides, as a pre-processing step for downstream image analysis.

Method

11,977 images from tissues banks of National Center for Tumor diseases (NCT) and University Medical Center Mannheim, Heidelberg University (UMM) were used to train and test a modified version of the Resnet18 convolutional neural network. Each of the images was manually annotated as one of three classes: tumour (colorectal cancer and stomach cancer epithelial tissue); stroma and muscle; adipose and mucus. The model was trained and tested using a 5-fold cross validation methodology.
The parameters used to train the model is shown in table 1.

Table 1: Parameters used in training algorithm

Parameters used	
Input Image Size	224 x 224
Initial learning rate	10^{-5}
L2 Regularisation rate	10^{-4}
Hot Layers	20
Learning rate factor	2
Max Epoch	10
Pixel Range Shear	5

The model was then evaluated on 750 colorectal images from a CRC clinical trial dataset. Each whole slide image was divided into 224x224 blocks and a prediction was made on each block. A binary mask for each class on the whole slide image is then generated: A block will be assigned 1 for one of the three classes with the highest prediction score generated by the algorithm and 0 for the other two classes. The tumour binary mask is then compared with existing pathologist binary mask for tumour. Similarity indexes including Jaccard similarity, Dice similarity and tumour detection accuracy are used to assess the accuracy of the model at picking up tumour when

compared with pathologists' annotation. The formulas used to calculate each of the indexes are shown in figure 1. Tumour to stromal ratio (TSR) was used to assess the clinical relevance of the model. The formula used to calculate TSR is shown in figure 2. All images were then classified into high TSR (TSR >0.5) and low TSR (TSR<0.5) and compared with pathologists' classification. A confusion matrix is then used to evaluate the accuracy.

- Jaccard similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Dice similarity

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

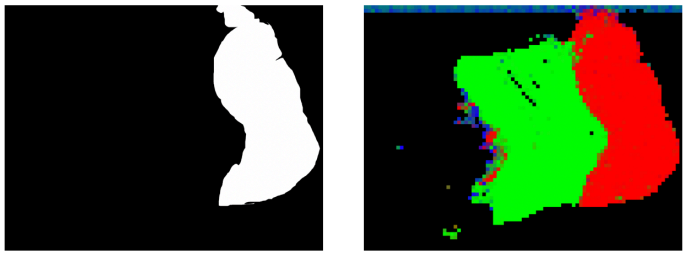
- Confusion Matrix

$$Accuracy = \frac{TP + TN}{All\ classes}$$

True Positive	False Positive
False Negative	True Negative

Figure 1: Formula to calculate similarity indexes

$$TSR = \frac{Pixel\ count\ of\ Tumour}{Pixel\ count\ of\ (Tumour + stroma)}$$



Pathologists' annotation of whole tumour

Epithelial tumour

stroma

Figure 2: Formula to calculate TSR

Results

The model attained excellent validation accuracy of with an Area Under the Curve (AUC) of lowest 0.989 as shown in figure 3. A confusion matrix showing the validation accuracy is shown in figure 4.

