Deep Learning Model based Ki-67 Index estimation with Automatically Labelled Data

Lakshmi S.[a], Kotra Venkata Sai Ritwik[a], Deepu Vijayasenan[a], Sumam David S.[a], Saraswathy Sreeram[b] and Pooja K Suresh[b]

***Abstract*— Ki-67 labelling index is a biomarker which is used across the world to predict the aggressiveness of cancer. To compute the Ki-67 index, pathologists normally count the tumour nuclei from the slide images manually; hence it is time-consuming and is subject to inter pathologist variability. With the development of image processing and machine learning, many methods have been introduced for automatic Ki-67 estimation. But most of them require manual annotations and are restricted to one type of cancer. In this work, we propose a pooled Otsu's method to generate labels and train a semantic segmentation deep neural network (DNN). The output is post-processed to find the Ki-67 index. Evaluation of two different types of cancer (bladder and breast cancer) results in a mean absolute error of 3.52%. The performance of the DNN trained with automatic labels is better than DNN trained with ground truth by an absolute value of 1.25%.**

I. Introduction

Ki-67 labelling index is a proliferation marker used for assessment of biopsies in cancer patients. It helps in predicting the tumour cell advance and the therapy responses. The Ki-67 index is defined as the percentage of immunopositive nuclei to the total nuclei [1]. Pathologists observe the stained slides under the microscope and count the immunopositive and immunonegative nuclei. The tissue sections are stained according to the Immunohistochemistry (IHC) protocol which results in brown-coloured immunopositive nuclei and blue-coloured immunonegative nuclei. The manual counts are performed on the regions in which Ki-67 staining is notably higher relative to the adjacent areas (hotspots) [2]. Pathologists take a mean time of 17 minutes per case to count manually, as reported in [3]. Thus the manual assessment of Ki-67 index is a slow process. Computer-aided image analysis is more reliable and fast. Moreover, automation also helps in eliminating inter pathologist variations.

Deep Neural Network (DNN) models were proposed for automated assessment of Ki-67 index [4], [5]. Training of such a DNN requires pixel-level ground truth images corresponding to all input images. Since the annotations for the ground truth are marked manually, it is a time-consuming process which limits the amount of training data available for any model. Thus it can be a potential constraint in the performance of the system. We also note that most of the existing systems are restricted to one type of cancer (e.g., Breast cancer [5]–[7], Neuroendocrine tumour [8]–[10], Nasopharyngeal cancer [11], Meningiomas [12]).

In this work, we aim to develop a DNN system for Ki-67 index estimation of two different kinds of cancer (bladder and breast cancer) with a set of automatically generated labels. The system performs an approximate segmentation based on Otsu's thresholding method for separating immunopositive and immunonegative nuclei. An overview of this approximate segmentation and a deep learning U-Net model for Ki-67 index estimation are described in the following section. Section III presents the comparison of automatically generated labels and manual labels.

II. Methodology

In the Ki-67 staining process, the slides are incubated with the Ki-67 antibody. The Ki-67 antibody binds with the Ki-67 mitotic antigen, which is present in all stages of the cell cycle except the resting phase. Slides are then treated with Diaminobenzidine (DAB) chromogen to develop the brown color after which counterstaining with Meyer's hematoxylin (H) stain is done. This results in brown immunopositive nuclei and blue immunonegative nuclei [13].

In order to generate an automated ground truth, we propose to use Otsu's thresholding of the DAB and H stain values. The ground truth generated by this method is used to train a semantic segmentation system that identifies the immunopositive and immunonegative nuclei. These steps are detailed as follows:

A. Otsu's thresholding

We first separate the H and DAB stain using color deconvolution [6]. The optical density matrix [14] used to convert RGB space to H, eosin(E) and DAB is given as:

$$\begin{pmatrix} H \\ E \\ DAB \end{pmatrix} = \begin{pmatrix} 1.88 & -0.07 & -0.60 \\ -1.02 & 1.13 & -0.48 \\ -0.55 & -0.13 & 1.57 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (1)$$

To segment the nuclei, we perform Otsu's thresholding on H and DAB stained individual images corresponding to hotspot regions. Otsu's algorithm determines a threshold to separate the two classes by minimizing the intraclass variance in the intensity histogram. Thresholding on H image results in all background pixels black and nuclei pixels white, whereas on DAB image only immunopositive nuclei are white. To generate the labels, we obtain the immunopositive mask from DAB and immunonegative mask from H by removing the DAB part. We refer these labels as *otsu-labels*.

