nnUNet meets pathology: Bridging the gap for application to whole slide images and computational biomarkers

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. While the self-configuring nnUNet framework for image segmentation tasks completely shifted the state-of-the-art in the radiology field, it has never been adapted to overcome its limitations for application on the pathology domain. Our study showcases the potential of nnUNet in computational pathology and bridges the gap that currently exists in utilizing nnUNet for pathology applications. Our proposed nnUNet for pathology framework has demonstrated its significance and potential to shift the state-of-the-art in the computational pathology field, as seen from the exceptional first-place segmentation ranking on the TIGER challenge's experimental leaderboard 1. Our framework includes critical hyperparameter adjustments and pathology-specific color augmentations, as well as an essential WSI inference pipeline and a novel inference uncertainty approach that proves helpful for biomarker development. We release our code of our accurate and workflow-friendly segmentation tool to promote and foster growth within the computational pathology community.[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), and the engineering of computational biomarkers in both radiology (Parekh et al., 2020) and histopathology images (Kulkarni et al., 2020). The advent of deep learning (LeCun et al., 2015) facilitated a substantial breakthrough in the development of accurate detection and segmentation models, powering systems that achieve and even surpass human performance on certain medical tasks (Esteva et al., 2017). In semantic image segmentation, the UNet architecture (Ronneberger et al., 2015) has increasingly been adopted and has become the state-of-the-art (SOTA) approach. However, hyperparameter tuning of Unet architectures is a crucial, non-trivial, and time consuming endeavor for training highly specialized medical imaging solutions.

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., 2020) introduced the self-configuring nnUNet framework, which utilizes well-established internal fingerprints, fixed parameters, empirical decisions and a cross-validation strategy to determine optimal hyperparameters to

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

© 2023 Anonymous.

train a UNet model. This self-configuring approach enabled nnUNet to surpass the SOTA performance in many radiology applications (Antonelli et al., 2022). While transformer-based segmentation approaches are entering the SOTA [**citations here**], Unet-based models with manually tweaked hyperparameters are still daily practice in the computational pathology domain [**citations here**]. Surprisingly, the challenge winning self-configuring nnUNet framework has never been adapted to suite the requirements for application to the pathology domain.

Despite the general purpose nature of nnUNet, several aspects of pathology data are intrinsically different from radiology. For example, loading entire gigapixel histopathology whole-slide images (WSIs) poses memory constraints during training and inference, resorting to customized dataset preparation and sliding window techniques that are not included in nnUNet's default framework. Additionally, multi-class semantic segmentation in histopathology involves a large variety of patterns, cell types and tissue morphologies, which may not be fully considered 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 smaller batch sizes, in combination with Instance normalization (IN). Additionally, capturing the full spectrum of tissue and cellular varieties the in training data is difficult, and deep learning models often make incorrect, overconfident predictions for out-of-training-distribution data. Extracting a model's uncertainty can therefore be useful in identifying poorly segmented regions when applied to unseen data (Lakshminarayanan et al., 2016). Finally, pathology data often exhibits variations in slide staining, even within the same lab on different days. To address this, pathology-specific data augmentation techniques are commonly used (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 identify and address pathology-specific workflow-related limitations of the framework, 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, and introduce a lightweight uncertainty extraction method to identify likely-incorrectly segmented regions.

To prove the versatility and robustness of the entire nnUNet framework for pathology data, we assess the proposed configuration's training and WSI-inference performance on the most recent public pathology benchmark for multi-class segmentation; the TIGER challenge [2]. This challenge involves a different organ and segmentation task, and therefore serves as a perfect performance readout of our work against optimized SOTA approaches.

Finally, we showcase use of the proposed nnUNet framework in biomarker development on a tumor-infiltrating lymphocytes (TILs) quantification task in a cohort of metastatic NSCLC cases treated with immunotherapy (Theelen et al., 2019). While initially intratumoral TILs ($_i$TILs) were deemed most relevant, (Salgado et al., 2015) proposed visual estimation in the stromal regions of the tumor ($_s$TILs), while excluding necrotic areas. Since this approach depends on the identification of multiple tissue types, it was adopted as a suitable proof of concept segmentation task for nnUNet's use in biomarker development.