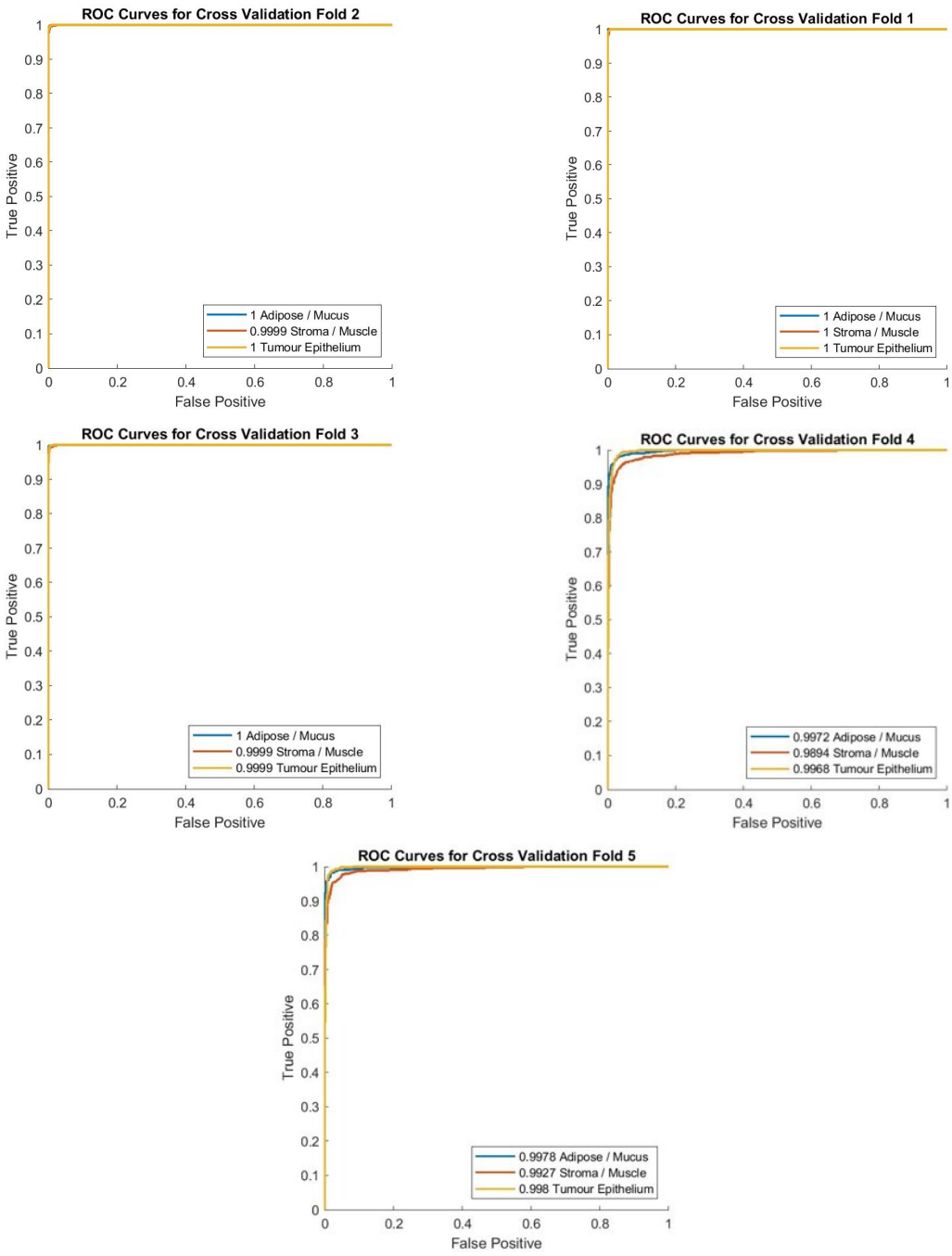


Figure 3: ROC curve for 5-fold cross validation

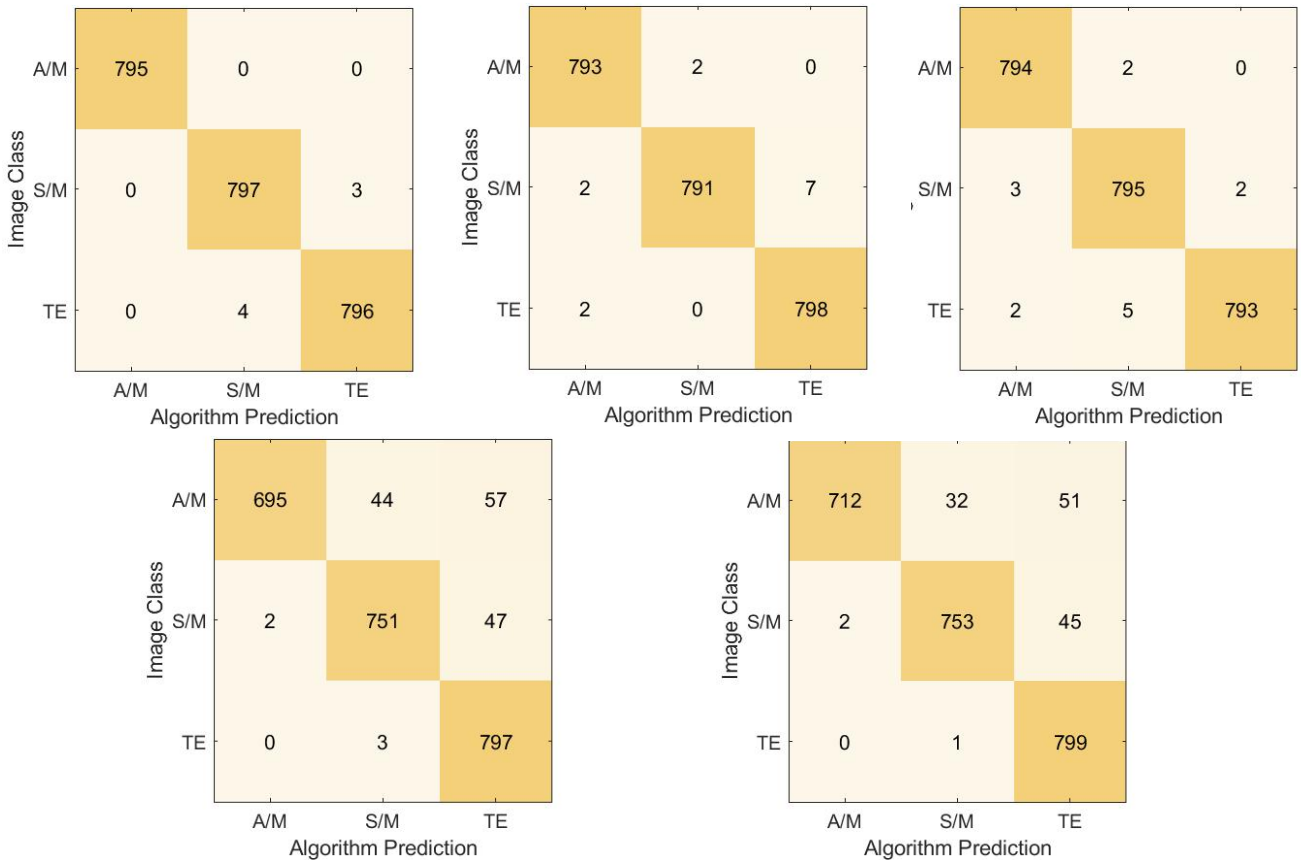


Figure 4: Confusion matrix for 5-fold validation accuracy

The evaluation results for similarity indexes are shown in table 2 with a specific example shown in figure 5. The model achieved the best results when evaluated by tumour detection accuracy with a median accuracy of 0.87.