[a]Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India
lakshmi1510@gmail.com, deepu.senan@gmail.com, sumam@nitk.edu.in
[b]Department of Pathology, Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Manipal, Karnataka, India
saraswathy.sreeram@manipal.edu, pooja.suresh@manipal.edu

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

However, the algorithm assumes that the foreground and background classes have similar variances. The resulting threshold could be incorrect otherwise. In our dataset, there are many examples where the number of immunopositive nuclei is low. Thresholding on such images results in background pixels falling in immunopositive class.

We address this issue by determining the thresholds from the combined histograms obtained from multiple training images. The combined histogram will have similar foreground and background variance instead of few images having skewed class distributions. Therefore we observe the thresholds obtained through this method provide better segmentation. Pooled Otsu's thresholds are computed for each type of cancer separately. We refer to labels generated by this method as *pool-otsu-labels*.

B. Deep Neural Network (DNN)

We train a DNN with the targets as *pool-otsu-labels*. We use UNet architecture [15] for the pixel-wise classification of images into the immunopositive/immunonegative and background classes. The architecture has a contracting and an expanding path. The contracting path reduces the spatial information and increases the feature information. Expanding path combines both information with up convolutions and concatenations from the contracting path. Our implementation uses separable convolutions [16] to reduce the number of parameters and speed up the computation.

C. Ki-67 index calculation

Trained DNN gives the probabilities corresponding to each class as output. We compute class-wise masks by considering the set of pixels that have the maximum probability for a particular class. These masks are then subjected to a sequence of morphological opening and closing operations with a circular kernel. We find the local maxima of the immunopositive class probability values within the same class mask. We then count the local maxima to determine the number of immunopositive nuclei. Nuclei which are not round in shape results in more than one local maxima. Therefore local maxima at a separation less than one nucleus size are avoided. The process is repeated for immunonegative nuclei to get the count. The Ki-67 index is computed using the equation 2,

$$\text{Ki-67 index } (\%) = \frac{N_{IP}}{N_{IP} + N_{IN}} \times 100 \qquad (2)$$

where N_{IP} is the total number of immunopositive nuclei and N_{IN} is the total number of immunonegative nuclei.

III. Experiments and Results

We perform all the experiments on a dataset of 180 Ki-67 stained images corresponding to 18 patients. Otsu's thresholding is performed to generate automatic labels for training the UNet. The Ki-67 indices are computed from the UNet output. The results are compared with manually computed Ki-67 indices by the expert pathologists.

A. Dataset

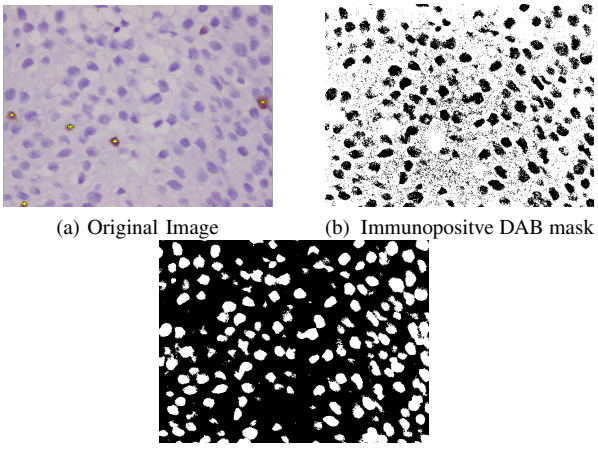
The dataset comprises of images from 10 breast cancer patients and 8 bladder cancer patients. Ten images are obtained for each patient from the hotspot regions. All images are of resolution 1920 $\times$ 1440 pixels at 40x magnification. The entire data is collected from Kasturba Medical College Mangalore India after ethical clearance (Ref.No: IEC KMC MLR 01-19/35).

