Ki67 Proliferation Index Quantification using Silver Standard Masks

Seyed Hossein Mirjahanmardi[*1] SHMIRJAHANMARDI@RYERSON.CA
Melanie Dawe[2] MELANIE.DAWE@UHNRESEARCH.CA
Anthony Fyles[2] ANTHONY.FYLES@RMP.UHN.CA
Wei Shi[2] W.SHI@UHNRESEARCH.CA
Fei-Fei Liu[2] FEI-FEI.LIU@RMP.UHN.CA
Dimitri Androutsos[1] DIMITRI@RYERSON.CA
Susan J. Done[2,4] SUSAN.DONE@UHN.CA
April Khademi[1] AKHADEMI@RYERSON.CA
[1] *Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, CAN*
[2] *Princess Margaret Cancer Centre, University Health Network, Toronto, ON, CAN*
[4] *Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, CAN*

Editors: Under Review for MIDL 2022

Abstract

Deep learning (DL) systems obtain high accuracy on digital pathology datasets that are within the same distribution as the training set, but when applied to unseen datasets there can be a reduction in performance due to differences in acquisition hardware/software and staining protocols/vendors. This is a barrier to translation since developed models cannot be readily deployed at new labs. To overcome this challenge, we present silver standard (SS) annotations as a method to improve the performance of deep learning architectures on unseen Ki67 pathology images. An unsupervised technique called IHCCH was used to generate SS masks for Ki67+ and Ki67- nuclei from the target lab. A previously validated architecture for Ki67, UV-Net, is trained with gold standard (GS) images and a combination of SS and GS masks to evaluate performance. It was found that adding SS masks from the unseen center to the training pool improved performance over clinically relevant PI ranges. The SS model with equal amounts of SS and GS (310 patches each) was shown to significantly improve PI estimation consistency over all PI ranges. Since SS masks are easy to generate, this method can be used for per-centre calibration to improve consistency and reliability of Ki67 quantification which is paramount for wide-scale adoption.

Keywords: Computational pathology, silver standards, Ki67, UV-Net, deep learning.

1. Introduction

Breast cancer is the second most commonly diagnosed cancer among women worldwide (Bray et al., 2018). Invasive ductal carcinoma (IDC) accounts for 80% of breast cancer cases, making it the most invasive type in breast cancers. If IDC is remained untreated, it can metastasize to other regions of the body (Mohan, 2018). Hence, the survival rate of patients suffering from this cancer type is heavily attributed to precise and timely diagnosis. An accurate diagnosis depends on analyzing important features such as tumor cell proliferation index (PI), tumor size, and its morphological features (Walters et al., 2013). Grading criteria such as the Modified Bloom-Richardson are used to evaluate nuclear grades and

mitotic index in breast cancer (Elston and Ellis, 1991). Immunohistochemical biomarkers can potentially enhance tumor characterization and response to therapy (Veronese et al., 1993). The MIKB (Ki67) biomarker has been gaining interest for assessing the proliferation and aggressiveness of breast cancer tumors. The accurate calculation of Ki67 PI can be used as a key metric for diagnosis and treatment planning (Dowsett et al., 2011; Jalava et al., 2006). While this biomarker can potentially improve patient care, manual counting of cells is time-consuming, costly, and laborious. With the advent of whole slide imaging (WSI) scanners along with advances in computational resources and artificial intelligence algorithms, these technologies can be used to deliver more objective, efficient and quantitative Ki67 scoring. Deep learning techniques show much potential for automated processing of different biomarkers in histopathology images (Srinidhi et al., 2020; Van der Laak et al., 2021; Amgad et al., 2021; Graham et al., 2019). Deep learning architectures such as piNet (Geread et al., 2021), KiNet (Xing et al., 2019), and UV-Net (Mirjahanmardi et al., 2022) have been specifically developed for Ki67 PI quantification. KiNet was developed for pancreatic cancer and PiNet and UV-Net have been developed for breast cancer. As was shown in (Geread et al., 2021) and (Mirjahanmardi et al., 2022), while piNet and UV-Net achieved high accuracy on multi-institutional datasets, there was reduced performance and consistency on unseen datasets. This is largely due to the fact that labs generate images from different acquisition systems and staining vendors/protocols (Van der Laak et al., 2021) which can create generalization issues for deep learning systems. Therefore, this causes deployment challenges and is a barrier to wide-scale adoption.

