Deep learning-based automated hot-spot detection and tumor grading in human gastrointestinal neuroendocrine tumor

Darshana Govind, Kuang-Yu Jen, Pinaki Sarder

Darshana Govind, Kuang-Yu Jen, Pinaki Sarder, "Deep learning-based automated hot-spot detection and tumor grading in human gastrointestinal neuroendocrine tumor," Proc. SPIE 11320, Medical Imaging 2020: Digital Pathology, 113200P (16 March 2020); doi: 10.1117/12.2548759

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Deep learning-based automated hot-spot detection and tumor grading in human gastrointestinal neuroendocrine tumor

Darshana Govind[1], Kuang-Yu Jen[2], and Pinaki Sarder[1,*]

[1]Department of Pathology and Anatomical Sciences,
University at Buffalo – The State University of New York

[2]Department of Pathology, University of California at Davis

[*]Address all correspondence to: Pinaki Sarder
Tel: 716-829-2265; E-mail: pinakisa@buffalo.edu

ABSTRACT

Ki-67 index is an important diagnostic factor in gastrointestinal neuroendocrine tumor (GI-NET). The current gold standard for grading GI-NETs involves the visual screening of histopathologically stained tissues, for hot-spots containing high amounts of proliferating tumor cells (stained with Ki-67 antibody). Subsequently, the Ki-67 index, i.e. the percentage of proliferating tumor cells within the hot-spot is manually obtained. To automate this subjective and time consuming process, we have developed an integrated pipeline, termed SKIE (synaptophysin-Ki-67 index estimator), combining double-immunohistochemical (IHC) staining for synaptophysin (stains tumor) and Ki-67, with whole slide image (WSI) analysis. The Ki-67 index for 50 human GI-NET WSIs were estimated by SKIE and compared with three pathologists' assessment, and the gold standard (exhaustive counting by a fourth pathologist) based on the double-stained image. All four pathologists unanimously graded 38 WSIs, among which, SKIE achieved 94.74% accuracy. One discrepant case was attributed to staining inconsistencies and the other to SKIE selecting a better hot-spot. The remaining 12 WSIs had discrepant grades among pathologists, and hence, the gold standard was chosen for comparison, wherein, 10 WSI grades matched with that of the gold standard, and SKIE assigned a lower and higher grade to two cases. Overall, SKIE agreed with the gold standard with a substantial linear weighted Cohen's kappa $\kappa = 0.622$ with CI [0.286, 0.958]. We further expanded our method to deep-SKIE, wherein, a deep convolutional neural network (DCNN) was trained and validated using 13,736 hotspot-sized tiles from 40 WSIs, each categorized into one of four classes (background, non-tumor, tumor grade 1, tumor grade 2) by SKIE and tested on 9 WSIs. Deep-SKIE achieved an accuracy of 91.63% with near-perfect agreement ($\kappa = 0.88$ with CI [0.87, 0.89]) with the gold standard.

Keywords: Whole slide image analysis, gastrointestinal neuroendocrine tumor (GI-NETs), SKIE, deep-SKIE, tumor grading

1. INTRODUCTION

Gastrointestinal neuroendocrine tumor (GI-NET) grading involves the manual screening for hot-spots[1, 2] (region with highest density of Ki-67 positive cells), and the subsequent enumeration of cells within the hot-spot, to obtain the Ki-67 index, i.e. the percentage of Ki-67 positive tumor cells within the selected hot-spot. This index is used to categorize the tissue into various grades of tumor: G1, G2 or G3, based on the WHO classification criteria 2010[1] (0 > Ki-67 index > 3%: G1; 3 > Ki-67 index

Medical Imaging 2020: Digital Pathology, edited by John E. Tomaszewski,
Aaron D. Ward, Proc. of SPIE Vol. 11320, 113200P · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2548759

>25%: G2; Ki-67 index >25%: G3). This technique is subjective, cumbersome and error prone[3, 4]. Additionally, the difficulty in distinguishing non-tumor cells from tumor cells results in the false elevation of the tumor grades[5]. Furthermore, the most widely used immunostain quantification tool, ImmunoRatio[6], requires the manual detection of hot-spots and there exists no provision to distinguish the tumor from non-tumor cells.

