Unsupervised Domain Adaptation for Ki-67 Scoring in Digital Pathology: A Step Closer to Widescale Deployment

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

The Ki-67 proliferation index (PI) is a predictive and prognostic biomarker of invasive breast cancer. Unfortunately, domain shifts between source and target data can create challenges for deep learning systems and automatic PI scoring. This study presents a self-supervised domain adaptation pipeline for automatic Ki-67 quantification using two Ki-67 scoring architectures (UV-Net and piNET). An image processing method is used to generate silver standard (SS) data from the target domain, which is used to augment the gold standard (GS) data from the source domain. Results reveal models trained on SS (pseudo-labelled) data from the target domain and fine-tuned on GS data from the source domain performed better than models trained on GS data alone. The PI estimation accuracy increased by $\sim 3.75\%$ and was more consistent when using the models pre-trained on SS data. All experiments were validated on large multi-institutional datasets (3 centers and 20970 patches). The proposed pipeline can be used to easily and efficiently incorporate data from the target domain, to reduce this generalization gap. The methodology has wide applications in many AI for medical imaging tasks and can accelerate adoption and translation opportunities.

Keywords: Self-supervised, Ki-67, digital pathology, unsupervised domain adaptation

1. Introduction

Breast cancer is the most diagnosed cancer and the leading cause of cancer death in women worldwide (Sung et al., 2022). Ki-67 immunohistochemistry (IHC) biomarker is gaining traction for evaluating the proliferation rate of invasive breast cancers (Davey et al., 2021; Dowsett et al., 2011). Ki-67 expression is related to prognosis, can identify high-risk early-stage breast cancers (Penault-Llorca and Radosevic-Robin, 2017; Dowsett et al., 2011), and is used to determine treatment modalities (Penault-Llorca and Radosevic-Robin, 2017; Davey et al., 2021; Dowsett et al., 2011). Ki-67 proliferation index (PI) is the score used to assess the proportion of Ki-67 positive tumour cells to the total number of tumour cells in a breast tissue section (Walters et al., 2013). Unfortunately, quantifying this biomarker is labour-intensive, time-consuming, and vulnerable to errors (Walters et al., 2013; Davey et al., 2021).

Ki-67 PI scores can be calculated with deep learning (DL) nuclei detection algorithms for more efficient and objective quantification. There have been a few tools addressing automated Ki-67 PI scoring in literature, such as piNET (Geread et al., 2021) and UV-Net (Mirjahanmardi et al., 2022) which are DL tools specifically developed for Ki-67 PI quantification in breast cancer. As automated artificial intelligence (AI) tools get more robust, there is a chance for translation and deployment. However, the challenge with widescale adoption is performance differences at deployed target sites which happens when the target data is

© 2023 Anonymous.

from a center that is not included in the (source) training set. It is especially evident in digital pathology given the variation in patient factors, specimen processing/staining protocols and acquisition devices across pathology laboratories. Annotations from target sites could be included in training sets, but generating gold standard (GS) ground truths is laborious and expensive for medical imaging.

Mitigating the domain shift has become a topic of extensive research (Wang and Deng, 2018; Guan and Liu, 2022; Zou et al., 2020) and unsupervised domain adaptation (UDA) is gaining considerable attention for this task. UDA methods seek to overcome the domain gap with unlabeled data from the target domain. Self-training (pseudo-label based methods) has emerged as a promising UDA solution (Natarajan et al., 2013; Zou et al., 2018). Self-training generates a set of pseudo-labels in the target domain and then re-trains a network based on the generated pseudo-labels. Self-training loss encourages cross-domain feature alignment by learning from the labelled source data and pseudo-labelled target data. Pseudo-labels can be quickly generated in large amounts for any number of datasets, which is cost-effective and reduces development time. However, perfect accuracy cannot be guaranteed for pseudo-labels which can lead to propagated errors when fine-tuning. Because pseudo-labels do not capture detailed features as well as GS labels we hypothesize that pre-training a network on pseudo-labels from the target data will allow a network to first learn basic, low-level features that are task- and domain-dependent. Fine-tuning with GS from the source data can then allow for more detailed features to be learned by the network. This work proposes a pipeline that (1) uses an unsupervised Ki-67 PI quantification algorithm to automatically generate pseudo-labels, which we call silver standard (SS) masks, in the unlabeled target domain, (2) pre-trains a network on SS masks, and (3) fine-tunes the network on GS data from the source domain. This pipeline can be used to calibrate automated DL-based medical imaging tools on a per-dataset basis, in an easy and unsupervised manner. We validate our method on 325 clinical tissue microarrays (TMAs) (20800 patches) from a target domain. Experimental results show the proposed approach achieves exceptional performance on the pixel-level and patient-level, therefore, providing UDA for robust and accurate Ki-67 PI estimation.

2. Methods

2.1. Deep Learning Models

Two DL architectures are used for experiments: UV-Net and piNET, both developed for Ki-67 PI quantification in breast cancer and validated on large multi-institutional datasets. piNET was built on the original U-NET architecture with an extra layer (see Figure 3). The UV-Net architecture was designed to preserve nuclear features of clustering or overlapping nuclei through dense 'V' blocks to retain the high-resolution details (Mirjahanmardi et al., 2022) (see Figure 4). UV-Net has shown high accuracy with strong generalization capabilities that outperform other networks such as U-Net, DenseU-Net, and MultiresU-Net (Mirjahanmardi et al., 2022). The output of piNET and UV-Net is a multi-channel probability map, with center locations of tumour nuclei detected for two classes of Ki-67$^-$ and Ki-67$^+$ cells. Otsu's thresholding is used to convert the probability maps into binary representations, median filtering is applied to remove small and irrelevant false positives, and water-shedding is applied to separate any possible overlapping predictions (see Figure 5). Experiments are conducted with an NVIDIA GeForce RTX 3070 Ti. 100 epochs are used

with an Adam optimizer, batch size of 4, and learning rate of 10^3. The Huber loss function is used for all architectures to regress and predict the centroids of nuclei. All models are validated on 170 GS patches and the epoch with the lowest validation loss was saved. Data augmentations such as horizontal/vertical flips and scaling were used.

2.2. Transfer Learning

Transfer learning (TL) (Pan and Yang, 2010; Weiss et al., 2016) has proven to be effective for many real-world applications by exploiting knowledge in labelled training data from a source domain. TL has made major contributions to medical image analysis as it overcomes the data scarcity problem as well as preserving time and hardware resources. By supplementing training with extensive amounts of labelled data from a related source domain, a model's ability to predict target instances can be improved. In this study, we propose a method of using an unsupervised Ki-67 nuclei detection scheme to generate SS masks in the unlabeled target domain. These SS masks are used to pre-train a base model, allowing it to learn low-level nuclei features and achieve optimal parameter initialization. GS masks are then utilized to fine-tune the model, resulting in the attainment of more precise feature details. We deploy two network architectures (UV-Net and piNET) and evaluate their performance when (1) training with GS Only, (2) training with SS Only, (3) training with a mixture of GS and SS, and (4) performing transfer learning with GS on pre-trained SS Only models.

2.3. Target Domain Silver Standard Masks

An unsupervised Ki-67 nuclei detection scheme called the immunohistochemical colour histogram (IHCCH) approach is used to generate SS masks in the target domain. IHCCH is a colour separation method that separates blue and brown objects automatically. 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 colour 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. All objects are detected with connected components and the nuclei radius is estimated based on a percentile of 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 6. Code for the IHCCH algorithm is available at `https://github.com/IAMLAB-Ryerson/IHCCH`.

2.4. Dataset

This study uses Ki-67 stained invasive breast cancer images obtained from three institutions. The dataset was partitioned into source domain (GS) and target domain (SS masks and unseen target datasets). Table 1 summarizes the Ki-67 datasets used for each model.

