Multi_Scale_Tools: A Python Library to Exploit Multi-Scale Whole Slide Images

Niccolò Marini[1,2]*, Sebastian Otálora[1,2], Damian Podareanu[3], Mart van Rijthoven[4], Jeroen van der Laak[4,5], Francesco Ciompi[4], Henning Müller[1,6] and Manfredo Atzori[1,7]

[1]Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland, [2]Centre Universitaire d'Informatique, University of Geneva, Carouge, Switzerland, [3]SURFsara, Amsterdam, Netherlands, [4]Department of Pathology, Radboud University Medical Center, Nijmegen, Netherlands, [5]Center for Medical Image Science and Visualization, Linkoping University, Linkoping, Sweden, [6]Medical Faculty, University of Geneva, Geneva, Switzerland, [7]Department of Neurosciences, University of Padua, Padua, Italy

Algorithms proposed in computational pathology can allow to automatically analyze digitized tissue samples of histopathological images to help diagnosing diseases. Tissue samples are scanned at a high-resolution and usually saved as images with several magnification levels, namely whole slide images (WSIs). Convolutional neural networks (CNNs) represent the state-of-the-art computer vision methods targeting the analysis of histopathology images, aiming for detection, classification and segmentation. However, the development of CNNs that work with multi-scale images such as WSIs is still an open challenge. The image characteristics and the CNN properties impose architecture designs that are not trivial. Therefore, single scale CNN architectures are still often used. This paper presents Multi_Scale_Tools, a library aiming to facilitate exploiting the multi-scale structure of WSIs. Multi_Scale_Tools currently include four components: a pre-processing component, a scale detector, a multi-scale CNN for classification and a multi-scale CNN for segmentation of the images. The pre-processing component includes methods to extract patches at several magnification levels. The scale detector allows to identify the magnification level of images that do not contain this information, such as images from the scientific literature. The multi-scale CNNs are trained combining features and predictions that originate from different magnification levels. The components are developed using private datasets, including colon and breast cancer tissue samples. They are tested on private and public external data sources, such as The Cancer Genome Atlas (TCGA). The results of the library demonstrate its effectiveness and applicability. The scale detector accurately predicts multiple levels of image magnification and generalizes well to independent external data. The multi-scale CNNs outperform the single-magnification CNN for both classification and segmentation tasks. The code is developed in Python and it will be made publicly available upon publication. It aims to be easy to use and easy to be improved with additional functions.

Keywords: multi-scale approaches, computational pathology, scale detection, classification, segmentation, deep learning

1 INTRODUCTION

The implicit multi-scale structure of digitized histopathological images represents an open challenge in computational pathology. Training machine learning algorithms that can simultaneously learn both microscopic and macroscopic tissue structures comes with technical and computational challenges that are not yet well studied.

As of 2021, histopathology represents the gold standard to diagnose many diseases, including cancer (Aeffner et al., 2017; Rorke, 1997). Histopathology images include several tissue structures, ranging from microscopic entities (such as single cell nuclei) to macroscopic components (such as tumor bulks). Whole Slide Images (WSIs) are digitized histopathology images that are scanned at high-resolution and are stored in a multi-scale (pyramidal) format. WSI resolution is related to the spatial resolution and the optical resolution used to scan the images (Wu et al., 2010). The spatial resolution is the minimum distance that the scanner can capture so that two objects are still distinguished, measured in terms of μm per pixel (Sellaro et al., 2013). The optical resolution (or magnification) is the magnification factor (x) of the lens within the scanner (Sellaro et al., 2013). Currently, the *de facto* standard spatial resolutions adopted to scan tissue samples (for example in The Cancer Genome Atlas) are usually 0.23–0.25 μm (magnification ×40) or 0.46–0.50 μm (magnification ×20). Tissue samples such as surgical resection samples (or specimens) are often approximately 20 mm × 15 mm in size[1], while samples such as biopsies are approximatively 2 mm × 6 mm in size. The size of the samples combined with the spatial resolution of the scanners leads to gigapixel images: image size can reach 200 000 × 200 000 pixels, meaning gigabytes of pixel data. The multi-scale WSI format (**Figure 1**) includes several magnification levels (with a different spatial resolution) of the sample, stored in a pyramid, usually varying between ×1.25 and 40x. The baseline image of the pyramid is the one at the highest resolution. The multi-scale structure of the images allows pathologists to analyze the image from the lowest to the highest magnification level. Pathologists analyze the images by first identifying a few regions of interest and zooming afterwards through them to visualize different details of the tissue (Schmitz et al., 2019). Each magnification level includes different types of information (Molin et al., 2016), since tissue structures appear in different ways according to their magnification level. Therefore, it is essential to detect an abnormality and detect it in a specific range of levels. The characteristics of microscopes and scanners often lead to a scale-dependent analysis. For example, at middle magnification levels (such as 5–10x) it is possible to distinguish between glands, while at the highest ones (such as 20–40x) it is possible to better resolve cells. **Figure 2** includes examples of tissues scanned at different magnification levels.

Computational pathology is the computational analysis of digital images obtained through scanning slides of cells and tissues (van der Laak et al., 2021). Currently, deep Convolutional Neural Networks
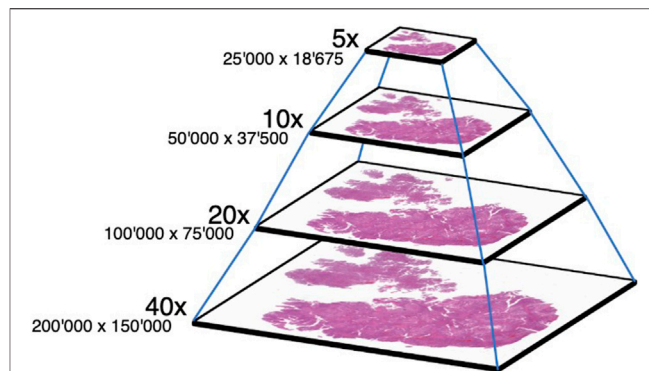

FIGURE 1 | An example of WSI format including multiple magnification levels. The size of each image of the pyramid is reported under the magnification level in terms of pixels.
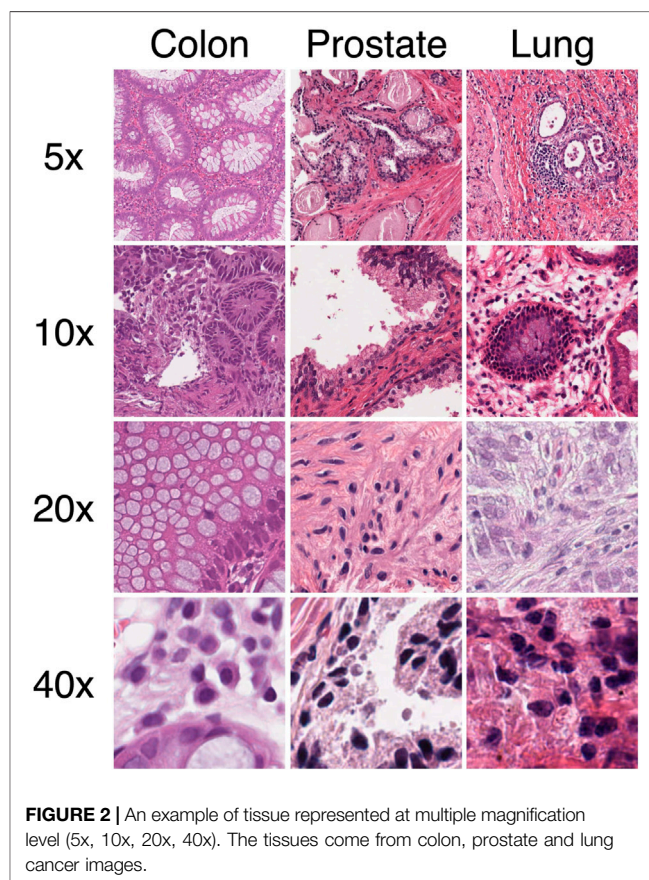

FIGURE 2 | An example of tissue represented at multiple magnification level (5x, 10x, 20x, 40x). The tissues come from colon, prostate and lung cancer images.