2. https://tiger.grand-challenge.org/

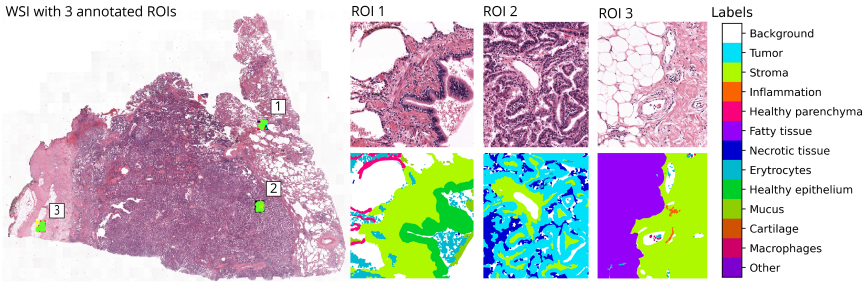


Figure 1: Visual example a WSI with 3 annotated ROIs; labels legend indicates names of annotated classes and their corresponding colors.

2. Methods

We first describe the data used during training, validation, and biomarker development.Then, we introduce nnUNet for pathology, where we focus on added technical and methodological contributions to the framework, and go over the experiments carried out in an ablation study to verify the proposed changes for enhanced segmentation performance over the default nnUNet framework. Finally, we detail how the proposed nnUnet framework was evaluated on the TIGER challenge, and explain the steps taken for TIL biomarker development.

2.1. Materials

Train and test data. For training, as set of 30 H&E stained WSI of NSCLC tissue samples (21 surgical resections, 9 biopsies) from ANONYMOUS medical center. Cases were selected to represent a balanced mix of adenocarcinoma (AD) and squamous cell carcinoma (SC) subtypes, tumor growth patterns, and morphologies (**Appendix X**). For each WSI, approximately n=5 regions of interest (ROI) of 2×2 *mm* were selected by a pathologist and manually annotated densely to cover 13 morphological patterns in the tissue (Fig. 1). Using the same selection and annotation procedure, we obtained an external test set containing 14 cases (7 AD, 7 SC) from the publicly available TCGA-LUAD and TCGA-LUSC datasets (respectively containing AD and SC cases). Additionally, test cases were sought out to contain varying stain color characteristics, to assess the generalizability of trained nnUNet models to staining variations across different medical centers. In total, 163 training ROIs and 55 test ROIs were annotated; an example of the data is depicted in Fig. 1.

TIGER challenge data. The challenge's publicly available training data contains 195 WSI from various medical centers with a total of 286 annotated ROIs. Each ROI contains up to 7 classes, and may contain un-annotated regions that should be excluded from loss calculation. The experimental test used for SOTA comparison of our proposed nnUNet framework contained 26 WSI with 130 ROIs. The final test set was completely withheld since the official challenge ended.

PEMBRO-RT trial data. The predictive value of TILs was assessd on a cohort of NSCLC treated with immunotherapy (Pembrolizumab) from the phase-2 clinical trial PEMBRO-RT (Theelen et al., 2019). This cohort consists of 68 tumor biopsies from patients with metastatic NSCLC, containing 25 biopsies from the lung and 43 from other metastatic sites. Since our train and test data solely contains lung tissue, we restricted our selection to the 25 lung 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. We developed a custom dataset preparation pipeline for the nnUNet framework that is compatible with pathology data. Our pipeline uses the WholeSlideData Python library[3] to fast and efficiently extract annotated ROIs and corresponding annotations from the WSIs on a set spacing of 0.5 μm/px. These extracted ROIs and annotation masks are then converted to the required nifti file format for compatibility with nnUNet's pre-processing pipeline. We ensured similar class distributions in the train and validation splits instead of using a random 5-fold cross-validation split.