One possible way to minimize the generalization gap for digital pathology and deep learning could be to obtain a small dataset from the institution of interest, and retrain the model to include some data from that centre. This can be considered as a per-center fine-tuning and calibration. We hypothesize that including images from the unseen centre into the training mix can improve performance and consistency of the tool for the target lab. However, generating large amounts of ground truth data, especially pixel-wise annotations for many cells is a time-consuming task, which also requires specialist expertise. To overcome these challenges, this paper evaluates the use of silver-standard (SS) ground truths from unseen centers for Ki67 PI quantification. Sliver standards are ground truths that are noisy or generated by an unsupervised automated or semi-automated tool, i.e. they are less than ideal compared to the "gold standard" (GS) which is exhaustive annotations performed by the pathologist. SS annotations can be quickly generated and in large amounts, which is more cost-effective and reduces development time significantly. They can be added to the GS training pool for retraining and fine-tuning on a lab basis, which would accelerate translation.

2. Dataset

This work uses Ki67 stained breast cancer images obtained from different institutions. A total of 500 patches with size 256×256 extracted whole slide images (WSI) from two sources are used, one from the St. Michael's Hospital (SMH) in Toronto, and an open-source "Deepslide" (Senaras, 2018) with ×20 Aperio AT Turbo and ×40 Aperio ScanScope scanners, respectively. The Deepslide images are down-sampled to ×20 to be compatible with other datasets. The images were annotated by marking $Ki67^-$ and $Ki67^+$ centroids (Geread

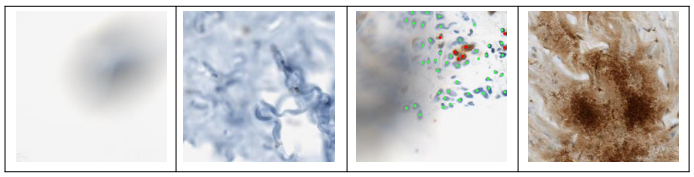


Figure 1: Patches with artifacts: dust, folding, blur/dust, overstaining.

et al., 2021). Centroid annotations were converted into a Gaussian kernel which includes surrounding regions to improve learning. This dataset is used primarily for training (62%), validation (20%), and testing (18%). Noise and artifacts including overstaining, background, folders, blur, and dust are common in tissue slides. Therefore, 15% of the training dataset includes tiles with artifacts to reduce false positives. Examples of such images are shown in Figure1. University Health Network (UHN) is the last dataset which contains 411 tissue microarrays (TMA) from 175 patients for a total of 26,304 patches. Each TMA is 2000×2000 and an expert PI estimate is available for each patient. 381 TMAs are used for testing and the remaining 30 were used to create SS ground truths.

3. Methods

This work evaluates the application of SS for Ki67 nuclei detection and PI quantification. An unsupervised method is used to generate the SS masks, and deep learning models are trained with GS masks, as well as GS and SS masks combined. SS are created from the unseen dataset (UHN) and performance on the held-out test set for all models is investigated.

3.1. Deep Learning Framework

The UV-Net architecture, developed for Ki67 PI quantification in breast cancer, and validated on a large multi-institutional dataset with high performance, is the model of interest in this work. The pipeline is shown in Figure 6. UV-Net has shown high accuracy with strong generalization capabilities that outperforms other networks such as U-Net, DenseU-Net, and MultiresU-Net (Mirjahanmardi et al., 2022). The model is shown in Figure 2 which has seven stages with multiple V-Blocks in the encoding and decoding arms. The V-Block connections are utilized to preserve high-resolution details related to nuclear features and each V-Block has an input with n channels and output with 2n channels (creating a "V" shape) through four successive stages. Two hyperparameters, f and k, are defined for each V-Block as the number of input channels, and the output channels at the end of each stage, respectively. The hyper-parameters f and k in the V-Blocks vary over each stage with their values shown in Figure 2. Figure 2b shows a V-Block wherein $f = 16$ and $k = 4$. In each stage, the input feature is processed by a 1×1 convolution with $f = 16$ filters, then transformed to the output with $k = 4$ filters. The output of this step is concatenated to the input, creating a matrix with 20 filters which are fed to the second stage. This process is repeated for a total of four times to generate an output with $2 \times f$ filters. A regression loss is used and the output channels contain background as well as the predicted $Ki67^-$ and $Ki67^+$ nuclei. Each channel is post processed (Otsu thresholding, median filtering, and

