



UvA-DARE (Digital Academic Repository)

Improving the speed and quality of cancer segmentation using lower resolution pathology images

Li, J.; Osseyran, A.; Hekster, R.; Rudinac, S.; Codreanu, V.; Podareanu, D.

DOI

[10.1007/s11042-023-15984-9](https://doi.org/10.1007/s11042-023-15984-9)

Publication date

2024

Document Version

Final published version

Published in

Multimedia Tools and Applications

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Li, J., Osseyran, A., Hekster, R., Rudinac, S., Codreanu, V., & Podareanu, D. (2024). Improving the speed and quality of cancer segmentation using lower resolution pathology images. *Multimedia Tools and Applications*, 83(4), 11999-12015. <https://doi.org/10.1007/s11042-023-15984-9>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Improving the speed and quality of cancer segmentation using lower resolution pathology images

Jieyi Li¹ · Anwar Osseyran¹ · Ruben Hekster² · Stevan Rudinac¹ · Valeriu Codreanu² · Damian Podareanu²

Received: 17 May 2022 / Revised: 13 April 2023 / Accepted: 4 June 2023 /

Published online: 29 June 2023

© The Author(s) 2023

Abstract

In this paper, we propose a pipeline to investigate the performance of semantic segmentation model that employs an encoder-decoder architecture with atrous separable convolution and spatial pyramid pooling, trained on multi-resolution whole slide breast pathological images with different patch sizes. Our segmentation model obtains the best performance on zoom level 2 (10× magnification) with AUC score 0.974 in terms of slide-level classification. This outperforms both the performance of the pathologist and other semantic segmentation models on the Camelyon16 dataset. By offering a larger field of view and reducing noise and detail, training a semantic segmentation model on the properly selected lower resolution pathology images can further improve the precision of pixel-wise cancer region segmentation. By contrast, the corresponding inference time is 14 times shorter than the inference time trained on the highest resolution patches, and it is also shorter than the time required by a pathologist with time constraints. Moreover, we prove that the model trained on lower resolution patches can still generate refined external polygons of cancer region on the highest resolution image. This study provides new insights into efficient gigapixel histopathology analysis that will make clinical adoption more likely.

Keywords Semantic segmentation · Breast cancer · Whole-slide imaging · Gigapixel histopathology · Pixel-based convolutional neural networks

1 Introduction

As the highest incidence rate of cancer among women and the second leading cause of cancer death worldwide, breast cancer can be seen nowadays as the most serious threat to women's health. In 2019, 111,710 new cases of breast cancer were diagnosed worldwide and 41,760

✉ Jieyi Li
j.li3@uva.nl

¹ Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, Amsterdam 1018 TV, The Netherlands

² High Performance Machine Learning group, SURF, Science Park 140, Amsterdam 1098 XG, The Netherlands

woman died from it [41]. However, with the development of pathology instrumentation and accessories, standardized pathological examination can essentially contribute to better diagnosis and treatment of breast cancer. Especially in recent years, the possibility to capture an entire slide and save it in a digital format named Whole-Slide Image (WSI) enables pathologists to examine the biopsy at different levels of magnification, which significantly increased the accuracy of diagnosis [38].

However, manual microscope image examination is a time-consuming and tedious task and is made more difficult by the fact that tumor cells often occur as small patches that are difficult to spot. Computer-assisted pathology image analysis is therefore needed to help pathologists detect small metastatic areas in a more accurate and efficient manner. Machine learning has been widely adopted in the computer vision domain [23, 31, 45]. For example, deep learning models and, particularly, Convolutional Neural Networks (CNNs), have been extensively utilized in medical image analysis [4, 11, 19, 20, 27, 29, 32, 34, 42, 47]. In order to obtain better model accuracy, the number of parameters that need to be trained has dramatically increased, as CNN architectures go wider and deeper. However, training such models has significantly higher computation and memory demands. Hence, High Performance Computing (HPC) facilities based on CPU [9], GPU [48] and FPGA [15] enable training such heavy models while significantly reducing training times, which also benefits the WSI analysis.

The WSIs are typically high resolution images with a relatively large memory footprint. The size of one WSI could be around 2 gigabytes (GB) considering the slide image dimensions of 200,000 by 100,000 pixels in the highest resolution mode (40× magnification). However, since loading an entire WSI is currently still computationally intractable [25], a typical approach to tackle this challenge is to extract smaller, manageable patches from the highest resolution WSI and perform binary classification training and prediction of tumors by using deep convolutional neural network methods [27, 46].

The Cancer Metastases in Lymph Nodes Challenge 2016 (i.e. CAMELYON16 or Camelyon16) was initiated by the IEEE International Symposium on Biomedical Imaging (ISBI) with the aim of detecting the presence of breast cancer metastases in lymph nodes [1]. It has since become one of the most widely-used public datasets for studying cancer detection using deep learning methods [18, 21, 22, 28, 30, 40, 44, 46, 51, 53]. The work presented in [46], which won the Camelyon16 challenge [1] with patches size 256×256 using GoogLeNet architecture [43], reports the best results on the highest zoom level (40×) images. However, this conclusion is based on patch-level classification, which has several limitations. Firstly, since the approach of patch level classification is predicting the whole patch as positive or negative, it ignores the situation where tumor and healthy tissue are coexisting in the same patch and can only generate prediction on very low resolution masks [28]. Secondly, in order to create relatively dense heatmap predictions, they have to extract relatively small patches such as 256×256 and 299×299 from the highest resolution [28, 46]. Such small patches on the highest resolution can not provide sufficient contextual information, which may lead to inconsistent predictions. Thus, there is a contradiction between patch size and dense prediction by using patch level classification methods. Some attempts have been made to consider the spatial correlations of neighboring patches such as introducing Conditional Random Field (CRF) [27]. This has only been tested on small numbers of neighboring patches so the contextual information that can be learned is still limited. Finally, the number of patches extracted from single WSI would be around 10,000 to 400,000 (median 90,000) [28], which is very time consuming and computationally intensive in terms of data preprocessing and prediction.