Images corresponding to 10 patients (5 breast cancer and 5 bladder cancer) are used for training and the rest are used for evaluation of Ki-67 index computation. We perform manual annotation to obtain the ground truth for a subset of the dataset (Referred to as *man-labels*). The subset includes entire training data and 2 patients in the test dataset. The *man-labels* are verified by the pathologists. This manual annotation is performed to have a baseline comparison for *pool-otsu-labels*.

B. Segmentation

We converted the RGB input images into H, E and DAB format and performed Otsu's segmentation on DAB and H components of individual images. Fig. 1 shows the masks generated for a bladder cancer image which gives a wrong segmentation for immunopositive nuclei. This is because the image had very few immunopositive nuclei, which resulted in the background to be predicted as the immunopositive class.

Hence histograms of all the training images are extracted and pooled together for each type of cancer. This histogram is used for Otsu's thresholding and *pool-otsu-labels* are generated. Fig. 2 shows the results of the segmentation of the same image in Fig. 1. We observe that the immunopositive nuclei are segmented properly.

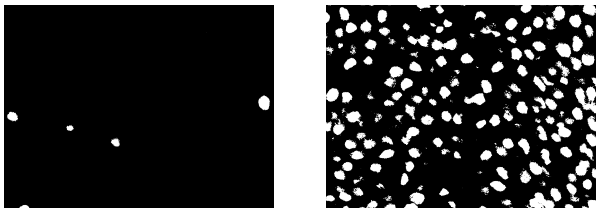


(a) Original Image (b) Immunopositve DAB mask

(c) Hematoxylin mask

Fig. 1: An example of wrong segmentation. Immunopositive nuclei are marked in yellow in the original image (a).

In order to asses the segmentation performance of individual and pooled Otsu's method we compute the dice scores of *otsu-labels* and *pool-otsu-labels* against *man-labels* in the training images. Table. I shows that there is 50% improvement in the dice score after pooling. We also



(a) Immunopositive DAB mask (b) Hematoxylin mask

Fig. 2: An example for pooled segmentation

use the pooled Otsu's output to generate method on the test data (Referred as Otsu) by using the training thresholds determined separately for bladder and breast cancer.

TABLE I: Evaluation of segmentation on Training data

Labels	Evaluation Measures	
	Dice score	*Dice score without background*
otsu-labels	0.47	0.41
pool-otsu-labels	0.94	0.70

C. DNN training

We use the training data of 10 patients to train a DNN model (UNet) using the *pool-otsu-labels* as ground truth (Referred as Otsu_DNN). We also train another system with *man-labels* as target values for comparison (Referred as GT_DNN). Each image is divided into 512 $\times$ 512 pixels, 18 overlapping patches. Two per cent of the training data is used exclusively for validation. We train both the DNNs for 100 epochs. The model weights corresponding to maximum validation accuracy is chosen for testing. The DNN is then used to generate the segmentation on the test data.

We evaluate the DNN segmentation on the subset (20 images) of test data where *man-labels* are available. Table. II shows the dice score, precision and recall obtained on the test data with respect to the ground truth. We observe that GT_DNN gives the maximum scores. It can be seen that the automatic methods also have very similar performance (5.6% absolute difference in dice score evaluated without considering the background).

TABLE II: Evaluation of segmentation on Test data

Method	Evaluation Measures			
	Precision	*Recall*	*Dice score*	*Dice score without background*
Otsu	0.89	0.89	0.94	0.71
GT_DNN	0.91	0.90	0.95	0.75
Otsu_DNN	0.89	0.89	0.94	0.71

D. Ki-67 index Evaluation

To determine the number of nuclei, we count the local peaks from the probabilities predicted by DNN. The local maxima which are nearer than 50 pixels (minimum size of one nucleus at 40x magnification) are removed as described in Section II-C. The Ki-index is then computed using equation 2.

Fig. 3 shows the bar plot of mean absolute error of Ki-67 from the different methods. First three cases correspond to bladder cancer, and the next five are of breast cancer. For low Ki-67 index values (i.e., less than 30%) Otsu_DNN predicted Ki67 error is less than 2%. Worst case error is 14%. The proposed method is even better than GT_DNN in 5 out of 8 cases.