(CNNs) are the state-of-the-art machine learning algorithms in computational pathology tasks, in particular for classification (del Toro et al., 2017; Arvaniti and Claassen, 2018; Coudray et al., 2018; Komura and Ishikawa, 2018; Ren et al., 2018; Campanella et al., 2019; Roy et al., 2019; Iizuka et al., 2020) and segmentation (Ronneberger et al., 2015; Paramanandam et al., 2016; Naylor et al., 2017; Naylor et al., 2018; Wang et al., 2019) of images. Their success relies on automatically learning the relevant

---

[1]http://dicom.nema.org/Dicom/DICOMWSI/. Retrieved 13th of November, 2020

features from the input data. However, usually, CNNs cannot easily handle the multi-scale structure of the images since they are not scale-equivariant by design (Marcos et al., 2018; Zhu et al., 2019) and because of WSI size. The equivariance property of a transformation means that when a transformation is applied, it is possible to predict how the representation will change (Lenc and Vedaldi, 2015; Tensmeyer and Martinez, 2016). This is not normally true for CNNs, because if a scale transformation is applied to the input data, it is usually not possible to predict its effect on the output of the CNN. The knowledge about the scale is essential for the model to identify diseases, since the same tissue structures, represented at different scales, include different information (Janowczyk and Madabhushi, 2016). CNNs can identify abnormalities in tissues, but the information and the features related to the abnormalities are not the same for each scale representation (Jimenez-del Toro et al., 2017). Therefore, the proper scale must be selected to train CNNs (Gecer et al., 2018; Otálora et al., 2018b). Unfortunately, scale information is not always available into images. This is the case, for instance, of pictures taken with standard cameras or processed in compression and resolution, such as images downloaded from the web or images included in scientific articles. Furthermore, modern hardware (Graphic Processing Units, GPUs) cannot easily handle WSIs, due to their large pixel size and the limited video random access memory space that has to temporally store it. The combination of different magnification levels leads to larger images, making it even harder to analyze the images.

The characteristics of the WSIs can lead to modification of CNNs in terms of architecture, both for classification (Jimenez-del Toro et al., 2017; Lai and Deng, 2017; Gecer et al., 2018; Yang et al., 2019; Hashimoto et al., 2020) and segmentation (Ronneberger et al., 2015; Li et al., 2017; Salvi and Molinari, 2018; Schmitz et al., 2019; van Rijthoven et al., 2020), such as multi-brach networks (Yang et al., 2019; Hashimoto et al., 2020; Jain and Massoud, 2020), multiple receptive fields convolutional neural networks (Han et al., 2017; Lai and Deng, 2017; Ullah, 2017; Li et al., 2019; Zhang et al., 2020) and U-Net based networks (Bozkurt et al., 2018; van Rijthoven et al., 2020). The modification of architectures to include multiple scales is prevalent in medical imaging, since it can allow to identify examples of architecture's modifications also from other modalities, such as MRI imaging (Zeng et al., 2021a) and Gold immunochromatographic strip (GIGS) images (Zeng et al., 2019; Zeng et al., 2021b).

The code library (called Multi_Scale_Tools) described in this paper contributes to alleviate the mentioned problems by presenting tools that allow handling and exploiting histopathological images' multi-scale structure end-to-end CNN architectures. The library includes pre-processing tools to extract multi-scale patches, a scale detector, a component to train a multi-scale CNN classifier and a component to train a multi-scale CNN for segmentation. The tools are platform-independent and developed in Python. The code is publicly available at https://github.com/sara-nl/multi-scale-tools. Multi_Scale_Tools is aimed at being easy to use and easy to be improved with additional functions.

2 METHODS

The library includes four components: a pre-processing tool, a scale detector tool, a component to train a multi-scale CNN classifier and a component to train a multi-scale segmentation CNN. Each tool is described in a dedicated subsection as follows:

- Pre-processing component, Sub-**section 2.1**
- Scale detector, Sub-**section 2.2**
- Multi-scale CNN for classification, Sub-**section 2.3**
- Multi-scale CNN for segmentation, Sub-**section 2.4**

2.1 Pre-Processing Component

The pre-processing component allows researchers to generate multi-scale input data. The component includes two parametric and scalable methods to extract patches from the different magnification levels of a WSI: the *grid* extraction and the *multi − center* extraction method. Both methods need a WSI and the corresponding tissue mask as input, and they both produce images and metadata as output. The *grid* extraction methods (Patch_Extractor_Dense_Grid.py, Patch_Extractor_Dense_Grid_Strong_Labels.py), allow to extract patches from one magnification level (**Figure 3**). The tissue mask is split in a grid of patches according to the following parameters: magnification level, mask magnification, patch size, and stride between the patches. The output of the method is a set of patches selected according to the parameters. The *multi − center* extraction methods (Patch_Extractor_Dense_Centroids.py, Patch_Extractor_Dense_Centroids_Strong_Labels.py) allow to extract patches from multiple magnification levels. According to the user's highest magnification level, the tissue mask is split into a grid (as done in the functions previously described). The patches within this grid are called centroids. Each centroid is used to generate the coordinates for a patch at a lower magnification level, so that the latter includes the centroid (the patch at the highest magnification level) in its central section. The method's output is a set of tuples, each one including patches at different magnification levels (**Figure 4**). Compared with other patch extraction methods, such as the one presented in (Lu et al., 2021), this pre-processing component has two main characteristics. The first one is that the component extracts patches from multiple magnification levels of the WSIs, pairing the patches coming from the same region of the image. The second one is that the component allows extracting patches from an arbitrary magnification level, despite the magnification level not being included in the WSI. Usually, patch extractor methods extract patches only from the magnification levels stored in the WSI format (M_a), such as 40x, 20x, 10x, 5x, 2.5x and 1.25x. This process is driven by the input parameters that include both the patch size (P_w) and the magnification wanted (M_w). The method extracts a patch of size P_a from a magnification stored in the WSI and afterwards the patch is resized to P_w.

$$P_w : M_w = P_a : M_a. \qquad (1)$$

In both methods, only patches from tissue regions are extracted and saved using tissue masks, distinguishing between
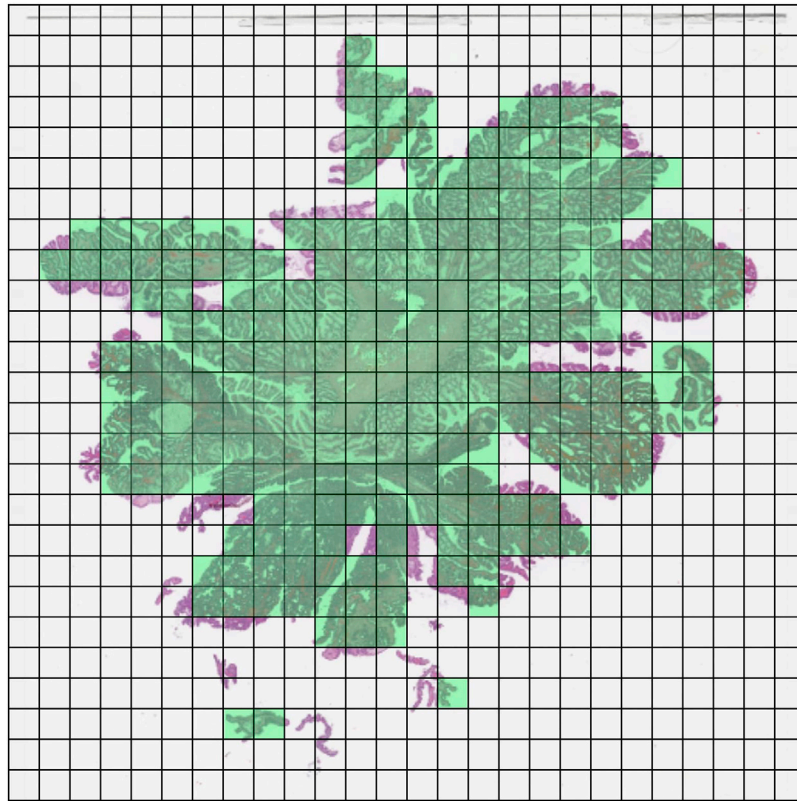
FIGURE 3 | An example of the *grid* extraction method. The patches in green are selected since they contain enough tissue.

