Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: an experiment on prostate histopathology image classification

Niccolò Marini*[a,b], Sebastian Otálora*[a,b], Henning Müller[a,c], Manfredo Atzori[a]

[a] *Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Technopôle 3, 3960 Sierre, Switzerland*
[b] *Centre Universitaire d'Informatique, University of Geneva, 1227 Carouge, Switzerland*
[c] *Medical faculty, University of Geneva, 1211 Geneva, Switzerland*

ARTICLE INFO

Article history:

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Computational Pathology, Deep Learning, Semi-supervision

ABSTRACT

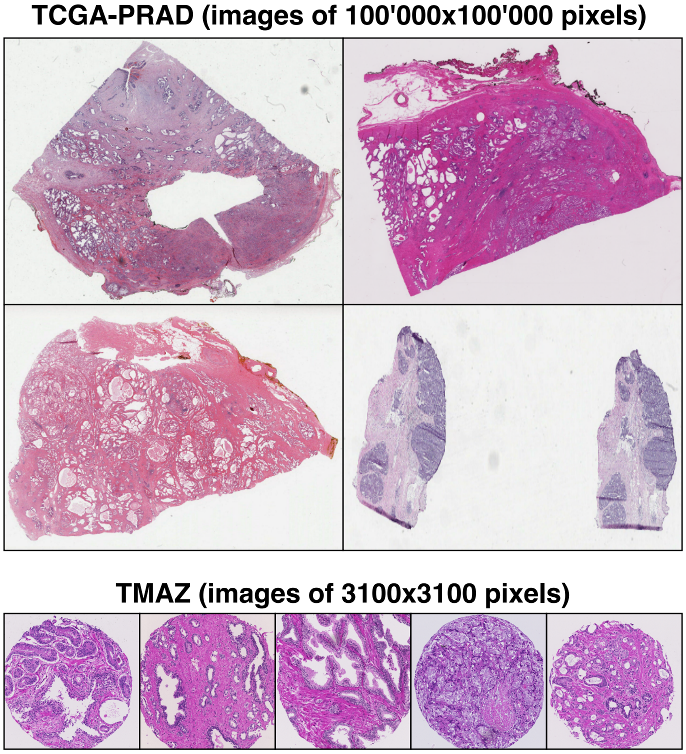
Convolutional neural networks (CNNs) are state-of-the-art computer vision techniques for various tasks, particularly for image classification. However, there are domains where the training of classification models that generalize on several datasets is still an open challenge because of the highly heterogeneous data and the lack of large datasets with local annotations of the regions of interest, such as histopathology image analysis. Histopathology concerns the microscopic analysis of tissue specimens processed in glass slides to identify diseases such as cancer.

Digital pathology concerns the acquisition, management and automatic analysis of digitized histopathology images that are large, having in the order of $100'000^2$ pixels per image. Digital histopathology images are highly heterogeneous due to the variability of the image acquisition procedures. Creating locally labeled regions (required for the training) is time-consuming and often expensive in the medical field, as physicians usually have to annotate the data. Despite the advances in deep learning, leveraging strongly and weakly annotated datasets to train classification models is still an unsolved problem, mainly when data are very heterogeneous. Large amounts of data are needed to create models that generalize well. This paper presents a novel approach to train CNNs that generalize to heterogeneous datasets originating from various sources and without local annotations. The data analysis pipeline targets Gleason grading on prostate images and includes two models in sequence, following a teacher/student training paradigm. The teacher model (a high-capacity neural network) automatically annotates a set of pseudo-labeled patches used to train the student model (a smaller network). The two models are trained with two different teacher/student approaches: semi-supervised learning and semi-weekly supervised learning. For each of the two approaches, three student training variants are presented. The baseline is provided by training the student model only with the strongly annotated data. Classification performance is evaluated on the student model at the patch level (using the local annotations of the Tissue Micro-Arrays Zurich dataset) and at the global level (using the TCGA-PRAD, The Cancer Genome Atlas-PRostate ADenocarcinoma, whole slide image Gleason score). The teacher/student paradigm allows the models to better generalize on both datasets, despite the inter-dataset heterogeneity and the small number of local annotations used. The classification performance is improved both at the patch-level (up to $\kappa = 0.6127 \pm 0.0133$ from $\kappa = 0.5667 \pm 0.0285$), at the TMA core-level (Gleason score) (up to $\kappa = 0.7645 \pm 0.0231$ from $\kappa = 0.7186 \pm 0.0306$) and at the WSI-level (Gleason score) (up to $\kappa = 0.4529 \pm 0.0512$ from $\kappa = 0.2293 \pm 0.1350$). The results show that with the teacher/student paradigm, it is possible to train models that generalize on datasets from entirely different sources, despite the inter-dataset heterogeneity and the lack of large datasets with local annotations.