The mean absolute error (MAE) of entire test data using various methods is mentioned in Table. III. It is observed that Otsu_DNN gave the least error (3.52%). The MAE corresponding to bladder cancer is 0.43% and breast cancer is 5.35%. We also observe that Otsu_DNN gives better results for bladder cancer data.

TABLE III: Evaluation of Ki-67 index

Ki-67 index mean absolute error (%)		
Otsu	GT_DNN	Otsu_DNN
5.56	4.77	3.52

Manual annotations are prone to errors due to personal biases. Automated labels resolve these errors. Moreover, labels are generated much faster in automated methods. Hence time consumed during manual ground truth annotation for a new data can be avoided. This method can scale up to large datasets.

Fig 4 shows the best prediction for the cancer case 2 (bladder cancer) and Fig 5 shows the worst prediction for the cancer case 5 (breast cancer) by Otsu_DNN method. It can be observed that we have a higher error in the Fig 5 because the small cells which are lymphocytes (highlighted as green) are counted as immunonegative nuclei. This can be resolved further by considering lymphocytes as a separate class during segmentation.

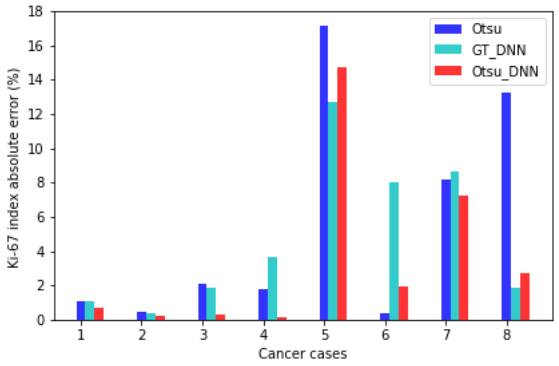
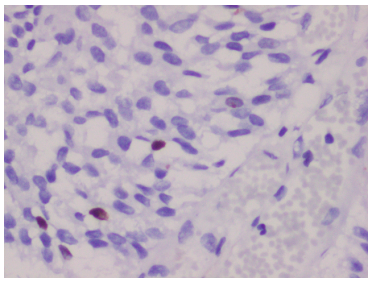


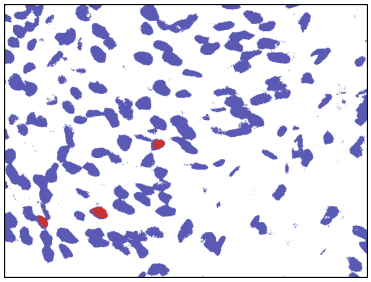
Fig. 3: Ki Index Prediction error for various cancer cases

IV. Conclusion

In this work, we proposed an automatic labelling scheme and trained a DNN across two different types of cancer. H and DAB stains were separated by color deconvolution from the original RGB image. Otsu's thresholding was then performed by combining the histograms obtained from multiple training images for each type of cancer. We generated labels from these thresholded images by obtaining the immunopositive nuclei from DAB mask and immunonegative nuclei from H mask that was exclusive of DAB. The generated labels

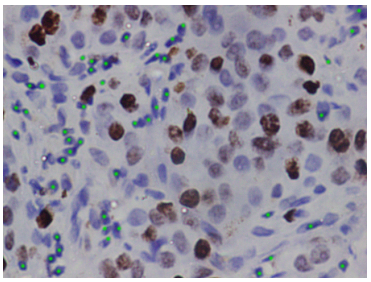


(a) Bladder cancer image

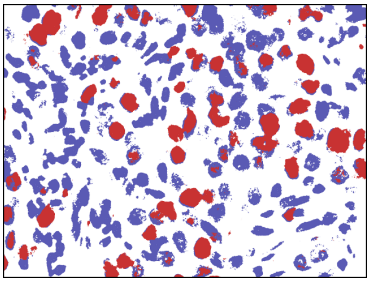


(b) Predicted Image

Fig. 4: Prediction of bladder cancer - best case



(a) Breast cancer image



(b) Predicted Image

Fig. 5: Prediction of breast cancer - worst case. Lymphocytes are marked in green in the original image (a).

are able to achieve a dice score of 0.94 in the training data. These labels are then used to train the Otsu_DNN system. Ki-67 index estimated by this system resulted in an MAE of 3.52%. The Otsu_DNN performed better than GT_DNN by an absolute value of 1.25%. Thus unlike previous works, we have developed a DNN system with automated labels which provided consistent performance across two different types of cancers.