patches from tissue regions and patches from the background. The methods are developed to work with masks including tissue and, in case they are available, with pixel-wise annotated masks. In the case of tissue masks, the tissue masks are generated using HistoQC tools (Janowczyk et al., 2019). The HistoQC configuration adopted is reported in the repository. In the case of pixel-wise annotations, the masks must be firstly converted to a RGB image.

Besides the patches, the methods save also metadata file (csv files). The metadata includes information regarding the magnification level where the patches are extracted and the x and y coordinates of the patches' upper left corner. The scripts are developed to be multi-thread, in order to exploit hardware architectures with multiple cores. In the **Supplementary Materials** section, the parameters for the scripts are described in more detail.

2.2 Scale Detector

The scale detector tool is a CNN trained to estimate the magnification level of a given patch or image. This task has been explored in the past Otálora et al. (2018a), Otálora et al. (2018b) in the prostate and breast tissue types. Similar approaches have been recently extended to different organs in the TCGA repository Zaveri et al. (2020). The tool involves the scripts related to the training of the models (the input data generation, the training and testing modules) and a module to use the detector as a standalone

component that performs the magnification inference for new images. The models are trained in a fully-supervised fashion. Therefore, the scripts to train them need a set of patches and the corresponding magnification level as input, which are provided into csv files, including the patch path and the corresponding magnification levels. Two scripts are developed to generate the input files, assuming that the patches are previously generated with the pre-processing components, described in the previous section. The first script is made to split the WSIs into partitions (Create_csv_from_partitions.py), which generates three files (the input data for training, validation and testing partitions) starting from three files (previously prepared by the user) including the names of the WSIs. The second script (Create_csv.py) generates an input data csv starting from a list of files. The model is trained (Train_regressor.py) and tested (Test_regressor.py) with several magnification levels that the user can choose (in this paper, 5x, 8x, 10x, 15x, 20x, 30x, 40x were used). Training the model with patches from a discrete and small set of scales can lead to regressors that are precise to estimate the magnifications close to input scales, and less precise when scales are far from them. Therefore, a scale augmentation technique was applied to patches and labels during the training (in addition to more standard augmentation techniques adopted such as rotation, flipping and color augmentation). In order to perform scale augmentation, the image is randomly cropped of a factor and resized to the original patch size. The factor is applied to perturbate also the magnification level. The scale
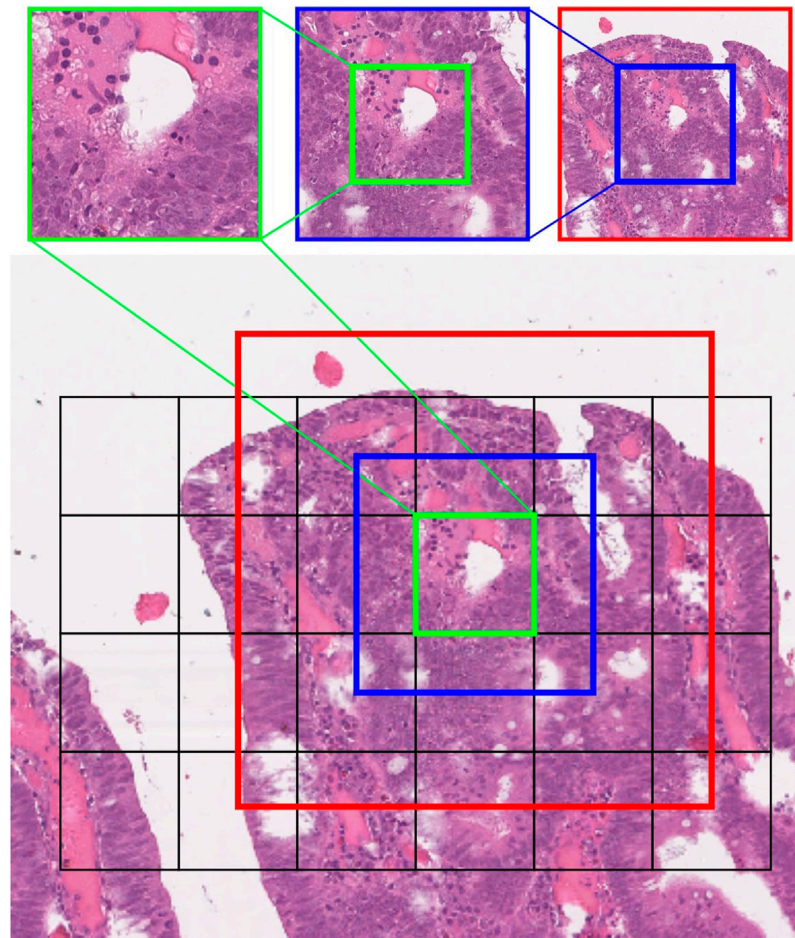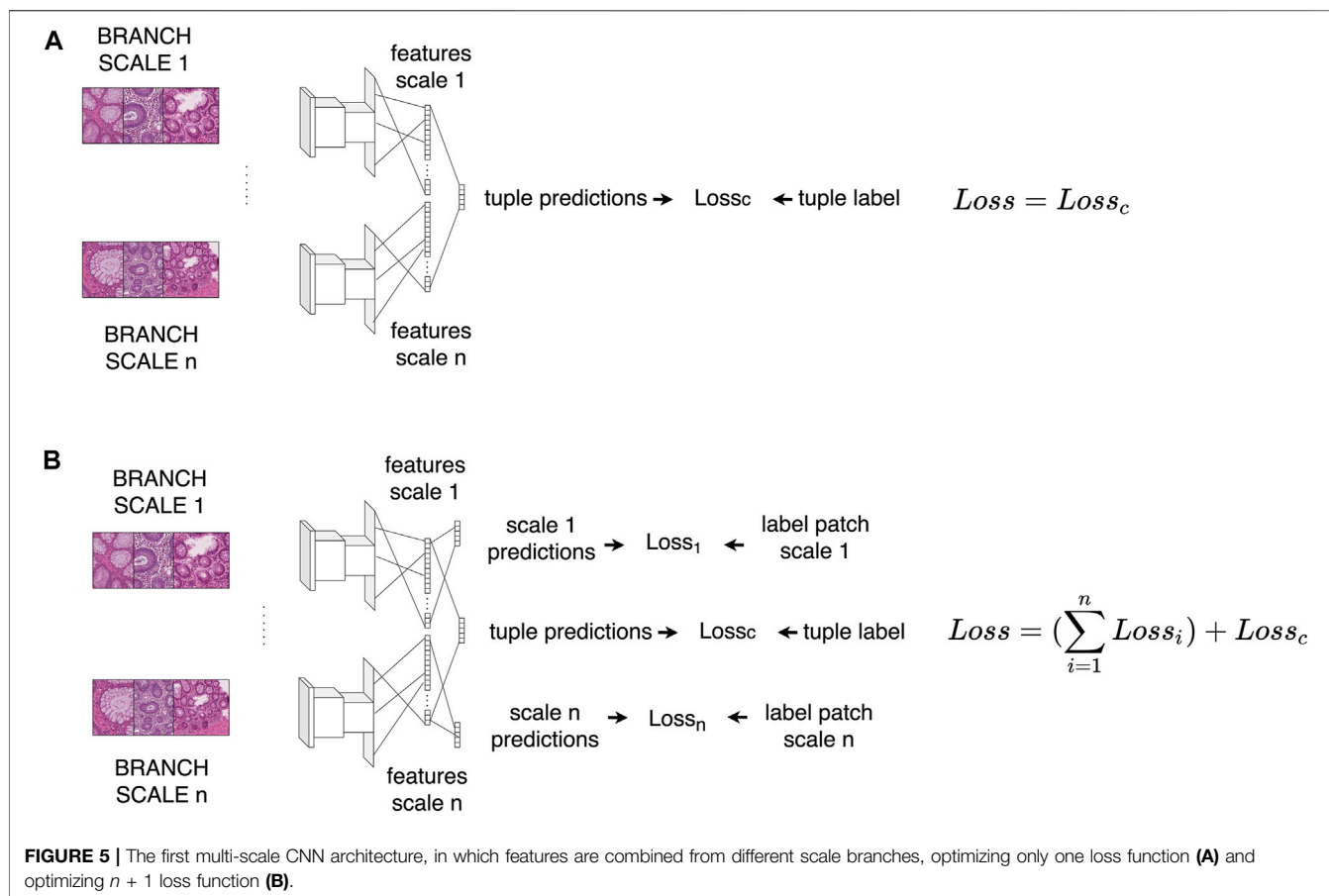
FIGURE 4 | An example of the *multi − center* extraction method. The grid is made according to the highest magnification level selected by the used. The patch is the centroid for patches at lower magnification levels.