2.4.1. Source Dataset

A total of 510 patches with size 256×256 pixels are extracted from whole slide images (WSI) from St. Michael's Hospital (SMH) in Toronto and an open-source database "Deepslide" (Senaras, 2018). The ×20 Aperio AT Turbo and ×40 Aperio ScanScope scanners were used, respectively. Deepslide images are down-sampled to ×20 for compatibility. Images were annotated by marking Ki-67$^-$ and Ki-67$^+$ centroids (Geread et al., 2021). Centroid annotations were converted into a Gaussian kernel to improve learning. 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. This dataset is referred to as the source and GS dataset interchangeably and is used for training ($66.\overline{6}$%) and pixel-level testing ($33.\overline{3}$%).

2.4.2. Target Dataset

The target dataset was provided by the University Health Network (UHN) and contains 411 tissue microarrays (TMA) from 175 patients. Each patient has 1 to 3 corresponding TMAs of 2000 × 2000 pixels in size and an expert PI estimate is available for each patient. 24 TMAs from 24 patients were used to create the SS masks. These 24 TMAs were tiled into patches of 256x256 pixels and 345 patches which contained ≥ 80% tumorous tissue were extracted and the remaining patches were discarded. The TMAs from patients used for SS mask generation were removed from our target dataset to prevent patient data leakage. 10 TMAs were randomly selected from the remaining pool and annotated by an anatomical pathology resident (N.N.J.N) and verified by a breast pathologist (S.D.) to produce pixel-wise nuclei annotations in the target domain to be used for testing. This test set is referred to as the target dataset and contains 325 TMAs (151 patients) with patient-level PI scores and 10 TMAs with nuclei annotations.

2.5. Evaluation Metrics

Nuclei detection is evaluated by comparing the Ki-67$^-$ and Ki-67$^+$ centroids between the prediction and GS ground truths through the F1 score. The F1 score is the harmonic mean of precision and recall which is dependent on the number of true positives (TP), false positives (FP), and false negatives (FN). A TP was detected whenever the Euclidean distance between an annotation centroid and a detected centroid was less than 6 µm. This value corresponds to the average radius of tumourous cells from the source dataset. All detected cells not within 6 µm of a ground truth annotation were considered FP. Multiple detections of an already counted cell were also counted as FP. All ground truth cells without a detection within 6 µm proximity are considered FN. The overall F1 was calculated for each TMA instead of the mean F1 per patch since patches without nuclei would skew this average. Tumour proliferation is measured by the proliferation index (PI):

$$PI = \frac{\#\ Ki-67^+\ tumour\ cells}{\#(Ki-67^+\ +\ Ki-67^-)\ tumour\ cells} \tag{1}$$

which is computed over the whole TMA based on the detected nuclei. The PI difference or PI error is used to investigate the error between predicted and actual PI values: $\Delta PI =$

$|PI_{actual} - PI_{predicted}|$. Bland Altman plots and correlations are used to investigate prediction errors of the top models. Pairwise one-way ANOVA is used to compare model performance.

2.6. Experimental Setup

Four model training configurations are used to study Ki-67 nuclei detection and PI estimation accuracy in the target domain for both piNET and UV-Net. The first configuration is GS only, which uses only the GS data from the source domain (340 patches). Both architectures are trained on the same 340 GS patches. The second configuration is SS only, which uses only the SS data generated by the unsupervised IHCCH algorithm from the target domain. SS Only models are trained with increments of 50 SS; SS50 is trained on 50 SS, SS100 is trained on 100 SS (which includes the previous 50 SS), etc. The third configuration is Mixed which includes both GS and SS in the training pool; Mixed50 is trained on 340 GS and 50 SS, Mixed100 is trained on 340 GS and 100 SS (which includes the previous 50 SS), etc. The last configuration is Tuned which uses SS for pre-training and GS for fine-tuning; Tuned50 is the SS50 network with all layers fine-tuned on 340 GS, Tuned50 is the SS100 network with all layers fine-tuned with 340 GS, etc.

3. Results

Example nuclei detection predictions in source and target domains are shown in Figure 16.

3.1. Source Domain: Nuclei Detection

For the source domain validation set (170 GS patches), the F1 score is calculated per stain and overall. Figure 7 shows F1 score per stain and overall F1 score in Figure 8. Tuned and Mixed UV-net and piNET models show equivalent or better nuclei detection performance on the source domain data, compared to the respective GS Only models for both stains (Hematoxylin and DAB). Overall F1 score follows the same trends. This suggests the addition of SS masks (pseudo-labels) from a different domain in the training pool empirically improves a model's prediction performance. The GS Only models were trained on the same distribution as the validation set and are outperformed by the Mixed and Tuned models. Table 2 shows the p-values for pairwise F1 performance comparisons of GS Only versus other model configurations. All SS only models were statistically different but had subpar nuclei detection performance (except for piNET SS150 for Hematoxylin stain which showed equivalent performance). The top performing models (highest overall F1 score) for each model configuration are as follows: SS200 and SS150 offer the best overall F1 performance for SS Only UV-Net and piNET models, respectively. Mixed300 and Mixed345 offer the best overall F1 performance for Mixed UV-Net and piNET models, respectively. Most notably, the Tuned300 models offer the best overall F1 performance across both architectures with a mean F1 increase of $\sim 4\%$ compared to GS Only performance. These top performing models will be analyzed further in section 3.4.

3.2. Target Domain: Nuclei Detection

The proposed models are tested on 10 TMAs with pixel-level annotations from the target domain. Figure 11 shows the mean nuclei detection performance on the target TMAs (dis-

tributions are shown in Figure 9). The most notable trends when comparing Tuned models to GS Only are a modest improvement of $\sim 10\%$ in detection for UV-Net nuclei, and a large improvement of $\sim 16\%$ in piNET Ki-67$^-$ nuclei. Figure 10 shows overall F1 scores on the target domain. The Mixed and Tuned models predominately achieve higher overall F1 scores (improve 6% to 15%) and lower variability when compared to GS Only models. Table 3 shows the p-values for pairwise F1 performance comparisons of GS Only versus other model configurations. Tuned and Mixed mainly show a significant improvement ($p < 0.05$) in Ki-67$^-$ (Hematoxylin) and overall F1 performance compared to GS Only models.

3.3. Target Domain: PI Computation

PI computation error is assessed on the target dataset that includes 151 patients (and 325 TMAs). Average PI absolute error is shown in Figure 1 (and distributions in Figure 12). GS Only models for both piNET and UV-Net have approximately the same (8%) error. There is a slight improvement (lower PI error) for the SS Only and Mixed models (except for piNET Mixed50), and a moderate improvement for the Tuned models when compared to GS Only models. Table 4 shows the p-values for pairwise PI performance comparisons of GS Only versus other model configurations. All Tuned models show a significant improvement ($p < 0.05$) in PI estimation performance (except for UV-Net Tuned50 which showed equivalent performance). Most notably, UV-Net Tuned300 offers a mean error reduction of $\sim 3.6\%$ and piNET Tuned100 offers a mean error reduction of $\sim 3.75\%$; reducing the absolute error by roughly $\frac{1}{2}$ compared to GS Only models. Figure 13 shows the average standard deviation (std) of the PI performance; all Tuned models display lower std (except for UV-Net Tuned50) which indicates the Tuned models provide more consistent and reliable predictions.

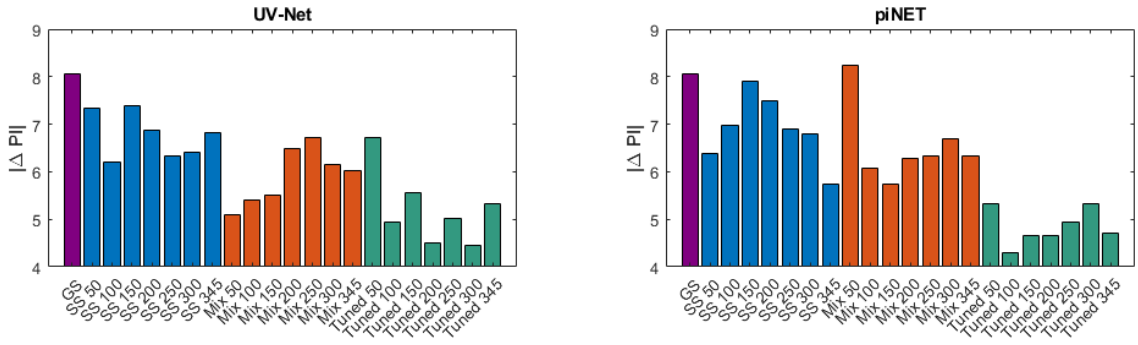

Figure 1: Patient-level mean PI absolute error piNET and UV-Net.