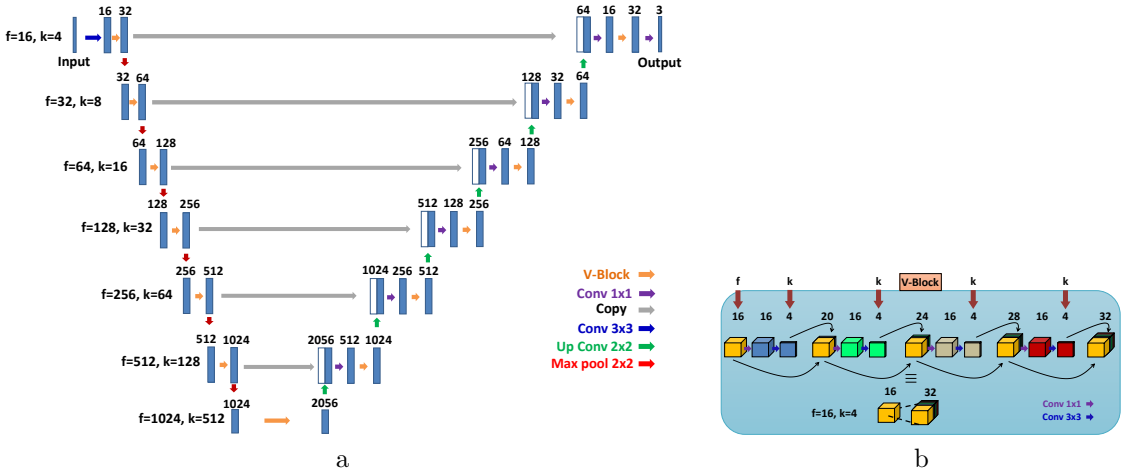


Figure 2: The UV-Net architecture that is made up of encoding and decoding arms. a. network b. One V-Block example where f=16, and k=4, composed of four stages.

morphological processing) to isolate Ki67$^-$ and Ki67$^+$ cells. To separate nuclei that are overlapping, the watershed algorithm is used. Centroids are compared with the GS ground truths (testing set) to evaluate the models with the F1-score:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{1}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively. To measure PI quantification performance, the PI is estimated

$$PI = \frac{\#\ Ki67^+\ cells}{\#(Ki67^+\ +\ Ki67^-)\ cells} \tag{2}$$

and the difference in PI between manual and automated is measured, $|\Delta PI|$. To investigate prediction consistency, the coefficient of variation (CoV) is examined for $|\Delta PI|$. The difference in CoV, $|\Delta CoV|$, for two PI ranges i, j, is computed to measure the variation between intervals, and the mean $|\Delta CoV|$ is computed.

3.2. Unsupervised Ki67 Nuclei Detection (IHCCH)

The immunohistochemical (IHC) color histogram (IHCCH) approach is the unsupervised method that is compared to the deep learning-based methods, and also used to generate SS masks. Preprocessing for this method is vector median filtering and background subtraction. The algorithm relies on an unsupervised thresholding method in the b* channel from the L*a*b* space. The L*a*b* space correlates well with human color perception and the b* channel strongly separates between blue and brown pixels, which correspond to the hematoxylin (H) and DAB stains (Geread et al., 2019). An adaptive b* threshold is found using the rolling ball method which operates on the b* histogram and finds local minima. The threshold is used to separate the IHC images into H and DAB channels and

nuclei detection is performed separately on each channel. To detect nuclei, first all objects are detected with connected components the the nuclei radius is estimated based on a percentile of the smaller objects. Nuclei are detected using the gradient and the estimated nuclei radius, more details can be found in (Geread et al., 2019). The block diagram for this algorithm is shown in Figure 7. Code for the IHCCH algorithm will be available at: at https://github.com/IAMLAB-Ryerson/IHCCH.