Otsu_DNN system performed better for the bladder cancer images (MAE- 0.43%) as compared to breast cancer images (MAE- 5.35%). This is because small cells such as lymphocytes are present in the breast cancer images. These are mistakenly counted as immunonegative nuclei. This issue will be addressed in future work by considering lymphocytes as a separate class.

REFERENCES

[1] W. Jonat and N. Arnold, "Is the ki-67 labelling index ready for clinical use?" *Annals of oncology: official journal of the European Society for Medical Oncology*, vol. 22, no. 3, p. 500, 2011.

[2] M. H. Jang, H. J. Kim, Y. R. Chung, Y. Lee, and S. Y. Park, "A comparison of ki-67 counting methods in luminal breast cancer: the average method vs. the hot spot method," *PloS one*, vol. 12, no. 2, p. e0172031, 2017.

[3] J. Cottenden, E. R. Filter, J. Cottreau, D. Moore, M. Bullock, W.-Y. Huang, and T. Arnason, "Validation of a cytotechnologist manual counting service for the ki67 index in neuroendocrine tumors of the pancreas and gastrointestinal tract," *Archives of pathology & laboratory medicine*, vol. 142, no. 3, pp. 402–407, 2018.

[4] S. Lakshmi, D. Vijayasenan, D. S. Sumam, S. Sreeram, and P. K. Suresh, "An integrated deep learning approach towards automatic evaluation of ki-67 labeling index," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2310–2314.

[5] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer," *Scientific reports*, vol. 7, no. 1, p. 3213, 2017.

[6] V. J. Tuominen, S. Ruotoistenmäki, A. Viitanen, M. Jumppanen, and J. Isola, "Immunoratio: a publicly available web application for quantitative image analysis of estrogen receptor (er), progesterone receptor (pr), and ki-67," *Breast cancer research*, vol. 12, no. 4, p. R56, 2010.

[7] M. K. K. Niazi, C. Senaras, M. Pennell, V. Arole, G. Tozbikian, and M. N. Gurcan, "Relationship between the ki67 index and its area based approximation in breast cancer," *BMC cancer*, vol. 18, no. 1, pp. 1–9, 2018.

[8] F. Xing, H. Su, J. Neltner, and L. Yang, "Automatic ki-67 counting using robust cell detection and online dictionary learning," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 859–870, 2014.

[9] L. H. Tang, M. Gonen, C. Hedvat, I. M. Modlin, and D. S. Klimstra, "Objective quantification of the ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods," *The American journal of surgical pathology*, vol. 36, no. 12, pp. 1761–1770, 2012.

[10] F. Xing, H. Su, and L. Yang, "An integrated framework for automatic ki-67 scoring in pancreatic neuroendocrine tumor," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 436–443.

[11] P. Shi, J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen, "Automated ki-67 quantification of immunohistochemical staining image of human nasopharyngeal carcinoma xenografts," *Scientific reports*, vol. 6, p. 32127, 2016.

[12] B. Grala, T. Markiewicz, W. Kozłowski, S. Osowski, J. Słodkowska, and W. Papierz, "New automated image analysis method for the assessment of ki-67 labeling index in meningiomas." *Folia Histochemica et Cytobiologica*, vol. 47, no. 4, pp. 587–592, 2009.

[13] H. Dong, C. Bertler, E. Schneider, and M. A. Ritter, "Assessment of cell proliferation by agnor scores and ki-67 labeling indices and a comparison with potential doubling times," *Cytometry: The Journal of the International Society for Analytical Cytology*, vol. 28, no. 4, pp. 280–288, 1997.

[14] A. C. Ruifrok, D. A. Johnston *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.