3.4. Target Domain: Post-hoc Analysis

In this section, the top performing models (highest mean F1 on source domain) determined in section 3.1 will be analyzed further. While average performance rates give us a global indication of differences in performance, correlation and Bland Altman plots can be used to deepen our understanding and add another dimension to differences between algorithms on a per-patient level; see Figure 14 and Figure 15. Assessing the UV-Net GS Only model,

there is a low correlation (0.32), large variability (bounds), small slope (0.53) and a high offset (10.7). Mixed and Tuned models show higher correlations (0.64, 0.64), and lower offsets (0.71, 2.71). UV-Net Tuned also presents a larger slope (0.68) indicating better performance in the target domain. For piNET, analogous trends were found (GS Only models have large offset, wide variability and low correlation). Both Mixed and Tuned models had lower offsets (0.86, 2.6) and higher correlations (0.65, 0.63). Although the piNET GS Only model had the slope closest to one (0.97), this improvement is largely offset by the 6.33 bias and wide variability. For both architectures, the Tuned models exhibit a lower standard deviation, near-zero mean, lower y-intercept (bias), and a slope closer to one. This indicates a strong relationship between the predicted PI and the ground truth in the target domain.

The patient-level PI performance was evaluated in intervals of 10% as shown in Figure 2. Tuned models maintain the lowest mean PI error across intervals between 10% and 30% for UV-Net and 10% and 40% for piNET which demonstrates optimal performance in clinically relevant ranges. Furthermore, Tuned models outperform GS Only models for intervals between 0% and 30%. For PI $<$ 10%, Mixed models outperform all models. The high error rate in the interval range 30% to 40% is likely due to the low cohort size (8 patients).

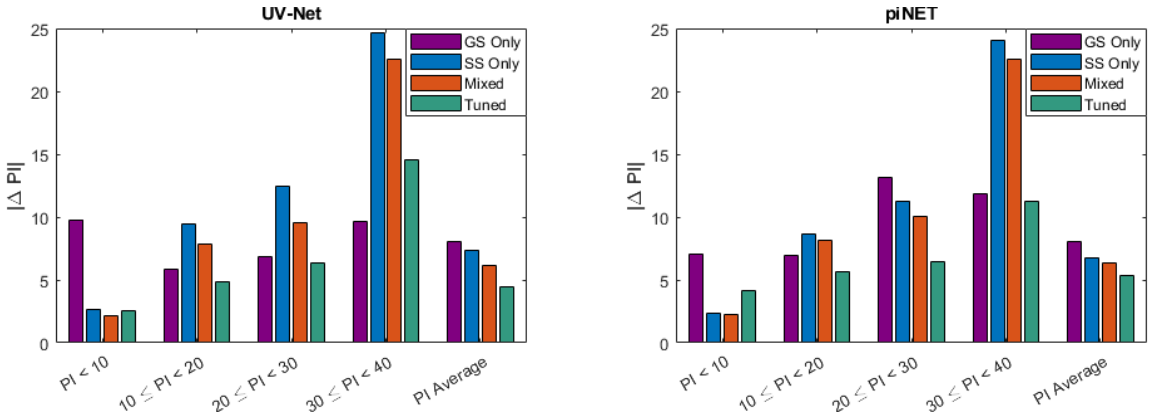

Figure 2: PI absolute error for top performing piNET and UV-Net models. Interval [0 10) has 72 patients, [10 20) has 52 patients, [20 30) has 19 patients, [30 40) has 8 patients. PI average is the mean performance across 151 patients.

4. Discussion

Ki-67 is a clinically relevant biomarker used in early-stage invasive breast cancer to estimate prognosis (Petrelli et al., 2015) and help clinicians decide whether adjuvant chemotherapy should be added to a patient's treatment plan (Nielsen et al., 2021). Various preanalytical, analytical and interpretation factors affect the scoring of Ki-67 and lead to high inter-rater variability (Nielsen et al., 2021) that limits clinical utility. An automated scoring tool for Ki-67 PI would reduce inter-rater variability and allow for a single reliable cut-off point for treatment decision-making. Despite DL methods achieving high performance, there are still technical challenges related to domain shift which limits widescale adoption.

To overcome these challenges and bring us a step closer to widescale deployment of Ki-67 scoring tools, this work proposes an unsupervised domain adaptation method that generates SS in the target domain quickly and efficiently. In contrast to self-training methods which generate pseudo-labels corresponding to high prediction scores in the target domain and re-training a network based on the generated pseudo-labels, our pipeline generates models from scratch with SS data from the target domain allowing networks to learn domain-specific and problem-specific global trends. We refine the models by fine-tuning with GS data (clean labels) to attain more precise feature details and improve pixel-level and patient-level performance. Additionally, existing self-training approaches produce sparse pseudo-labels because only confident predictions are used, this is critical because the insufficient training signals lead to suboptimal, error-prone models (Shin et al., 2020). Our pipeline generates SS from the target site efficiently and can be deployed to any number of laboratories with ease. It was shown that training with the addition of SS from a different domain achieves the same or increased nuclei detection performance on the source domain. More importantly, in the target domain, there was a statistically significant increase in PI computation accuracy, increased correlation to patient-level ground truths, lower prediction variability and lower prediction bias. We suspect this is attributed to the statistically significant increase in nuclei detection performance in the Hematoxylin channel (Ki-67^{-}) on the target domain. The PI divisor (Equation 1) is based on the Hematoxylin nuclei count and differences can strongly impact the PI score. Our proposed pipeline allows for learning the distribution of the target centre without manual annotations; this holds the potential to greatly simplify widescale deployment not only for Ki-67 but for many more applications.

In our work, we evaluated three training configurations on two network architectures for DA prediction tasks and showed increased performance compared to GS models. We note there is much room to improve the current performance and deepen our understanding of the influence of SS. For further improvements, we can increase the number of SS beyond the GS dataset to determine the impact when the SS dominate the training ratio. We can consider refining the patch selection for the SS generation and removing sub-optimal SS. Increasing the patient cohort for SS generation can also enhance the dataset variability as patient diversity would improve domain adaptation outcomes. For future investigations, we will test the SS framework for per-site calibration on different sites and image types.

5. Conclusion

In this study, we address the problem of DA by proposing a self-supervised pipeline for automatic Ki-67 quantification in invasive breast cancer. The results showed that models trained on SS from the source domain and fine-tuned on GS data from the target domain improved in performance (for source and target domains) compared to models trained solely on GS data. Extensive experimental results demonstrated that the inclusion of self-supervised training improved a model's robustness to domain variations and enhanced its adaptation across unseen data distributions. Our pipeline enhances the reliability of AI tools when deployed to new domains and further, solves the problem of lacking labelled data and the expensive cost of manual annotations. These findings can help increase the widescale clinical adoption of automated quantification tools in digital pathology.

Acknowledgments

Acknowledgments withheld.

References

Matthew G. Davey, Sean O. Hynes, Michael J. Kerin, Nicola Miller, and Aoife J. Lowery. Ki-67 as a prognostic biomarker in invasive breast cancer. *Cancers (Basel)*, 13(17):4455, 2021.

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.

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.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, Mar 2022. doi: 10.1109/tbme.2021.3117407.

Seyed Hossein Mirjahanmardi, Melanie Dawe, Anthony Fyles, Wei Shi, Fei-Fei Liu, Susan Done, and April Khademi. Preserving dense features for Ki67 nuclei detection. In John E. Tomaszewski, Aaron D. Ward, and Richard M. Levenson M.D., editors, *Medical Imaging 2022: Digital and Computational Pathology*, volume 12039, page 120390Y. International Society for Optics and Photonics, SPIE, 2022. doi: 10.1117/12.2611212.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Torsten O Nielsen, Samuel C. Y Leung, David L Rimm, Andrew Dodson, Balazs Acs, Sunil Badve, Carsten Denkert, Matthew J Ellis, Susan Fineberg, Margaret Flowers, Hans H Kreipe, Anne-Vibeke Laenkholm, Hongchao Pan, Frédérique M Penault-Llorca, Mei-Yin Polley, Roberto Salgado, Ian E Smith, Tomoharu Sugie, John M. S Bartlett, Lisa M McShane, Mitch Dowsett, and Daniel F Hayes. Ki-67 as a prognostic biomarker in invasive breast cancer. *J Natl Cancer Inst*, 113(7):808–819, 2021.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Frederique Penault-Llorca and Nina Radosevic-Robin. Ki-67 as a prognostic biomarker in invasive breast cancer. *Pathology*, 49(2):166–171, 2017.