3.3. Experimental Setup

The GS data from Deepslide/SMH (310 patches) was used to train the models and the SS masks were generated from the unseen UHN dataset (310 patches). To select SS images, TMAs were tiled into patches of 256x256 in size, and patches with at least ∼80% tumorous tissue were selected, which took less than five minutes. SS masks were generated by the unsupervised IHCCH algorithm and nuclei centroids were marked up by a Gaussian circular kernel. UV-NET and comparison models are trained with GS and SS masks in an incremental manner and the performance is analyzed. A total of 50, 140, 186, and 310 SS generated masks are added to the 310 GS patches and the deep learning models are retrained each time. Each model is then tested on 90 held-out patches from Deepslide/SMH and F1 score is measured. To examine generalization capabilities, the models are tested on the 381 heldout TMAs from UHN and performance is evaluated with respect to PI quantification accuracy (ΔPI) and consistency ($CoV, \Delta CoV$). The comparison architectures include MultiResU-Net (Ibtehaz and Rahman, 2020), U-Net(Ronneberger et al., 2015), and piNet (developed specifically for breast Ki67 quantification) and they are trained with GS data. Experiments are conducted on the same machine with an NVIDIA GeForce RTX 2080 Ti. A total of 100 epochs are used with an Adam optimizer, batch size=16, and learning rate=10^{-3}. Huber loss function was used for all architectures to predict nuclei centroids. Data augmentations such as horizontal/vertical flips and scaling are used.

4. Results and Discussion

Nuclei detection performance, quantified by F1 score, for the SMH/Deepslide test set with nuclei annotations is shown in Table 1 for all experiments (Figure 5 shows the corresponding F1 distributions). UV-Net_310GS-50SS is the naming convention used to indicate UV-NET was trained with 50 SS and 310 GS masks. As compared to the traditional models (UNET, PiNET, MultiResU-Net), UV-NET has the highest performance over all GS models. The lowest F1 score is achieved by the IHCCH method with a mean F1 of 63.5% on Ki67$^-$ and 61.8% on Ki67$^+$. UV-NET with 50 SS shows reduced performance compared to the GS-only model for both cell types. However, increasing the number of SS masks improves performance. In particular, when 310SS are added, the model achieves the top F1 score on the Ki67$^-$ channel (F1=83.7%) and the model with 140SS shows the highest performance, with a relatively large margin, for the Ki67$^+$ channel (F1=84.5%).

The corresponding models are tested on the held-out UHN TMA data set (381 TMAs) to further analyze the effect of adding SS from the same data distribution. The PI difference between automated and expert PI estimates, $|\Delta PI|$ is shown in Figure 3. The mean

Table 1: Mean F1-score for SMH/Deepslide test data.

Architecture	# of SS masks	F1(Ki67$^-$)	F1(Ki67$^+$)
IHCCH	-	0.635	0.618
piNet_GS	-	0.749	0.783
MultiResU-Net_GS	-	0.0.777	0.0.807
U-Net_GS	-	0.747	0.782
UV-Net_GS	-	0.833	0.820
UV-Net_310GS-50SS	50	0.739	0.801
UV-Net_310GS-140SS	140	0.818	**0.845**
UV-Net_310GS-186SS	186	0.811	0.827
UV-Net_310GS-310SS	310	**0.837**	0.817

performance of models with SS+GS are similar to the GS only model, with the best performance seen by UV-Net_GS with an average PI difference of 4.86%. MultiResU-Net and U-Net show the lowest performance over all PI intervals. As shown by Figure 3b, the PI differences are lowest in the clinically important ranges ([10-20] and [20-30]%) for the SS model UV-Net_310GS-186SS. A meta-analysis (Petrelli et al., 2015) shows high Ki-67 PI levels ($> 10\%$) are associated with $> 50\%$ risk of death among patients with early breast cancer, particularly in those with ER(estrogen receptors)+ disease. For any 10% increase of Ki-67 level, there is a significant 19% increase in mortality (Petrelli et al., 2015). A PI $> 25\%$ is associated with a greater risk of death (Petrelli et al., 2015). A Ki-67 threshold of $> 25\%$ is associated with the most powerful outcome when prognostication is of concern (Petrelli et al., 2015). Therefore, the 186SS model has optimal performance in these clinically relevant ranges which could positively impact patient care. Shortly behind the 186SS model, is the 310SS model, with similar accuracy in associated ranges.