Table 2: Similarity index comparison results

	Jaccard Similarity	Dice Similarity	Tumour detection accuracy
Median	0.4278	0.5992	0.86817
Maximum	0.8542	0.9213	0.97752
Minimum	0	0	0.51871
Standard deviation	0.2390	0.2637	0.0796

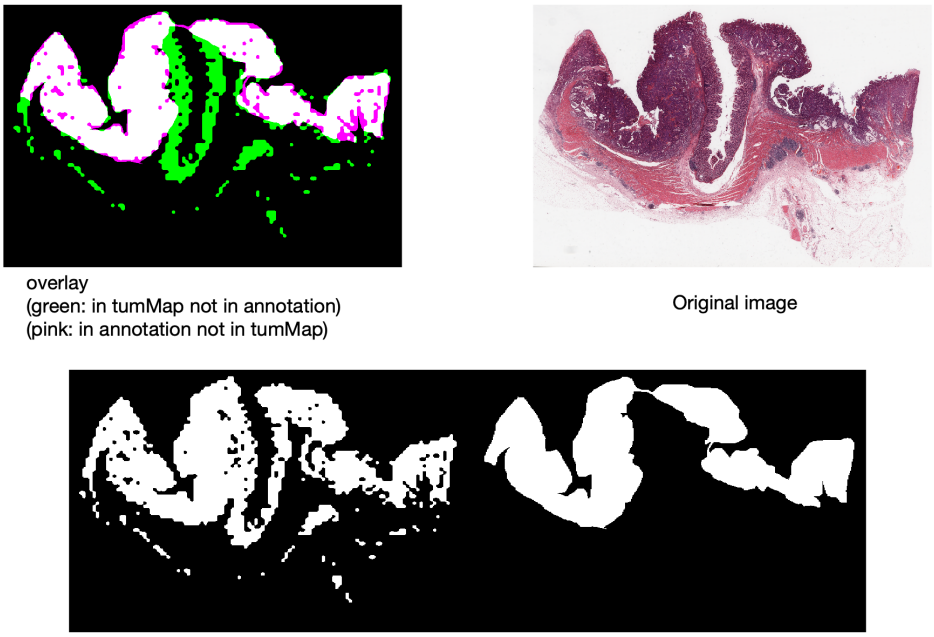


Figure 5: Comparison between ground truth and AI prediction (115965.svs)