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.

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

Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 532–548, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58601-0.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal For Clinicians*, 71(3):209–249, 2022.

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.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.05.083.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), 2016. doi: 10.1186/s40537-016-0043-6.

Danbing Zou, Qikui Zhu, and Pingkun Yan. Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3291–3298. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/455. Main track.

Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 297–313, Cham, 2018. Springer International Publishing.

6. Appendix

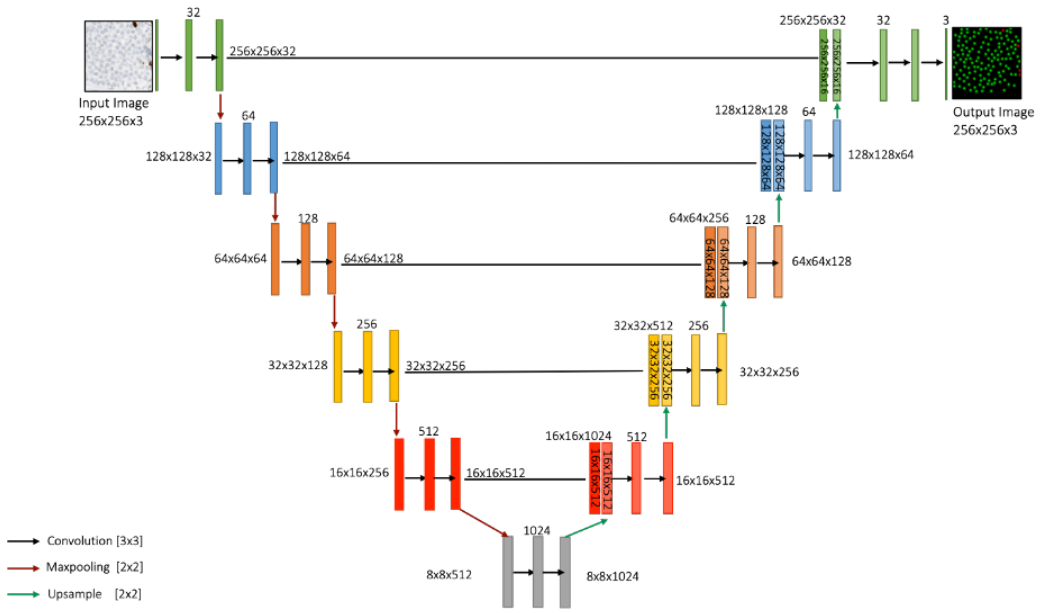

Figure 3: The piNET architecture. Ki-67$^+$ / Ki-67$^-$ detection is performed through a U-NET model.

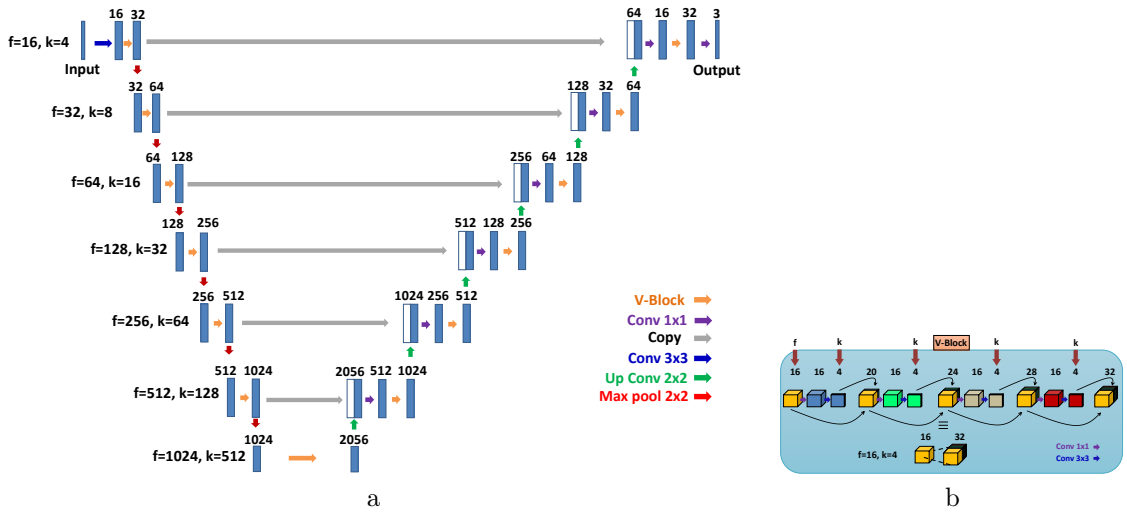

Figure 4: The UV-Net architecture is made up of encoding and decoding arms. a. network b. One V-Block example where f=16, and k=4, composed of four stages.

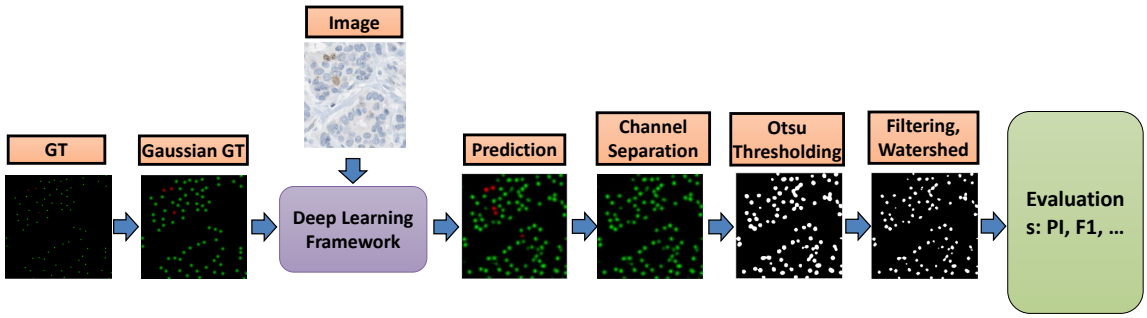

Figure 5: Processing pipeline. The process includes three steps, pre-processing (GT, Gaussian GT), deep learning framework, and post-processing (channel separation, Otsu thresholding, median filter, and watershed).

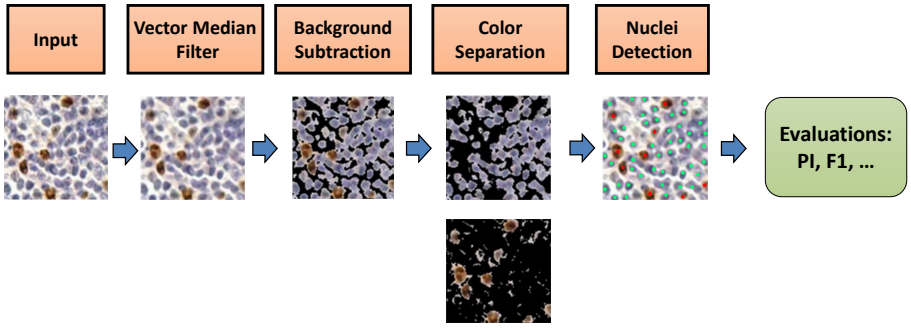

Figure 6: IHCCH unsupervised processing pipeline. The process includes vector median filtering, background subtraction, channel separation, and adaptive radius nuclei detection. No ground truth mask is needed here. More information can be found in (Geread et al., 2019).