To analyze consistency in the PI ranges, the CoV of $|\Delta PI|$ is shown in Figure 4a and the difference in CoV $|\Delta CoV|$ for two neighboring PI ranges is explored in Figure 4b. The consistent models score low CoV over different PI intervals. The model with 310 SS masks consistently has lower variability compared to UV-Net with GS masks demonstrating that adding SS masks has improved consistency over all PI ranges. When considering the CoV difference, $|\Delta CoV|$, highly consistent methods would have low differences in CoV between PI ranges. This indicates that the error rate and the prediction performance is comparable over all PI ranges and is therefore, more reliable and repeatable. UV-Net_310GS-310SS scores the highest consistency by achieving $|\Delta CoV|$ near zero for over clinically relevant ranges ($PI > 10\%$), which was not achieved by UV-Net_GS. Thus, with the addition of SS masks, there is an increase in performance consistency (see Figure 4b), in that the algorithm behaves the same over all disease levels with a significant improvement comparing to UV-Net_GS. This is at the expense of a slightly lower PI estimation performance. Therefore, one suggestion might be to match the number of GS masks with that of SS ones, although this needs to be tested further in the future.

Ideally, a clinical tool should have high and similar performance, i.e. produces consistent results, over different clinically relevant ranges to ensure reliability and repeatably. If a tool is consistent, the same diagnostic accuracy and management would be offered to patients,

regardless of the PI. However, if the tool has variable performance over different PI ranges, there is variability in the predictions which reduces quality of care. Deep learning is known to suffer from generalization issues for images that are out-of-distribution. This can manifest in different ways, including unwanted variability in the predictions. For medical imaging applications, this can be a significant barrier for practical deployment at new labs, which have different scanners and staining protocols. Therefore, this work proposes a way to mitigate these challenges for Ki-67 PI estimation. For qualitative purposes Figure 8 shows an example TMA, processed by UV-Net_GS and UV-Net_310GS-186SS. Additionally, Figure 9 shows three patches across all architectures.

In the future, there are a number of ways to improve on the work. First, additional tests will be conducted to determine the optimal of number of SS for the tasks, perhaps by considering both consistency and accuracy at the same time, only in the clinically relevant intervals. However, we believe that the relationship will not be a linear one, due to the non-linearity of the systems, as well as the fact that there are separate predictions for both Ki67+ and Ki67- nuclei. Additionally, it is possible to increase the number of SS images in smaller increments and also include a total amount that is larger than the GS dataset to determine the impact on prediction performance when the SS masks dominate the training mix. It is also worth consider refining the patches that were selected for the SS mask generation, or manual removal of poor SS masks as well as other unsupervised tools to improve SS mask generation (although there are minimal works in the literature for Ki67). Other ideas include sampling more patients to enhance the dataset variability since currently, we only used 31 TMAs to develop the 310 patches. Perhaps more patient-diversity would improve generalization further. Lastly, a final goal of this work will be to test the SS framework for per-site calibration for whole slide images, over multiple labs. If time permits, we may compare this to style-transfer and GAN-based domain adaptation techniques.

5. Conclusion

This paper introduces the integration of silver standard (SS) masks from an unseen center, generated by an unsupervised Ki67 nuclei detection algorithm (IHCCH), along with gold standard (GS) masks from a different site, generated by expert pathologists for Ki67 PI quantification. It was found that adding SS masks from the unseen center to the deep-learning model UV-Net improved performance over clinically relevant PI ranges. The architecture with 310 SS shows a remarkable consistency for PI ranges above 10% while a slight degradation of F1-score accuracy. Since the SS masks are simple and fast to generate, it is possible to use this method to fine-tune algorithms and at new centers that are deploying AI algorithms to improve consistency and reliability of Ki67 quantification. This will be key for clinical translation.