In contrast, semantic segmentation models can generate refined predictions by performing pixel-level classification and do not need to consider the trade-off between patch size and

prediction quality. In recent years, several attempts have been made to investigate the training of semantic segmentation models on the highest resolution WSI such as [18]. However, the performance of the model with the largest patch size 1024×1024 (AUC 0.95) reported from the paper is still lower than that of the pathologist without time constraints (AUC 0.966) with relatively long inference time (79 mins/slide) using one GTX 1080Ti GPU, which may hinder clinical adoption. Using similar-sized patches extracted from lower resolution WSI offers a larger field-of-view (FoV), since the amount of extracted patches for prediction is significantly smaller compared to the highest resolution one. This may contribute to better model performance and shorter inference time. Therefore, it is worth investigating the performance of semantic segmentation models on different zoom levels using HPC resources, as this facilitates pixel-wise classification instead of patch level classification.

In this paper, we introduce a framework to evaluate the performance of a semantic segmentation model (DeepLabV3+) on various lower zoom levels, comparing them to the highest resolution patches ($40\times$ magnification), to explore the feasibility of training deep convolutional neural network with lower resolution images. We also conduct a comparison of model performance among different patch sizes trained on lower resolution. To our knowledge, no research has focused primarily on the trade-off between different zoom levels, segmentation models performance, and corresponding training and prediction time, which would offer insight into efficient gigapixel histopathology image analysis and its potential for clinical adoption.

The main contributions of this work are the following:

- Properly selecting lower resolution pathology images ($10\times$ magnification) can contribute to better semantic segmentation model training and prediction. It overcomes the limitation of patch-level classification using the highest resolution WSI, which generates low-resolution predictions by predicting the whole patch as positive or negative and ignoring coexisting tumor and healthy tissue within the same patch.
- The semantic segmentation pipeline we proposed succeeds in outperforming pathologists. Additionally, the inference time is 14 times shorter than both the model trained on the highest resolution patches, and is also shorter than the time required by pathologists with time constraints.
- Our model, trained on lower resolution patches ($10\times$ magnification), generates more precise segmentation than the same model trained on higher resolution patches. It accurately produces external boundaries of cancer regions on the highest resolution image ($40\times$ magnification), indicating its potential as a computer-assisted diagnostic and annotation tool to help pathologists reduce labeling time and identify cancerous regions.

2 Dataset and evaluation metrics

To compare our workflow with state-of-the-art strategies, as a testbed for this study we adopt well-known Camelyon16 dataset, a publicly available WSI dataset of lymph node section [1]. We further adopt the evaluation strategy from the Camelyon16 challenges with regard to slide-level classification. The details are discussed in the following subsections.

2.1 Camelyon16 dataset

The Camelyon16 dataset has 270 pixel-annotated WSIs obtained from the sentinel lymph nodes of breast cancer patients. The objective of the competition is to develop an algorithm

capable of automatically detecting the scope of breast cancer metastasis in lymph node slices stained with Hematoxylin and Eosin (H&E). These contain 110 metastases collected in the Radboud University Medical Center and the University Medical Center Utrecht. The average memory footprint of a WSI in the Camelyon16 dataset is around 1.9 GB, with the smallest WSI being 522MB and the largest 3.8GB. At the highest magnification level, WSIs have an average width of 109,666 pixels, ranging from 45,050 to 221,184 pixels, and an average height of 173,012 pixels, ranging from 28,672 to 221,696 pixels. The WSIs are stored in a 3-channel encoded TIFF format, which involves multiple levels of downsampling, to address the different zoom levels in one image. The largest resolution can be obtained at 40× magnification, which is also considered level 0, while the magnification of 20× corresponds to level 1, 10× corresponds to level 2 etc.

2.2 Evaluation metrics

In this study, we use the first evaluation strategy (i.e. Slide-based Evaluation) of the Camelyon16 challenge to measure the performance of our predictions¹. It enables us to compare our model results with the Public Leaderboard 1 in Camelyon16 challenge in terms of WSI classification task. This metric evaluates the performance of classification models at discriminating between WSIs which contain metastasis and normal slides. In the challenge, participants submit the list of probabilities which indicate the likelihood of each slide containing tumor and the area under the receiver operator curve (AUC) will be calculated to measure the performance of the model [2]. We use percentile bootstrap method and construct 95% confidence intervals for AUC score calculation [13].

3 Method

In this section, we further describe the testbed used for our study, including the complete proposed workflow, model details, and hardware configuration. Figure 1 illustrates the cancer metastasis detection workflow in our study, which can be divided into training and prediction pipelines. The training pipeline encompasses patch extraction from WSI tissue areas with multiple magnification levels and corresponding DeeplabV3+ model training. The prediction pipeline involves patch extraction, inference, pixel-wise tumor probability heatmap generation, cancer regions segmentation, and slide-based classification. The details are explained in the following subsections.

3.1 Data preprocessing

In this section, we describe how we removed the background area from the whole-slide images (WSI) to extract only the tissue area and generate non-overlapping patches. We also outline our data preparation and augmentation steps for model training. The details are discussed in the following subsections.