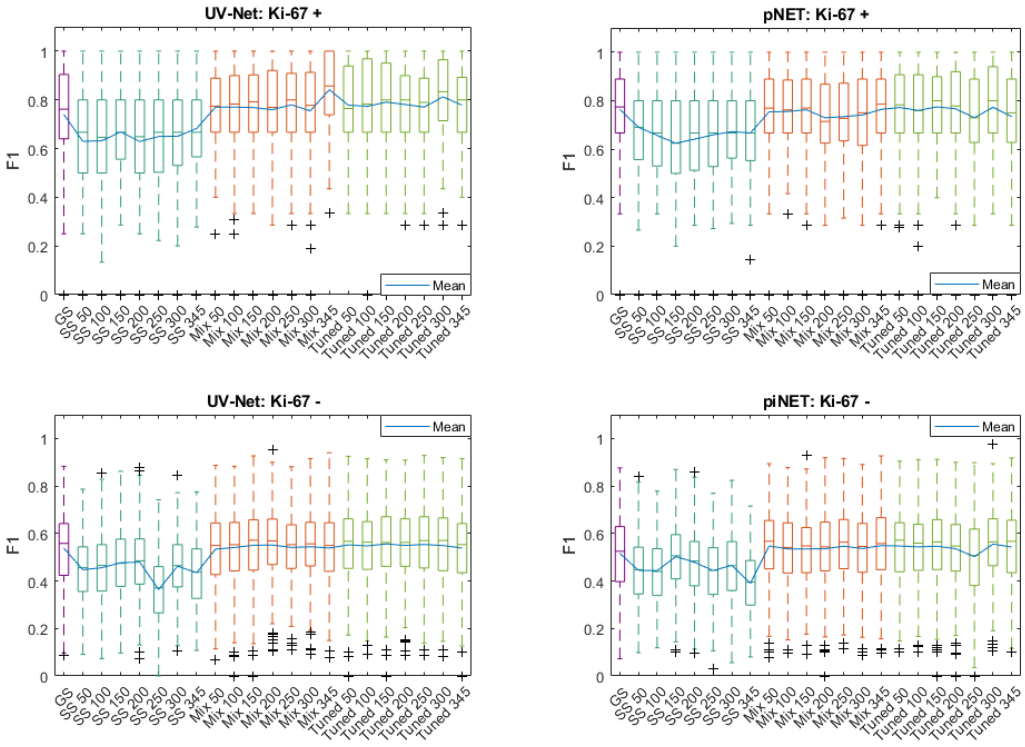

Figure 7: F1 scores for piNET and UV-Net on the GS dataset (SMH and Deepslides) per stain Ki-67$^-$ (Hematoxylin) and Ki-67$^+$ (DAB).

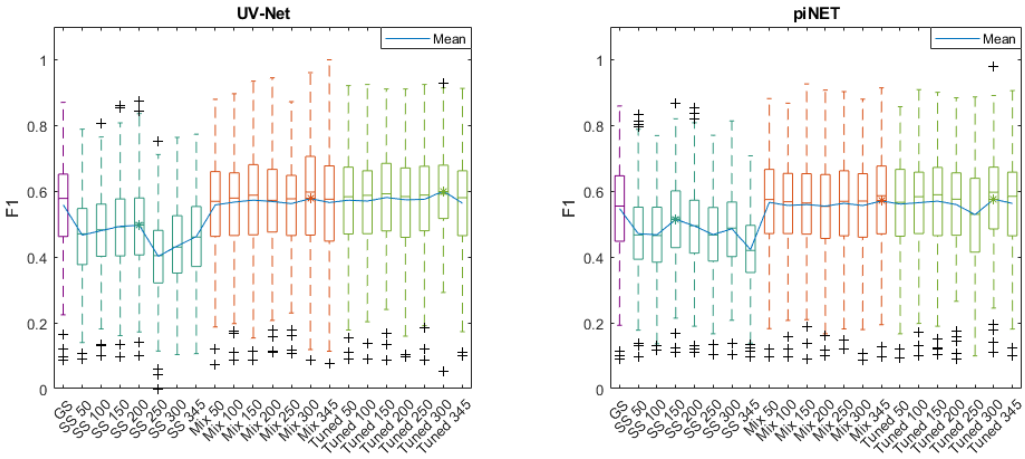

Figure 8: Overall F1 scores for piNET and UV-Net on the GS dataset (SMH and Deepslides). Models from each training configuration with the highest mean F1 performance are marked with an asterisk.

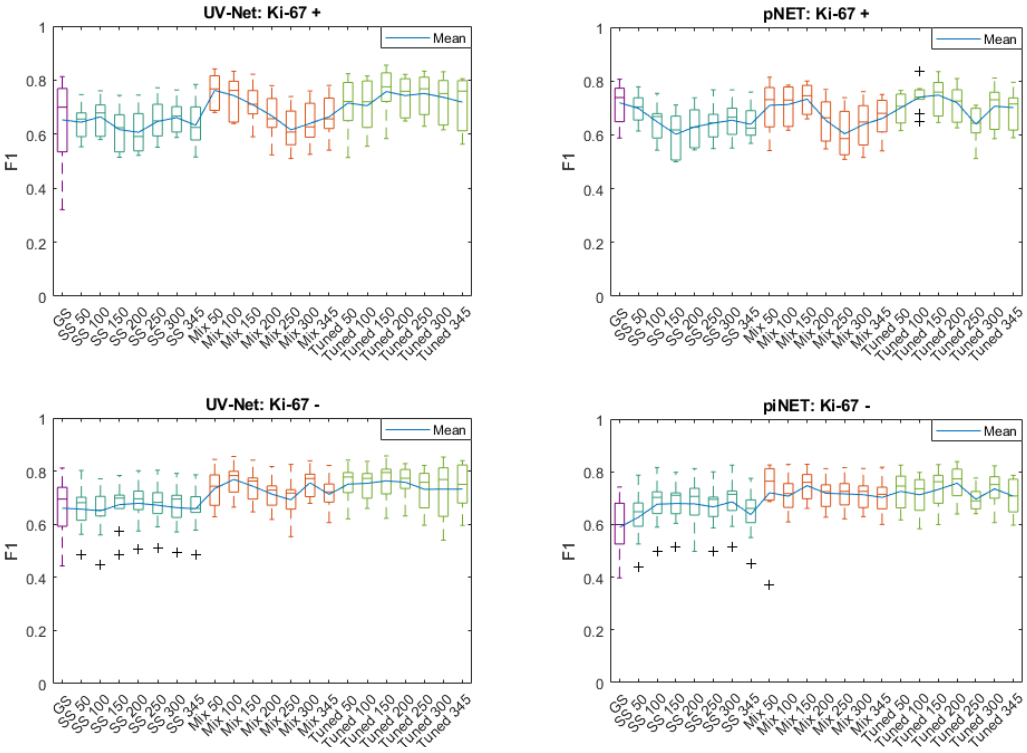

Figure 9: F1 scores for piNET and UV-Net on the target TMAs (UHN) per stain Ki-67^{-} (Hematoxylin) and Ki-67^{+} (DAB).

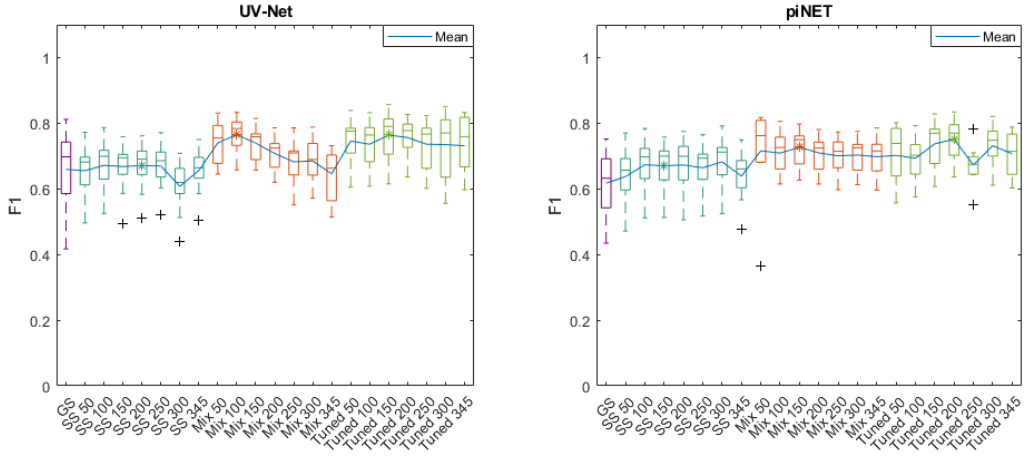

Figure 10: Overall F1 scores for piNET and UV-Net on the target TMAs (UHN). Models from each training configuration with the highest mean F1 performance are marked with an asterisk.