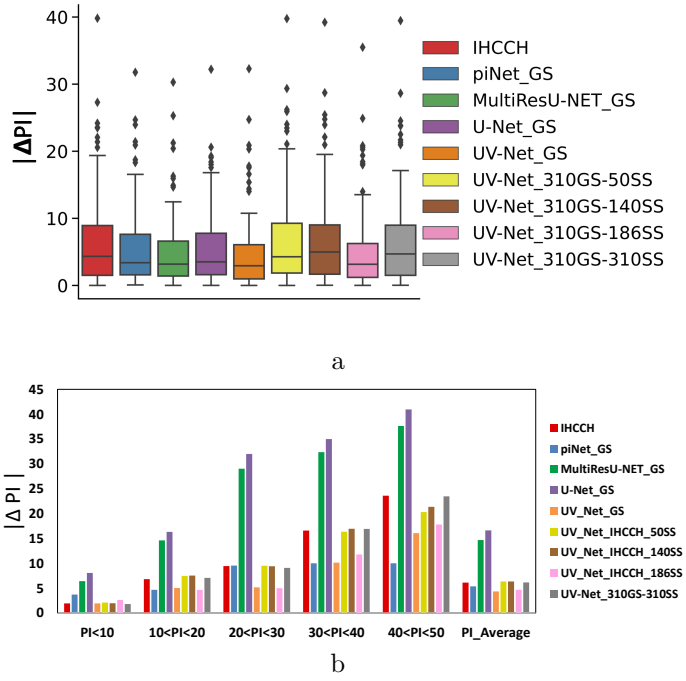


Figure 3: UHN TMA dataset: a. PI difference across all architectures and b. PI ranges.

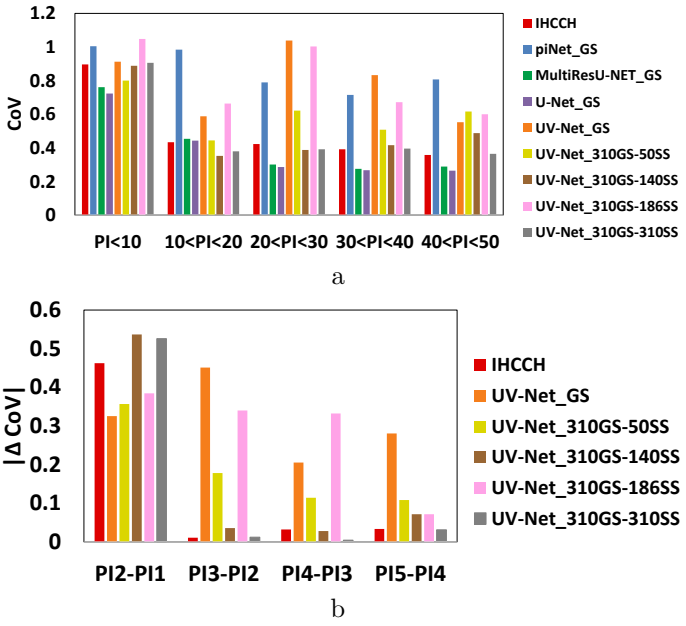


Figure 4: Coefficient of variation. a. across different ranges b. CoV difference, $|\Delta CoV|$. PI1-PI2 refers to the calculated $|\Delta CoV|$ across [0-10] and [10-20]%

References

Mohamed Amgad, Lamees Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha Elsebaie, Ahmed Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid Elmatboly, Philip Pappalardo, et al. Nucls: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. *arXiv preprint arXiv:2102.09099*, 2021.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer journal For Clinicians*, 68(6):394–424, 2018.

Mitch Dowsett, Torsten Nielsen, Roger A'Hern, John Bartlett, Charles Coombes, Jack Cuzick, Matthew Ellis, Lynn Henry, Judith Hugh, Tracy Lively, et al. Assessment of ki67 in breast cancer: recommendations from the international ki67 in breast cancer working group. *Journal of the National Cancer Institute*, 103(22):1656–1664, 2011.

Christopher Elston and Ian Ellis. Pathological prognostic factors in breast cancer. I. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

Rokshana S Geread, Peter Morreale, Robert D Dony, Emily Brouwer, Geoffrey A Wood, Dimitrios Androutsos, and April Khademi. Ihc color histograms for unsupervised Ki67 proliferation index calculation. *Frontiers in bioengineering and biotechnology*, 7:226, 2019.

Rokshana Stephny Geread, Abishika Sivanandarajah, Emily Brouwer, Geoffrey Wood, Dimitrios Androutsos, Hala Faragalla, and April Khademi. Pinet–an automated proliferation index calculator framework for Ki67 breast cancer images. *Cancers*, 13(1):11, 2021.

Simon Graham, Quoc Dang Vu, Shan Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020.

P Jalava, T Kuopio, L Juntti-Patinen, T Kotkansalo, P Kronqvist, and Y Collan. Ki67 immunohistochemistry: a valuable marker in prognostication but with a risk of misclassification: proliferation subgroups formed based on ki67 immunoreactivity and standardized mitotic index. *Histopathology*, 48(6):674–682, 2006.