To address the above limitations, we have developed an integrated approach (tested on 50 whole slide images (WSIs) of GI-NET biopsies), termed SKIE (synaptophysin-Ki-67 index estimator), to accurately assess the Ki-67 index using WSIs of synaptophysin/Ki67 double immunostains. Our work automates the clinical procedure of hot-spot detection and Ki-67 estimation, while being faster (~ 600 times) and more accurate, thereby improving clinical workflow. Additionally, SK (where synaptophysin stains the tumor regions) ensures the accurate estimation of Ki-67 index as it eliminates the accidental inclusion of non-tumor cells. We further expanded our pipeline to deep-SKIE, by the incorporation of a deep convolutional neural network (DCNN) (Inception V3[7]), to automatically predict the tumor grade and generate a heat-map displaying the tumor distribution across the WSI.

2. METHODOLOGY

2.1 Data acquisition

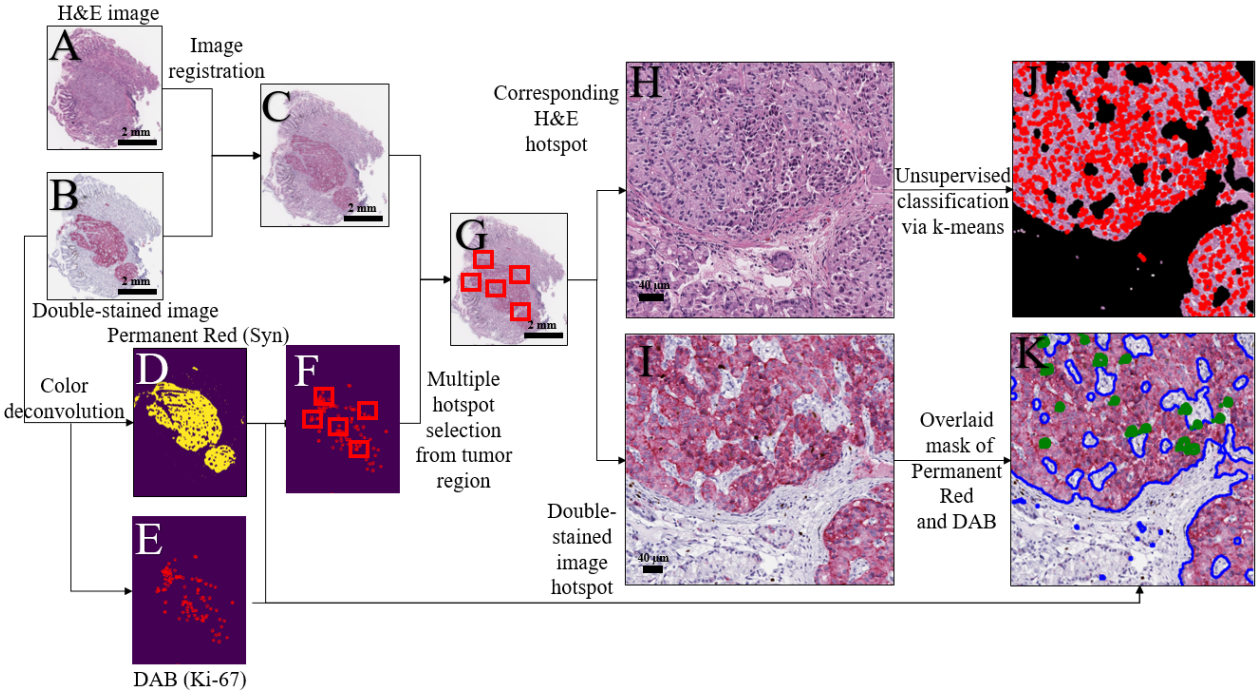

Fig 1. Synaptophyin-Ki-67 Index Estimator (SKIE) computational pipeline. GI-NET tissue section stained with (A) hematoxylin and eosin, and (B) synaptophysin (red) and Ki-67 (brown). (C) Result of landmark-based image registration. (D-E) Binary mask of synaptophysin positive region and Ki-67 positive cells, respectively, obtained via color deconvolution. (F-G) Automated detection of five candidate hot-spots. (H-I) Extracted hot-spot from 2A and 2B. (J) Overlay of nuclei mask obtained via unsupervised classification of pixels via k-means clustering of 2H to obtain all cells within tumor regions. (K) Overlay of the hot-spot using masks from 2D and 2E to obtain Ki-67 positive cells (green) within tumor regions (blue).