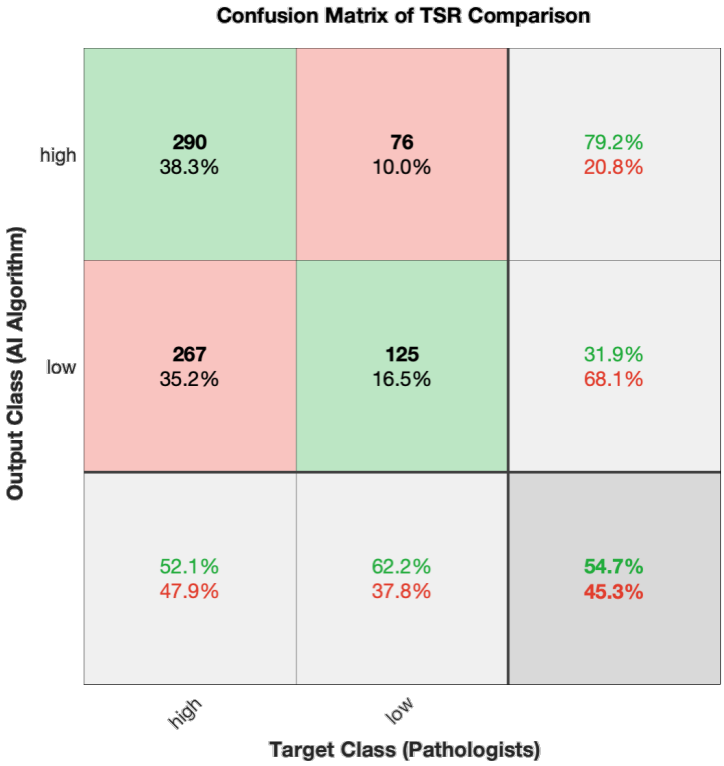


Figure 6: TSR evaluation results

For algorithm's prediction of TSR, it achieved an overall accuracy of 54.7%. However, it is worth noting that the algorithm has a 79.2% specificity at picking out high TSR images.

Limitation

When looking at images that scores 0 for similarity indexes, it was found that the algorithm misclassified epithelial tumours that look like stroma as shown in figure 6. The clinical trial dataset also suffers from problems of poor staining and poor image qualities which might have contributed to the low accuracy when evaluated by jaccard and dice similarity.

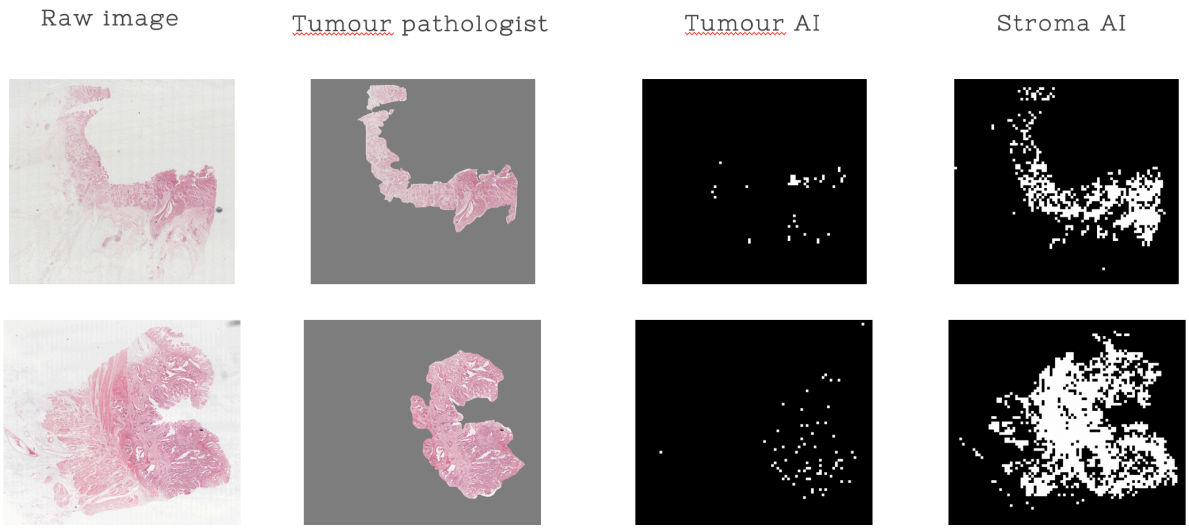


Figure 7: Limitation of algorithm at picking at stroma like epithelial tumours

To develop this algorithm further, we would use smaller image input size than 224x224 to Reduce the possibility of images containing multiple tissue types. We would also use better quality images to train and evaluate the model. Additionally, different AI model may be explored such as SegNet and more classes may be included to improve on model accuracy.

Conclusion

These preliminary findings show that when DL algorithms are trained on datasets that are free from variation caused by routine lab practices, the resulting models are more rigid than previous research has indicated. The validation results highlighted an issue with the evaluation methods used in this experiment, in that the ground truth annotations did not contain normal epithelial tissue and the training dataset did not distinguish between them. Either changing the training data to incorporate normal and cancer epithelium or using regions of interest that contained both types of epithelial tissue will make the evaluation more appropriate and should also improve the results.
Robust automatic detection of colorectal cancer will allow for higher throughput of patient samples, allowing pathologists to make more comprehensive treatment decisions based on consistent and reliable measurements.

Acknowledgement

I would like to extend my gratitude towards BDIAP for the kind support. I would also like to thank my supervisors, Alexander Wright and Darren Treanor for all the training and advice without which this project wouldn't have been possible.