1. Introduction

One of the current challenges in medical imaging and particularly in computational pathology is the management of the highly-heterogeneous data available to train robust deep learning models Cheplygina et al. (2019) and to overcome the lack of locally-annotated (strongly-annotated) datasets.

This lack of data persists despite the increasing amount of publicly available datasets, such as TCGA (The Cancer Genome Atlas) (Tomczak et al., 2015) or TCIA (The Cancer Imaging Archive) (Prior et al., 2013). Deep Convolutional Neural Network (CNN) models are currently the backbone of the state-of-the-art methods to analyze Whole Slide Images (WSIs) (Jimenez-del Toro et al., 2017; Litjens et al., 2017). Convolutional neural networks often need large locally-annotated datasets to train models that generalize to unseen new data (Komura and Ishikawa, 2018). Local annotations (strong labels) are pixel-wise annotations, made by a pathologist. This type of annotation is expensive and time-consuming to be produced. On the other hand, global annotations (weak labels) are less expensive to be collected. They usually involve the image diagnosis (or staging/grading), that refers to the whole image, without any information about the region of interest that leads to the diagnosis. Creating large sets of annotated images is a challenge for researchers in computational pathology since performing very domain-specific annotations is essential. For instance, Campanella et al. (2019) used 24’859 whole slide images (not publicly available) with global annotations and Arvaniti and Claassen (2018) used 886 tissue micro-arrays (publicly available) with local annotations. However, large locally annotated datasets are scarce and only a few are publicly available, such as the Camelyon dataset (Litjens et al., 2018). The scarcity is a consequence of the annotation process: it is expensive, time-consuming and usually requires experts (sometimes a consensus), so highly trained pathologists. The Prostate cANcer graDe Assessment (PANDA) Challenge dataset[1] is to the best of our knowledge the largest available prostate image dataset with local annotations. The dataset was released as part of a prostate cancer grading challenge in digitized images, proposed during the MICCAI 2020 conference. The training set includes 11’000 digitized whole slide images, eight times larger than the dataset proposed in the CAMELYON challenge. The images originate from two medical centers (Radboud and Karolinska) and are annotated at the pixel-level by uro-pathologists. As of early 2021, this dataset can not be used for publications, as it is under embargo until the paper describing the data is published.


TCGA-PRAD (images of 100'000x100'000 pixels)

TMAZ (images of 3100x3100 pixels)

Fig. 1. Example slides from The Cancer Genome Atlas-PRostate ADenocarcinoma dataset (TCGA-PRAD, above) and the Tissue Micro Array Zurich cores, (TMAZ, below). The slides from the TCGA-PRAD dataset are in the order of $100'000^2$ pixels, while the slides from TMAZ dataset are images of $3'100^2$ pixels