This study was approved by the Institutional Review Board at the University of California Davis Medical Center. Case selection, tissue section preparation and staining were previously detailed[5]. The 50 GI-NET cases used for analysis were derived from stomach (*n*=8), small bowel (*n*=13), appendix (*n*=5), colon (*n*=3), rectum (*n*=16), and pancreas (*n*=5). One double-stained, one Ki-67-only and one H&E-stained adjacent sections were generated for each case. The slides were scanned with Aperio AT2 scanner at 20X magnification.

2.2 Quantification of Ki-67 index by pathologists

Details on the participating pathologist's quantification of the Ki-67 index were previously

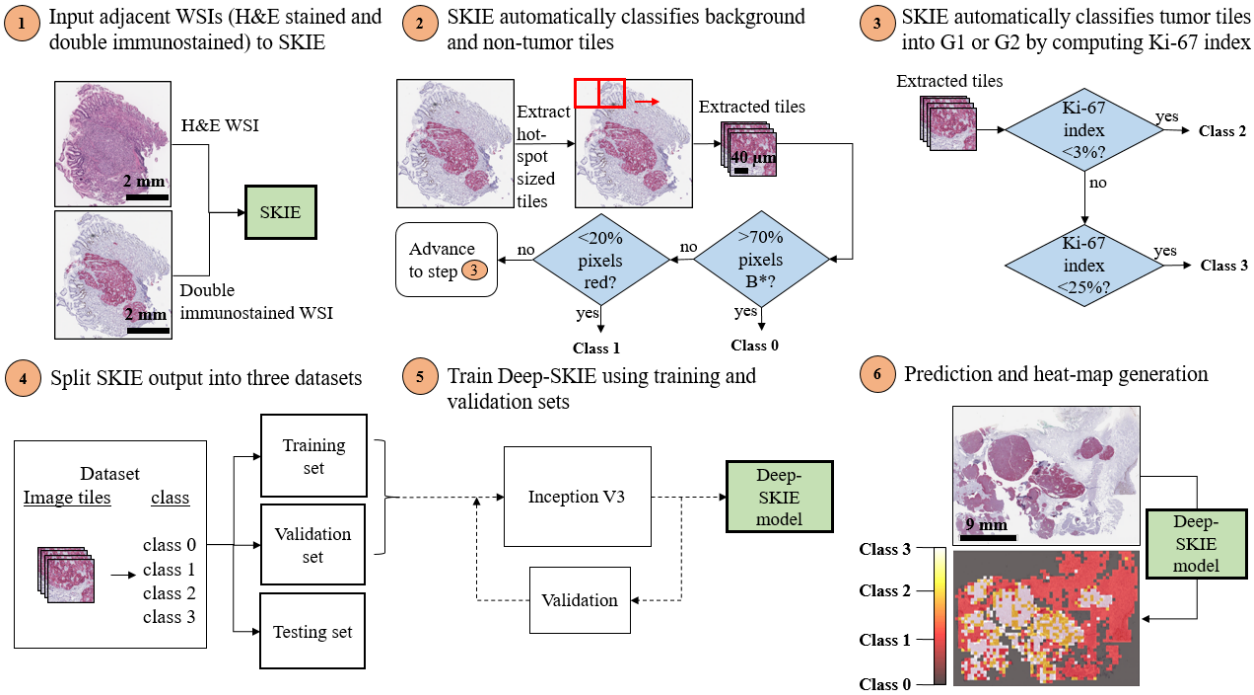

Fig 2. Deep-SKIE computational pipeline. SKIE classifies hot-spot-sized tile into class 0 (>70% background (B* in step 2)), class 1 (<20% synaptophysin), class 2 (G1) or class 3 (G2). The extracted tiles and labels are used to train, validate and test the Inception V3 network. The test set predictions are displayed as a heat-map highlighting the tumor distribution across the WSI.

described[5]. Each pathologist subjectively selected a hot spot and enumerated the Ki-67 positive and negative tumor cells until at least 500 total tumor cells were counted. As gold standard, a fourth pathologist measured the Ki-67 index by exhaustive counting (i.e. manually ticking off each Ki-67-positive and negative tumor cell).

