nnUNet meets pathology: Model analysis and use in quantifying immune infiltrate to predict NSCLC immunotherapy response

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

Editors: Under Review for MIDL 2023

Abstract

Image segmentation is at the core of many tasks in medical imaging, including the engineering of computational biomarkers. Deep learning-based segmentation models can be difficult to train, especially in the case of multi-class segmentation. The self-configuring nnUNet framework has achieved state-of-the-art performance in image segmentation tasks for radiology but has not yet been substantially explored in pathology applications. Here we explore nnUNet for pathology: our contributions are (1) developed additions to the framework to prepare pathology training data for nnUNet, and apply inference on WSI; (2) a new proposed nnUNet configuration and added pathology color augmentations to improve tissue segmentation performance; (3) a lightweight uncertainty extraction approach that proves helpful in biomarker development. We validated our proposed nnUNet configuration and showcase its application on a proof of concept TIL quantification task for immunotherapy response prediction in NSCLC biopsies. We identified critical improvements in the data normalization approach, related batch and patch size configurations, and added pathology-specific color augmentations. All code is made available at ANONYMOUS LINK.[1].

Keywords: nnUNet, computational pathology, NSCLC, TIL biomarker, uncertainty.

1. Introduction

Computer vision tasks such as semantic segmentation and object detection are the foundation of automated image analysis and understanding. In medical imaging, these tasks function at the core of computer-aided detection and diagnosis systems (Gao et al., 2019), as well as building blocks to engineer computational biomarkers in both radiology (Parekh et al., 2020) and histopathology images (Kulkarni et al., 2020). Approaches to develop accurate computer systems for detection and segmentation have been an open question for decades and have been largely investigated by academia and industry. The advent of deep learning (LeCun et al., 2015) has enabled a substantial leap in such a development, powering systems that can reach and even surpass human performance at some medical tasks (Esteva et al., 2017). In semantic image segmentation, the encoder-decoder UNet model (Ronneberger et al., 2015) has increasingly been adopted and has become the state-of-the-art approach. But the devil is in the details: hyperparameters tuning is vital for training accurate UNet models.

In recent years, Auto-ML approaches have been proposed to address this problem (He et al., 2021). In the field of radiology, Isensee et al. introduced the nnUNet framework

1. anonymous url GitHub; full link will be provided upon acceptance of this paper

© 2023 Anonymous.

(Isensee et al., 2020), which relies on self-configuration via internal fingerprints, fixed parameters, empirical decisions and a cross-validation scheme to search for optimal hyperparameters to train a UNet model. As a result, nnUNet outperformed the state of the art across many 3D radiology applications (Antonelli et al., 2022).

Despite the general purpose nature of the framework, nnUNet was mainly designed for radiology applications, while several aspects of pathology data are intrinsically different from radiology. First, the gigapixel image data format of histopathology whole-slide images (WSIs) poses a challenge when entire images need to be loaded to memory for training and inference, resorting to custom dataset preparation and sliding window approaches that are lacking in nnUNet's default framework. Second, multi-class semantic segmentation in histopathology involves a large variety of patterns, cell types and tissue morphologies, which may not be fully taken into account in nnUNet's radiology-oriented default configurations. To accommodate for the high morphological variety, segmentation models for pathology tasks typically use Batch Normalization (BN) with an appropriate batch size, in contrast with nnUNet's preference for large patch sizes and typically smaller batch sizes, in combination with Instance normalization (IN). Additionally, it is hard to capture the full spectrum of tissue and cellular varieties in the training data, while deep learning models tend to over-confidently predict out-of-training-distribution data incorrectly. Therefore, the ability to extract a model's uncertainty can be a beneficial tool to identify patches with poor segmentation performance, when the model is applied on unseen data (Lakshminarayanan et al., 2016). Finally, domain shift is intrinsically part of pathology data, where slide staining varies largely across different pathology labs and even within the same lab in different days of the week; to mitigate this effect, data-specific augmentation techniques are usually adopted (Tellez et al., 2019).

In this paper, we build upon the existing nnUNet framework and, to the best of our knowledge, propose the first application of nnUNet to digital pathology data. We (1) identify and address pathology application specific workflow-related shortcomings of the framework, (2) critically assess nnUNet's model configurations on a multi-class tissue segmentation task on hematoxylin and eosin (H&E)-stained non-small cell lung cancer (NSCLC) WSIs, (3) implement a lightweight uncertainty extraction method to identify likely incorrectly segmented regions.

We assess the validity of the proposed framework on a tumor-infiltrating lymphocytes (TILs) quantification task in a cohort of metastatic non-small cell lung cancer (NSCLC) cases treated with immunotherapy from a phase-2 trial (Theelen et al., 2019). While immunotherapy treatment response prediction is still largely dependent on the suboptimal PD-L1 based stratification, TILs are a potential predictive biomarker which has gained increased attention lately (Park et al., 2022). While initially intra-tumoral TILs ($_i$TILs) were deemed most relevant, (Salgado et al., 2015) proposed a TIL scoring protocol for breast cancer, where TIL densities are visually estimated in the stromal regions of the tumor ($_s$TILs), while excluding necrotic areas. Since this approach depends on the identification of multiple tissue types, we adopted it as a good proof of concept segmentation task for nnUNet's use in biomarker development.