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 color in RGB space. The default nnUNet pre-processing method, z-score normalization, normalizes each image channel separately with a mean of 0 and a standard deviation of 1. However, this method may result in color shifts in the pathology data's RGB channels. To mitigate this issue, we evaluated the performance of 0-1 scaling, which scales 8-bit images with values ranging from 0 to 255 to a new range of 0-1 without causing color shifts.

Feature normalization & patch and batch size. We noticed experimentally that the default settings of nnUNet can lead to a 'label switching' patch effect (Fig. 5 A). 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 evaluated the segmentation performance and the impact on this patch effect 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 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

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

(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. By default, nnUNet trains five models through a 5-fold cross-validation approach. These models enable the extraction of pixel uncertainty, which reflects the internal disagreement among the 5 folds, where greater disagreement corresponds to more uncertain pixel predictions. To calculate a fold's disagreement on a pixel, we use 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 obtained by averaging the disagreements from the 5 folds, following a method similar to that proposed by Lakshminarayanan et al. (Lakshminarayanan et al., 2016). Given that uncertain predictions tend to be more incorrect, we use the uncertainty value as a measure of incorrectness. We examine the uncertainty value distribution for each predicted output class on the full test set, splitting TP and FP predictions' uncertainty values for each class. We establish a cutoff point in the uncertainty values for each class individually using an ROC curve's optimal operation point, to distinguish 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 $_{s}$TILs), and nnUNet's tumor segments (for $_{i}$TILs). Here, stromal segments are defined as the union of nnUNet's 'stroma' and (stromal) 'inflammation' class. Additionally, we investigate the use of our proposed uncertainty filtering approach to exclude likely-FP regions from the nnUNet's 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.

2.4. TIGER challenge segmentation

We adapted the proposed and default configurations to suit the labels of the TIGER dataset, trained the 5-folds, and subsequently applied them though our WSI inference pipeline on the experimental test set on the Grand Challenge platform, returning the final dice scores of the proposed and default configurations.

3. Results

nnUNet for pathology Out-of-the-box application of nnUNet to pathology data showed promising quantitative results 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. By conducting an ablation study, we identified several critical changes and beneficial additions that are necessary to improve 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 normalization can introduce color shifts in adjacent or overlapping patches, which may contribute to the label-switching patch effect. However, using 0-1 scaling did not resolve the issue (Appendix Fig. 5 B). Although the performance of both methods on patch segmentation level is comparable (Table 1 A and B), 0-1 scaling is preferred since it avoids color channel disturbance 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 Table 1 (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 from the default solved the identified 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. 9), yielded the highest overall performance in all averaged f1 scores, and and is thus proposed as the final nnUNet configuration for pathology. 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 demonstrates two visual examples of uncertainty extracted from the 5-fold disagreement, revealing that uncertain predictions are more likely to be incorrect. In Appendix Fig. 8, the complete result of our approach to differentiate TP from FP predicted class pixels using the class' uncertainty values is depicted. The TP and FP predictions for each class exhibit clear differences in the uncertainty distributions, with TP predictions having mostly low uncertainty values and FP predictions having high uncertainty values. The distinction in uncertainty distributions was utilized to

determine an optimal split point to distinguish likely-FP from likely-TP pixel predictions through ROC analysis. Fig. 3 A shows the uncertainty distributions and the defined cutoff value from the classes used for TIL biomarker development. Fig. 3 B provides a visual example of the effectiveness of uncertainty filtering in removing large incorrectly predicted tumor segments, which leads to a more accurate area for $_i$TIL density calculation.

