Canterbury Medical Research Foundation

Final report for Grant in Aid GIA2019-2

Grant recipient

Associate Professor Ramakrishnan Mukundan
University of Canterbury
Computer Science and Software Engineering

mukundan@canterbury.ac.nz
033692201
Private Bag 4800, Christchurch. 8014

Grant details

GRANT TYPE	Grant in Aid	FUNDING ROUND	2019 Grant In Aid
GRANT REFERENCE	GIA2019-2	GRANT AMOUNT	$5,000

Final report

1. Scientific Assessing Committee report

Automatic Nuclei Segmentation and Tumour Cellularity Assessment in Breast Cancer Histopathological Slides

Ramakrishnan Mukundan (Principal Investigator)
Huyuan Shangguan (Research Assistant)
Department of Computer Science and Software Engineering
University of Canterbury, Christchurch, New Zealand.

Summary

The primary goals of the project were (i) to develop image analysis algorithms to process whole slide images of breast cancer tissue specimens to automatically segment nuclei regions (ii) to classify nuclei as belonging to benign or malignant epithelial cells (iii) to use the result of the classification algorithm to identify tumour regions, (iv) to use statistical measures on tumour regions containing malignant cell nuclei to obtain the cellularity category, and (v) to perform experimental validation of the developed methods. The CMRF Grant in Aid was helpful in conducting a detailed study and development of image segmentation and classification algorithms for automated tumour cellularity assessment. All project goals mentioned above were achieved. The project was carried out in stages as described in the following sections.

1. *Data sets*

For this project, we used the following two datasets:

BreastPathQ Dataset [1]: The dataset was provided by the Sunnybrook Health Sciences Centre, Toronto, as part of the online challenge on tumour cellularity estimation. It comprises 96 whole slide images (WSI) at 20x magnification (0.5 m/pixel). WSIs were extracted from 64 patients with residual invasive breast cancer. The training set consists of 2394 image patches extracted from the WSIs, each with size 512x512 pixels with the tumour cellularity scored by one expert pathologist. The validation set consists of 185 image patches. The dataset also contains additional useful data for distinguishing between malignant and epithelial nuclei. 153 regions of interest (ROI) containing annotations of lymphocytes, malignant epithelial and normal epithelial cell nuclei, as well as coordinates of nuclei centroids are provided. The widths of ROIs range from 411 to 1008 pixels, and heights range from 373 to 775 pixels. This is the main dataset used in our project for developing a neural network for nuclei classification and for evaluating the neural network's accuracy.

Kaggle Dataset [2]: This dataset is particularly useful for nuclei segmentation. It contains 1095 images of size 256x256 pixels containing nuclei regions and their corresponding masks of cell nucleus present in the images. Several image intensity and shape parameters (such as colour, cell type, magnification) vary across dataset. This is useful for testing the developed algorithms under a variety of conditions. We used this dataset for developing a neural network for nuclei segmentation. The segmented data is used in the classification algorithm based on the BreastPathQ dataset.

MICCAI-2018 Dataset [3]: This dataset contains whole slide images of H&E images at 40x magnification from the Cancer Genome Atlas [4]. Additionally the dataset provides around 30 images with around 22,000 nuclear boundary annotations. This dataset is useful for developing machine learning algorithms for tumour assessment. We used this dataset only to perform some

initial validation experiments comparing the results with the Kaggle dataset.

2. *Methodology*

We used a Convolutional Neural Network (CNN) known as U-Net [5] and trained it using the data and segmented nuclei masks in the Kaggle dataset. The trained U-Net is then used to predict the masks of images in BreastPathQ dataset. The output masks are then matched using closest centroid points with cell annotations given in the BreastPathQ dataset. Thus every segmented nuclei in the input images is assigned a cell type label. This data derived from the BreastPathQ dataset is used to train a neural network similar CNN for nuclei classification into lymphocytes, benign and epithelial nuclei. Finally the cellularity of the BreastPathQ images is calculated based on the area covered by the identified malignant figures.

The processing pipeline can be broadly divided into two stages: nuclei segmentation stage and cellularity computation stage.

3. *Nuclei Segmentation*

The first stage of the processing pipeline is a U-Net network that performs the image segmentation task, identifying nuclei regions in the input image and outputting an image mask. UNet treats image segmentation task as pixel-level classification task. It adopts a downsampling-upsampling style network structure to classify the pixels in an image into different classes so that all pixels are classified into background or foreground. To keep all the information of the input images, UNet usually adopts skip layers. There are many varieties of UNet architecture. The UNet architecture used in this project is U-Net+ResNet152. The loss function used is the sum of dice coefficient and cross-entropy loss. Plots of the accuracy and loss function for 40 epochs are given in Figure 1. After the network is trained using the Kaggle dataset, it is used to predict the nuclei masks of images in the BreastPathQ dataset. The centres of the mask regions are calculated and matched with the nearest nuclei centroids given in the BreastPathQ dataset. From this matching, we also obtain a label of the nuclei from the dataset that tells us whether a nucleus belongs to a lymphocyte, benign or malignant epithelial cell. A set of sample results obtained from this computational stage is shown in Figure 2.