2. Methods

We first introduce the data used during training, validation, and downstream TILs analysis. Then, we introduce nnUNet for pathology. We focus on added technical and methodological

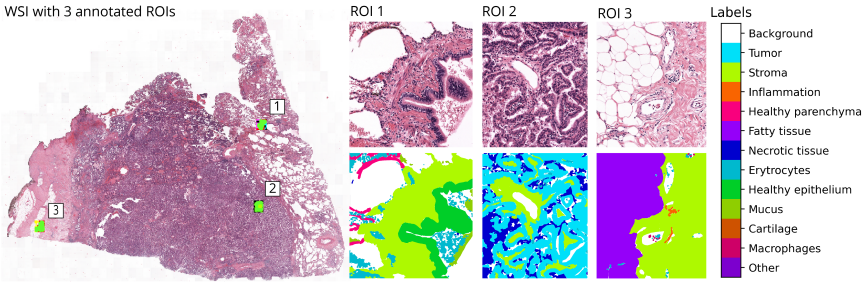

Figure 1: Visual example a WSI with 3 annotated ROIs; labels legend indicates names of annotated classes and their corresponding colors.

contributions to the framework, and go over the experiments carried out in an ablation study, to verify the proposed changes result in an enhanced segmentation performance over the default nnUNet framework. Finally, we detail how the proposed nnUnet framework was used to develop predictive TIL biomarkers of immunotherapy response in NSCLC.

2.1. Materials

Training data. We collected a set of n=30 H&E stained whole-slide images of non-small cell lung cancer tissue samples (21 surgical resections, 9 biopsies) from ANONYMOUS medical center, which were scanned at 40× magnification (0.25 μm/px spacing) using a 3DHISTEC P1000 scanner. Cases were selected to contain a balanced amount of NSCLC subtypes (56% adenocarcinoma, AD; 44% squamous cell carcinoma, SC), AD growth patterns (including lepidic, acinar, papillary, micropapillary, and solid AD morphologies), as well as keratinizing and non-keratinizing cases for SC. For each WSI, approximately n=5 regions of interest (ROI) of 2×2 mm were selected by a pathologist and manually annotated densely (i.e., all pixels were labeled) to cover 13 morphological patterns in the tissue: background, tumor, stroma, inflammation, healthy parenchyma, fatty tissue, necrotic tissue, erythrocytes, healthy epithelium, mucus, cartilage, macrophages, and other. In total, n=163 ROIs were annotated; an example of training WSI is depicted in Fig. 1.

Test data. We defined an external test set to evaluate the multi-class segmentation performance of nnUNet, containing n=14 cases (n=7 AD, n=7 SC) from the publicly available TCGA-LUAD and the TCGA-LUSC datasets. In line with the training set, the test set cases were selected to cover several NSCLC growth patterns and morphological characteristics, as well as varying stain color characteristics, to assess the generalizability of trained nnUNet models to staining variations across different medical centers. Manual annotations were made in the same way as done for the training set, resulting in a total of n=55 annotated ROIs in the full test set.

PEMBRO-RT trial data. As a proof of concept biomarker development application, we assess the predictive value of TILs in a cohort of NSCLC treated with immunotherapy (Pembrolizumab) from the phase-2 clinical trial PEMBRO-RT (Theelen et al., 2019). This cohort consists of n=68 tumor biopsies from patients with metastatic NSCLC, containing biopsies from the lung (n=25) and from other metastatic sites (n=43). Since our nnUNet model is solely trained on lung tissue, we restricted our selection to 25 lung tissue tissue biopsies. Furthermore, we excluded two cases due to insufficient biopsy quality, assessed by a resident pathologists, resulting in n=23 cases. As a readout for immunotherapy treatment response, we defined response as stable disease (as defined by the RECIST1.1 guidelines (Eisenhauer et al., 2009)) or better at 12 weeks after treatment.

2.2. nnUNet for pathology

From whole-slide images to nnUNet and back. The default nnUNet depends on nifti file formats that are uncommon for pathology data. Furthermore, due to their sheer possible size, entire WSIs cannot always be fed into memory. We therefore developed a custom dataset preparation pipeline, which extracts WSI's annotated ROIs and their corresponding annotations on a set spacing of 0.5 μm/px, using the WholeSlideData Python library[2] that enables efficient, fast, and easy patch sampling from WSI. The extracted ROIs and annotation masks are then converted to the required nifti file formats for compatibility with nnUNet's pre-processing pipeline. Instead of using a random 5-fold cross-validation split of the training ROIs, the split was based on the pixel class distributions within each ROI, ensuring similar class distributions in the train and validation splits.

