Supplementary Material B

Related Work on Automated IHC Scoring

Automated image analysis is observed as a solution [1,2] to overcome the inter- and intra-observer variations found in conventional assessment of tissue slides. Hence, the automated scoring of routine H&E and IHC stained slides has received huge interest in recent years. In literature, several classical machine learning approaches [3–5] have been presented but recently deep learning based approaches have been profoundly employed for H&E and IHC histology image analysis [6,7].

In literature, a wide range of handcrafted features was proposed for IHC scoring algorithms [4,5]. For instance, Choudhury *et al*. [8] proposed an averaged threshold measure (ATM) for scoring of digitized images of IHC stained tissue microarrays. A set of arbitrary chosen thresholds was selected, whereby an optimal threshold using the ATM is used for calculating the percentage of stained area. The proposed ATM statistic presented as a generalization of the HSCORE [9] statistic for scoring IHC slides. Reyes-Aldasoro *et al*. [10] presented an alternative approach for automated segmentation of microvessels in IHC tumor slides. For segmentation, distinguishing hues of stained vascular endothelial nuclei and tissue regions were explored to extract the seeds for a 'region-growing' model. Their post-processing of segmented microvessels from CD31 immunostaining contained three steps, closing morphological objects from tumour margins, combining isolated objects, and splitting objects into individual vessels with having multiple lumina. Although the thresholding approaches perform well on a specific dataset, they are likely to fare not as well on an unseen dataset as distinctive hues can be significantly varying. A potential reason of such variation lies in staining process, as the histology slides normally stained at different occasions with inconsistent concentrations often exhibit large variations in colour and appearance. Such differences in slide preparation make the colour and morphological appearance of tissue components more unpredictable.

Kuse *et al*. [11] used local isotropic phase symmetry measure as a significant feature for beta cell detection and lymphocytes. By calculating the peak of median phase energy after stain normalization but due to heterogeneous appearance and often-clumped structure makes nuclei segmentation a non-trivial task. Khan *et al*. [5] used stain quantization for the scoring of Estrogen Receptor (ER) and Progesterone Receptor (PR) by determining the amount of chromatin material and protein content from IHC stained WSIs. Ali *et al*. [12] used astronomical algorithms for the scoring of ER on IHC stained images of breast cancer. However, in this contest the classical machine learning approaches have been outperformed by deep learning approaches. Most of the published algorithms are based on different approaches with different dataset whereas this contest provides a platform where participants can develop and validate the performance of their algorithms on same dataset.

Description of Automated Methods

The concise description of automated methods employed by top-ranked teams are described below.

Team Indus

In this approach, a deep convolutional neural network (CNN) was employed for predicting the Her2 score whereas for estimating the percentage of complete membrane staining, a set of handcrafted morphological features were extracted from H&E and IHC stained slides.

Pre-processing: The patches with average edge strength lies higher then certain threshold were selected for training CNN.

Her2 Score Prediction: The presented CNN architecture contains five convolutional layers, one concatenation layer with following two fully connected and one classification layer. After each convolution and fully connected layer, a ReLu activation was performed whereas for classification layer a softmax activation was placed. After convolution layers a concatenation layer was positioned. The concatenation layer combines the activation maps from the convolution layers and the average control tissue intensity for the corresponding WSI from which the patches were originated. The weights for training CNN were initialized using H&E normal initializations [13] and updated using mini batch gradient descent (learning rate = 0.00015, weight decay = 10^{-6}, Nesterov momentum = 0.95, batch size = 32). The CNN was trained over 41K patches generated each of size 224x224 from 52 training WSIs for 65 epochs.

During testing, the trained network assigned a score to each patch of a WSI and to aggregate the patch scores into a single Her2 score following criteria was proposed. Let n_0, n_1, n_2 and n_3 be the number of patches scored as 0, 1+, 2+ and 3+ respectively and N be the total number of patches generated from a WSI.

```
If $n_3/N > 0.08$:
    predict 3+
else if $n_2/N > 0.4$:
    predict 2+
else if $n_1/N > 0.14$:
    predict 1+
else:
    predict 0
```

Percentage of Complete Membrane Staining (PCMS): To estimate the PCMS, first tumor regions were identified by extracting the morphological features from tumor and normal regions of H&E images.