One of the most prominent challenges in digital pathology is handling data heterogeneity, especially in data from various sources. The heterogeneity of digital pathology images is a consequence of the sample acquisition procedures. The acquisition procedure concerns the devices and the staining (Schulte, 1991) applied to the tissue before creating the actual image. The tissue samples are often stained with hematoxylin and eosin (H&E), considered the gold standard for staining in many situations (Fischer et al., 2008; Titford, 2005). However, the staining procedure is not fully standardized. The lack of standardization can easily lead to inter-dataset heterogeneity. This heterogeneity makes it difficult for the models to generalize on external datasets. Models trained on a dataset often show a decrease in the performance when tested on data originating from a different source (Ström et al., 2019; Tellez et al., 2019; Otálora et al., 2019). An example of data heterogeneity is shown in Figure 1 with samples from two publicly available datasets. While there is an increasing amount of available digital pathology data, it is still challenging to find reliable annotations accompanying these data. Valuable public datasets do exist: for instance, the Camelyon dataset for breast cancer (Litjens et al., 2018) and The Cancer Genome Atlas TCGA[2] includes several datasets containing up to 500 whole slide images for individual organs, such as prostate[3]. The challenges in digital pathology are well represented in the TCGA datasets: images are usually without

*Both authors contributed equally to this work. Corresponding author: Niccolò Marini. Tel.: +41-027-606-9033

e-mail: niccolo.marini@hevs.ch (Niccolò Marini*), juan.otaloramontenegro@hevs.ch (Sebastian Otálora*), henning.mueller@hevs.ch (Henning Müller), manfredo.atzori@hevs.ch (Manfredo Atzori)

[1] https://www.kaggle.com/c/prostate-cancer-grade-assessment/overview. Retrieved 17th of January, 2021

[2] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[3] https://portal.gdc.cancer.gov/projects/TCGA-PRAD. Retrieved 1st of January, 2020

local annotations and are highly heterogeneous.

Despite the lack of annotations and the highly heterogeneous data, recently proposed methods have shown partial success in medical image analysis when trained with small sets of annotations (Bulten et al., 2020; Madabhushi et al., 2020; Shaw et al., 2020; Cheplygina et al., 2019; Campanella et al., 2019; Otálora et al., 2017; Tajbakhsh et al., 2016; Litjens et al., 2017; Arvaniti et al., 2018), using techniques such as semi-supervised learning, even though the generalization of models on heterogeneous datasets is still an open challenge. Semi-supervised learning algorithms have recently shown their potential, leveraging large weakly-annotated and unlabeled datasets to create new annotated data, given the small amount of locally-annotated data. Semi-supervised learning algorithms reach sometimes better performance than state-of-the-art supervised models on the large ImageNet dataset (Yalniz et al., 2019).

This paper is a novel approach in computational pathology, focusing on prostate cancer, as it trains classification models that generalize on datasets collected from several sources despite the highly heterogeneous data and the lack of locally annotated data. This paper describes two semi-supervised teacher/student learning approaches applied to the digital pathology task of prostate cancer classification, partly presented in Otálora et al. (2020a); Marini et al. (2020). Two heterogeneous prostate cancer datasets (the Tissue Micro-Arrays Zurich, TMAZ and The Cancer Genome Archive-PRostate ADenocarcinoma, TCGA-PRAD) are used. The TCGA-PRAD dataset (without local annotations) includes an equivalent number of pixels to the 588 images of the TMAZ dataset (locally-annotated).