detector component includes also a module to import and use the model in the code (regression.py). The component works both as a standalone module (with the required parameters) but it is also possible to load the functions from the python module. The **Supplementary Materials** section includes a more thorough description of the parameters for the scripts.

2.3 Multi-Scale CNN for Classification

The Multi-scale CNN component includes scripts to train a multi-scale CNN for classification, in a fully supervised fashion. Two different multi-scale CNN architectures and two training variants are proposed and compared with a single-scale CNN. The multi-scale CNN architectures are composed of multiple branches (one for each magnification level) trained with patches that come from several magnifications. Each branch is fed with patches from a specific magnification level. The first architecture of multi-scale CNN combines each CNN branch features (the output of the convolutional layers). The scripts developed to train and test the models are Fully_supervised_training_combine_features_multi.py and Fully_supervised_testing_combine_features_multi.py The second architecture of multi-scale CNN combines the classifier

predictions (the output of each CNN's fully-connected layers). The scripts developed to train and test the models are Fully_supervised_training_combine_probs_multi.py and Fully_supervised_testing_combine_probs_multi.py Both architectures are presented in two variants, optimizing respectively one and multiple loss functions. In the first variant (one loss function), the input is a set of tuples of patches from several magnification levels (one patch for each level), generated using the *multi − center* extraction tool (presented in **Section 2.1**). The input tuples are generated with a script (Generate_csv_multicenter.py) that exploits the coordinates of the patches (stored in the metadata) to generate the tuples (stored in a csv file). The tuple label corresponds to the class of the centroid patch (the patch from the highest level within the tuple). Therefore, the model outputs only the class of the tuple. Only one loss function is minimized in this variant, i.e. the categorical cross-entropy between the CNN output and the patch ground truth. **Figure 5** summarizes the CNN architecture. In the second variant (multiple loss functions), the input is a set of tuples of patches from several magnification levels (one patch for each level), previously generated using the *grid* extraction method (presented in **Section 2.1**). The input tuples are generated with a script

FIGURE 5 | The first multi-scale CNN architecture, in which features are combined from different scale branches, optimizing only one loss function **(A)** and optimizing $n + 1$ loss function **(B)**.

(Generate_csv_upper.py) that exploits the coordinates of the patches (stored in the metadata) to generate the tuples (stored in a csv file). The tuple labels correspond to the classes of the patches. The model has $n + 1$ outputs: the class for each of the n magnification levels and the whole tuple class. In this variant, $n + 1$ loss functions are minimized (n representing the number of magnification levels considered). The n loss functions are the categorical cross-entropy between the output for each of the scale branches and the tuple labels. The other loss term is the categorical cross-entropy between the output of the network (after the combination of the features or the predictions of the single branches) and the tuple labels. **Figure 6** summarizes the CNN architecture. The Supplementary Materials section includes a more thorough description of the parameters.

2.4 Multi-Scale CNN for Segmentation

This component includes HookNet (van Rijthoven et al., 2020), a multi-scale CNN for semantic segmentation. HookNet combines information from low-resolution patches (large field of view) and high-resolution patches (small field of view) to semantically segment the image, using multiple branches. The low-resolution patches come from lower magnification levels and include context information, while the high-resolution patches come from higher magnification levels and include more fine-grained information. The network is composed of two branches of encoder-decoder models, the context branch (fed with low-resolution patches) and the target

branch (fed with high-resolution patches). The two branches are fed with concentric multi-field-view multi-resolution (MFMR) patches (284 × 284 pixels in size). Although they have the same architecture, the branches do not share their weights (an encoder-decoder CNN based on U-Net). Hooknet is thoroughly described in a dedicated article (van Rijthoven et al., 2020).
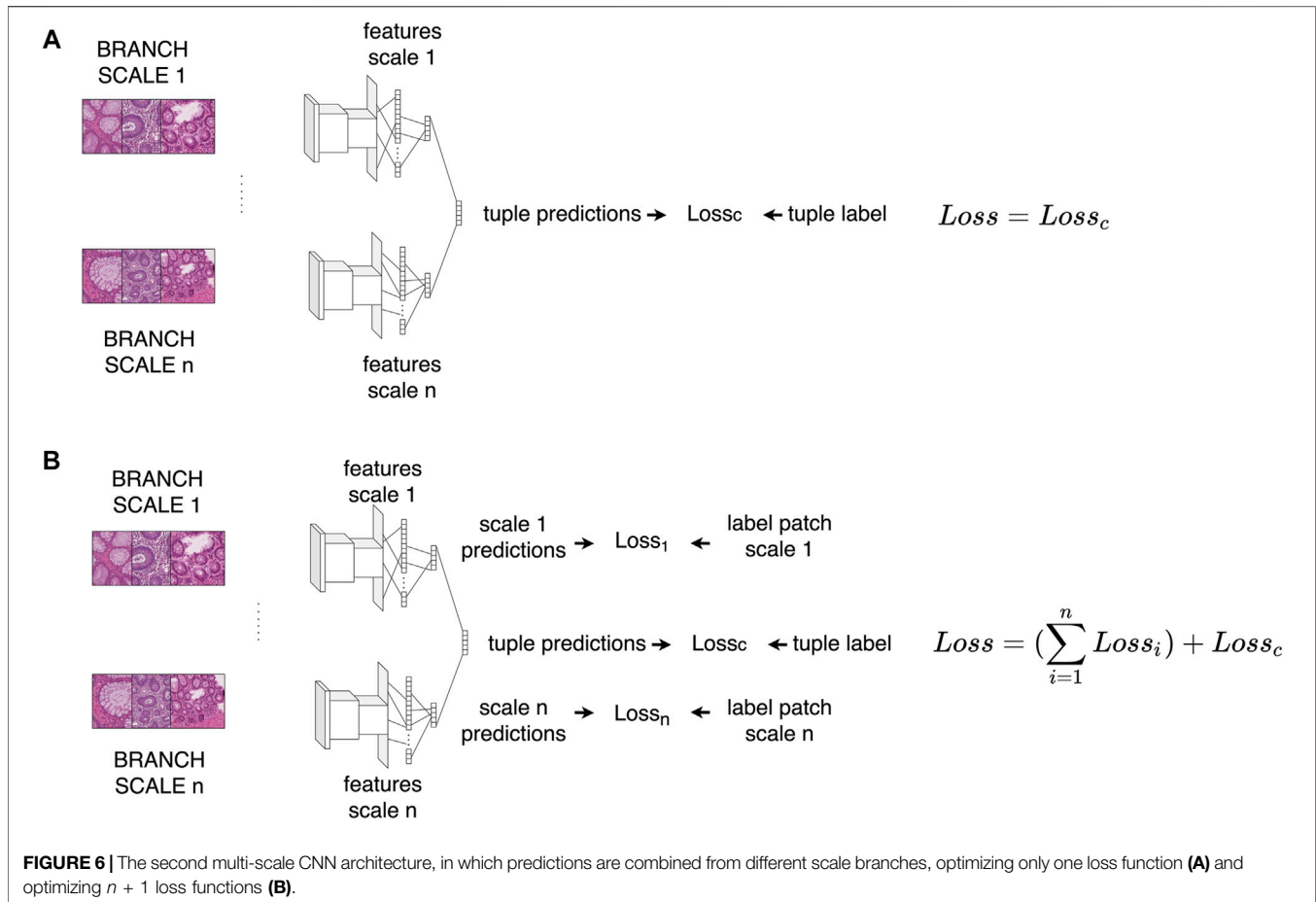
2.5 Datasets