Additionally, we developed a WSI inference approach as an essential feature for nnUNet for pathology. The pipeline (1) accepts a WSI and its tissue/background mask (Bándi et al., 2019), (2) samples patches of the WSI, using a sliding window on patches containing tissue, (3) applies nnUNet's default sliding window approach with half overlap within the sampled patches, (4) only writes the fully overlapped region of the sampled patches in a pyramidal TIFF file structure, and (5) outputs the model output mask, and uncertainty mask (further described in the 'Inference uncertainty extraction' section). Further details of the WSI inference pipeline are discussed in the Appendix.

WSI segmentation with nnUNet. Evaluated configurations of the nnUNet model are: types of input pre-processing, feature normalization methods, batch size and patch size configurations, and the use of pathology-specific color augmentations. Performance of the evaluated changes was measured using the f1-score, which combines the precision and recall into a single metric by taking the harmonic mean.

Input pre-processing. In contrast with radiology image pixels representing signal intensities, pathology image pixels represent the actual color in RGB space. While nnUNet uses z-score normalization as default for input pre-processing, which normalizes each image channel separately with a mean of 0 and a standard deviation of 1, this method may result in color shifts in the pathology data's RGB channels. We therefore evaluated the performance impact of using 0-1 scaling, which statically scales 8-bit images with values ranging within 0-255 to a new range within 0-1 and therefore does not cause color shifts.

Feature normalization & patch and batch size. The default nnUNet strongly favours large patch sizes over large batch sizes, which supports the use of IN that, in contrast with BN, still works well with extremely small batch sizes. We assessed the impact of these configurations on pathology data by evaluating the performance of both normalization methods on a model with patch size 1024 (ps1024) and batch size 2 (bs2) (configured by default nnUNet), and patch size 512 (ps512) and batch size 8 (bs8).

Data augmentation. We assessed the use of Hematoxylin-Eosin-DAB (HED) and Hue-Saturation-Value (HSV) augmentation for nnUNet. HED decomposes an RGB image into the staining intensities of the individual hematoxylin, eosin and DAB staining, augments their intensities, and converts them back to RGB. Similarly, HSV first converts the RGB images to HSV space, augments the images' hue, saturation and value, and converts the augmented image back to RGB space. We evaluated the addition of HED, and the sequential

2. https://github.com/DIAGNijmegen/pathology-whole-slide-data

addition of HED and HSV, as the first color-augmenting transformations in the default data augmentation pipeline. From the HED and HSV augmentation settings proposed in (Tellez et al., 2019), we used the HED-light setting, and an adjusted HSV-light setting, with additional value augmentation in the same intensity rage as the hue and saturation.

Next to augmentations during training, nnUNet also encompasses test time augmentation (TTA), which mirrors the input image along the two axes during inference, and reconstructs and averages the predictions on the augmented images into a final output. The effect of TTA was evaluated on all before-mentioned nnUNet variants.

Inference uncertainty extraction. nnUNet by default trains five individual models using a 5-fold cross-validation method. Using these 5 models, we can extract the uncertainty of every pixel by measuring the internal disagreement of the 5 folds, where a larger disagreement indicates a more uncertain pixel output. Here, a fold's disagreement on a pixel is calculated as the soft Cross-Entropy loss over the fold's pixel log(softmax) values and the mean pixel log(softmax) values of the 5 folds. The final uncertainty value per pixel is calculated as the average of the 5 folds' disagreements for that pixel, similar to the method proposed by Lakshminarayanan et al. (Lakshminarayanan et al., 2016).

Given that more uncertain predictions also tend to be more incorrect, we can use a pixel's uncertainty value as a measure of incorrectness. We therefore investigate the uncertainty value distribution for each predicted output class (defined using argmax on the pixel's softmax output) on the full test set. We then separate the class' uncertainty values by splitting the True Positive (TP) and False Positive (FP) predictions' uncertainty values. Finally, we define a cutoff point in the uncertainty values for each class, using an ROC curve's optimal operation point, to separate TP from FP predictions based on the uncertainty value.

2.3. TIL biomarker development

As a proof of concept biomarker development application for nnUNet, we investigate different TIL densities on each of the 23 lung biopsies of metastatic NSCLC from the PEMBRO-RT trial (Theelen et al., 2019). The HoVerNet algorithm, a nuclei segmentation and classification model which is trained on multiple organ tissues (Graham et al., 2019), was used to detect inflammatory cells as a proxy for TILs; we validated its use as TIL detector using manual annotations (see Appendix Fig. 6). TIL densities were computed within the whole tissue ($_{tissue}$TILs), nnUNet's stromal sections for stromal-TILs ($_{s}$TILs) and nnUNet's tumor segments for intra-tumoral TILs ($_{i}$TILs). Here, stromal segments are defined as the union of nnUNet's stroma and HoVerNet's 'inflammation' class. Additionally, we investigate the use of our proposed uncertainty FP filtering approach to exclude likely-FP regions from the nnUNet tumoral and stromal regions. The derived TIL densities were used to stratify responders using ROC analysis. Welsh test statistics were used to evaluate the significance of the TIL density distributions from responders and non-responders originating from different populations.