Prostate cancer is one of the most common cancers worldwide[4] and the gold standard adopted for its diagnosis is the Gleason score Pierorazio et al. (2013). Prostate cancer is the fourth most common cancer and in 2018 there were 1.28 million new diagnoses worldwide[5]. Prostate cancer is a highly heterogeneous disease displaying several tumour features in glands and having high inter-rater variability among pathologists Berg et al. (2011); Arvaniti et al. (2018); Tolkach et al. (2020); Nagpal et al. (2019). The Gleason score is used in clinical practice as a standard protocol when assessing prostate adenocarcinoma. It is required for deciding on the treatment and for predicting the patient's prognosis. The Gleason scoring system describes the abnormality of cancer cells and the deformation of glands within a prostate needle biopsy or a biopsy after radical prostatectomy. Pathologists evaluate it based on the tissue structures observed on microscopic or digital images of biopsies. The Gleason score is based on two steps: recognizing the relevant Gleason patterns and evaluating the Gleason score, that is computed from the two most prominent patterns. The Gleason patterns aims to quantify the tumour aggressiveness and disease prognosis to plan the treatment. The tissue structures, the cell abnormality and the gland deformation allow distinguishing between Gleason patterns. The Gleason patterns vary from 1 to 5 (Chen and Zhou, 2016). Lower patterns are related to a more favourable condition, where the cells and the glands are better differentiated. Higher patterns are related to poor conditions, where the cells and the glands are poorly-differentiated. Gleason patterns 1 and 2 are rarely identified within a core biopsy, as in these cases biopsies are rarely taken. Gleason pattern 1 presents well-differentiated small glands and small regions of stroma between the glands. Gleason pattern 2 presents larger glands and more stroma that is present between the glands. Gleason pattern 3 presents distinct glands of variable sizes and cells that start to infiltrate the surrounding tissue. Gleason pattern 4 presents poorly-differentiated glands and cells that invade the surrounding tissue. Gleason pattern 5 presents no recognizable glands and layers of cells within the surrounding tissue. The Gleason score is calculated by summing up the two most prominent Gleason patterns (GP) observed in the tissue slide. The Union for International Cancer Control[6] and the World Health Organization/International Society of Urological Pathology[7] describe the guidelines for the Gleason pattern evaluation (Montironi et al., 2005). In malignant prostate cancer, the Gleason score generally varies from 6 to 10 (Epstein et al., 2016).

The classification procedure described in this article is based on the semi–supervised classification approach presented in Yalniz et al. (2019). The approach is based on two models in sequence: the *teacher* model and the *student* model. The teacher is a large model that generates pseudo-labeled examples from unlabeled regions in the TCGA-PRAD dataset. It uses a high-capacity ResNext model pre-trained with a dataset of one billion natural images retrieved from Instagram and fine-tuned with both weakly annotated images (from TCGA) and strongly-annotated tissue micro-array images. The student is a smaller model that is subsequently trained with the pseudo-labeled examples provided by the teacher and then fine-tuned with the available local annotations from a separate dataset. The strategy is compared against a fully-supervised approach as well as other variants of semi-supervised training.

1.1. Related work

Data heterogeneity and lack of large annotated datasets are still open challenges in computational pathology. In addition to fully supervised approaches, several methods can help to tackle these challenges, such as active learning, weakly-supervised learning, transfer learning and semi-supervised learning. Table 1 summarizes the state-of-the-art methods used for prostate cancer classification tasks.

Fully-supervised learning: Fully supervised learning includes methods to train machine learning models using datasets where each of the samples is locally labeled. Classification tasks require pixel-level labels. The lack of large datasets with local annotations is common in the medical domain and not limited to computational pathology. A few large, locally-annotated datasets exist for computational pathology and these

[4] https://www.cancer.net/cancer-types/prostate-cancer/statistics. Retrieved 24th of July, 2020

[5] https://www.who.int/en/news-room/fact-sheets/detail/cancer. Retrieved 16th of March, 2020

[6] https://www.uicc.org/topics/prostate. Retrieved 13th of July, 2020

[7] https://isupweb.org/isup/. Retrieved 13th of July, 2020

Table 1. State of the art approaches regarding deep learning models for Gleason grading and Gleason scoring. For each of the reference articles, we present the task evaluated, the training, validation and testing partition used to evaluate the models, the results reached and the annotations used in the training are reported. The tasks are: tumour detection (benign vs. malignant tissue), Gleason grading (Benign, GP 3, GP4, GP5), Gleason scoring (GS6, GS7, GS8, GS9, GS10), ISUP Gleason scoring (GS6, GS7=3+4, GS7=4+3, GS8, GS9-10), low vs. high scoring (GS6,GS7 vs GS8,GS9,GS10).