The following datasets are used to develop the Multi_Scale_Tools components:

- Colon dataset, Sub-**section 2.5.1**, used in the Pre-processing component, the Scale detector and the Multi-scale CNN for classification
- Breast dataset, Sub-**section 2.5.2**, used in the Multi-scale CNN for segmentation
- Prostate dataset, Sub-**section 2.5.3**, used in the Scale detector
- Lung dataset, Sub-**section 2.5.4**, used in the Scale detector and the Multi-scale CNN for segmentation

2.5.1 Colon Dataset

The colon dataset is a subset of the ExaMode colon dataset. This subset includes 148 WSIs (provided by the Department of Pathology of Cannizaro Hospital, Catania, Italy), stained with

FIGURE 6 | The second multi-scale CNN architecture, in which predictions are combined from different scale branches, optimizing only one loss function **(A)** and optimizing $n + 1$ loss functions **(B)**.

Hematoxylin and Eosin (H&E). The images are digitized with an Aperio scanner: some of the images are scanned with a maximum spatial resolution of 0.50 µm per pixel (20x), while the others are scanned with a spatial resolution of 0.25 µm per pixel (40x). The images are pixel-wise annotated by a pathologist. The annotations include five classes: cancer, high-grade dysplasia, low-grade dysplasia, hyperplastic polyp and non-informative tissue.

2.5.2 Breast Dataset
The breast dataset (provided by Department of Pathology of Radboud University Medical Center, Nijmegen, Netherlands) is a private dataset including 86 WSIs, stained with H&E. The images are digitized with a 3DHistech scanner, with a spatial resolution of 0.25 µm per pixel (40x). The images are pixel-wise annotated by a pathologist. 6,279 regions are annotated, with the following classes: ductal carcinoma *in-situ* (DCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC) benign epithelium (BE), other, and fat.

2.5.3 Prostate Dataset
The prostate dataset is a subset of the publicly available database offered by The Cancer Genome Atlas (TCGA-PRAD), that includes 20 WSIs, stained with H&E. The images come from several sources and are digitized with different scanners, with a

spatial resolution of 0.25 µm per pixel (40x). The images come without pixel-wise annotations.

2.5.4 Lung Dataset
The Lung dataset is a subset of the public available database offered by The Cancer Genome Atlas Lung Squamous Cell carcinoma dataset (TCGA-LUSC), including 27 WSIs stained with H&E. The images come from several sources and are digitized with different scanners, with a spatial resolution of 0.25 µm per pixel (40x). Initially, the images come without pixel-wise annotation from the repository, but a medical expert from Radboudc Hospital pixel-wise annotated them with four classes: tertiary lymphoid structures (TLS), germinal centers (GC), tumor, and other.

3 EXPERIMENTS AND RESULTS

The Section presents the assessment of the components of the library Multi_Scale_Tools in dedicated subsections as follows:

- Pre-processing component assessment, Sub-**section 3.1**
- Scale detector assessment, Sub-**section 3.2**
- Multi-scale CNN for classification assessment, Sub-**section 3.3**

TABLE 1 | Number of patches extracted with the grid extraction method (above) and with the multi-center method (below), at different magnification level.

			Grid		
Dataset/Magnification	5x	10x	20x	40x	Total
Colon dataset (148 WSIs)	15,514	67,592	279,964	1,127,190	1,490,260
Prostate dataset (20 WSIs)	11,468	46,676	187,254	743,583	988,981
Lung dataset (27 WSIs)	22,124	90,307	365,398	886,298	1,364,127

	Multicenter		
Dataset/Magnification	5x/10x	5x/10x/20x	Total
Colon dataset	135,184	839,892	975,076
Prostate dataset	93,352	561,762	655,114
Lung dataset	180,614	1,096,194	1,276,808

TABLE 2 | Time needed to extract the patches (in seconds), varying the amount of threads, using the *grid* extraction method (above) and using the *multi − center* method (below). The method is evaluated on colon dataset (148 WSIs). The number of patches extracted from each method is reported in **Table 1**.

Magnification/N_threads	10	20	30	40	50
grid extraction					
5x	408 ± 3	317 ± 7	285 ± 5	255s ± 5	238 ± 10
10x	553 ± 5	429 ± 6	389 ± 5	384 ± 8	371 ± 8
20x	1,295 ± 100	969 ± 113	876 ± 69	872 ± 41	869 ± 15
multi − center extraction					
5x/10x	1,662 ± 30	1,180 ± 119	1,071 ± 50	1,039 ± 18	1,022 ± 14
5x/10x/20x	6,604 ± 104	5,745 ± 45	5,283 ± 161	4,814 ± 137	4,549 ± 82

- Multi-scale CNN for segmentation, Sub-**section 3.4**
- Library organization, Sub-**section 3.5**

3.1 Pre-Processing Tool Assessment

The pre-processing component allows to extract a large amount of patches from multiple magnification levels, guaranteeing scalable performance. The patch extractor components (grid and multi-center methods) are tested on WSIs scanned with Aperio (.svs), 3DStech (.mrxs) and Hamamatsu (.ndpi) scanners, on data coming from different tissues (colon, prostate and lung) and datasets. **Table 1** includes the number of patches extracted. The upper part of the table includes the number of patches extracted with the grid extraction method, considering four different magnification levels (5x, 10x, 20x, 40x). The lower part of the Table includes the number of patches extracted with the multi-center extraction method, considering two possible combinations of magnification levels (5x/10x, 5x/10x/20x). In both cases, patches are extracted with a patch size of 224 × 224 pixels without any stride. Methods performance are evaluated in terms of scalability, since the methods are designed to work on multi-core hardwares. **Table 2** includes the time results obtained with the grid method (upper part) and with the multi-center method (lower part). The evaluation is made considering the amount of time needed to extract the patches from the colon dataset, using several threads. The results show that both the methods benefit from multi-core hardwares, reducing the time needed to pre-process data.