3. Results

nnUNet for pathology Out-of-the-box application of nnUNet to pathology tissue segmentation resulted in promising quantitative results already on patch segmentation level.

However, we identified a 'label-switching' patch effect on adjacent or overlapping patches (Appendix Fig. 5 A) inherent to nnUNet's default configuration, which is detrimental to WSI inference performance. Using an ablation study design we identified several critical changes and beneficial additions to nnUNet's default configurations for its application to pathology data. An overview of the evaluated nnUNet configurations is shown in Table 1.

Table 1: nnUNet experiments table

	nnUNet configuration				F1 score (F1 score without TTA)		
	input scale	**feature norm**	**batch/patch size**	**DA**	**Micro avg**	**Macro avg**	**Weighted avg**
A	z-score	IN	2/1024	default	0.697 (0.692)	0.378 (0.374)	0.655 (0.649)
B	0-1	IN	2/1024	default	0.695 (0.691)	0.382 (0.379)	0.652 (0.648)
C	0-1	IN	8/512	default	0.728 (0.728)	0.406 (0.408)	0.690 (0.690)
D	0-1	BN	2/1024	default	0.752 (0.750)	0.486 (0.485)	0.725 (0.723)
E	0-1	BN	8/512	default	0.765 (0.754)	0.511 (0.510)	0.735 (0.733)
F	0-1	BN	8/512	+HED	**0.768 (0.766)**	**0.522 (0.520)**	**0.748 (0.746)**
G	0-1	BN	8/512	+HED,HSV	0.766 (0.764)	**0.522 (0.520)**	0.746 (0.743)

Intensity scaling. nnUNet's default z-score input scaling may cause color shifts in the RGB channels, and in adjacent or overlapping patches. While this could be a possible cause of the label switching patch effect, using 0-1 scaling did not solve this issue (Appendix Fig. 5 B)). Table 1 A and B shows the performance of z-score normalization and 0-1 scaling respectively. Although both methods scored equally well on patch segmentation level, 0-1 scaling is preferred since it does not disturb color channels, and is required for HED and HSV augmentation.

Feature normalization & batch size, patch size. The performance of the evaluated models with different feature normalization approaches, and batch and patch sizes are depicted in Table1 (B, C, D, and E). Both replacing IN with BN, and using bs8 ps512 instead of bs2 ps1024 yielded superior results. Replacing the feature normalization approach yielded a major performance increase compared to changing the batch and patch size. A combination of both adjustments yielded a synergistic performance increase on all computed f1 scores, with a macro averaged f1 score of 0.510 compared to 0.379 for both default settings. All evaluated variations, in contrast with the default configuration, solved the 'label-switching' patch effect (Appendix Fig. 5).

Data augmentation. Both HED, and the combination of HED and HSV improved segmentation results compared to nnuNet's default data augmentations (1 E, F, G). Although very close to the configuration with HED and HSV augmentation, the addition of HED only (example augmentations in Appendix Fig. 8), yielded the highest overall performance in all averaged f1 scores, and is used in our TIL biomarker development task. A WSI inference comparison example of nnUNet's default configuration versus our proposed configuration is shown in Fig. 2.

Inference uncertainty extraction. Appendix Fig. 7 shows 2 visual examples of the extracted uncertainty that depict that uncertain predictions are more likely to be incorrect. In Fig. 3 we depict the complete result of our endeavor to stratify TP from FP predicted class pixels using the class' uncertainty values. These results show a clear difference in the uncertainty distributions of TP (orange) and FP (blue) predictions for each class, where TP predictions are less uncertain (distribution with mostly low uncertainty values) and FP predictions are more uncertain (distribution with high uncertainty values). This clear distinction in uncertainty distributions was used to define an optimal split point to discriminate likely-FP from likely-TP pixel predictions using ROC analysis.

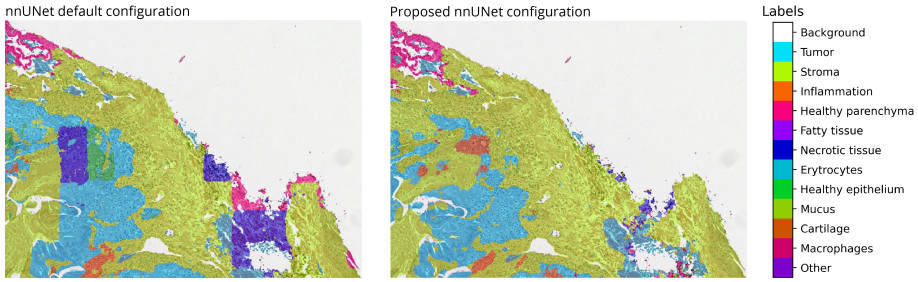

Figure 2: Comparison of WSI inference of the default nnUNet configuration, with the pronounced 'label switching' patch effect, and our proposed nnUNet configuration with correct WSI inference.