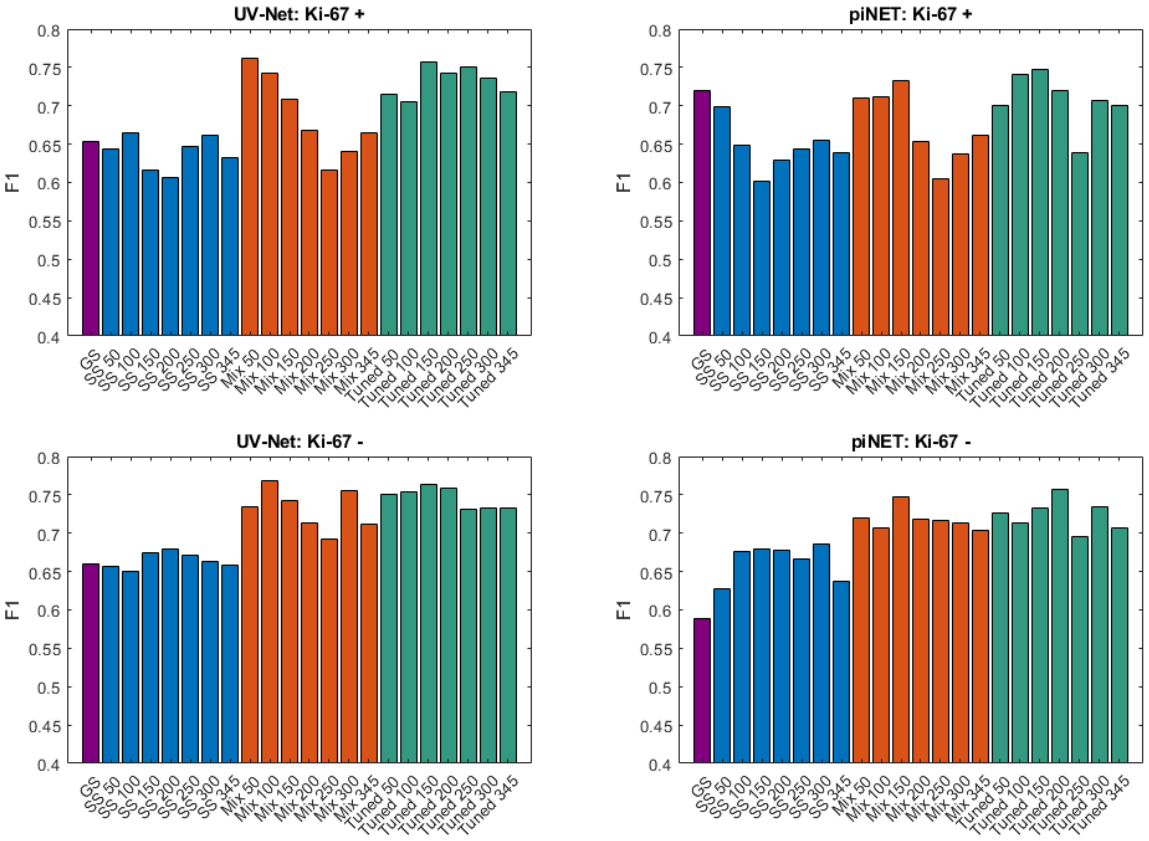

Figure 11: Mean F1 scores for piNET and UV-Net on target TMAs (UHN) per stain Ki-67^{-} (Hematoxylin) and Ki-67^{+} (DAB).

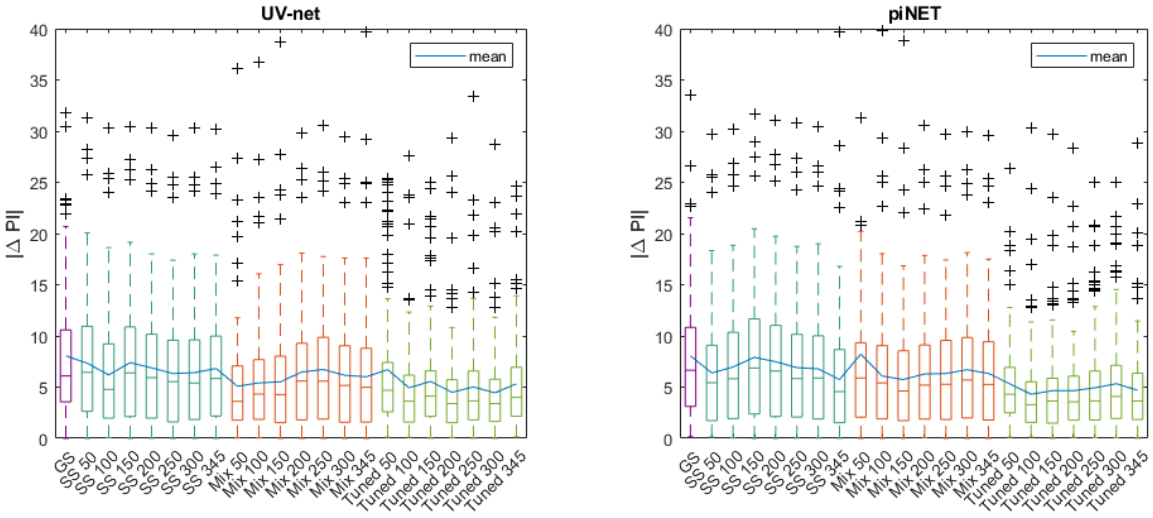

Figure 12: Patient-level PI absolute error for piNET and UV-Net architectures.

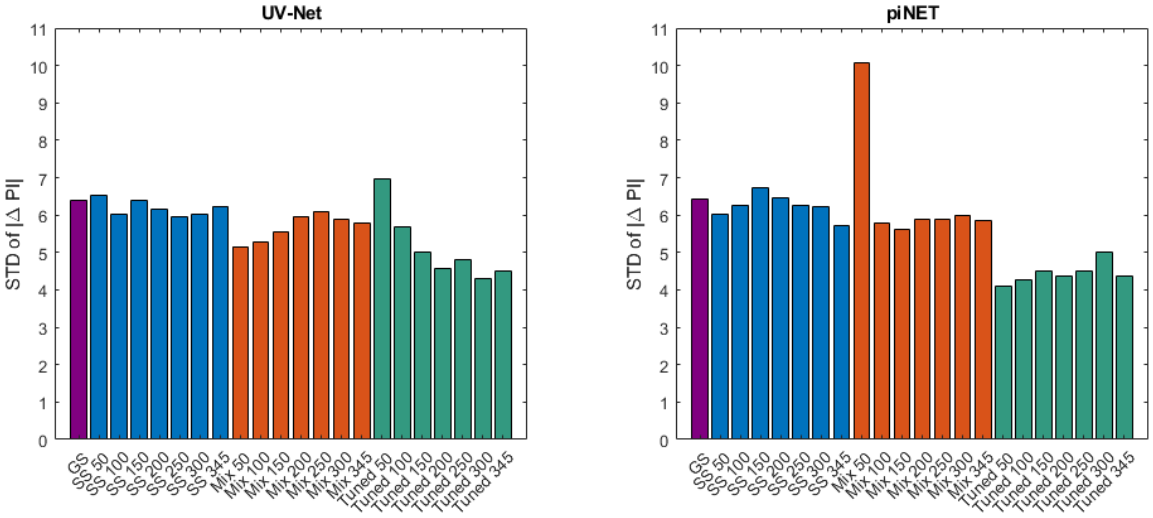

Figure 13: Standard deviation of patient-level PI absolute error for piNET and UV-Net architectures.