3.2 Scale Detector Assessment

The scale detector shows high performance in estimating the magnification level of patches that come from different tissues. The detector is trained with patches from the colon dataset and it is tested with patches from three different tissues. The performance of the models is assessed with the coefficient of determination (R^2), the Mean Square Error (MSE), the Cohen's κ-score (McHugh, 2012) and the balanced accuracy. While the experimental setup and the metrics descriptions are presented in detail the supplementary material, **Table 3** summarizes the results. The higher performance is reached on the colon test partition, but the scale detector shows high performance also on the other tissues. The scale detector makes almost perfect scale estimations in the colon dataset (data come from the same medical source and include the same tissue type), in both the regression and the classification metrics. The scale detector makes reasonably good scale estimations also on the prostate data, in both the regression and the classification metrics, and in lung dataset, where the performance is the lowest though. The fact that the regressor shows exceptionally high performance in colon data and good performance in other tissues means that it has learnt to distinguish the colon morphology represented at different magnification level very well and that the learnt knowledge can generalize well to other tissues too. Even though tissues from different organs share similar structures (glands, stroma, etc.), the morphology of the structures is different in the organs, such as prostate and colon glands. Training the regressor with patches from several organs may allow to close this gap, guaranteeing extremely high performance for different types of tissue.

TABLE 3 | Performance of the scale detector, evaluated on three different tissue dataset. The scale detector is evaluated in: coefficient of determination (R^2), Mean squared error (MSE), balanced accuracy, Cohen's κ-score.

Dataset/Metric	R^2	MSE	Balanced accuracy	κ-score
Colon dataset	0.9997 ± 0.0001	0.0250 ± 0.0155	0.9859 ± 0.0086	0.9991 ± 0.0004
Prostate dataset	0.8013 ± 0.0798	19.34 ± 7.78	0.9094 ± 0.0268	0.8515 ± 0.0589
Lung dataset	0.6682 ± 0.1549	32.13 ± 15.01	0.7973 ± 0.0458	0.8743 ± 0.0571

TABLE 4 | Performance of the multi-scale CNNs architectures, compared with CNNs trained with patches from only one magnification level, evaluated in κ-score and balanced accuracy. Both the multi-scale architectures are presented (combine features and combine predictions from multi-scale branches) and both the training variants (one loss function and $n + 1$ losses). The values marked in bold highlight the method that reaches the highest performance, respect to the metric.

Magnification/metric	κ-score	Balanced-accuracy
Single scale CNNs		
5x	0.7127 ± 0.0988	0.6558 ± 0.0903
10x	0.6818 ± 0.0940	0.6200 ± 0.0780
20x	0.6005 ± 0.1106	0.5744 ± 0.0804
Multi-scale CNNs (combine features)		
5x/10x (One loss)	0.6955 ± 0.1013	0.6529 ± 0.0859
5x/10x ($n + 1$ losses)	**0.7167 ± 0.1060**	**0.6813 ± 0.0942**
5x/10x/20x (One loss)	0.6630 ± 0.1090	0.6508 ± 0.1089
5x/10x/20x ($n + 1$ losses)	0.6871 ± 0.1110	0.6364 ± 0.1046
Multi-scale CNNs (combine probabilities)		
5x/10x (One loss)	0.6901 ± 0.1136	0.6582 ± 0.0973
5x/10x ($n + 1$ losses)	0.7026 ± 0.0988	0.6626 ± 0.0897
5x/10x/20x (One loss)	0.6678 ± 0.0973	0.6239 ± 0.0860
5x/10x/20x ($n + 1$ losses)	0.6784 ± 0.0995	0.6355 ± 0.0835

3.3 Multi-Scale CNN for Classification Assessment

The multi-scale CNNs show higher performance in the fully supervised classification compared to the single-scale CNNs. Several configurations of the multi-scale CNN architectures are evaluated. They involve variations in optimization strategy (one or multiple loss functions), in the magnification levels (combinations of 5x, 10x, 20x) and in how information from the scales is combined (combining the single-scale predictions or the single-scale features). **Table 4** summarizes the results obtained. The CNNs are trained and tested with the colon dataset, that come with pixel-wise annotations made by a pathologist. The performance of the models is assessed with the Cohen's κ-score and the balanced accuracy. More detailed descriptions of the experimental setup and the metrics adopted are presented in the Supplementary material. In the presented experiment, the best multi-scale CNN architecture is the one that combines features from 5/10x magnification levels and is trained optimizing $n + 1$ loss functions. It outperforms the best single-scale CNN (trained with patches acquired at 5x) in terms of balanced accuracy, while the κ-score of the two architectures is comparable. The characteristics of the classes involved can explain the fact that CNNs trained combining patches from 5/10x reach the highest results. These classes show morphologies including several alterations of the gland structure. Glands can be usually identified at low magnification levels, such as 5/10x, while at 20x the cells are visible. For this reason, the CNNs show high performance with patches from magnification 5/10x, while

including patches from 20x decreases the performance. The fact that the discriminant characteristics are identified in a range of scales may explain why the combination of the features shows higher performance than the combination of the predictions.

3.4 Multi-Scale CNN for Segmentation Assessment

The multi-scale CNN (HookNet) shows higher tissue segmentation performance than single-scale CNNs (U-Net). The model is trained and tested with breast and lung datasets, comparing it with models trained with images from a single magnification level. The performance of the models is assessed with the F1 score and the macro F1 score. More detailed descriptions of the experimental setup and the metrics adopted are presented in the **Supplementary Material**. **Table 5** and **Table 6** summarize the results obtained respectively on the breast dataset and on lung dataset. In both the tissues, HookNet shows an higher overall performance, while some of the single scale U-Nets have better performance for some segmentation tasks (such as breast DCIS or lung TLS). This result can be interpreted as a consequence of the characteristics of the task, therefore the user should choose the proper magnification levels to combine, depending of the problem.

3.5 Library Organization

The source code for the library is available on GIT[2], while the HookNet code is available here[3]. The library is available can be deployed as Python package directly from the repository or as Docker container that can be downloaded from[4] (the multiscale folder). Interaction with the library is done through a model class and an Inference class[5]. The model instantion depends on the choice of algorithms. For a more detailed explanation about the hyperparameters and other options please make sure to browse the Readme file[6]. An example can be found here[7]. The Python libraries used to develop Multi_Scale_Tools are reported in **Supplementary Materials**.

---

[2]https://github.com/sara-nl/multi-scale-tools. Retrieved 11th of January, 2021

[3]https://github.com/DIAGNijmegen/pathology-hooknet. Retrieved 19th of June, 2021

[4]https://surfdrive.surf.nl/files/index.php/s/PBBnjwzwMragAGd. Retrieved 11th of January, 2021

[5]https://github.com/computationalpathologygroup/hooknet/blob/fcba7824ed982f663789f0c617a4ed65bedebb85/source/inference.py#L20. Retrieved 11th of January, 2021

[6]https://github.com/sara-nl/multi-scale-tools/blob/master/README.md. Retrieved 11th of January, 2021

[7]https://github.com/DIAGNijmegen/pathology-hooknet/blob/master/hooknet/apply.py. Retrieved 19th of June, 2021

TABLE 5 | Performance of the U-Net (above) and HookNet (below) on the breast dataset. The architectures are compared on the F1 score, for each tissue type (description of the tissue type in the **Supplementary Material**). The overall macro-F1 score is reported. The values marked in bold highlight the method that reaches the highest performance, respect to the task.