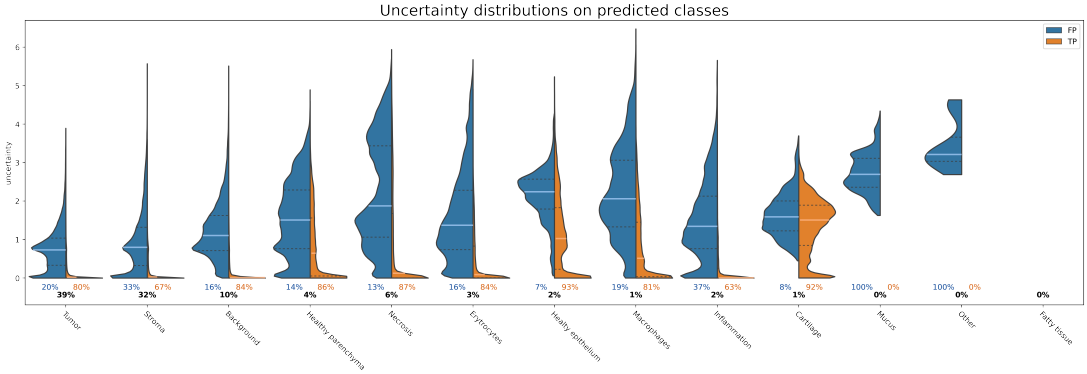

Figure 3: Pixel uncertainty distributions on predicted classes. (TP: orange, FP: blue). Percentages depict the sizes of the distributions.

TILs to predict immunotherapy response ROC curves on response stratification performance of our evaluated TIL density features are shown in Fig. 4. The overall $_{tissue}$TIL density feature showed an AUC of 0.727 (Fig. 4 A). Both nnUNet derived $_s$TIL and $_i$TIL density features showed improved discrimination compared to the $_{tissue}$TIL densities with AUCs of 0.843 and 0.799 respectively when the TIL densities were calculated in nnUNet's full tumor and stromal segments (Fig. 4 B and C). The exclusion of FP regions of the tumor and stromal segments, using our uncertainty approach, showed superior discriminative performance in both nnUNet derived TIL density features, with AUCs of 0.876 and 0.826 for $_s$TIL and $_i$TIL densities respectively. All evaluated TIL density features showed Welsh test statistics with p-values <0.05. However, with Bonferroni correction, only the $_s$TIL density features remained significant.

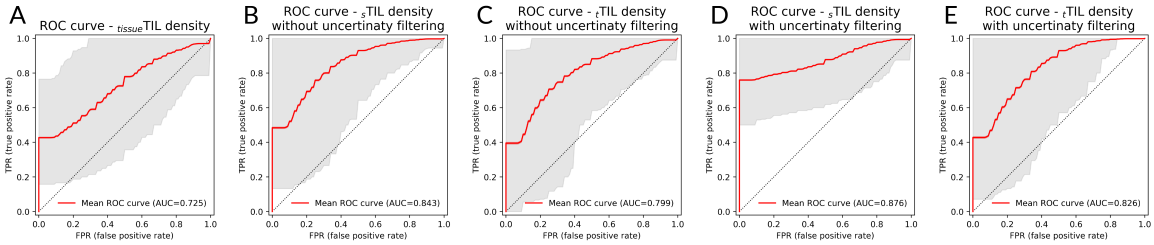

Figure 4: ROC curves with confidence interval and AUCs from different TIL density features

4. Discussion

In this study, we explored nnUNet's novel application to the computational pathology domain. We identified several shortcoming in the framework's application to pathology data, during both training and inference. While the current training data preparation pipeline suits our needs, it can be further improved by connecting a WSI batch iterator to nnUNet directly, to allow dynamic ROI sampling from the full WSI without the need for padding in spatial transformations (Fig. 8).

One of the primary contributions is the full WSI inference pipeline as an essential pathology tool that now makes the nnUNet framework applicable to pathology data. By deployment on WSI, we go beyond the limit of patch-level analysis, and revealed the systematic presence of the 'label swtiching' patch effect, inherent to nnUNet's default configuration. Next to improving overall segmentation performance, solving this patch effect problem is an essential step for achieving proper WSI inference performance.

In our ablation study, we show that both switching from IN to BN and increasing the batch size, and the synergistic combination of both, yielded superior segmentation performance and solved the 'label-switching' patch effect. In pathology data, the high inter and intra-class cell morphology variability in histopathology WSI may cause extreme variance within a batch with cell-level patches. We consider this as a possible reason why nnUNet for pathology applications benefits from more from BN and a larger batch size.

To assess generalization of our proposed nnUNet, we trained our nnUNet configurations on data from a single medical center, and evaluated the segmentation performance on cases from the TCGA LUAD and LUSC datasets, originating from a variety of medical centers with different staining procedures. We showed that next to the aforementioned improvements, the addition of pathology-specific color augmentations further improved the segmentation performance on external data. Altogether, the configuration with 0-1 input scaling, BN, patch size 512, batch size 8, and additional HED augmentation yielded the best performance on all averaged f1 scores, and was used for TIL biomarker development.