Model	Source/Target	SS Train	GS Train	GS Tune	Source/Target Test
GS	Source	0	340	0	170/20800
SS50	Target	50	0	0	170/20800
SS100	Target	100	0	0	170/20800
SS150	Target	150	0	0	170/20800
SS200	Target	200	0	0	170/20800
SS250	Target	250	0	0	170/20800
SS300	Target	300	0	0	170/20800
SS345	Target	345	0	0	170/20800
Mixed50	Source/Target	50	340	0	170/20800
Mixed100	Source/Target	100	340	0	170/20800
Mixed150	Source/Target	150	340	0	170/20800
Mixed200	Source/Target	200	340	0	170/20800
Mixed250	Source/Target	250	340	0	170/20800
Mixed300	Source/Target	300	340	0	170/20800
Mixed345	Source/Target	345	340	0	170/20800
Tuned50	Source/Target	50	0	340	170/20800
Tuned100	Source/Target	100	0	340	170/20800
Tuned150	Source/Target	150	0	340	170/20800
Tuned200	Source/Target	200	0	340	170/20800
Tuned250	Source/Target	250	0	340	170/20800
Tuned300	Source/Target	300	0	340	170/20800
Tuned345	Source/Target	345	0	340	170/20800

Table 1: Ki-67 breast cancer digital pathology datasets and the number of patches used for model training, tuning and testing.

Nuclei Detection: Source Domain						
	UV-Net			piNET		
Model	Ki-67$^+$	Ki-67$^-$	Overall	Ki-67$^+$	Ki-67$^-$	Overall
GS vs SS50	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
GS vs SS100	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
GS vs SS150	**<0.01**	**0.03**	**<0.01**	**<0.01**	0.44	**0.04**
GS vs SS200	**<0.01**	**0.04**	**<0.01**	**<0.01**	**0.03**	**<0.01**
GS vs SS250	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
GS vs SS300	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
GS vs SS345	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
GS vs Mixed50	0.71	0.33	0.49	0.64	0.09	0.25
GS vs Mixed100	0.73	0.18	0.23	0.68	0.30	0.56
GS vs Mixed150	0.81	0.07	0.12	0.96	0.29	0.44
GS vs Mixed200	0.86	0.06	0.19	0.10	0.29	0.65
GS vs Mixed250	0.41	0.16	0.35	0.14	0.10	0.33
GS vs Mixed300	0.69	0.12	0.10	0.28	0.26	0.57
GS vs Mixed345	**<0.01**	0.21	0.30	0.98	0.08	0.17
GS vs Tuned50	0.44	0.06	0.13	0.67	0.09	0.40
GS vs Tuned100	0.62	0.09	0.17	0.85	0.14	0.27
GS vs Tuned150	0.16	**0.03**	**0.04**	0.57	0.11	0.18
GS vs Tuned200	0.37	0.07	0.11	0.88	0.28	0.46
GS vs Tuned250	0.74	**0.04**	0.10	0.11	0.46	0.27
GS vs Tuned300	<0.01	0.08	**<0.01**	0.65	**0.03**	0.08
GS vs Tuned345	0.44	0.22	0.32	0.16	0.14	0.33

Table 2: One-way ANOVA p-values for the pairwise comparison of the F1 scores for piNET and UV-Net on the source domain (SMH and Deepslides patches).

Nuclei Detection: Target Domain						
	UV-Net			piNET		
Model	Ki-67$^+$	Ki-67$^-$	Overall	Ki-67$^+$	Ki-67$^-$	Overall
GS vs SS50	0.88	0.94	0.37	0.50	0.44	0.64
GS vs SS100	0.83	0.84	0.19	**0.04**	0.07	0.18
GS vs SS150	0.53	0.78	0.23	**<0.01**	0.06	0.19
GS vs SS200	0.42	0.71	0.20	**0.01**	0.07	0.20
GS vs SS250	0.92	0.81	0.20	**0.03**	0.10	0.25
GS vs SS300	0.88	0.97	0.82	0.06	**0.05**	0.12
GS vs SS345	0.73	0.96	0.34	**0.02**	0.31	0.60
GS vs Mixed50	0.06	0.11	**<0.01**	0.81	**0.03**	0.08
GS vs Mixed100	0.12	**0.02**	**<0.01**	0.84	**0.01**	**0.02**
GS vs Mixed150	0.32	0.07	**<0.01**	0.65	**<0.01**	**0.01**
GS vs Mixed200	0.80	0.24	**0.02**	0.07	**0.01**	**0.02**
GS vs Mixed250	0.51	0.50	0.11	**0.00**	**0.01**	**0.04**
GS vs Mixed300	0.82	**0.04**	0.10	**0.03**	**0.01**	**0.03**
GS vs Mixed345	0.84	0.25	0.49	0.09	**0.02**	**0.04**
GS vs Tuned50	0.30	0.07	**0.01**	0.54	**0.01**	0.07
GS vs Tuned100	0.38	**0.05**	**0.01**	0.48	**0.01**	0.06
GS vs Tuned150	0.08	**0.04**	**<0.01**	0.39	**0.01**	**0.01**
GS vs Tuned200	0.12	**0.04**	**<0.01**	1.00	**0.00**	**<0.01**
GS vs Tuned250	0.09	0.14	**0.01**	**0.02**	**0.02**	0.15
GS vs Tuned300	0.15	0.18	**0.02**	0.72	**<0.01**	**0.01**
GS vs Tuned350	0.28	0.15	**0.02**	0.59	**0.01**	**0.03**

Table 3: One-way ANOVA p-values for the pairwise comparison of the F1 scores for piNET and UV-Net on the target domain (UHN TMAs).

PI Computation: Target Domain		
Model	UV-Net	piNET
GS vs SS50	0.33	**0.02**
GS vs SS100	0.01	0.14
GS vs SS150	0.36	0.86
GS vs SS200	0.10	0.45
GS vs SS250	**0.02**	0.12
GS vs SS300	**0.02**	0.09
GS vs SS345	0.09	**<0.01**
GS vs Mixed50	**<0.01**	0.85
GS vs Mixed100	**<0.01**	**0.01**
GS vs Mixed150	**<0.01**	**<0.01**
GS vs Mixed200	**0.03**	**0.01**
GS vs Mixed250	0.06	**0.02**
GS vs Mixed300	**0.01**	0.06
GS vs Mixed345	**<0.01**	**0.02**
GS vs Tuned50	0.08	**<0.01**
GS vs Tuned100	**<0.01**	**<0.01**
GS vs Tuned150	**<0.01**	**<0.01**
GS vs Tuned200	**<0.01**	**<0.01**
GS vs Tuned250	**<0.01**	**<0.01**
GS vs Tuned300	**<0.01**	**<0.01**
GS vs Tuned345	**<0.01**	**<0.01**

Table 4: One-way ANOVA p-values for the pairwise comparison of the PI absolute error for piNET and UV-Net in UHN target domain (151 patients).