2.3 SKIE–computational pipeline

Fig. 1 shows the computational pipeline for SKIE. Manually selected landmark points chosen within H&E and double-immunostained WSI by the user is used for landmark-based image registration[9] (Fig. 1A-C). The binary masks of the tumor region and the Ki-67-positive nuclei (Fig. 1D-E) were obtained via color deconvolution[10]. Five high-density Ki-67 positive locations (500 x 500 µm) were automatically selected (Fig. 1F) from the tumor region (using a 2D histogram of the Ki-67 positive cells) as candidate hot spots (Fig. 1G). The Ki-67 negative tumor nuclei stained with hematoxylin in the double-immunostained images appeared to be obscured in certain locations due to low contrast

against Permanent Red (synaptophysin). To avoid inconsistencies in the detection of Ki-67 negative tumor nuclei due to this poor contrast, the enumeration was performed on the registered H&E-stained section. Thus, the binary mask of the tumor region was applied to the H&E-stained counterpart of the candidate hot spot (Fig. 1H-I). The non-proliferating tumor cells were detected via unsupervised classification of pixels using k-means clustering algorithm (with $k = 3$, representing hematoxylin, eosin and the background). The cluster representing hematoxylin was chosen to obtain the mask for all nuclei (red contours in Fig. 1J) within the tumor region. The overlap of the tumor region and Ki-67 positive nuclei (Fig. 1D-E) were extracted to obtain the Ki-67 index (Fig. 1K).

2.4 Deep-SKIE computational pipeline

Fig. 2 shows the deep-SKIE pipeline. A sliding window is used to extract all hot-spot-sized tiles from the WSI, which are then classified by SKIE into background (class 0) or non-tumor (class 1), if the tile has > 70% background pixels, and < 20% synaptophysin stain, respectively. The remaining tiles are processed to compute their Ki-67 index, which are used to categorize them into class 2 (G1) and 3 (G2). 80% of WSIs (40 WSIs), i.e. 13,736 (250 x 250) image tiles were used to train and validate a deep convolutional neural network (DCNN) (Inception V3), and ~20% of WSIs (9 WSIs, 1 WSI with staining inconsistencies was removed) were used to test the model.

3. RESULTS

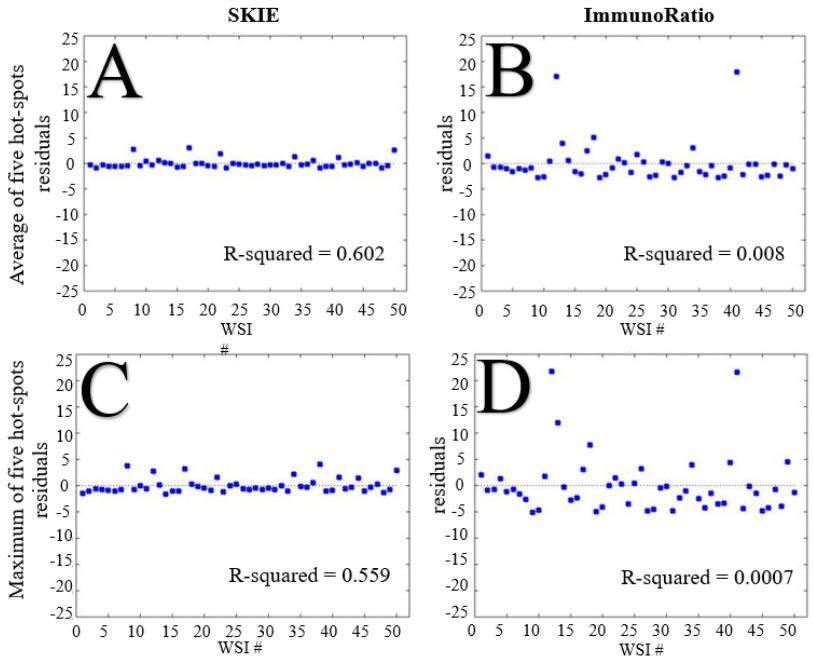

Fig 3. Residual error comparison between SKIE and ImmunoRatio. (A) The residual error for the average of five candidate hot-spots picked by SKIE versus (B) ImmunoRatio on the same fields as compared to the gold standard of exhaustive manual counting of the pathologist selected hot-spot. (C) The residual error of the maximum hot-spot picked by SKIE versus (D) ImmunoRatio on the same field as compared to the gold standard of exhaustive manual counting of the pathologist selected hot-spot