As a proof of concept, we showcased the use of nnUNet for segmenting tumor and stromal regions, while excluding necrotic regions, to compute $_s$TILs and $_i$TIL density features, on a small subset of the PEMBRO-RT cohort. All nnUNet derived features yielded improved AUCs on treatment response, compared to the overall $_{tissue}$TIL density feature. Additional filtering of likely-FP tissue segments using pixel uncertainty cutoffs showed to yield improved AUC values, which supports the application and design of our inference uncertainty approach although in a small cohort. Using Welsh test statistics we found that the $_s$TIL density values significantly separated responders from non-responders.

5. Conclusion

In this paper, we built on nnUNet's framework and tailored it to application to pathology. We've (1) developed additions to the framework to prepare pathology training data for nnUNet, and apply inference on WSI; (2) identified critical configuration changes and added pathology color augmentations to improve tissue segmentation performance; (3) added a lightweight uncertainty extraction approach, and (4) validated our proposed nnUNet configuration and showcase its application for biomarker development. To encourage use of nnUNet for pathology applications, we release the code for our proposed nnunet-for-pathology, and part of the data open source to promote and foster developments within the computational pathology community.

Acknowledgments

Acknowledgments withheld.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022.

Péter Bándi, Maschenka Balkenhol, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. *PeerJ*, 7:e8242, December 2019.

Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Jun Gao, Qian Jiang, Bo Zhou, and Daozheng Chen. Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: an overview. *Mathematical Biosciences and Engineering*, 16(6):6536–6561, 2019.

Simon Graham, Quoc Dang Vu, Shan E 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. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101563. URL https://www.sciencedirect.com/science/article/pii/S1361841519301045.

Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, jan 2021. doi: 10.1016/j.knosys.2020.106622. URL https://doi.org/10.1016%2Fj.knosys.2020.106622.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, December 2020. doi: 10.1038/s41592-020-01008-z. URL https://doi.org/10.1038/s41592-020-01008-z.

Prathamesh M Kulkarni, Eric J Robinson, Jaya Sarin Pradhan, Robyn D Gartrell-Corrado, Bethany R Rohr, Megan H Trager, Larisa J Geskin, Harriet M Kluger, Pok Fai Wong, Balazs Acs, et al. Deep learning based on standard h&e images of primary melanoma tumors identifies patients at risk for visceral recurrence and deathdeep learning–based prognostic biomarker for melanoma. *Clinical Cancer Research*, 26(5):1126–1134, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016. URL https://arxiv.org/abs/1612.01474.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Vishwa S Parekh, Katarzyna J Macura, Susan C Harvey, Ihab R Kamel, Riham El-Khouli, David A Bluemke, and Michael A Jacobs. Multiparametric deep learning tissue signatures for a radiological biomarker of breast cancer: Preliminary results. *Medical physics*, 47(1): 75–88, 2020.

Sehhoon Park, Chan-Young Ock, Hyojin Kim, Sergio Pereira, Seonwook Park, Minuk Ma, Sangjoon Choi, Seokhwi Kim, Seunghwan Shin, Brian Jaehong Aum, Kyunghyun Paeng, Donggeun Yoo, Hongui Cha, Sunyoung Park, Koung Jin Suh, Hyun Ae Jung, Se Hyun Kim, Yu Jung Kim, Jong-Mu Sun, Jin-Haeng Chung, Jin Seok Ahn, Myung-Ju Ahn, Jong Seok Lee, Keunchil Park, Sang Yong Song, Yung-Jue Bang, Yoon-La Choi, Tony S. Mok, and Se-Hoon Lee. Artificial intelligence–powered spatial analysis of tumor-infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non–small-cell lung cancer. *Journal of Clinical Oncology*, 40(17):1916–1928, June 2022. doi: 10.1200/jco.21.02010. URL https://doi.org/10.1200/jco.21.02010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F.L. Baehner, F. Penault-Llorca, E.A. Perez, E.A. Thompson, W.F. Symmans, A.L. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D.L. Rimm, K.H. Allison, J.S. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo, and S. Loi. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs working group 2014. *Annals of Oncology*, 26(2):259–271, February 2015. doi: 10.1093/annonc/mdu450. URL https://doi.org/10.1093/annonc/mdu450.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, December 2019. doi: 10.1016/j.media.2019.101544. URL https://doi.org/10.1016/j.media.2019.101544.

Willemijn S. M. E. Theelen, Heike M. U. Peulen, Ferry Lalezari, Vincent van der Noort, Jeltje F. de Vries, Joachim G. J. V. Aerts, Daphne W. Dumoulin, Idris Bahce, Anna-Larissa N. Niemeijer, Adrianus J. de Langen, Kim Monkhorst, and Paul Baas. Effect of Pembrolizumab After Stereotactic Body Radiotherapy vs Pembrolizumab Alone on Tumor Response in Patients With Advanced Non–Small Cell Lung Cancer: Results of

the PEMBRO-RT Phase 2 Randomized Clinical Trial. *JAMA Oncology*, 5(9):1276–1282, 09 2019. ISSN 2374-2437. doi: 10.1001/jamaoncol.2019.1478. URL https://doi.org/10.1001/jamaoncol.2019.1478.