Reference	Task	Train dataset	Test dataset	Results	Annotations
Nagpal et al. (2019)	Tumour detection	912 WSIs	752 WSIs	ACC = 0.94	Strong
Tolkach et al. (2020)	Tumour detection	389 WSIs	279 WSIs	ACC = 0.97	Strong, Weak
Campanella et al. (2019)	Tumour detection	24’859 WSIs	1’784 WSIs	AUC = 0.986	Weak
Ström et al. (2019)	Tumour detection	6’682 WSIs	1’631 WSIs	AUC = 0.997	Strong
Arvaniti et al. (2018)	Gleason grading	641 TMAs	245 TMAs	$\kappa = 0.55$	Strong
Ström et al. (2019)	Gleason grading	6’682 WSIs	1’631 WSIs	$\kappa = 0.67$	Strong
Otálora et al. (2021)	Gleason grading	641 TMAs, 255 WSIs	245 TMAs	$\kappa = 0.55$	Strong, Weak
This work	Gleason grading	641 TMAs, 255 WSIs	245 TMAs	$\kappa = 0.61$	Strong, Weak
Arvaniti and Claassen (2018)	Low vs High Scoring	641 TMAs, 447 WSIs	245 TMAs	AUC = 0.882	Strong, Weak
del Toro et al. (2017)	Low vs High Scoring	235 WSIs	46 WSIs	ACC = 0.78	Weak
Nagpal et al. (2019)	Low vs High Scoring	580	498 WSIs	ACC = 0.97	Strong
Arvaniti et al. (2018)	Gleason scoring	641 TMAs	245 TMAs	$\kappa = 0.75$	Strong
Bulten et al. (2020)	Gleason scoring	1’143 WSIs	245 TMAs	$\kappa = 0.71$	Strong, Weak
Nagpal et al. (2019)	Gleason scoring	580	498 WSIs	ACC = 0.71	Strong
Otálora et al. (2021)	Gleason scoring	641 TMAs, 255 WSIs	245 TMAs	$\kappa = 0.69$	Strong, Weak
This work	Gleason scoring	641 TMAs, 255 WSIs	245 TMAs	$\kappa = 0.76$	Strong, Weak
Otálora et al. (2020b)	ISUP Gleason scoring	290 WSIs	51 WSI	$\kappa = 0.44$	Weak
Bulten et al. (2020)	ISUP Gleason scoring	1’143 WSIs	100 WSIs	$\kappa = 0.91$	Strong, Weak
This work	ISUP Gleason scoring	641 TMAs, 255 WSIs	46 WSIs	$\kappa = 0.45$	Strong, Weak

were used in fully supervised approaches (Arvaniti et al., 2018; Ström et al., 2019; Nagpal et al., 2019). The work of Arvaniti et al. (2018) presents a CNN trained using the publicly available Tissue-Micro Array Zurich dataset that was pixel-wise annotated by two pathologists. The CNN is trained to predict Gleason patterns, reaching κ=0.55 as best result (κ=0.67 is the agreement reached by the pathologists, where κ=0 would mean to only have by chance agreement). The predictions reported within a core are aggregated summing them up, in order to obtain the corresponding Gleason scores (GS6, GS7, GS8, GS9, GS10), reaching κ=0.75 as best result (κ=0.71 is the agreement reached by the pathologists). The work of Ström et al. (2019) presents an ensemble of deep CNNs, trained using a cohort of the private Stockholm 3 (STHLM3) dataset (Grönberg et al., 2015), including 6’682 WSIs in the training partition and 1’631 WSIs in the test partition, pixel-wise annotated by 23 expert pathologists. The ensemble of networks is trained to classify benign and tumour tissue, reaching an $AUC = 0.99$. The ensemble of networks is also trained to predict Gleason patterns at the patch level (Benign, GP3, G4, G5) and to aggregate them to evaluate the Gleason score (groups stated by the International Society of Urological Pathology (ISUP): GS6, GS7=3+4, GS7=4+3, GS8,GS9-GS10), reaching κ=0.83. The work of Nagpal et al. (2019) presents a CNN, trained using private datasets from four medical centers, pixel-wise annotated by 19 expert pathologists.The CNN is trained to classify benign and tumour tissue, reaching an accuracy = 0.94. The CNN is also trained to predict Gleason patterns at the patch level (Benign, GP3, G4, G5), reaching an accuracy = 0.71. The Gleason score is assigned providing the percentage for each of the Gleason patterns. The performance is evaluated considering [GS6-GS7] vs. [GS-8,9,10] as the task, reaching an accuracy = 0.92.