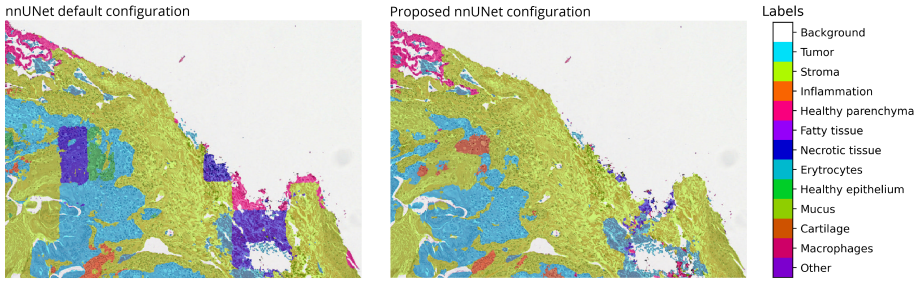
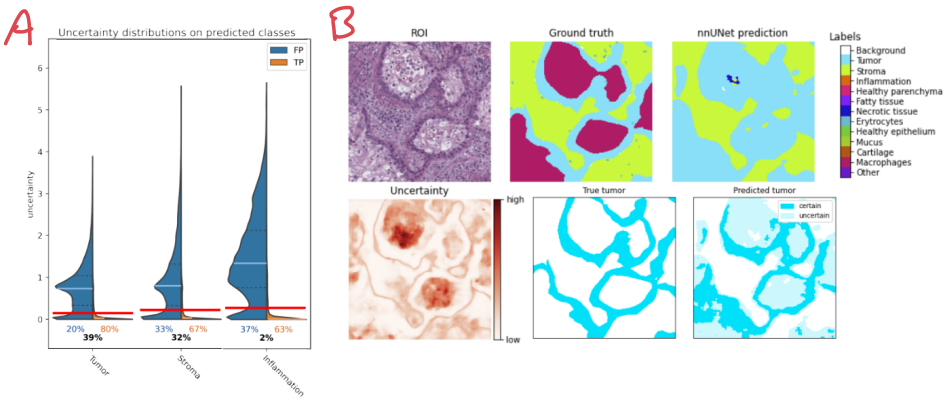


Figure 2: Comparison of WSI inference of the default nnUNet configuration, with the pronounced 'label switching' patch effect, and our proposed nnUNet configuration with correct WSI inference.



~~**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. These results demonstrate that nnUNet-derived sTIL and iTIL features outperformed the overall tissue TIL density feature, with AUCs of 0.843 and 0.799, respectively (Fig. 4 B and C). Furthermore, using our uncertainty approach to exclude likely-FP regions in the tumor and stromal segments improved the discriminatory power of both nnUNet-derived TIL density features, with AUCs of 0.876 and 0.826 for sTIL and iTIL, respectively. All TIL density features showed significant differences in Welsh test statistics (p-values < 0.05), but only the $_s$TIL density features remained significant after Bonferroni correction.

TIGER challenge segmentation performance On our first try, our proposed nnUNet for pathology framework ranked first on the segmentation leaderboard, with a dice score of 0.8083, beating the best participant's dice score of 0.7997. The default configuration ranked [**RANK**], with a dice score of [**DICE**], depicting the significance of our proposed changes.

will be combined into a single block to save space

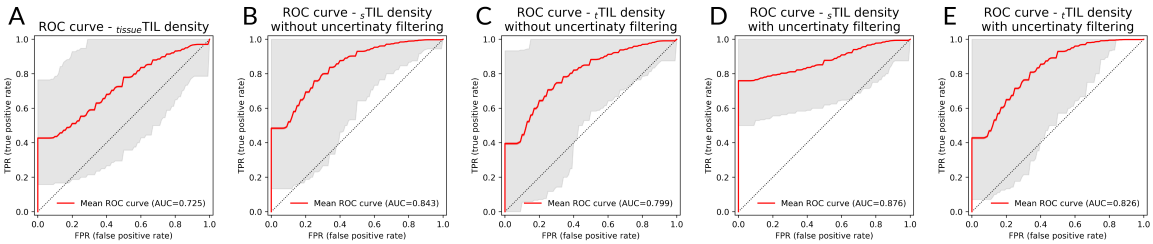


Figure 4: ROC curves with confidence interval and AUCs from different TIL density features

4. Discussion

Our study demonstrates the potential of nnUNet in computational pathology by identifying and addressing its limitations in training and inference. The improvements proposed in this study, including the switching from IN to BN and increasing the batch size by sacrificing the patch size, are effective in addressing the high variability in cell morphology in histopathology WSI. Furthermore, by training on data from a single medical center and testing on external datasets, we validated our proposed configurations and showed the success of pathology-specific color augmentations in improving the segmentation performance on external datasets with different medical centers with varying staining procedures.