Appendix A. WSI inference pipeline

We developed a WSI inference approach as an essential feature for nnUNet for pathology. At the core of the pipeline is a batch iterator from our WholeSlideData library, that enables users to sample patches from the data efficiently, fast, and easily. The Batch iterator loops patch-wise over the WSI, and writes the model's argmax output on the 5-folds softmax values, and pixel-wise uncertainty map into 2 separate tiff files. The pipeline efficiently samples patches containing tissue by using a tissue/background mask (Bándi et al., 2019). It samples patches bigger than the model input size, making use of nnUNets gaussian sliding window approach, and only saves the inner part with full 4x overlap. To speed up the inference time, large parts of background are stripped of the patch (while preserving constant overlap) before inference. nnUNet's TTA is turned off with minimal/no performance drop.

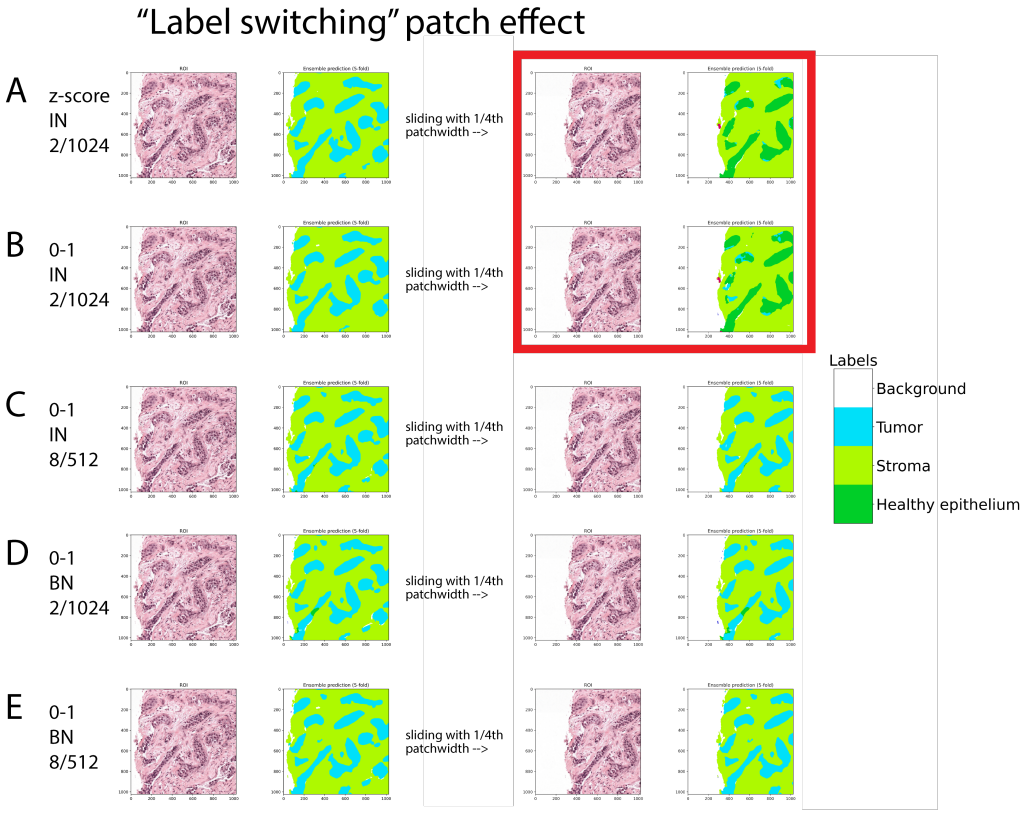

Figure 5: Example of the 'label-switching' problem that occurs in nnUNets default configuration (A), and persists when only the input normalization approach is adjusted to 0-1 scaling. All experiments with a bigger batch size or BN instead of IN do not exhibit this effect (C, D, E)

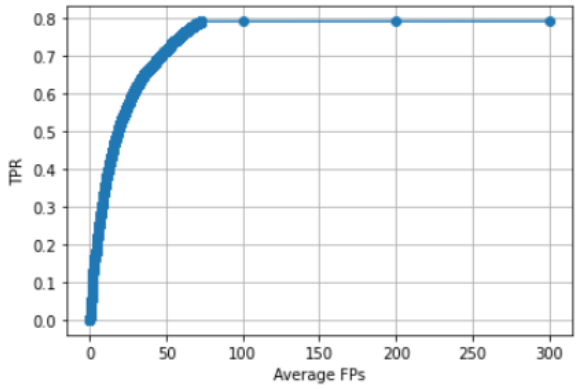

Figure 6: FROC analysis on HoVerNet's inflamed class as a proxy for TILs. We measure an FROC of 0.66, similar or even surpassing TIL detections FROCs in the TIGER challenge (https://tiger.grand-challenge.org/evaluation/segmentation-and-detecton-final/leaderboard/)