¹ <https://camelyon16.grand-challenge.org/Evaluation/>

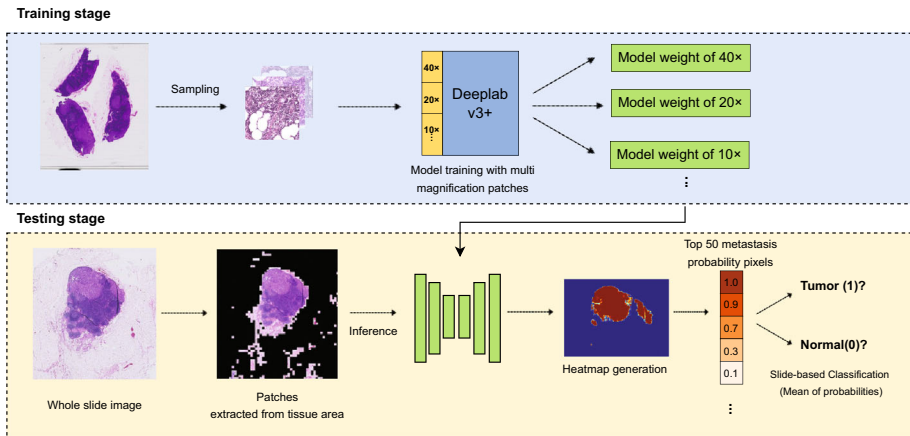


Fig. 1 Overview of the proposed pipeline

3.1.1 Patch sampling

In our data preprocessing pipeline, we use Openslide [17], a C library that provides a simple interface for reading whole-slide images, to open the Camelyon16 pathology image files. To prevent generating many white (empty) patches, similar to [24] and inspired by Otsu's seminal work [37], we convert WSIs from BGR to HSV color-space and threshold the three channels of HSV image in the range of 0 to 200 to obtain the tissue area. We then perform morphological open and closing operations to eliminate any noise present. Next, we use the `cv2.findContours` and `cv2.boundingRect` functions to generate contours of the areas containing cell tissues and to draw a bounding box outside the contours in preparation for the patch extraction process. Subsequently, we extract non-overlapping patches of size (768×768) using sliding windows inside the bounding boxes of these contours only. After generating a patch, a post-processing step is performed looking at the standard deviation of the patch in combination with the amount of black and white pixels, to investigate whether pixels in the patch are part of the artefacts in the background (e.g. some WSI's have black background). If that is the case, the patch is discarded. Using the same method, we extract the tissue area, calculate the percentage of tissue in each patch, and conduct the sampling as follows. For negative patch sampling, i.e. sampling of patches that do not contain any cancerous tissue, we only extract negative patches from normal WSI to avoid the interference of rough annotation. In addition, we select patches which contain more than 5% tissue area. As for positive patch sampling, patches containing more than 0.5% cancer area from tumor WSI are considered as positive patches. The ratio of positive and negative patches used for training is 1:3, which is different from some related work, such as [26, 27], that sample the same number of tumor and normal patches. The reason for increasing the ratio of negative patches in the training set is the following: unlike patch-level classification, semantic segmentation models enable more precise prediction but also significantly increase the chance of wrong predictions, and in particular false positives, since they perform pixel level classification. Such wrong predictions will highly impact the WSI classification results. Contrary to balancing the classes by performing under-sampling of the majority category as in [26, 27], we conjecture that a higher ratio of negative patches will allow the model to better capture variety in the data by avoiding loss of potentially useful information. We confirm this assumption empirically

by obtaining a higher WSI classification accuracy when training the model with ratio 1:3 instead of 1:1.

3.1.2 Patch augmentation

Since the slides of Camelyon16 are collected from two medical centers, staining variability is often present in the training dataset, which may be caused by different product of hematoxylin and eosin-stains or subtle difference in chemical staining procedures. Several publications indicate high staining variations will negatively affect the model performance [12, 26]. In general, there are two approaches to tackle this problem. The first approach is to implement stain-color normalization [33, 52]. The second approach is to force the model to ignore the color variation in patches by adopting hard color augmentation [49, 50]. In order to force the model to ignore staining variability, we adopt hard data augmentation due to the time-consuming nature of stain color normalization, which needs to be implemented during both the training and inference processes. We apply random hue, saturation, brightness and contrast as described in Table 1. In addition, we also add random flip to our data augmentation process.

3.2 Semantic segmentation model

For the semantic segmentation task we use the DeeplabV3+ architecture developed by Google [6], which proved to be highly effective in semantic segmentation tasks on the benchmark datasets, such as PASCAL VOC 2012 [14] and Cityscapes [10]. DeeplabV3+ architecture is shown in Fig. 2. The whole network can be considered as an Encoder-Decoder structure. For the encoder part, it mainly uses the architecture proposed in the DeeplabV3 paper [5], which includes dilated convolutional operators and Atrous Spatial Pyramid Pooling (ASPP) capable of extracting features at different scales of receptive fields. To ensure that the memory footprint of the activation maps and therefore model remains within the hardware boundaries, this scaling is done by dilating convolutional operators. The multi-scale traits of the architecture fit the hypothesis about achieving better tumor segmentation by enlarging the patch size, thus taking in account more information due to the larger scale. The formal definition of the atrous convolution operation on the input feature map x is given in Equation 1. Here f_i is the output feature map f at location i , \mathbf{c} is the convolutional filter with length ℓ and the dilation rate d determines the stride of the input \mathbf{x} [6].

The decoder part is for the first time introduced in DeeplabV3+ architecture. In the former generation of Deeplab model (DeeplabV3) [5], the encoder generates the feature maps with output stride 16, bilinearly upsampled to the original patch size. However, this method is

Table 1 Patch augmentation details

Methods	Details	Triggering Probability
Random Hue	maximum delta of 0.2	0.66
Random Saturation	random saturation factor in [0.5, 1.5]	0.66
Random Brightness	maximum delta of 0.5	0.66
Random Contrast	random contrast factor in [0.7, 1.3]	0.66
Flip	Left/right flip	0.66

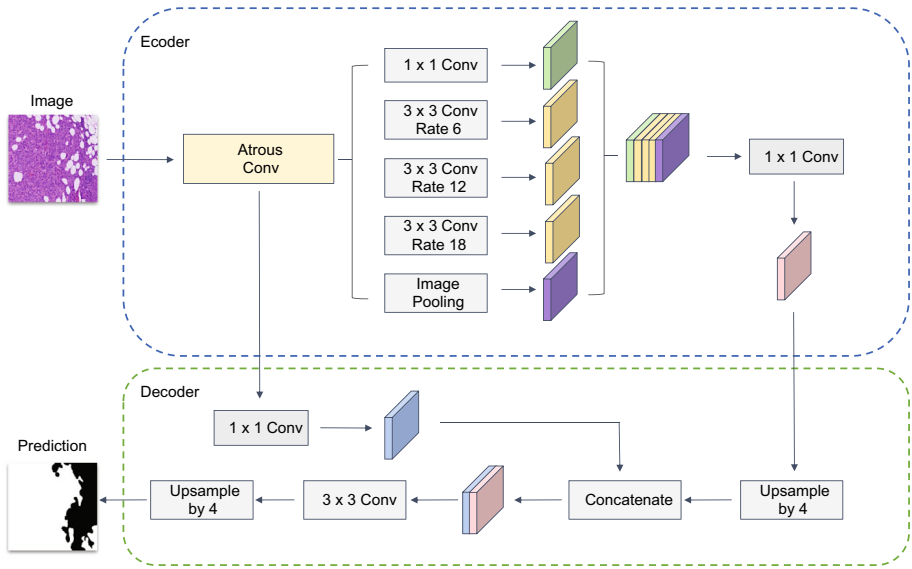


Fig. 2 An overview of DeeplabV3+ architecture

not effective in terms of obtaining detailed object boundaries. Therefore, the decoder first upsamples the feature map generated from the encoder by a factor of 4. It then concatenates this with the low-level features from the backbone with corresponding spatial resolution after 1×1 convolution. Subsequently, the merged feature map will be passed through another 3×3 convolution and upsampled again by a factor of 4 to the original input size.

$$f_i = \sum_{\ell=1}^{\ell} \mathbf{x}(i + d \cdot \ell) \mathbf{c}(\ell) \tag{1}$$

In addition, DeeplabV3+ uses atrous separable convolution to reduce the computation complexity of the proposed model while maintaining similar (or better) performance. It performs atrous convolution operations on each channel of input features separately, followed by a pointwise convolution to combine the output from atrous depthwise convolutions across channels. In our training process, we loaded the pre-trained PASCAL-VOC 2012 weights on the Xception backbone [8]. The Xception backbone in DeeplabV3+ has been modified as follows: (1) More layers are added. (2) All max pooling layers have been replaced with depthwise separable convolutions with strides. (3) Batch normalization and ReLU activation were introduced after each 3×3 depthwise separable convolution.