The nnUNet configuration encompassing 0-1 input scaling, BN, ps512, bs8, and additional HED augmentation yielded the best performance on all averaged f1 scores, and defines our proposed nnUNet model configuration. Moreover, the success of this proposed nnUNet model configuration and the full WSI inference pipeline is further validated by its exceptional first-place ranking in the TIGER challenge's experimental segmentation leaderboard 1.

The development of the WSI inference pipeline represents a crucial contribution to the application of nnUNet in pathology. This essential tool is instrumental in bridging the practical gap that currently exists in utilizing nnUNet for pathology applications. Our successful resolution of the 'label switching' patch effect through proposed changes enabled accurate WSI inference performance, rendering nnUNet more appealing for use in pathology.

As a proof of concept for nnUNet's application in biomarker development, we successfully segmented tumor and stromal regions, yielding improved $_s$TIL and $_i$TIL density features compared to the overall $_{tissue}$TIL density. Our novel inference uncertainty implementation was supported by the improved AUC values achieved by filtering likely-FP tissue segments using pixel uncertainty cutoffs. Although being applied on a small cohort, the $_s$TILs demonstrated significant of responder stratification, as determined by Welsh test statistics, even after bonferroni correction. These results demonstrate the potential of nnUNet and our inference uncertainty approach as a powerful tools for the development of novel biomarkers.

5. Conclusion

Our proposed nnUNet for pathology framework has demonstrated its significance and potential to change the SOTA in the field of computational pathology. The results from our paper, including the exceptional first-place ranking on the TIGER challenge segmentation leaderboard, validate the effectiveness of our proposed configurations and improvements. Our approach provides a more accurate and workflow-friendly segmentation tool, which has

the potential to transform the field of pathology and improve diagnosis and treatment of diseases. The development of our WSI inference pipeline represents a crucial contribution to the application of nnUNet in pathology, and our novel inference uncertainty approach shows promise for the development of novel biomarkers. We hope that the release of our code and open-sourced data will encourage and facilitate further developments in the field of computational pathology.

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 EI-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.

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.