Bland Altman and Correlation Plots: UV-Net

GT vs GS

GT vs Mixed

GT vs Tuned

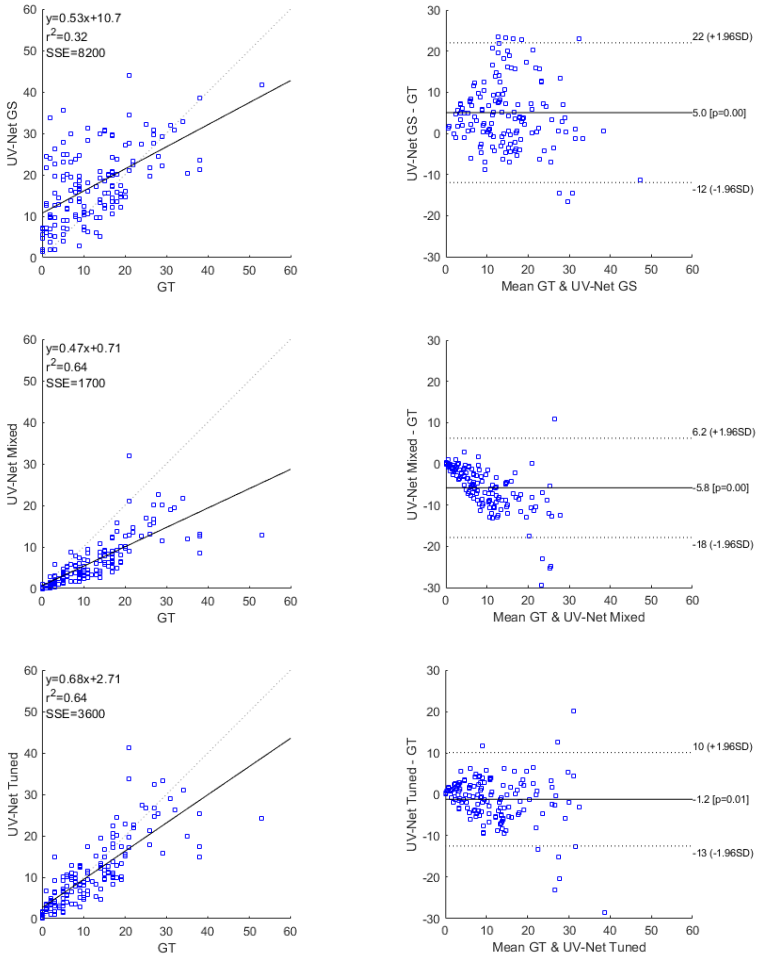

Figure 14: Correlation (left) and Bland Altman (right) plots for average PI error across UV-Net models. To models are used highest overall mean F1 performance.

Bland Altman and Correlation Plots: piNET

GT vs GS

GT vs Mixed

GT vs Tuned

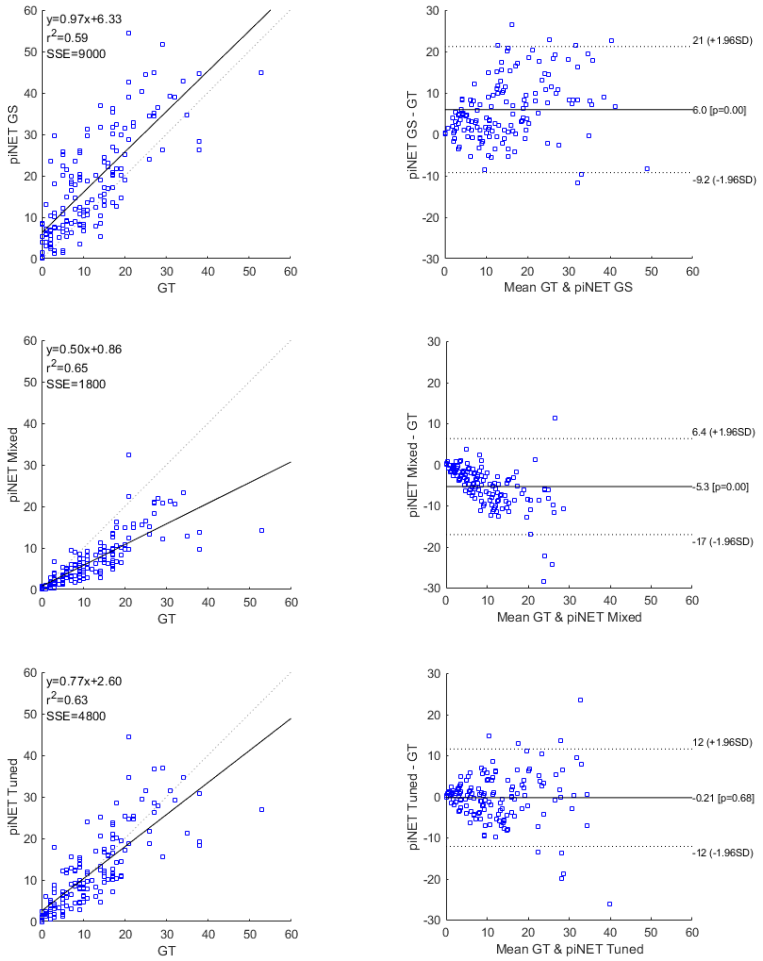

Figure 15: Correlation (left) and Bland Altman (right) plots for average PI error across piNET models. To models are used highest overall mean F1 performance.

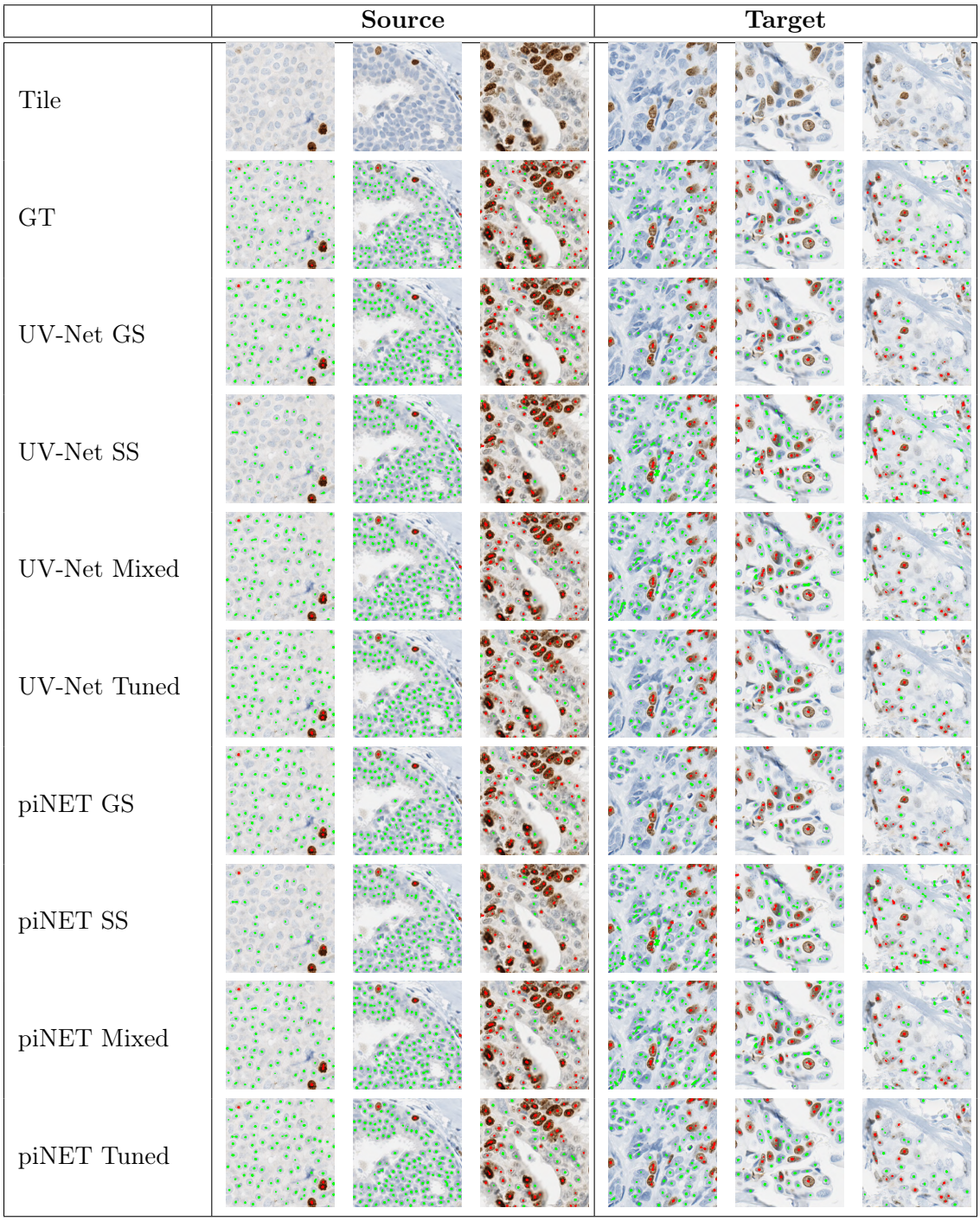

Figure 16: Model predictions on source (SMH/Deepslides) and target (UHN) domains. Green: Ki-67^{-}, Red: Ki-67^{+}. Models selected using the highest overall mean F1 performance.