3.3 Pipeline of WSI prediction

In order to evaluate the performance of our model trained on multiple zoom levels and patch sizes, we set up the WSI prediction pipeline, which is also illustrated in the testing stage of Fig. 1. Similar to patch sampling, non-overlapping patches from different zoom levels are sequentially extracted from the bounding boxes of tissue area and fed into the trained model for inference. For example, 768×768 patches extracted from level 1 are sent to model trained on level 1 and so on. In order to accelerate the inference process, for level 0 and 1, we predict

the patches which contain more than 20% tissue area in order to avoid irrelevant predictions on background or fat tissue. As for level 2 and level 4, since the extracted patches have a relatively large field of view (FoV), we reduce the threshold of tissue area from 20% to 10% in order to avoid missing tissue area during the patch prediction. Figure 3 shows an example of 768×768 patches extracted from the test set from level 0 to 4. Note that the extracted patches cover the tissue area of WSI. Similar to [27], the patch prediction will be resized to the corresponding size on level 6 and stitched on a level 6 blank mask for post-processing. Since semantic segmentation inevitably leads to false positives, we first list the top 50 pixels from the WSI prediction based on the metastasis probability in descending order. Then, the mean of tumor probabilities is calculated and used as the likelihood of containing cancer for a single WSI.

3.4 Hardware configuration

The training exercise of this study was implemented on the GPU nodes from the LISA Compute Cluster at the Dutch National Supercomputing Centre SURFsara². The training with patch sizes from 384 to 1024 on each level is performed on two GPU nodes, each consisting of one Intel Xeon Gold 5118 Processor and four NVIDIA Titan RTX GPUs with 24 GB of GDDR6 memory. For the inference step, we still used NVIDIA Titan RTX GPU node but only one GPU was needed for prediction because the inference process is not very time consuming. Therefore, the figures of prediction time presented in Table 3 are based on using a single NVIDIA Titan RTX GPU.

4 Results

We conduct a series of experiments to answer the following research questions:

1. How does the model performance change when training with multi resolution pathology images? Is it possible to achieve even a better result at lower magnification levels?
2. Can such training protocol on WSIs yield a comparable performance to pathologist and state-of-the-art automatic approaches working with high-resolution images?
3. What is the difference in inference time between models trained on different WSI resolutions and pathologist performance?
4. Can the presented model trained with lower resolution images still generate refined external polygons of cancer region on the highest resolution?

4.1 Performance of the model trained on different zoom levels and patch sizes

The results of our model trained on different zoom levels and patch sizes performing the Camelyon16 WSI Classification task are shown in Fig. 4. The left part of the figure presents the ROC curves of DeeplabV3+ trained from zoom level 0 to 4 with same patch size 768×768 . AUC indicates the area under the receiver operating characteristic curve. We observe that the model trained on level 2 patches achieves the best AUC score of 0.9713. The figure on the right demonstrates the classification performance of models trained with four different patch sizes (i.e. 384, 512, 768, and 1024) on level 2. The model trained on 1024×1024 patch size achieves the best result (0.9741), and we observed a general trend suggesting that

² <https://userinfo.surfsara.nl/systems/lisa/description>

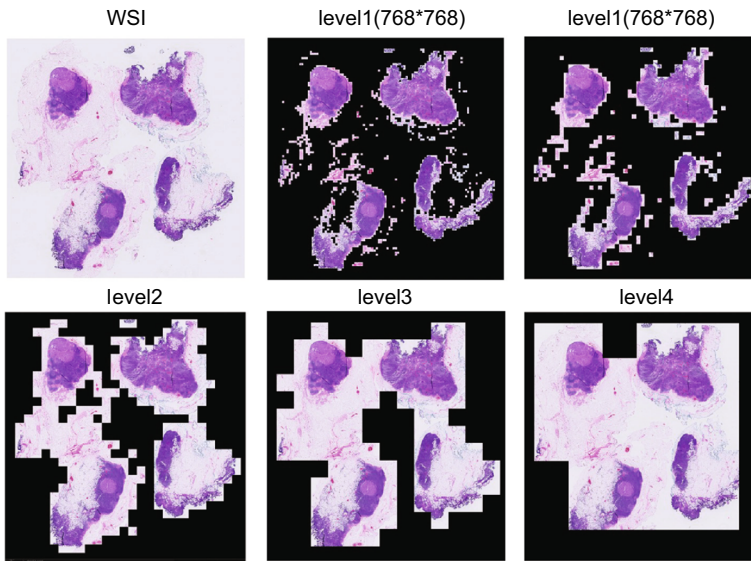


Fig. 3 Patches extracted for prediction from different zoom levels

larger patches lead to better model performance. In addition, interesting comparison between the model trained with 384×384 patches on level 2 and 768×768 patches on level 1 can be made because they contain the same FoV. The result shows that the model trained with 384×384 patches on level 2 achieves better results (0.9591) than 768×768 patches on level 1 (0.9564), which indicates that even in the condition of same FoV, the model trained on level 2 still gets better performance.