After performing stain normalisation [14], the hematoxylin channel was extracted to segment the nuclei using Otsu thresholding. Further, nuclei contours were fit around each individual structure and filtered on basis of area and eccentricity. This resulted in tumor identification regions by detecting the tumour nuclei based on their roundness and size. In order to estimate the extent of membrane staining, the morphological features were extracted from an IHC image. In addition, a contagious chicken-wire pattern was observed for complete membrane stained regions whereas other tissue components result in a fragmented/broken-up skeleton. Further, by filling holes in the chicken-wire skeleton and by measuring similarity with the original binary image the extent of membrane staining was estimated.

The PCMS is estimated by calculating the ratio between extent of membrane staining and tumor identification regions as given below.

$$PCMS = \frac{\text{extent of membrane staining}}{\text{tumor identification regions}}\ x\ 100$$

MUCS

In this submission, the well-known neural networks Alexnet [15]and GoogLeNet [16] were adapted by adjusting the layer specific parameters, such as kernel size, stride, and padding. There were three submissions from the MUCS team with two submissions using Alexnet (MUCS-1 and MUCS-2) and one using GoogLeNet (MUCS-3).

Training: The training dataset was obtained by hand-picking the regions of interest from 52 training IHC images that were considered to contain the most representative samples from each class. The regions were selected from the low resolution (0.625×) and mapped to the highest resolution (40×) whereupon each region was divided into 128 x 128 pixel patches.

The MUCS-1 trained network had four output classes with corresponding Her2 scores from 0 to 3+. MUCS-2 and MUCS-3 had an additional output class for the background. The background class contained the regions with texture having only a weak appearance of nuclei (without blueish or brownish colour). The training dataset for MUCS-2 was extended by data augmentation (rotation and mirroring) and by adding the hand-picked regions from test images (without knowing the classification of the slide it originated from). The total patches for MUCS-1, MUCS-2 and MUCS-3 were 29000, 319000 and 33500, respectively. The training images were divided between actual training data (75%) and validation data (25%). For all three submissions, the base learning rate was set to 0.001, and the learning rate was dropped every one-third of the maximum iterations by a factor of 10 (γ=0.1). The mean pixel value was subtracted from the training dataset.

Classification: For testing, the common regions from H&E and IHC were selected at a low resolution and those regions were mapped to maximum resolution to generate the patches for testing. Further, adaptive thresholding was applied to each patch, with an offset of 10, to produce a binary image. If the proportion of ones in the binary image was smaller than a factor of 0.9, then patch was classified with the trained neural network model, otherwise the patch was marked as background and therefore did not require classification. The Her2 score for a WSI was determined using the classified patches as follows:

- Score 3+, if patches with class 3 was greater than or equal to 10% of total patches
- Score 2+, if patches with class 2 was greater than or equal to 10%, or patches with class 3 was between 1% and 10%, of total patches
- Score 1+, if patches with class 1 was greater than or equal to 10% of total patches
- Score 0, otherwise

The confidence value for each WSI was calculated by averaging the confidence values of each patch. PCMS was calculated by summing the number of Score 3+ and 2+ patches and dividing the sum by total number of patches (excluding the background) as

$$PCMS = 100(n_2 + n_3)(\sum_{s=0}^{3} n_s)^{-1} \qquad (2)$$

where n is the number of patches given score s, s ∈ {0,1,2,3}

MTB NLP

A CNN was trained to predict the Her2 score for 128 x 128 patches of the WSI. Furthermore, as a post-processing step, a Random Forest model was trained to aggregate an estimated Her2 score and percentages of cell membrane for the WSI.

Pre-Processing: In the first, tissue regions were manually annotated at 40× by drawing regions from IHC stained slide images. A class label was assigned to each annotated region that corresponds to WSI GT score. In total, there were 272 annotated regions with an average size of 800 x 800.

Patch Classification: The architectures similar to Alexnet [15] and VGG-16 [17] were trained to predict the Her2 scores but the results were only submitted for the architecture similar to

Alexnet. The annotated regions were separated at case level by using 207 regions for training and the remaining 65 for validation. Each patch was randomly flipped and rotated to increase the training dataset and a dropout layer [18] was positioned to prevent the overfitting. The model was trained with a total of 8,575,000 patches and with cross entropy loss.

For testing, each WSI split in to non-overlapping patches of 128 x 128 and fed in to the trained network for predictions. Further, connected component analysis based approach was carried out to merge 128 x 128 patches into clusters. For each of the class labels, aggregate metrics were computed for the WSI that captured the percent of the slide pixels.