Magnification	DCIS	IDC	ILC	Benign	Other	Fat	Overall
Model: U-Net							
20x	0.47	0.55	0.85	0.75	0.95	0.99	0.76
10x	0.67	0.69	0.79	**0.87**	0.98	1.00	0.83
5x	0.79	0.83	0.79	0.84	0.98	1.00	0.87
2.5x	0.83	0.85	0.63	0.73	0.96	1.00	0.83
1.25x	**0.86**	0.81	0.20	0.66	0.96	1.00	0.75
Model: HookNet							
20x (target)-5x (context)	0.62	0.75	0.82	0.82	0.98	1.00	0.83
20x (target)-1.25x (context)	0.84	**0.89**	**0.91**	0.84	0.98	1.00	**0.91**

TABLE 6 | Performance of the U-Net (above) and HookNet (below) on the lung dataset. The architectures are compared on the F1 score, for each tissue type (description of the tissue type in the **Supplementary Material**). The overall macro-F1 score is reported. The values marked in bold highlight the method that reaches the highest performance, respect to the task.

Magnification	TLS	GC	Tumor	Other	Overal
Model: U-Net					
20x	0.81	0.28	**0.75**	0.87	0.70
10x	**0.86**	0.44	0.71	0.86	0.72
5x	0.84	**0.49**	0.67	0.85	0.71
2.5x	0.80	0.37	0.56	0.80	0.63
1.25x	0.78	0.35	0.39	0.77	0.57
Model: HookNet					
20x (target)-5x (context)	0.84	0.48	0.72	0.87	**0.73**

4 DISCUSSION AND CONCLUSION

Multi_Scale_Tools library aims at facilitating the exploitation of multi-scale structure in WSIs with code that is easy to use and easy to be improved with additional functions. The library currently includes four components. The components are a pre-processing tool to extract multi-scale patches, a scale detector, two multi-scale CNNs for classification and a multi-scale CNN for segmentation. The pre-processing component includes two methods to extract patches from several magnification levels. The methods are designed to be scalable on multi-core hardware. The scale detector component includes a CNN allowing to regress the magnification level of a patch. The CNN obtains high performance in patches that come from the colon (the tissue used to train it) and it reaches good performance on other tissues such as prostate and lung too. Two multi-scale CNN architectures are developed for fully-supervised classification. The first one combines features from multi-scale branches, while the second one combines predictions from multi-scale branches. The first architecture obtains better performance and outperforms the model trained with patches from only one magnification level. The HookNet architecture for multi-scale segmentation is also included into the library, fostering its usage and making the library more complete. The tests show that HookNet outperforms single scale U-Net in the considered tasks. The presented library allows to exploit the multi-scale structure of WSIs efficiently. In any case, the user remains a fundamental part of the system for several components, such as identifying the scale that can be more relevant for a specific problem. The comparison between the single-scale CNNs and the multi-scale CNN is an example of this. The CNN is trained to classify between cancer, dysplasia (both high-grade and low-grade), hyperplastic polyp and non-informative tissue. In the classification task, the highest performance is reached using patches of magnification 5x and 10x, while patches from 20x lead to lower classification performance. This can likely be related to the fact that the main feature related to the considered classes is the structure of the glands, therefore high magnifications (e.g. 20x) limitedly introduce helpful information into the models. The importance of the user to select the proper magnification levels is highlighted even more in the segmentation results. Considering low magnifications, the models show good performance in ductal carcinoma *in-situ* and invasive ductal carcinoma segmentation since these tasks need context about the duct structures in the breast use case. Considering higher magnifications, the models perform well in invasive lobular carcinoma and benign tissue segmentation, where the details are more important. The methods identified to pair images from several magnification levels can pave the way to multi-modal combination of images too. The combination may increase the information included in the single modality, increasing the performance of the CNNs. Some possible applications can be the combination of WSIs stained with different reagents, such H&E and immunohistochemical (IHC) stainings, the application in Raman spectroscopy data, combining information about tissue morphologies and architectures with protein biomarkers, and the combination of patches from different focal planes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

NM: design of the work, software, analysis, original draft SO: design of the work, revised the work DP: software, revised the

work MR: software, analysis JL: revised the work, approval for publication FC: revised the work, approval for publication HM: revised the work, approval for publication MA: revised the work, approval for publication.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.684521/full#supplementary-material

REFERENCES

Aeffner, F., Wilson, K., Martin, N. T., Black, J. C., Hendriks, C. L. L., Bolon, B., et al. (2017). The Gold Standard Paradox in Digital Image Analysis: Manual versus Automated Scoring as Ground Truth. *Arch. Pathol. Lab. Med.* 141, 1267–1275. doi:10.5858/arpa.2016-0386-ra

Arvaniti, E., and Claassen, M. (2018). *Coupling Weak and strong Supervision for Classification of Prostate Cancer Histopathology Images.* Medical Imaging meets NIPS Workshop, NIPS 2018. arXiv preprint arXiv:1811.07013.

Bozkurt, A., Kose, K., Alessi-Fox, C., Gill, M., Dy, J., Brooks, D., and Rajadhyaksha, M. (2018). A Multiresolution Convolutional Neural Network with Partial Label Training for Annotating Reflectance Confocal Microscopy Images of Skin. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018 (Springer), 292–299. doi:10.1007/978-3-030-00934-2_33

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., et al. (2019). Clinical-grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat. Med.* 25, 1301–1309. doi:10.1038/s41591-019-0508-1

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and Mutation Prediction from Non-small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* 24, 1559–1567. doi:10.1038/s41591-018-0177-5

del Toro, O. J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., et al. (2017). "Convolutional Neural Networks for an Automatic Classification of Prostate Tissue Slides with High-Grade gleason Score," in *Medical Imaging 2017: Digital Pathology* (Bellingham, WA: International Society for Optics and Photonics), 10140, 101400O. doi:10.1117/12.2255710

Gecer, B., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., and Elmore, J. G. (2018). Detection and Classification of Cancer in Whole Slide Breast Histopathology Images Using Deep Convolutional Networks. *Pattern recognition* 84, 345–356. doi:10.1016/j.patcog.2018.07.022

Han, D., Kim, J., and Kim, J. (2017). Deep Pyramidal Residual Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, July 21-26 2017 (IEEE) 5927–5935. doi:10.1109/cvpr.2017.668

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., et al. (2020). Multi-scale Domain-Adversarial Multiple-Instance Cnn for Cancer Subtype Classification with Unannotated Histopathological Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14–19 June 2020 (IEEE) 3852–3861. doi:10.1109/cvpr42600.2020.00391

Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., and Tsuneki, M. (2020). Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci. Rep.* 10, 1504–1511. doi:10.1038/s41598-020-58467-9

Jain, M. S., and Massoud, T. F. (2020). Predicting Tumour Mutational burden from Histopathological Images Using Multiscale Deep Learning. *Nat. Mach Intell.* 2, 356–362. doi:10.1038/s42256-020-0190-5

Janowczyk, A., and Madabhushi, A. (2016). Deep Learning for Digital Pathology Image Analysis: A Comprehensive Tutorial with Selected Use Cases. *J. Pathol. Inform.* 7, 29. doi:10.4103/2153-3539.186902

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., and Madabhushi, A. (2019). Histoqc: an Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin. Cancer Inform.* 3, 1–7. doi:10.1200/cci.18.00157

Jimenez-del-Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., et al. (2017). "Analysis of Histopathology Images," in *Biomedical Texture Analysis* (Cambridge, MA: Elsevier), 281–314. doi:10.1016/b978-0-12-812133-7.00010-7

Komura, D., and Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* 16, 34–42. doi:10.1016/j.csbj.2018.01.001

Lai, Z., and Deng, H. (2017). Multiscale High-Level Feature Fusion for Histopathological Image Classification. *Comput. Math. Methods Med.* 2017, 7521846. doi:10.1155/2017/7521846

Lenc, K., and Vedaldi, A. (2015). Understanding Image Representations by Measuring Their Equivariance and Equivalence. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA, 7–12 June 2015 (IEEE) 991–999. doi:10.1109/cvpr.2015.7298701

Li, J., Sarma, K. V., Chung Ho, K., Gertych, A., Knudsen, B. S., and Arnold, C. W. (2017). A Multi-Scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies, *AMIA Annu. Symp. Proc.*. In AMIA Annual Symposium Proceedings, Washington, DC, 4–8 November 2017 (American Medical Informatics Association), vol. 2017, 1140, 1148.