Active learning: The research area that aims to reduce the labeling effort by introducing a human (or oracle) in the loop of training machine learning models is called active learning (Settles, 2009, 2011). A typical goal for an active learning system is to select the most relevant patches or images for training, therefore, avoiding unnecessary labeling costs. Usually, the learning algorithm is provided with a pool of unlabeled samples from which a model selects the samples to be annotated. The labels can be provided by a human expert and subsequently requested to be annotated and then used for training the models. The samples are ranked according to informative measures (Settles, 2009), such as entropy, or to query strategies, aiming to discard uninformative patches. Applications of active learning in medical imaging have focused on optimally selecting the patches that are annotated for the training of CNNs (Otálora et al., 2017), reducing user interactions for image segmentation Veeraraghavan and Miller (2011) and incremental fine-tuning of the CNN models (Zhou et al., 2017).

Weakly-supervised learning: The line of research that investigates how to best use image-level diagnostic labels and other less expensive annotations, known in machine learning literature as *weakly supervised learning*, has recently shown promising results in computational pathology (del Toro et al., 2017; Arvaniti and Claassen, 2018; Li et al., 2019; Otálora et al., 2020b; Campanella et al., 2019; Katharopoulos and Fleuret, 2019; van der Laak et al., 2019). Weak labels are often readily available in digital pathology via the pathology reports but are less specific than local region annotations. Usually, weak labels only summarize the pathologist’s main findings when analyzing the tissue slide or WSI, i.e., without any specific location or delineation of the regions used for the diagnosis. Weak labels usually refer to general categories such as cancer slide, benign tumour slide, or a score in a grading system for a specific organ, e.g. the Gleason grade in prostate cancer (del Toro et al., 2017). Detailed spatial specificity is often lacking in pathology reports since the relevant areas' exact location is usually not given. The work of del Toro et al. (2017) presents a CNN, trained using a cohort of the publicly available TCGA-PRAD dataset, labeled with global annotations. The CNN is trained to classify [GS6-7] vs [GS8-9-10], assigning global labels to the relevant patches, selected using the blue ratio within the WSI. The CNN reaches an accuracy = 0.78 for this binary problem. The work of Otálora et al. (2020b) compares several weakly supervised strategies for the fine-grained task of Gleason grading, reporting that the use of class-wise data augmentation and a DenseNet architecture using transfer learning lead to a $\kappa = 0.44$ in a set of 341 WSIs from the TCGA-PRAD dataset. The work of Arvaniti and Claassen (2018) presents a CNN architecture that combines weak and strong supervision for the task of low (GS6,7) vs. high (GS8,9,10) Gleason score classification. Two publicly available datasets are used: TMAZ (Arvaniti et al., 2018) and a cohort of TCGA-PRAD dataset. The model penalizes the weak supervision predictions, by weighing them using the predicted probability and the weak label. Therefore, the patches classified with a Gleason score that do not correspond with the weak label contribute less to the model's gradient updates. The results showed an accuracy of 0.848 for the binary cancer detection task using 447 WSIs. In the work of Campanella et al. (2019), the authors use transfer learning and a massive dataset of more than 44'000 WSIs (from breast, prostate and skin) with report-level labels to train weakly supervised binary CNN classifiers to distinguish between cancer and non-cancer slides. The ImageNet pre-trained classifiers were trained using a multiple instance learning paradigm, using bags in which the assigned label referred only to a non-empty subset of elements in the bag, accounting for the inherent label noise. Even though their results are a starting point for building screening tools that help the pathologist to discard non-cancer slides, their generalization to clinical scenarios (where data are highly-heterogeneous and Gleason score classification is evaluated instead of the binary tumour detection presented in the paper) has not been confirmed, yet.