4. *Nuclei Classification and Cellularity Estimation*

The previous stage yielded a training set containing images of the BreastPathQ dataset with corresponding nuclei classification masks and labels as shown in Figure 2. Nuclei belonging to a lymphocyte, benign or malignant epithelial cell are represent by red, green and blue colours respectively. These labelled images are used as ground truth for the training of the U-Net nuclei classification network. The training process of nuclei classification network is the same as the training of nuclei segmentation method. After the training is completed, we use the trained network to predict the label and the mask of the images in “train” images in BreastPathQ folder. For a given input test image, the individual nuclei figures are classified to the three known classes and the malignant epithelial candidates are identified.

5. *Summary of Results*

The test data consisted of 2394 patches of size 512x512 pixels. The cellularity of any given image patch is then calculated as the area covered by malignant figures divided by the area of the image patch. The image patches are classified into four categories: 0 ~ 25 as normal, 26 ~ 50 as low, 51 ~ 75 as medium, and 76 ~ 100 as high. Figure 3 gives the scatter plot and the confusion matrix showing the agreement between ground truth and predicted values of tumour cellularity. For the Normal, Low, Medium and High cellularity categories, the precision values were found to be 0.95, 0.35, 0.44, 0.88 respectively. The corresponding recall values were 0.67, 0.57, 0.68, 0.62. The interclass correlation coefficient was 0.763 and overall accuracy was 65%.

6. Current Development and Future Work

We are currently working on the tuning of hyperparameters for the neural network, adjusting the learning rate and the optimizer with the goal of further improving the accuracy of the estimated cellularity values. We are also in the process of creating an extended dataset for nuclei segmentation using data augmentation methods such as rotation, cropping and resizing. The method we have adopted is novel in that it uses a U-Net to obtain simultaneous nuclei segmentation,

and labelling useful for the cellularity classification algorithm. We plan to publish our work at the Medical Image Understanding and Applications conference (MIUA-2020) to be held in Oxford, UK.

References

1. SPIE. BreastPathQ Cancer Cellularity Challenge. https://breastpathq.grand-challenge.org/. Retrieved 10.07.2019.
2. Kaggle. Data Science Bowl 2018. https://www.kaggle.com/c/data-science-bowl-2018/data . Retrieved 12.07.2019.
3. MoNBuSeg. MICCAI-2018. https://monuseg.grand-challenge.org/Data/. Retrieved 12.07.2019
4. TCGA. Cancer Genome Atlas Program. https://portal.gdc.cancer.gov/. Retrieved 15.08.2019.
5. O. Ronneberger, P. Fischer, T. Box. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (2015): 234–241.

3. Attachments

View an attachment by double clicking the icon to the left of the file name. Icons are not displayed and attachments are not accessible when this PDF is viewed in a web browser; you must open it in PDF reader software.