Figure 5 illustrates the comparison between our model trained on multiple zoom levels with a patch size of 768×768 pixels and the performance of pathologists, both without time constraints (WOTC) and with time constraints (WTC) [1]. It firstly shows a clear trend on how model performance changes with the increasing zoom levels. Namely, the model performance improves in the WSI classification task when the zoom level increases from 0 to 2, but dramatically drops at levels 3 and 4. We assume that the reason for lower performance on zoom levels 3 and 4 lies in excessive down-sampling losses, which make it for the segmentation

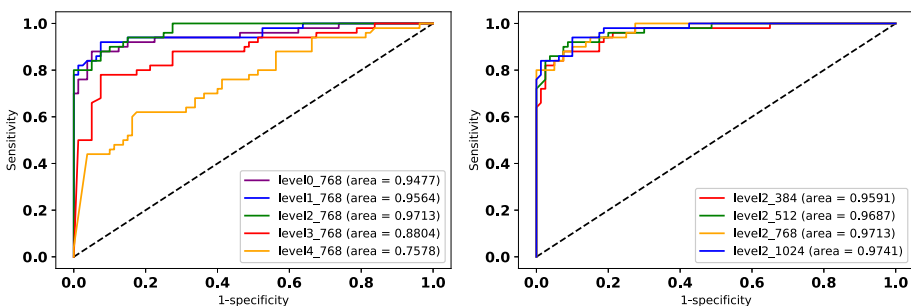


Fig. 4 Comparison results of our models trained on different zoom levels and patch sizes

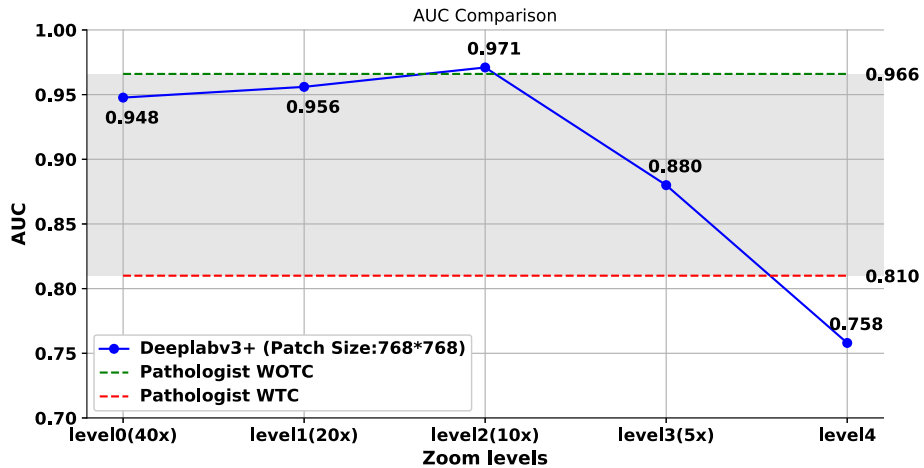


Fig. 5 Comparison results of our methods with pathologist performance

models very hard to distinguish tumor area from normal tissue. In addition, our model trained on zoom level 2 achieves better results than the pathologist WOTC.

4.2 Comparison with state-of-the-art

In addition, we compared our results with the official leaderboard of the Camelyon16 challenge and state-of-the-art papers using Camelyon16 dataset (Table 2). Our pipeline based on DeeplabV3+ model trained with 1024×1024 patches on level 2, achieved third place in Whole-slide-image classification public Leaderboard 1 [39]. It should be noted that a remarkable AUC score of 0.994 can be achieved by employing additional computationally demanding procedures, such as stain normalization and a smaller inference stride. However, this outcome is less noteworthy and presents considerable obstacles to clinical implementation due to the substantially increased computational expense. Ideally, the diagnostic process should be executed promptly with minimal resource utilization. Since 2019, several papers studied the performance of more recent segmentation models on Camelyon16 dataset. The study presented in [21] used ConcatNet (Four U-nets based on four histological features) and achieved an overall AUC 0.924. Another study [28] applied Unet and EffiNet, yielding an overall AUC of 0.935. The research presented in [18] implemented a classification model Inception-v3, followed by the use of a semantic segmentation model DCNN for enhanced segmentation, leading to an overall AUC of 0.966. However, those results didn't exceed the pathologist performance WOTC. Our model trained on level 2 patches obtained the best result (AUC 0.9742) for semantic segmentation on Camelyon16 dataset in terms of WSI classification task.

4.3 Training and prediction time comparison

Next to prediction accuracy, another important factor influencing clinical adoption is the average inference time on a single WSI. We only use one NVIDIA Titan RTX GPU to perform inference. Table 3 shows the Training and inference time comparison among our models trained on multiple zoom levels and patch sizes. The model trained on level 2 with

Table 2 Comparison of the results obtained by our model trained on different zoom levels and state-of-the-art methods; we make a distinction between patch level classification and semantic segmentation prediction approaches

Author	Method	Prediction method	AUC
Pathologist [1]	N/A	N/A	0.966
Wang et al. [46]	GoogLeNet	Patch level classification	0.994
Liu et al. [28]	ResNet	Patch level classification	0.976
Zhang et al. [53]	DPN, Swin-Transformer, SVM	Patch level classification	0.961
Shen et al. [40]	Pathology Deformable Conditional Random Field	Patch level classification	0.920
Yu et al. [51]	Bayesian Collaborative Learning	Patch level classification	0.956
Tourniaire et al. [44]	Mixedly Supervised-CLAM	Patch level classification	0.982
Khaliliboroujeni et al. [22]	ResNet50, spatial and channel attentions	Patch level classification	0.970
Quincy Wong [1]	SegNet	Semantic Segmentation	0.865
Jin et al. [21]	ConcatNet	Semantic Segmentation	0.924
Liu et al. [30]	Unet, EffiNet	Semantic Segmentation + Patch level classification	0.935
Guo et al. [18]	v3_DCNN_1280 (level0)	Patch level classification + Semantic Segmentation	0.966
Our method	DeeplabV3+_1024 (level2)	Semantic Segmentation	0.974
Our method	DeeplabV3+_768 (level2)	Semantic Segmentation	0.971
Our method	DeeplabV3+_512 (level2)	Semantic Segmentation	0.969

a patch size of 1024×1024 has the shortest inference time (42 s per slide), which is lower than that needed by a pathologist WTC (56 s per slide) [1]. This is because the number of patches extracted from lower resolution WSI for prediction is much lower than in case of the highest resolution. Based on the experimental results above, we propose that zoom level 2 ($10\times$ magnification) would be the appropriate magnification level to perform semantic segmentation on.