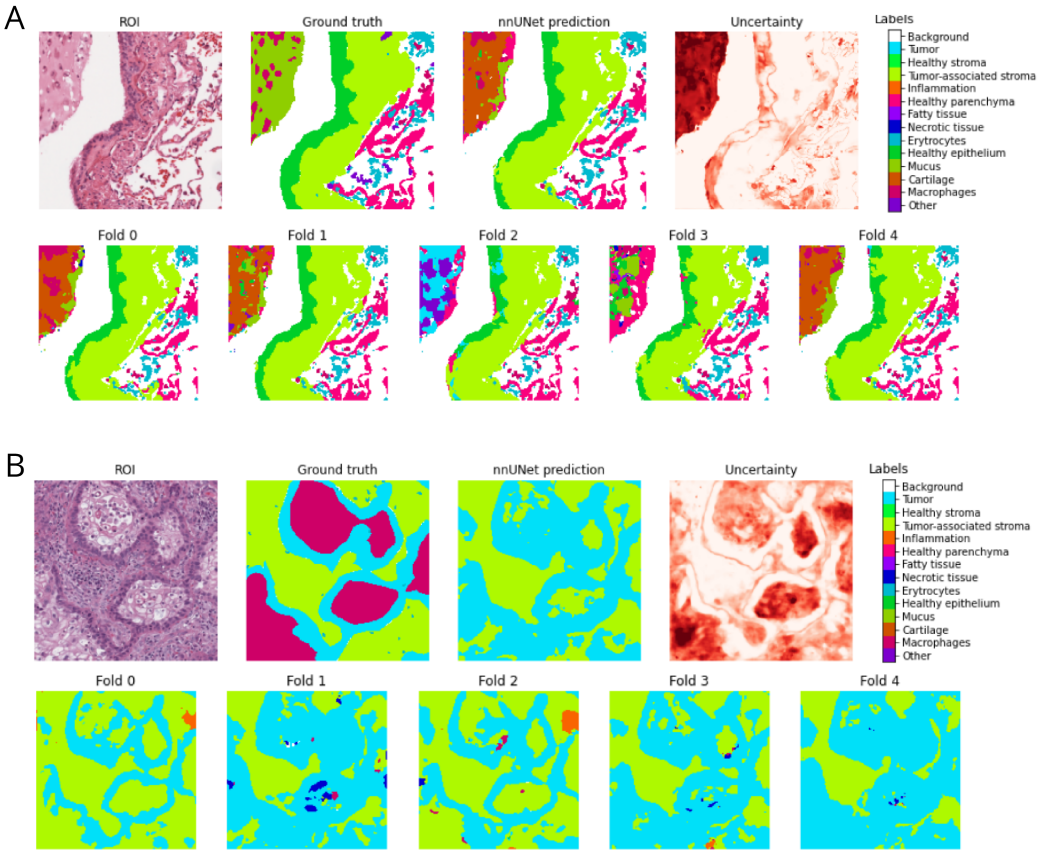

Figure 7: Visual examples of nnUNet's 5-fold inference output and the thereof extracted uncertainty. This image depicts uncertain regions that correlate with incorrect predictions.

Appendix B. TIL feature Welsh tests

TIL feature	p-value	Bonferroni corrected p-value
$_{tissue}$TIL	**0.040**	0.201
$_{s}$TILs without uncertainty filtering	**0.010**	**0.050**
$_{i}$TILs without uncertainty filtering	**0.033**	0.163
$_{s}$TILs with uncertainty filtering	**0.002**	**0.008**
$_{i}$TILs with uncertainty filtering	**0.019**	0.095

Appendix C. Exemplar pathology augmentations of nnUNet's default augmentations with additional HED

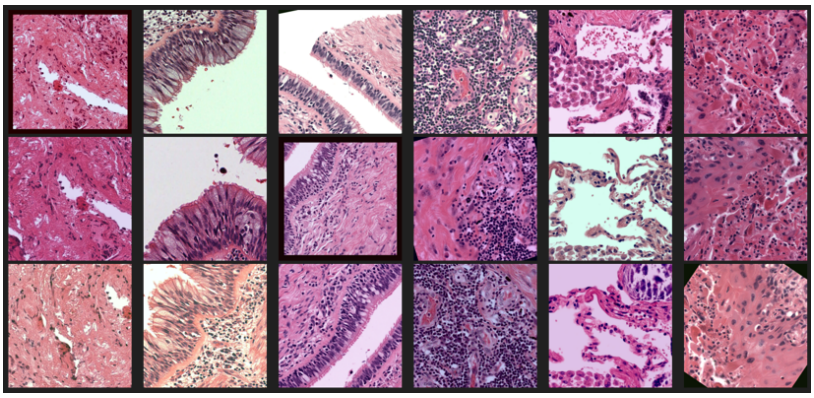

Figure 8: HED light augmentations