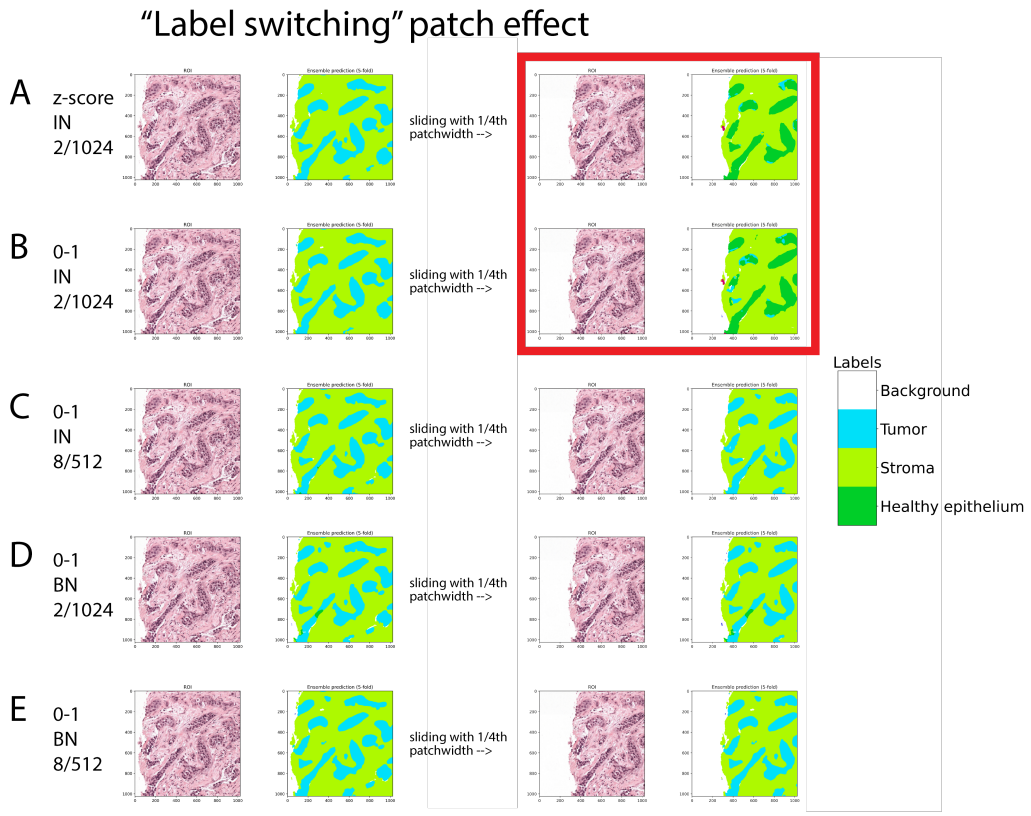


Figure 5: Example of the 'label-switching' problem that occurs in nnUNets default configuration (A), and persists when only the input normalization approach is adjusted to 0-1 scaling. All experiments with a bigger batch size or BN instead of IN do not exhibit this effect (C, D, E)

(will elaborate on rows and columns

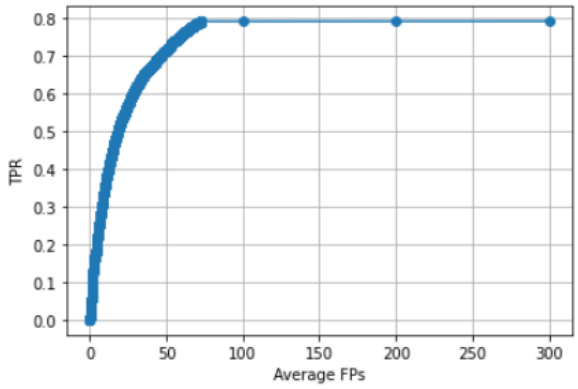


Figure 6: FROC analysis on HoVerNet's inflamed class as a proxy for TILs. We measure an FROC of 0.66, similar or even surpassing TIL detections FROCs in the TIGER challenge (https://tiger.grand-challenge.org/evaluation/segmentation-and-detecton-final/leaderboard/)