Aggregate Her2 Score and Percentage: To predict the Her2 score and PCMS process the aggregated metrics were computed during the patch classification. These metrics were used as predictors for a Random Forest classifier that produces that final class probabilities for each of the WSI. The same process was repeated using a Random Forest regressor to estimates for the percentage of cells that contained staining.

The 5-fold cross validation was done on all of the 52 training images. In each fold, all of the test images were scored and the predicted scores and percentage estimates were averaged over all folds to produce the final estimates.

VISILAB

In this method, the state-of-the-art GoogLeNet [16] was implanted to predict the Her2 score and the percentage of complete cell membrane.

Data Preparation: A handcrafted dataset was built. For this purpose, a set of representative patches of the four Her2 scoring classes were extracted from the ground truth WSIs. Additionally, an extra class was employed to collect background samples. These extracted patches from training WSIs were 68 x 68 pixels size each. A total of 5750 patches were selected with an average of 1150 patches per class. The dataset was further split in to training (75%) and validation (25%) dataset.

Training: Among several state-of-art CNNs, GoogLeNet was finally selected for submission according to the results on validation dataset. The prepared dataset was used for training, by selecting 0.01 as base learning rate, with a decreasing policy over 50 epochs, using the Stochastic Gradient Descent.

Classification: The algorithm takes a WSI and applies a grid technique to obtain the corresponding patches, with a similar size than the ones from the training dataset. These are later classified with the trained model, whose output is a class prediction and a percentage of confidence over that decision.

Her2 Scoring: Once every single patch is classified, a single class score is provided for the WSI. The decision rule takes into account the percentage of patches that belongs to each class (omitting the background, which was treated as a separate class) using the following criteria: starting from class 3+ to class 1+, the first one to achieve at least 10% of patches is chosen as final decision. Regarding the percentage of cells with full membrane staining, an expert rule was developed. The knowledge basis came from the alternative techniques that were also developed, such as the calculation of the staining density for the nuclei. As a result, a relationship between the classes percentage distribution and the percentage of membrane cell staining was discovered.

UCCSSE

This method is based on characteristics curves, a novel feature descriptor for predicting the Her2 score. In pre-processing phase, five regions of interest (ROI) were extracted from each WSI, each of size 1800 x 1200 at 20×. The only condition for selecting the ROIs was to select those regions that should not contain more than 30 % pixels as background.

The segmentation step consists of identifying the tissue portion including the IHC stained membrane. The selected ROIs were first segmented in HSB and CIELab colour spaces. In addition, some colour filters and neighbourhood masks were used to segment the connective tissues and fat lobules that should be separated before calculating the PCMS.

The essential part for the classification algorithm was the extraction of a characteristics curve for selected ROIs. The percentage-saturation characteristics curve was generated by varying the saturation limits from [0.1, 1] to [0.5, 1] in 20 steps by keeping the hue fixed. To plot the characteristics curve the percentage of stained region was calculated for each step by taking the ratio between segmented pixels to the number of pixels in an ROI. The characteristics curves have high discriminative appearance as shown in Fig 1. The curve always represents a smooth polynomial curve that can be accurately modelled using a cubic polynomial (best fit).

It was also observed during experimental analysis that when the Her2 score is 1+, the starting region of the curve always starts above the 10% mark depicting the presence of weak and incomplete membrane staining of regions. For 3+ score, the curves were lying above the 30% mark that shows the existence of an intense and uniform membrane staining areas.

RumRocks

In this approach, the two-dimensional (2D) CNN [15,19] models were trained for pre-processing and classification. First, as pre-processing step each WSI processed using deconvolution neural network (DCNN) and following by a CNN_1 to select the desired patches. Furthermore, the selected patches were processed through a CNN_2 to predict the Her2 score and the PCMS.