Seyed Hossein Mirjahanmardi, Melanie Dawe, Anthony Fyles, Wei Shi, Fei-Fei Liu, Susan Done, and April Khademi. Preserving dense features for ki67 nuclei detection. *Accepted to SPIE Medical Imaging*, 2022.

Harsh Mohan. *Textbook of pathology*. Jaypee Brothers Medical Publishers, 2018.

Fausto Petrelli, G Viale, M Cabiddu, and S Barni. Prognostic value of different cut-off levels of ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast cancer research and treatment*, 153(3):477–491, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.

C. Senaras. Deepslides dataset. *Zenodo, CERN: Meyrin, Switzerland*, 2018.

Chetan Srinidhi, Ozan Ciga, and Anne Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, page 101813, 2020.

Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5):775–784, 2021.

SM Veronese, M Gambacorta, O Gottardi, F Scanzi, M Ferrari, and P Lampertico. Proliferation index as a prognostic marker in breast cancer. *Cancer*, 71(12):3926–3931, 1993.

S Walters, C Maringe, J Butler, B Rachet, P Barrett-Lee, JPWVD Bergh, J Boyages, P Christiansen, M Lee, Fredrik Wärnberg, et al. Breast cancer survival and stage at diagnosis in australia, canada, denmark, norway, sweden and the uk, 2000-2007: a population-based study. *British Journal of Cancer*, 108(5):1195–1208, 2013.

Fuyong Xing, Toby C Cornish, Tell Bennett, Debashis Ghosh, and Lin Yang. Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in ki67 images. *IEEE Transactions on Biomedical Engineering*, 66(11):3088–3097, 2019.

6. Appendix

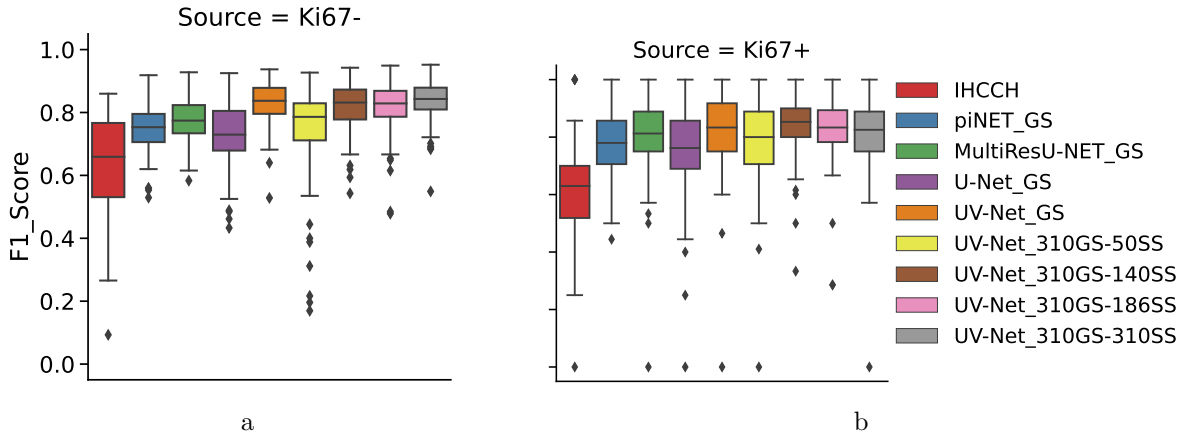


Figure 5: F1 distribution of trained architectures when different amount of SS masks are added to GS data and tested on SMH/Deepslide dataset. a. Ki67^{-} b. Ki67^{+}.

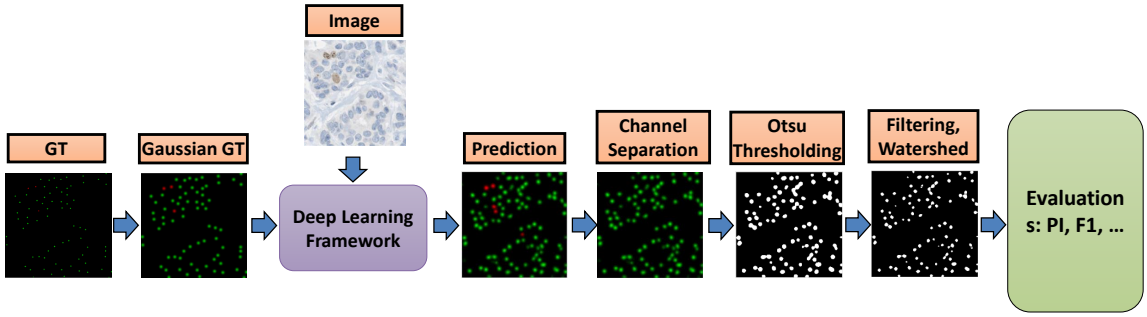


Figure 6: Processing pipeline. The process includes three steps, pre-processing (GT, Gaussian GT), deep learning framework, and post-processing (channel separation, Otsu thresholding, median filter, and watershed).