4.4 Segmentation result presentation

As we described previously, one of many advantages of performing semantic segmentation compared with patch-level classification is that it can offer refined cancer regions segmentation on high resolution WSIs, which can assist pathologists in identifying small metastasis area. This method could also be very useful when considered as a computer-aided annotation tool, which can significantly reduce the time of annotation for a pathologist. Therefore, it is worth evaluating the real segmentation performance of our model on both higher and lower resolution patches. In this study, we select our model trained on level 2 with a patch size of 768×768 , as well as a model trained on level 1 with the same patch size, to generate binary predictions from heatmaps using a threshold of 0.5. Subsequently, we use `cv2.drawContours` function from OpenCV³ library to draw external polygons of cancer

³ <https://opencv-python-tutroals.readthedocs.io/>

Table 3 Training and prediction time comparison of our models trained on different zoom levels and patch sizes

Zoom level	Training set size	Batch size	Iterations	Training (h)	Prediction (h)	Prediction/ slide (m)
level 0(768 × 768)	56,632	4	25,000	10.12	21.31	9.83
level 1(768 × 768)	34,524	4	25,000	9.72	4.75	2.19
level 2(1024 × 1024)	8,494	2	50,000	28.42	1.52	0.70
level 2(768 × 768)	10,420	4	25,000	9.58	1.55	0.71
level 2(512 × 512)	23,256	4	25,000	6.2	2.12	0.97
level 2(384 × 384)	37,688	4	25,000	4.92	3.1	1.43
level 3(768 × 768)	4,184	4	25,000	9.96	0.58	0.26
level 4(768 × 768)	1,992	4	25,000	9.23	0.23	0.18

Pathologist WOTC: 13.95 m/slide
Pathologist WTC: 0.93 m/slide

area. The segmentation result is shown in Fig. 6. Both models trained on level 1 and 2 can generate decent external polygons of the cancer area. However, we observe that the model trained on level 2 performs more precise segmentation than the one trained on level 1. The yellow arrows in Fig. 6 show that some subtle normal tissue and blank areas between tumor areas have been detected by the level 2 model but not by the level 1 one. In addition, several publications have discussed the issue of inaccurate labeling in histopathological imaging, including the Camelyon16 dataset [7, 16, 35]. Manually annotating large whole slide images on the pixel level of the highest magnification is an inherently challenging task, making it difficult to avoid unreliable labeling in practice. Notably, our model has the ability to accurately identify and label these areas. The fourth row of Fig. 6 illustrates an example in which some adipocytes were mistakenly labeled as tumors in the original mask, but our model trained on level 2 patches was able to accurately identify and correct them. Moreover, column (d) of Fig. 6 shows that even at the highest resolution, the polygons generated by our model trained on level 2 can still fit well with the tumor boundary, which indicates the interesting potential of our method in assisting pathologists with annotating high resolution images.

We conjecture that the reason why the model on level 2 obtains the best result could be two-fold. Firstly, similar-sized patches extracted from lower resolution WSI offer a larger field-of-view (FoV), which contributes to better model performance in terms of segmentation and slide-level classification [18]. Secondly, although semantic segmentation models highly rely on the quality of annotation when performing supervised learning, it is nearly impossible for the pathologist to annotate pixel by pixel on the high resolution image. Therefore, training segmentation model on the highest magnification level would enable the model to obtain more details of tissue but at the same time, rough annotations also have a strongly negative impact on the performance of the model. Appropriate downsampling can make the rough annotation on the highest resolution not obvious, therefore increasing the robustness of the segmentation model. Similar effect can be observed in using Gaussian filter to smooth the image and reduce the level of noise, which trades for better segmentation model performance [3, 36].