Transfer learning: Transfer learning includes a set of techniques adopted to apply models on a task (or dataset), even if they were previously trained on another task (or dataset). Two main reasons motivate the adoption of transfer learning approaches. The first reason is that the generic features, previously learnt, can be re-used for different tasks or datasets (Bengio, 2012; Otálora et al., 2021). The second reason is the accelerated learning process since the models converge faster. Two approaches are mainly used in medical image analysis (Litjens et al., 2017; Mormont et al., 2018; Otálora et al., 2021). In the first approach, a model pre-trained on a dataset is used as the initialization for another model and then *fine-tuned* on a new task or on a new dataset. The approach is usually adopted in digital pathology, where models trained on natural image datasets (often ImageNet) are fine-tuned. In the work of Otálora et al. (2021), the authors use transfer learning to combine strongly-annotated and weakly-annotated data from heterogeneous datasets to classify between prostate Gleason grading, reaching performance comparable to the pathologists. In the second approach, the models previously trained are used for extracting feature vectors from the data and a classifier is built and trained on the top of this vector. In computational pathology the first approach often shows better results than the second one (Mormont et al., 2018).

Semi-supervised learning: Semi-supervised learning can be defined as being in between unsupervised learning (training with datasets that do not have any label) and fully-supervised learning (training with datasets in which each example has a label associated). Recent work found that unlabeled data can significantly improve generalization performance when used in conjunction with a small amount of labeled data. Obtaining unlabeled datasets in the medical image analysis field is reasonable task, since hospitals generate them routinely and there are also many public data sources without annotations. For such tasks, semi-supervised learning can be of great practical value (Foucart et al., 2019). In the work of Yalniz et al. (2019), the authors train a large ResNet model with 1 billion natural images from Instagram. The trained model achieved state-of-the-art performance in classification of the ImageNet dataset. In the work of Bulten et al. (2020), the authors train a CNN with a semi-supervised strategy for Gleason score classification. The CNN is trained using a private dataset including 1243 WSIs. A tumour detection CNN and a tumour segmentation CNN are used to generate cancer masks in pure Gleason pattern images (i.e. images where the primary and the secondary Gleason pattern are the same). The regions detected within masks are labeled with the Gleason pattern from the corresponding report. The CNN is evaluated for Gleason scoring using an internal and an external test partition (the external test partition proposed by Arvaniti et al. (2018)). On the internal test set, the CNN reaches κ=0.91. On the external test set, the CNN reaches κ=0.71.

In Tolkach et al. (2020), the authors train a CNN with a semi-supervised strategy for Gleason pattern classification. The CNN is trained using a private dataset including 389 WSIs as training partition, pixel-wise annotated by three pathologists and several cohorts from private datasets as test set. The CNN is trained for two tasks: tumour detection (benign vs. tumour) and Gleason grading. The CNN training for Gleason grading is semi-supervised. The model is first trained with strongly-annotated patches that originate from pure-GP WSIs and then it is used to annotate regions within WSIs where the primary Gleason pattern differs from the secondary Gleason pattern (considered as complex images by the authors). The regions annotated are then used to fine-tune the CNN. The CNN reaches $\kappa = 0.96$ in tumour detection and $\kappa = 0.74$ in Gleason grading.