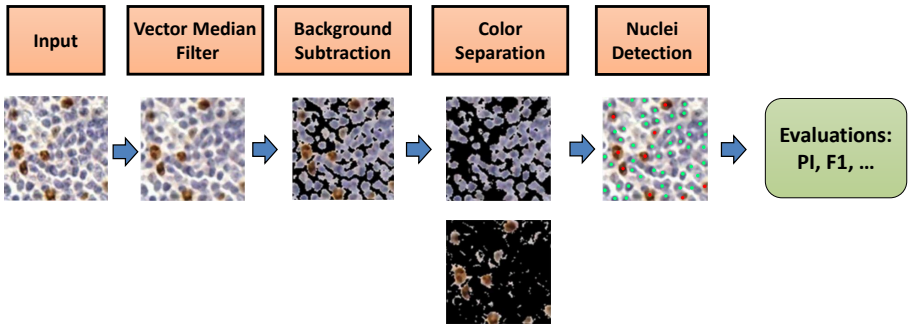
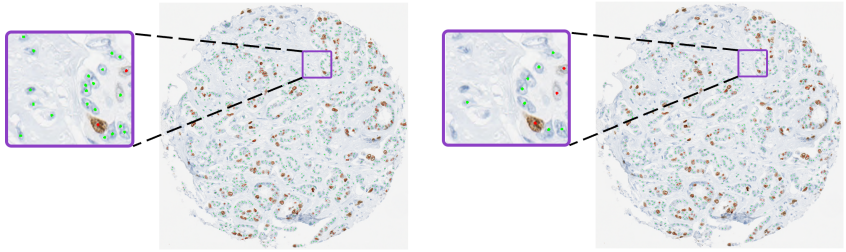


Figure 7: IHCCH unsupervised processing pipeline. The process includes vector median filtering, background subtraction, channel separation, and adaptive radius nuclei detection. No ground truth mask is needed here. More information can be found in (Geread et al., 2019).



a b

Figure 8: An example of one TMA image with ground truth PI=25. a. UV-Net_GS b. UV-Net_310GS-186SS. Ki67$^-$ nuclei are marked with green and Ki67$^+$ with red.

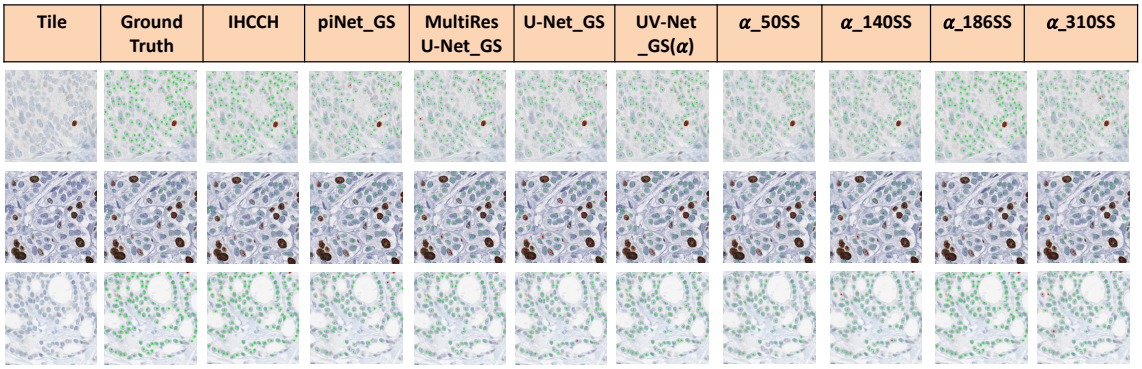


Figure 9: Qualitative results obtained on multiple tiles from SMH/Deepslide dataset. α stands for UV-Net_310GS.