Patch Selection: A low resolution representation of a WSI (0.3125×) was selected and passed through a DCNN to segment the tissue components. Next, the detected regions were divided into patches with only condition that selected patches should contain 50% or more region from area of interest. The subsampled patch coordinates were translated to 10× resolution for further processing. The CNN$_1$ trained to accept or reject a subsampled patch based on its morphological appearance. The overview of neural network architectures are as given below

$$DCNN_1 = \{D_1, D_1, \dots\dots, D_6\ - U_1, U_2, \dots\dots, U_5 - C_{2D}3\backslash1\ -\ Sg\ \}$$

$$CNN_1 = \{D_1, D_1, \dots\dots, D_7\ - Reshape -\ FC -\ Sg\ \} \qquad (3)$$

The notation of the architecture is as follows, D_1 represents the down-sampling convolutional whereas U_1 represents the up-sampling convolutional layer with 3 as kernel size and 1 as stride. After convolutional operations batch normalization, ReLu and max pooling operations were applied. FC represents fully connected layers and Sg represents sigmoid function.

Classification: For predicting the Her2 score and PCMS, a CNN$_2$ with combination of residual layers [20] was employed. The batch dimensions were exploited in order to feed in multiple patches from the same WSI simultaneously. Instead of combining the prediction of individual patches through averaging or aggregating metrics, a tensor was reshaped to a vector once the spatial size has been significantly reduced and forward it through a 1D convolution layer. The overview of architecture CNN$_2$ is given below

$$CNN_2 = \{C_{2D}3\backslash1 -\ resB_1, \dots\dots, resB_7\ - flatten - C_{1D}1\backslash1, - FC_1 - FC_2 - Sg\ \} \qquad (4)$$

The CNN models were trained using the mean squared error loss function and the Adam stochastic gradient decent optimization method with initial learning rate of 10^{-3}. The learning rate was reduced every 15,000 iterations by a factor of 1.5 and trained each network for between 200,000 – 300,000 iterations. The average was calculated for each networks prediction to form an ensemble based score.

FSUJena

The algorithm for automated Her2 scoring was based on Alexnet [15] CNN. In this method, an activation matrix was extracted after convolution layers to compute the bilinear filters for predicting the Her2 score and PCMS.

At the first, ROIs were manually probed and patches of size 227 x 227 were randomly extracted at 20×. The pre-trained version of Alexnet was used from ImageNet dataset for further training on contest dataset. For each patch in the training dataset, an activation matrix was extracted after convolutional layers. The activations can be represented as a tensor $x \in \mathbb{R}^{w \times h \times d}$ comprised of d-dimensional vectors in a $w \times h$ spatial grid. The bilinear features [21,22] were further computed as the Gramian G matrix by summing up dyadic products along the spatial dimensions: $G = \sum_{i,j} x_{i,j} \; , x_{i,j}^{T}$. The matrix G contains the second-order statistics of the CNN features and have been found to be extremely useful for fine-grained recognition tasks. Then the square root and L_2normalization of G were employed to increase the numerical stability of further processing steps [22]. To differentiate among four scoring classes a multi-class logistic regression was used. It was also observed that using a pre-trained network on ImageNet dataset is also beneficial to avoid the overfitting issues. In preliminary results the bilinear features approach outperformed the conventional CNN activations.

For testing a WSI and to predict the Her2 score, an average was calculated for all the random crops patches. To predict the PCMS the mean tumour cell percentage seen in the training set of for a particular class as an estimate.

Huang's Method (Huangch)

In this approach, a range of handcrafted features extracted from the IHC stained slides after performing the stain deconvolution. The handcrafted features were then fed in to a model of multi-class AdaBoosted decision trees.

Sampling: At the first, control tissue was extracted to developed a pseudo color space for stain deconvolution [23] to obtain the two staining vectors. Further, mean filtering was performed to record the local maximal points. The patches were selected from each WSI on the basis of local maximal points as they were representing the strongest Her2 stained over-expression signals

Feature Extraction and Classification: A combined but numerically independent features vector space constructed by including Gabor Filtering, Features of Fractal Dimension by Differential Box-Counting [23], multi-wavelet methods, histogram statics methods, gray-level (over all colour channels) co-occurrence based methods [24,25] etc.

For predicting the Her2 score and the PCMS, a model of multi-class AdaBoosted decision-trees was employed to map the features vector of each patch to a predicted value. This model is known as Stagewise Additive Modeling using a Multi-class Exponential [26] loss function (SAMME). The model composed by a series of decision-trees by assigning a weight to each decision-tree. Whereas while training, a pool of decision-trees generated and after each iteration the best decision-tree was selected with its corresponding weight. After certain iterations, a group of decision-trees was selected for testing phase.