Fig1.png
172.3 KiB

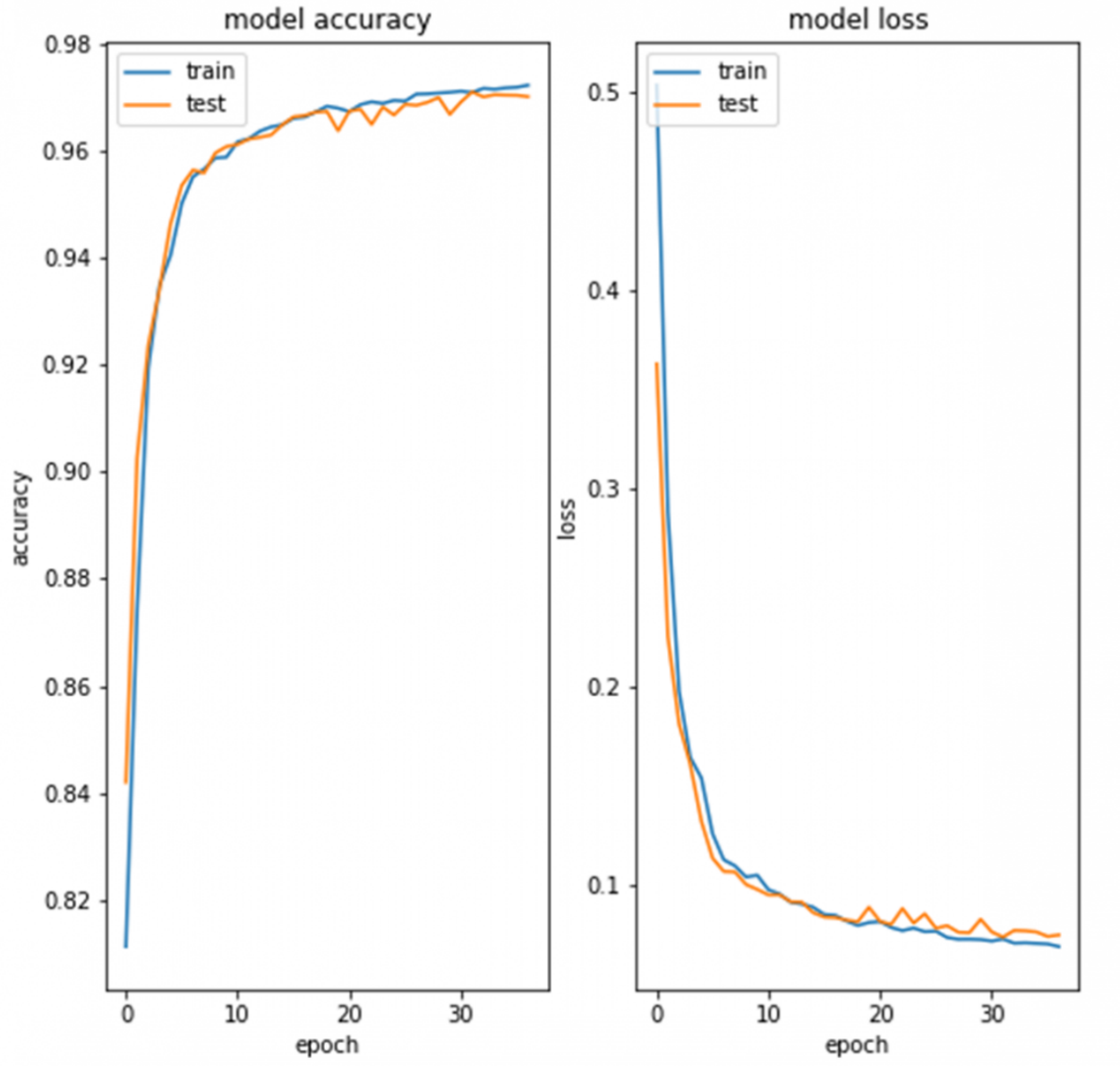


Figure 1. Plots of accuracy and loss function for the U-Net model developed for nuclei segmentation.