3.1 SKIE-based tumor grade and Ki-67 index performance evaluation

SKIE had a substantial agreement with the gold standard with a linear weighted Cohen's kappa[8] $\kappa = 0.62$ with 95% confidence interval (CI) [0.29, 0.96], with an accuracy of 92% and an index error of $0.80 \pm 0.96\%$ (absolute difference between the gold standard and SKIE index). The three pathologists and the gold standard unanimously agreed upon tumor grades for 38 of the 50 cases; 94.74% of these cases were accurately classified by SKIE. One discrepant case was attributed to staining artifacts, and the other to SKIE selecting a hot-spot of higher index. For the remaining 12 WSIs with varying grades assigned by different pathologists, the gold standard

was chosen for comparison, wherein, 10/12 WSIs matched with the gold standard and one discrepant case was assigned a lower grade (SKIE detected a higher number of Ki-67 negative cells than the pathologist) and the other, a higher grade than gold standard . SKIE roughly needed 1.4 ± 0.47 seconds to compute the index from a single hot-spot, while the gold standard pathologist needed approximately 15 minutes.

3.2 Comparison with ImmunoRatio

With the five-candidate hot-spots, two different metrics were selected: 1) the maximum and 2) average of these five candidates. The same hot-spots from adjacent sections stained with Ki-67 only were processed with ImmunoRatio. SKIE outperformed ImmunoRatio by approximately three-fold, with the average of the five-candidate hot-spots as detected by SKIE being the closest to the gold standard (Fig. 3). The residual error plot of the three pathologists, SKIE, and ImmunoRatio in comparison to the gold standard is shown in Fig. 4. SKIE generated comparable results with that of the pathologists', while being more reproducible, faster, and having a lower index error.

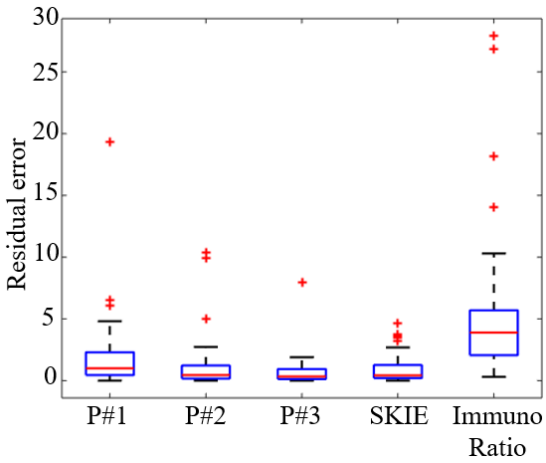

Fig 4. Ki-67 index residual error for participating pathologists, SKIE, and ImmunoRatio compared to exhaustive manual counting. Boxplots show the error rate of SKIE compared to three pathologists (P1, P2, and P3) and that of ImmunoRatio as compared to the gold standard of exhaustive manual counting.

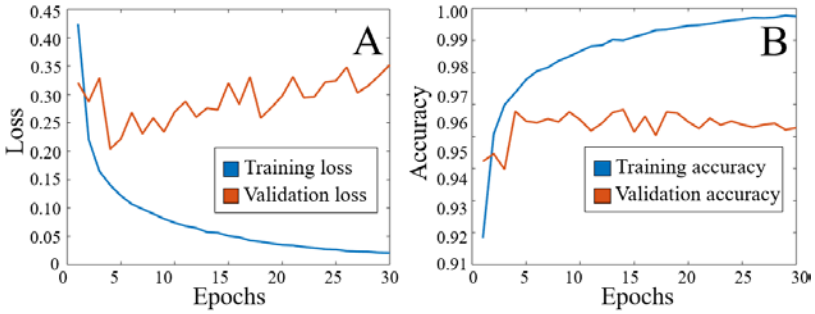

Fig 5. Training and validation of deep-SKIE. (A) Training and validation (A) loss and (B) accuracy over 30 epochs.