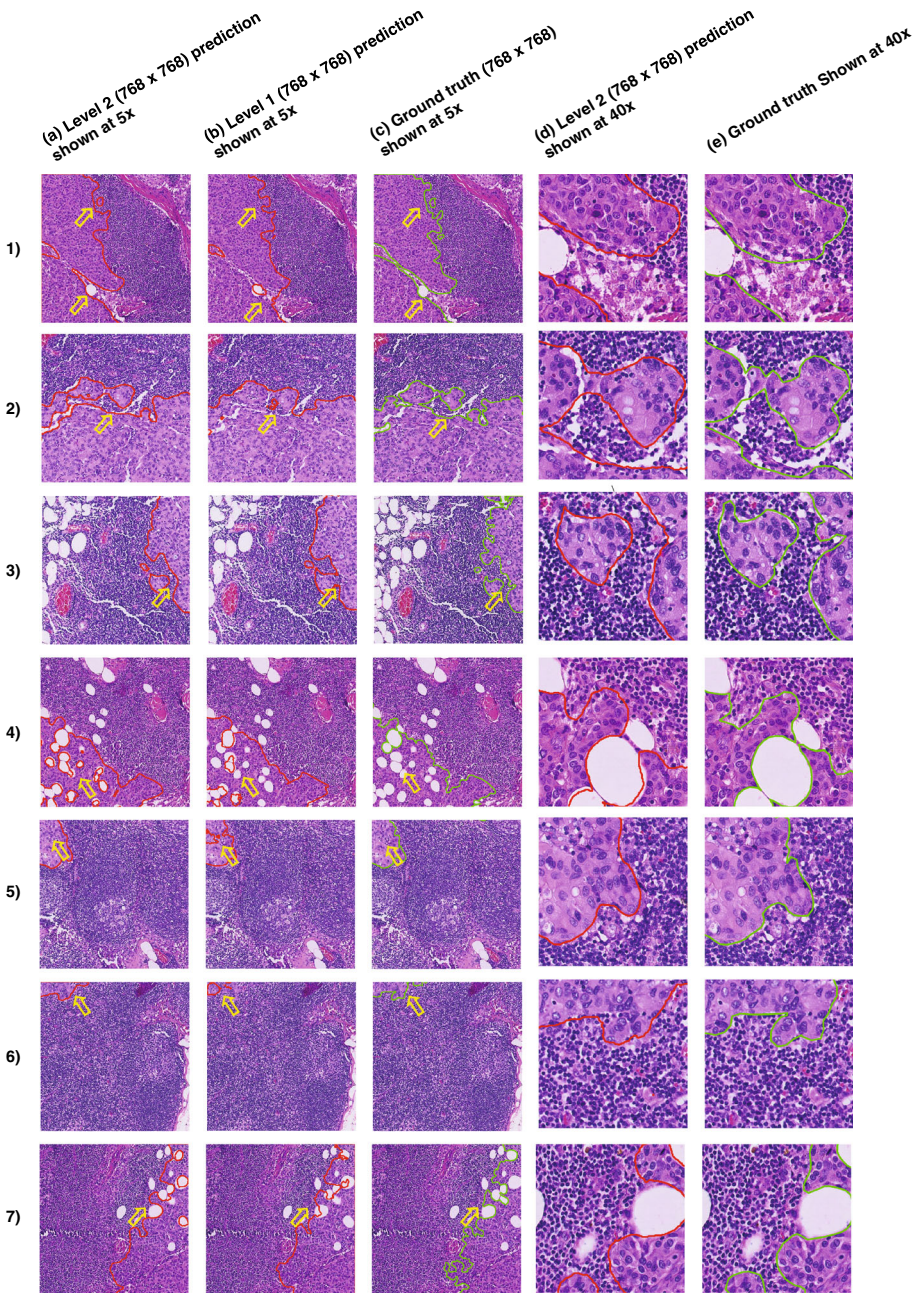


Fig. 6 External polygons of tumor regions generated by our models. From left to right, (a) is the prediction of model trained on level 2 shown at level 3 (5×) patches, (b) is the prediction of model trained on level 1 shown at level 3 (5×) patches, (c) is the Ground truth shown at level 3 (5×), (d) is the prediction of model trained on level 2 shown at level 0 (40×) patches, (e) is the ground truth shown at level 0 (40×) patches

5 Conclusion

In this paper, we study the impact of adopted resolution level and patch size on the performance of cancer region segmentation based on whole-slide images (WSIs) of sentinel lymph node. Firstly, we observe that the model achieves the best performance in the WSI classification task on zoom level 2 under the condition of same patch size and same field-of-view. Such results indicate that zoom level 2 (10×) constitutes an appropriate magnification level for training and prediction of the semantic segmentation model in cancer metastasis detection. This interesting finding shows the possibility of efficiently training deep convolutional neural networks with lower resolution images. Secondly, the model trained on zoom level 2 patches achieved better performance than pathologist WOTC and, at same time, lower inference time than the pathologist WTC, which may help potential clinical adoption. Conclusively, our model trained on zoom level 2 patches can perform more precise segmentation than the same model trained on the higher-resolution patches, while still being able to generate refined external polygons of cancer regions on the highest resolution image. It indicates the potential of using these techniques as the basis for computer-aided annotation tools that help pathologists reduce the time taken by labeling. Further research on WSI analysis could include training models on zoom level 2 with even larger patch sizes, which can possibly lead to the training based on the entire WSI using more HPC capacity. The study could help identify the patch size for which the model performance becomes saturated and also show the limitations of current HPC facilities in terms of memory capacity. Moreover, it would be also interesting to explore the effects of using both data-parallelism and model parallelism strategies to process even larger patch sizes with the same model and hardware constraints.⁴

Acknowledgements This study is supported by SURFsara national HPC center through Examode project within EU Horizon 2020 framework. Examode stand for Extreme-scale Analytics via Multimodal Ontology Discovery and Enhancement, which focuses on performing fast weakly supervised knowledge discovery of exascale heterogeneous data with limited human interaction. We also would like to thank Axel Berg, PhD, provided general support and critical reading of this manuscript

Funding Financial support for this study was provided in part by Atos through the HPC, AI and Quantum Life Sciences Centre of Excellence (CEPP); as well as by SURF, the collaborative organization for IT in Dutch education and research. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Jieyi Li, Anwar Osseyran, Ruben Hekster, Stevan Rudinac, Valeriu Codreanu, Damian Podareanu declare that they do not have any financial or personal relationships with other people or organizations that could have inappropriately influenced this study.

Data Availability Statement The datasets generated during and/or analysed during the current study are available in the Camelyon16 dataset repository, <http://gigadb.org/dataset/100439>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

⁴ <https://www.examode.eu/>

References

1. Bejnordi BE, Veta M, Van Diest PJ et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association* 318(22):2199–2210. <https://doi.org/10.1001/jama.2017.14585>
2. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
3. Cai K, Yang R, Chen H, Li L, Zhou J, Ou S, Liu F (2017) A framework combining window width-level adjustment and gaussian filter-based multi-resolution for automatic whole heart segmentation. *Neuro-computing* 220:138–150
4. Chanchal AK, Lal S, Kini J (2022) Deep structured residual encoder-decoder network with a novel loss function for nuclei segmentation of kidney and breast histopathology images. *Multimedia Tools and Applications* 81(7):9201–9224
5. Liu S, Guo C, Al-Turjman F, Muhammad K, de Albuquerque VHC: Reliability of response region: a novel mechanism in visual tracking by edge computing for iiot environments. *Mechanical systems and signal processing* 138, 106537 (2020)
6. Liu X, He J, Song L, Liu S, Srivastava G: Medical image classification based on an adaptive size deep learning model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17(3s), 1–18 (2021)
7. Liu X, Song L, Liu S, Zhang Y: A review of deep-learning-based medical image segmentation methods. *Sustainability* 13(3), 1224 (2021)
8. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH: Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, pp 2424–2433 (2016)
9. Li Y, Ping W: Cancer metastasis detection with neural conditional random field. In: *Medical Imaging with Deep Learning* (2018)
10. Wang J, Xu Z, Pang Z-F, Huo Z, Luo J: Tumor detection for whole slide image of liver based on patch-based convolutional neural network. *Multimedia Tools and Applications* 80(11), 17429–17440 (2021)
11. Deepa BG, Senthil S (2022) Predicting invasive ductal carcinoma tissues in whole slide images of breast cancer by using convolutional neural network model and multiple classifiers. *Multimedia Tools and Applications* 81:8575–8596. <https://doi.org/10.1007/S11042-022-12114-9/FIGURES/18>
12. Dimitriou N, Arandjelović O, Caie PD: Deep Learning for Whole Slide Image Analysis: An Overview. *Frontiers in Medicine* 6 (2019). <https://doi.org/10.3389/fmed.2019.00264>
13. Singh S, Kumar R: Breast cancer detection from histopathology images with deep inception and residual blocks. *Multimedia Tools and Applications* 81(4), 5849–5865 (2022)
14. Murtaza G, Shuib L, Mujtaba G, Raza G: Breast cancer multi-classification through deep neural network and hierarchical classification approach. *Multimedia Tools and Applications* 79(21), 15481–15511 (2020)
15. Chanchal AK, Lal S, Kini J: Deep structured residual encoder-decoder network with a novel loss function for nuclei segmentation of kidney and breast histopathology images. *Multimedia Tools and Applications* 81(7), 9201–9224 (2022)
16. Codreanu V, Podareanu D, Saletore V: Large minibatch training on supercomputers with improved accuracy and reduced time to train. In: *2018 IEEE/ACM Machine Learning in HPC Environments (MLHPC)*, pp 67–76 (2018). IEEE
17. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M (2013) Openslide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics* 4(1):27. <https://doi.org/10.4103/2153-3539.119005>
18. Guo Z, Liu H, Ni H, Wang X, Su M, Guo W, Wang K, Jiang T, Qian Y: A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images. *Scientific Reports* 9(1) (2019)
19. Gupta I, Nayak SR, Gupta S, Singh S, Verma K, Gupta A, Prakash D: A deep learning based approach to detect idc in histopathology images. *Multimedia Tools and Applications*, 1–22 (2022)
20. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH: Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016)
21. Jin YW, Jia S, Ashraf AB, Hu P (2020) Integrative data augmentation with u-net segmentation masks improves detection of lymph node metastases in breast cancer patients. *Cancers* 12(10):2934
22. Khalilboroujeni S, He X, Jia W, Amirgholipour S (2022) End-to-end metastasis detection of breast cancer from histopathology whole slide images. *Computerized Medical Imaging and Graphics* 102:102136
23. Khan AI, Al-Habshi S (2020) Machine learning in computer vision. *Procedia Computer Science* 167: 1444–1451

24. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B (2021) A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports* 11(1):1–14
25. Komura D, Ishikawa S (2018) Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal* 16:34–42. <https://doi.org/10.1016/J.CSBJ.2018.01.001>
26. Tourniaire P, Ilie M, Hofman P, Ayache N, Delingette H: Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Analysis* 85, 102763 (2023)
27. Khalilboroujeni S, He X, Jia W, Amirgholipour S: End-to-end metastasis detection of breast cancer from histopathology whole slide images. *Computerized Medical Imaging and Graphics* 102, 102136 (2022)
28. Jin YW, Jia S, Ashraf AB, Hu P: Integrative data augmentation with u-net segmentation masks improves detection of lymph node metastases in breast cancer patients. *Cancers* 12(10), 2934 (2020)
29. Liu S, Ren J, Chen Z, Hu K, Xiao F, Li X, Gao X: Effdiag: an efficient framework for breast cancer diagnosis in multi-gigapixel whole slide images. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp 663–669 (2020). IEEE
30. Guo Z, Liu H, Ni H, Wang X, Su M, Guo W, Wang K, Jiang T, Qian Y: A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images. *Scientific Reports* 9(1) (2019)
31. Liu S, Guo C, Al-Turjman F, Muhammad K, de Albuquerque VHC (2020) Reliability of response region: a novel mechanism in visual tracking by edge computing for iiot environments. *Mechanical systems and signal processing* 138:106537
32. Liu X, Song L, Liu S, Zhang Y (2021) A review of deep-learning-based medical image segmentation methods. *Sustainability* 13(3):1224
33. Efron B: Bootstrap methods: another look at the jackknife. In: *Breakthroughs in Statistics* vol. 501, pp. 569–593 (1979). <https://doi.org/10.2307/2958830>
34. Murtaza G, Shuib L, Mujtaba G, Raza G (2020) Breast cancer multi-classification through deep neural network and hierarchical classification approach. *Multimedia Tools and Applications* 79(21):15481–15511
35. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B: A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports* 11(1), 1–14 (2021)
36. Nasor M, Obaid W (2021) Segmentation of osteosarcoma in mri images by k-means clustering, chan-veye segmentation, and iterative gaussian filtering. *IET Image Processing* 15(6):1310–1318
37. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9(1):62–66. <https://doi.org/10.1109/TSMC.1979.4310076>
38. Pantanowitz L (2010) Digital images and the future of digital pathology. *Journal of Pathology Informatics* 1(1):15. <https://doi.org/10.4103/2153-3539.68332>
39. Magee D, Treanor D, Crellin D, Shires M, Smith K, Mohee K, Quirke P: Colour normalisation in digital histopathology images. In: *Proc Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, vol 100, pp 100–111 (2009). Citeseer
40. Shen Y, Shen D, Ke J (2022) Identify representative samples by conditional random field of cancer histology images. *IEEE Transactions on Medical Imaging* 41(12):3835–3848
41. Xu L, Walker B, Liang P-I, Tong Y, Xu C, Su YC, Karsan A: Colorectal cancer detection based on deep learning. *Journal of Pathology Informatics* 11(1), 28 (2020)
42. Singh S, Kumar R (2022) Breast cancer detection from histopathology images with deep inception and residual blocks. *Multimedia Tools and Applications* 81(4):5849–5865
43. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 801–818 (2018)
44. Tourniaire P, Ilie M, Hofman P, Ayache N, Delingette H (2023) Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Analysis* 85:102763
45. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E, et al: Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018 (2018)
46. Chen L-C, Papandreou G, Schroff F, Adam H: Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587* (2017)
47. Wang J, Xu Z, Pang Z-F, Huo Z, Luo J (2021) Tumor detection for whole slide image of liver based on patch-based convolutional neural network. *Multimedia Tools and Applications* 80(11):17429–17440
48. Wang T, Lei S, Jiang Y, Chang C, Snoussi H, Shan G, Fu Y (2022) Accelerating temporal action proposal generation via high performance computing. *Frontiers of Computer Science* 16(4):1–10
49. Wu Y, Koyuncu CF, Toro P, Corredor G, Feng Q, Buzzy C, Old M, Teknos T, Connelly ST, Jordan RC et al (2022) A machine learning model for separating epithelial and stromal regions in oral cavity squamous cell carcinomas using h&e-stained histology images: A multi-center, retrospective study. *Oral Oncology* 131:105942

50. Xu L, Walker B, Liang P-I, Tong Y, Xu C, Su YC, Karsan A (2020) Colorectal cancer detection based on deep learning. *Journal of Pathology Informatics* 11(1):28
51. Yu J-G, Wu Z, Ming Y, Deng S, Wu Q, Xiong Z, Yu T, Xia G-S, Jiang Q, Li Y: Bayesian collaborative learning for whole-slide image classification. *IEEE Transactions on Medical Imaging* (2023)
52. Zanjani FG, Zinger S, de With PHN, Bejnordi BE, van der Laak JAWM: Histopathology stain-color normalization using deep generative models, 1–11 (2018)
53. Zhang X, Liu C, Li T, Zhou Y (2023) The whole slide breast histopathology image detection based on a fused model and heatmaps. *Biomedical Signal Processing and Control* 82

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.