Fig2.png
1011.4 KiB

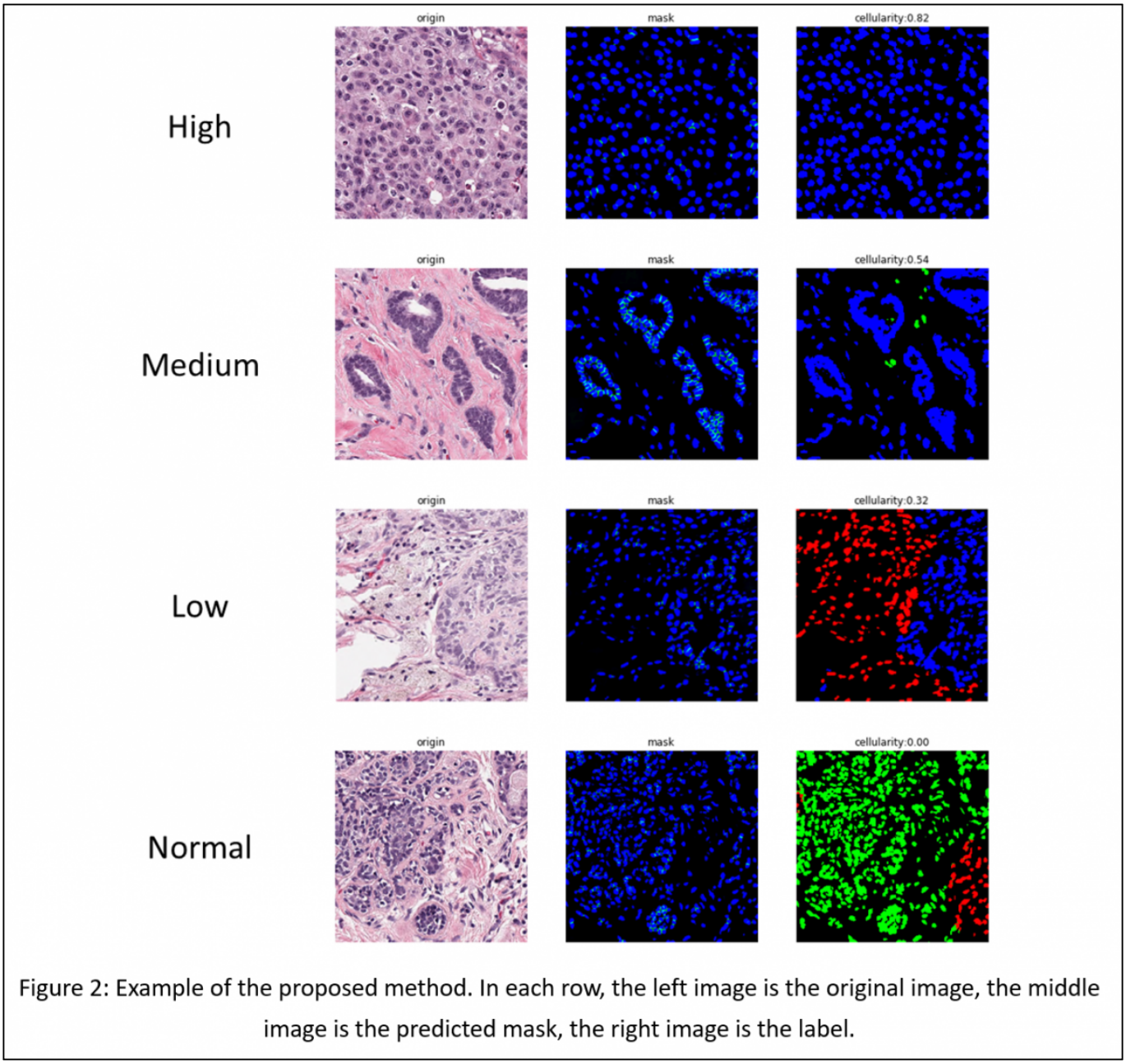
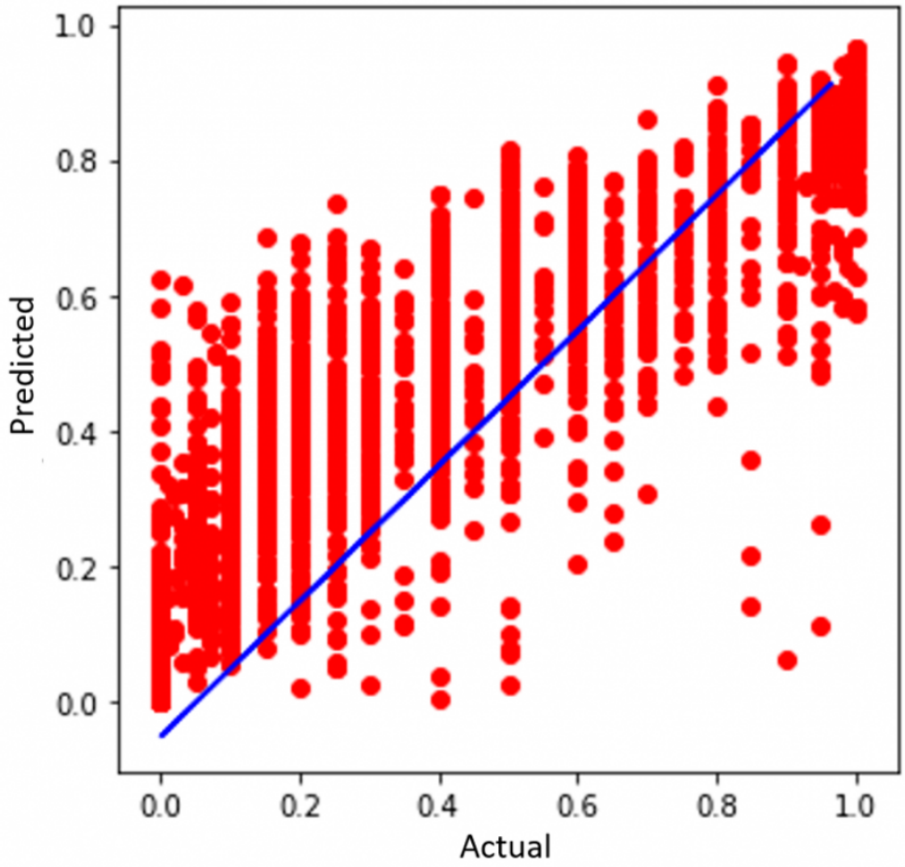


Figure 2: Example of the proposed method. In each row, the left image is the original image, the middle image is the predicted mask, the right image is the label.

Fig3.png
274.2 KiB



		Predicted			
		Normal	Low	Medium	High
Actual	Normal	850	358	52	0
	Low	32	234	141	0
	Medium	8	75	254	38
	High	4	6	125	217

Figure 3. Scatter plot and confusion matrix showing the agreement between the actual and predicted cellularity scores for 2394 image patches.

4. Feedback

Publication

Date

11/11/2019