3.3 Deep-SKIE performance evaluation

Deep-SKIE yielded a training and validation accuracy of 99.75% and 96.27%, respectively (Fig. 5), on 13,736 image patches from 40 WSIs and a testing accuracy of 91.63% on 7922 image patches from 9 WSIs. Deep-SKIE had a near perfect agreement ($\kappa = 0.88$ with CI [0.87, 0.89]) with the gold standard.

4. DISCUSSION

The accurate assessment of the Ki-67 index is crucial for tumor grading of GI-NETs, which determines the patient's prognosis and outcome. In this study, we developed an integrated approach (SKIE) to compute the Ki-67 index by combining a double-immunostain technique with an automated computational pipeline. This powerful approach allows for both automated hot spot selection and the Ki-67 index quantification. Furthermore, the hot-spots generated by SKIE, when re-analyzed by the pathologist lead to the revision of grades in three cases, highlighting SKIE's superior hot-spot selection capability and ability to improve clinical workflow. Additionally, we have expanded our pipeline to

deep-SKIE by incorporating a DCNN to automatically predict the tumor grade and generate a heat-map displaying the tumor distribution across the WSI. The proposed framework will aid in not only faster and more reproducible generation of the Ki-67 index, but will also be helpful in understanding the development of GI-NETs.

5. CONCLUSION AND FUTURE WORK

This study shows that SKIE has a better hot-spot selection and diagnostic capability of GI-NETs than standard the Ki-67 index based automation. Moreover, deep-SKIE enables the automation of tumor grading using double immunostained GI-NET tissue images. The proposed pipeline is faster, robust and more accurate than existing automated and manual methods. Future work will aim at studying the DCNN filters to better understand the features used by the network to predict tumor grades. Additionally, the concept of virtual staining will be further explored, i.e. the use of deep learning to virtually stain the tumor regions from H&E-stained tissue images. This exploration would not only eliminate the need for additional stains, but also aid in the development of novel digital biomarkers for GI-NET diagnosis and help better understand the disease progression.

6. ACKNOWLEDGEMENT

The project was supported by the faculty startup funds from the Jacobs School of Medicine and Biomedical Sciences, University at Buffalo. We thank NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research (NVIDIA, Santa Clara, CA).

REFERENCES

[1] G. Rindi, “Nomenclature and classification of neuroendocrine neoplasms of the digestive system,” WHO classification of tumours of the digestive system, 13-14 (2010).

[2] F. T. Bosman, F. Carneiro, R. H. Hruban *et al.*, [WHO classification of tumours of the digestive system] World Health Organization, (2010).

[3] Y. R. Chung, M. H. Jang, S. Y. Park *et al.*, “Interobserver variability of Ki-67 measurement in breast cancer,” Journal of pathology and translational medicine, 50(2), 129 (2016).

[4] L. H. Tang, M. Gonen, C. Hedvat *et al.*, “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, 36(12), 1761-1770 (2012).

[5] K. Matsukuma, K. A. Olson, D. Gui *et al.*, “Synaptophysin-Ki67 double stain: a novel technique that improves interobserver agreement in the grading of well-differentiated gastrointestinal neuroendocrine tumors,” Modern Pathology, 30(4), 620 (2017).

[6] V. J. Tuominen, S. Ruotoistenmäki, A. Viitanen *et al.*, “ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67,” Breast cancer research, 12(4), R56 (2010).

[7] C. Szegedy, V. Vanhoucke, S. Ioffe *et al.*, "Rethinking the inception architecture for computer vision." 2818-2826.

[8] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” Psychological bulletin, 70(4), 213 (1968).

[9] J. Modersitzki, [FAIR: flexible algorithms for image registration] SIAM, (2009).
[10] A. C. Ruifrok, and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” Analytical and quantitative cytology and histology, 23(4), 291-299 (2001).