References

1 Webster JD, Dunstan RW. Whole-Slide Imaging and Automated Image Analysis. *Vet. Pathol.* 2014; **51**; 211-223.

2 Gurcan MN, Boucheron LE, Can A, et al. Histopathological Image Analysis: A Review. *IEEE Rev. Biomed. Eng.* 2009; **2**; 147-171.

3 Qaiser T, Sirinukunwattana K, Nakane K, et al. Persistent Homology for Fast Tumor Segmentation in Whole Slide Histology Images. *Procedia Comput. Sci.* 2016; **90**; 119-124.

4 Akbar S, Jordan LB, Purdie CA, et al. Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays. *Br. J. Cancer* 2015; **113**; 1075-1080.

5 Khan AM, Mohammed AF, Al-Hajri SA, et al. A novel system for scoring of hormone receptors in breast cancer histopathology slides. In 2nd Middle East Conference on Biomedical Engineering. IEEE, 2014; 155-158.

6 Chen R, Jing Y, Jackson H. Identifying Metastases in Sentinel Lymph Nodes with Deep Convolutional Neural Networks. August 2016.

7 Sirinukunwattana K, Raza SEA, Tsang Y-W, et al. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans. Med. Imaging* 2016; **35**; 1196-1206.

8 Choudhury KR, Yagle KJ, Swanson PE, et al. A Robust Automated Measure of Average Antibody Staining in Immunohistochemistry Images. *J. Histochem. Cytochem.* 2010; **58**; 95-107.

9 Hatanaka Y, Hashizume K, Nitta K, et al. Cytometrical image analysis for

immunohistochemical hormone receptor status in breast carcinomas. *Pathol. Int.* 2003; **53**; 693-699.

10 REYES-ALDASORO CC, WILLIAMS LJ, AKERMAN S, et al. An automatic algorithm for the segmentation and morphological analysis of microvessels in immunostained histological tumour sections. *J. Microsc.* 2011; **242**; 262-278.

11 Kuse M, Kalasannavar V, Rajpoot N, et al. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *J. Pathol. Inform.* 2011; **2**; 2.

12 Ali HR, Irwin M, Morris L, et al. Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. *Br. J. Cancer* 2013; **108**; 602-612.

13 He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. February 2015.

14 Vahadane A, Peng T, Sethi A, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Trans. Med. Imaging* 2016; **35**; 1962-1971.

15 Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira F, Burges CJC, Bottou L et al., eds. Advances in Neural Information Processing Systems 25. Curran Associates, Inc., 2012; 1097-1105.

16 Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions. September 2014.

17 Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. September 2014.

18 Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. July 2012.

19 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**; 436-444.

20 He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. December 2015.

21 Gao Y, Beijbom O, Zhang N, et al. Compact Bilinear Pooling. November 2015.

22 Lin T-Y, RoyChowdhury A, Maji S. Bilinear CNN Models for Fine-grained Visual Recognition. April 2015.

23 Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, 2009; 1107-1110.

24 Po-Whei Huang, Cheng-Hsiung Lee. Automatic Classification for Pathological Prostate Images Based on Fractal Analysis. *IEEE Trans. Med. Imaging* 2009; **28**; 1037-1050.

25 DiFranco MD, O’Hurley G, Kay EW, et al. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Comput. Med. Imaging Graph.* 2011; **35**; 629-645.

26 Ji Zhu , Hui Zou SR and TH. Multi-class AdaBoost. *Stat. its Interface 2.3* 2009; 349-360.

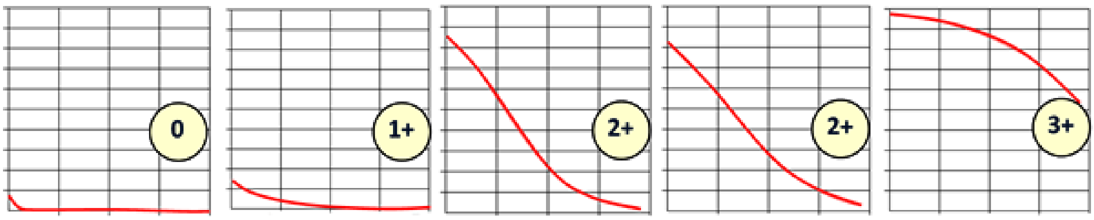

Fig 1: Characteristics curves and the corresponding Her2 score. The x-axis denotes range of the saturation value whereas y-axis denotes the calculated percentage from saturation limits. The predicted Her2 scores are also shown for each curve.