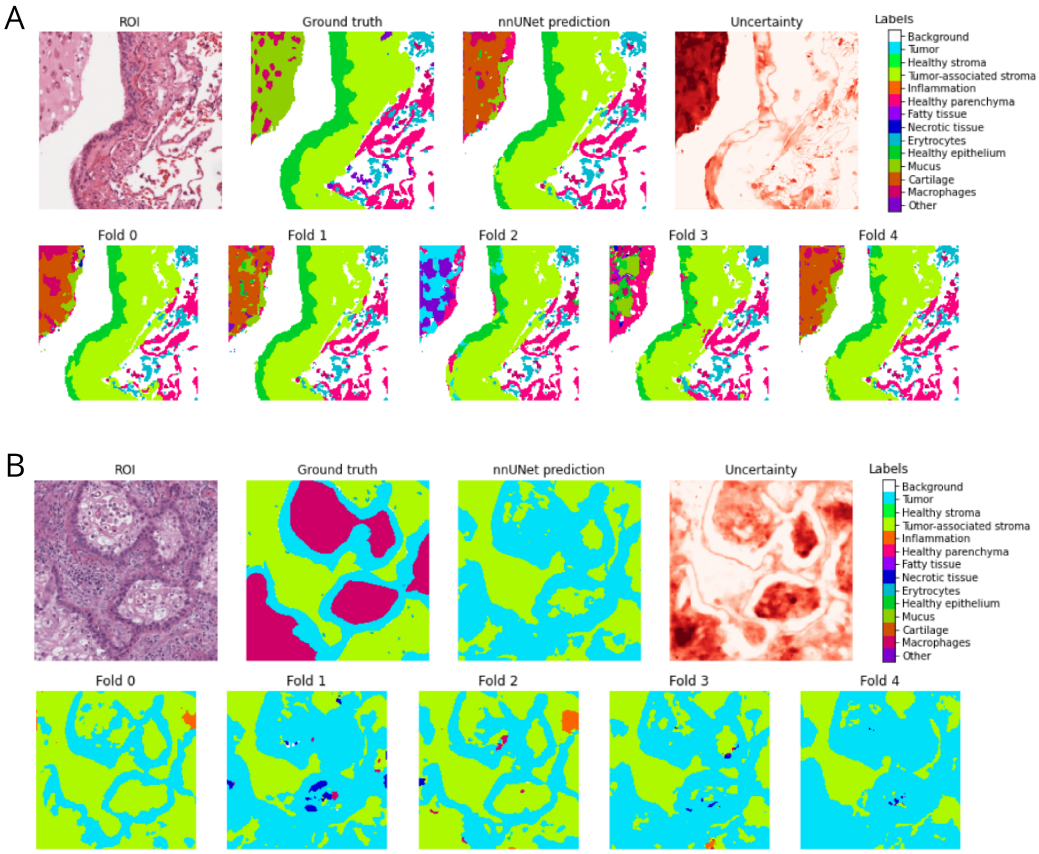


Figure 7: Visual examples of nnUNet's 5-fold inference output and the thereof extracted uncertainty. This image depicts uncertain regions that correlate with incorrect predictions.

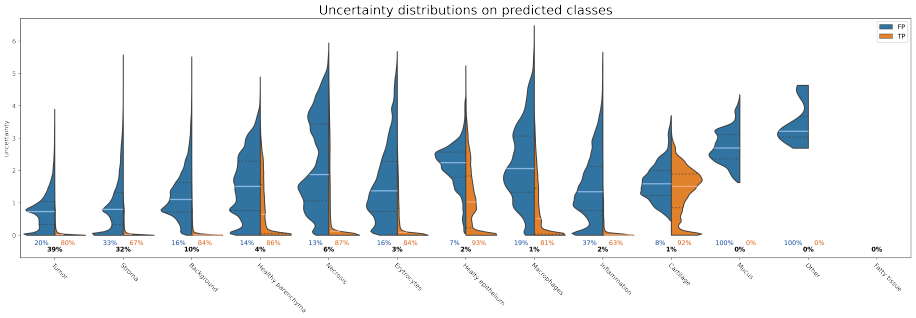


Figure 8: Pixel uncertainty distributions on predicted classes. (TP: orange, FP: blue). Percentages depict the sizes of the distributions. we will also add the FN distributions

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

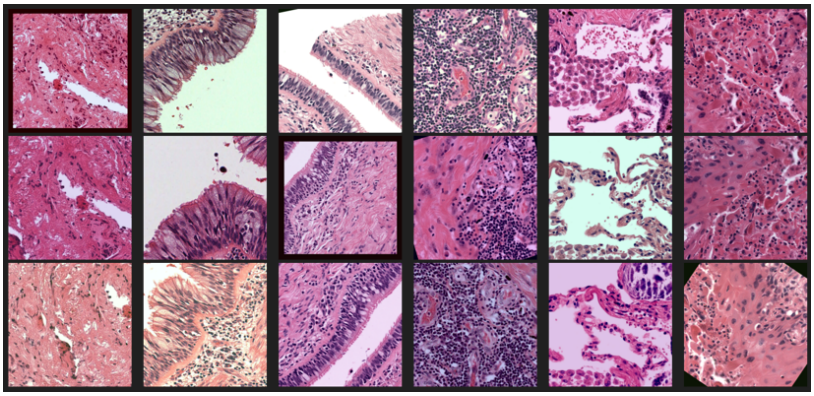


Figure 9: HED light augmentations