Li, S., Liu, Y., Sui, X., Chen, C., Tjio, G., Ting, D. S. W., and Goh, R. S. M. (2019). Multi-instance Multi-Scale Cnn for Medical Image Classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019 (Springer), 531–539. doi:10.1007/978-3-030-32251-9_58

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nat. Biomed. Eng.*, 1–16. doi:10.1038/s41551-020-00682-w

Marcos, D., Kellenberger, B., Lobry, S., and Tuia, D. (2018). *Scale Equivariance in Cnns with Vector fields*. ICML/FAIM 2018 workshop on Towards learning with limited labels: Equivariance, Invariance, and Beyond (oral presentation). arXiv preprint arXiv:1807.11783.

McHugh, M. L. (2012). Interrater Reliability: the Kappa Statistic. *Biochem. Med.* 22, 276–282. doi:10.11613/bm.2012.031

Molin, J., Bodén, A., Treanor, D., Fjeld, M., and Lundström, C. (2016). *Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization*. arXiv preprint arXiv:1610.04141.

Naylor, P., Laé, M., Reyal, F., and Walter, T. (2018). Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map. *IEEE Trans. Med. Imaging* 38, 448–459. doi:10.1109/TMI.2018.2865709

Naylor, P., Laé, M., Reyal, F., and Walter, T. (2017). Nuclei Segmentation in Histopathology Images Using Deep Neural Networks. In 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017 (IEEE), 933–936. doi:10.1109/isbi.2017.7950669

Otálora, S., Atzori, M., Andrearczyk, V., and Müller, H. (2018a). "Image Magnification Regression Using Densenet for Exploiting Histopathology Open Access Content," in *Computational Pathology and Ophthalmic Medical Image Analysis* (New York, USA: Springer), 148–155. doi:10.1007/978-3-030-00949-6_18

Otálora, S., Perdomo, O., Atzori, M., Andersson, M., Jacobsson, L., Hedlund, M., et al. (2018b). Determining the Scale of Image Patches Using a Deep Learning Approach. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, 4–7 Aprile 2018. (IEEE), 843–846. doi:10.1109/isbi.2018.8363703

Paramanandam, M., O'Byrne, M., Ghosh, B., Mammen, J. J., Manipadam, M. T., Thamburaj, R., et al. (2016). Automated Segmentation of Nuclei in Breast Cancer Histopathology Images. *PloS one* 11, e0162053. doi:10.1371/journal.pone.0162053

Ren, J., Hacihaliloglu, I., Singer, E. A., Foran, D. J., and Qi, X. (2018). Adversarial Domain Adaptation for Classification of Prostate Histopathology Whole-Slide Images. International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018 (Springer), 201–209. doi:10.1007/978-3-030-00934-2_23

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical image computing and computer-assisted intervention, Munich, Germany, 5-9 October 2015 (Springer), 234–241. doi:10.1007/978-3-319-24574-4_28

Rorke, L. B. (1997). Pathologic Diagnosis as the Gold Standard. *Cancer* 79, 665–667. doi:10.1002/(sici)1097-0142(19970215)79:4<665::aid-cncr1>3.0.co;2-d

Roy, K., Banik, D., Bhattacharjee, D., and Nasipuri, M. (2019). Patch-based System for Classification of Breast Histology Images Using Deep Learning. *Comput. Med. Imaging Graphics* 71, 90–103. doi:10.1016/j.compmedimag.2018.11.003

Salvi, M., and Molinari, F. (2018). Multi-tissue and Multi-Scale Approach for Nuclei Segmentation in H&E Stained Images. *Biomed. Eng. Online* 17, 89. doi:10.1186/s12938-018-0518-0

Schmitz, R., Madesta, F., Nielsen, M., Krause, J., Werner, R., and Rösch, T. (2019). *Multi-scale Fully Convolutional Neural Networks for Histopathology Image Segmentation: From Nuclear Aberrations to the Global Tissue Architecture*. *Medical Image Analysis* 70, 101996.

Sellaro, T. L., Filkins, R., Hoffman, C., Fine, J. L., Ho, J., Parwani, A. V., et al. (2013). Relationship between Magnification and Resolution in Digital Pathology Systems. *J. Pathol. Inform.* 4, 21. doi:10.4103/2153-3539.116866

Tensmeyer, C., and Martinez, T. (2016). *Improving Invariance and Equivariance Properties of Convolutional Neural Networks* ICLR 2017 conference.

Ullah, I. (2017). *A Pyramidal Approach for Designing Deep Neural Network Architectures* PhD thesis. Available at: https://air.unimi.it/handle/2434/466758#.YQEi7FMzYWo.

van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep Learning in Histopathology: the Path to the Clinic. *Nat. Med.* 27, 775–784. doi:10.1038/s41591-021-01343-4

van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., and Ciompi, F. (2020). Hooknet: Multi-Resolution Convolutional Neural Networks for Semantic Segmentation in Histopathology Whole-Slide Images. *Med. Image Anal.* 68, 101890. doi:10.1016/j.media.2020.101890

Wang, S., Yang, D. M., Rong, R., Zhan, X., and Xiao, G. (2019). Pathology Image Analysis Using Segmentation Deep Learning Algorithms. *Am. J. Pathol.* 189, 1686–1698. doi:10.1016/j.ajpath.2019.05.007

Wu, Q., Merchant, F., and Castleman, K. (2010). *Microscope Image Processing*. New York, USA: Elsevier.

Yang, Z., Ran, L., Zhang, S., Xia, Y., and Zhang, Y. (2019). Ems-net: Ensemble of Multiscale Convolutional Neural Networks for Classification of Breast Cancer Histology Images. *Neurocomputing* 366, 46–53. doi:10.1016/j.neucom.2019.07.080

Zaveri, M., Hemati, S., Shah, S., Damskinos, S., and Tizhoosh, H. (2020). Kimia-5mag–a Dataset for Learning the Magnification in Histopathology Images. In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), 9–11 November 2020 (IEEE), 363–367.

Zeng, N., Li, H., and Peng, Y. (2021a). A New Deep Belief Network-Based Multi-Task Learning for Diagnosis of Alzheimer's Disease. *Neural Comput. Appl.*, 1–12. doi:10.1007/s00521-021-06149-6

Zeng, N., Li, H., Wang, Z., Liu, W., Liu, S., Alsaadi, F. E., et al. (2021b). Deep-reinforcement-learning-based Images Segmentation for Quantitative Analysis of Gold Immunochromatographic Strip. *Neurocomputing* 425, 173–180. doi:10.1016/j.neucom.2020.04.001

Zeng, N., Wang, Z., Zhang, H., Kim, K.-E., Li, Y., and Liu, X. (2019). An Improved Particle Filter with a Novel Hybrid Proposal Distribution for Quantitative Analysis of Gold Immunochromatographic Strips. *IEEE Trans. Nanotechnology* 18, 819–829. doi:10.1109/tnano.2019.2932271

Zhang, Q., Heldermon, C. D., and Toler-Franklin, C. (2020). Multiscale Detection of Cancerous Tissue in High Resolution Slide Scans. In International Symposium on Visual Computing, San Diego, USA, 5–7 October 2020 (Springer), 139–153. doi:10.1007/978-3-030-64559-5_11

Zhu, W., Qiu, Q., Calderbank, R., Sapiro, G., and Cheng, X. (2019). *Scale-equivariant Neural Networks with Decomposed Convolutional Filters*. arXiv preprint arXiv:1909.11193.