In the recent work of Shaw et al. (2020), the authors use a chain of teacher-student models based on the approach of Yalniz et al. (2019) to annotate patches of colorectal cancer. The authors show that by using only a small fraction of the annotated data the model could automatically use the trained students to annotate the unlabeled patches. Their model achieves a comparable performance to the fully-supervised learning approach. Nevertheless, their approach is tested using a homogeneous dataset only, leaving the question to the generalization of teacher-student models to unseen centres open.

In a concurrent work Cheng et al. (2020), the authors propose a teacher-student model for segmentation of breast cancer lesions in the Camelyon dataset. Their teacher model learns from embeddings of spatially similar patches. Their model does not use the pseudo-labels generated from their teacher-student paradigm as ground truth for unlabeled samples but rather to counteract noisy labels in the ground-truth. The differences between the mentioned approaches and the one described in this article is presented in the Discussion section.

1.2. Contribution

As discussed in the above paragraphs, active, weakly and semi-supervised CNN models show feasible solutions to tackle classification tasks in computational pathology, particularly with the help of transfer learning approaches. The question that we address in this paper is: how can weakly annotated samples help to create strong region annotations? Semi-supervised and weakly-supervised learning are areas of active research in computational pathology with promising results (Bulten et al., 2020). This study aims to provide novel strategies to train models that generalize on heterogeneous datasets, using CNNs trained with pseudo-labeled data. Specifically, our contributions in this paper are the following

- We improve performance of fully–supervised models on the dataset proposed by Arvaniti et al. (2018), using pseudo–labeled examples from a weakly-annotated dataset, i.e. without requiring additional manual annotations on unseen heterogeneous data.
- We improve the generalization of CNN models on heterogeneous datasets in a context of few annotations.
- We study the overfitting in transfer learning when using several data sources for the supervision of the CNN models.
- We propose and evaluate three training variants of deep CNN models using both strongly and weakly-annotated datasets for the task of Gleason grading.

The rest of the paper is organized as follows: In Section 2, the teacher-student model is presented and the datasets used in the experimental evaluation are described in detail. In Section 3, the experimental results of the strategies are shown, as well as the evaluation varying the number of pseudo-labeled data included in the training of the student models. In Section 4 we discuss the results and in Section 5 concluding remarks finish the paper.

2. Methods

2.1. Datasets

Two openly accessible datasets are used for evaluating the teacher/student approaches. The datasets are highly heterogeneous, so they are more similar to a real scenario, as shown in Figure 1. The datasets are the Tissue MicroArray dataset Zurich (TMAZ) (Arvaniti et al., 2018) and a subset of the TCGA-PRAD dataset[8]. Among all the articles shown in Table 1, these two datasets are the only ones that are publicly available, even though only TMAZ is already split in partitions. The TMAZ dataset (Arvaniti et al., 2018) is composed of 886 prostate TMA cores that were scanned at the University Hospital of Zurich (NanoZoomer-XR Digital slide scanner, Hamamatsu). Each core is 3100^2 pixels, scanned at magnification 40x (0.23 microns per pixel). The dataset includes pixel-wise local annotations of pathologists. It is partitioned into a training partition with 508 TMA cores, a validation partition with 133 cores and a testing partition of 245 cores. The composition of the dataset is shown in Table 3. The partitions are the ones presented in Arvaniti et al. (2018): the training partition includes cores from three different TMA arrays (ZT111, ZT199, ZT204), while the validation partition and the test set include both only one array (respectively ZT76 and ZT80). This small amount of arrays implies a visual homogeneity between the TMAZ cores, mostly regarding the stain colour and tissue structures, as shown in Figure 1. TCGA-PRAD includes 449 WSIs (Formalin-Fixed Paraffin-Embedded, FFPE), scanned from several centers. The WSIs are scanned at magnification 40x and they can easily be in the order of $100'000^2$ pixels per image. WSIs are provided with the corresponding pathologist reports, but without any annotation. A subset of 301 WSIs was manually labelled, as the label extraction process is time-consuming. The labels, the primary and the secondary Gleason patterns, are used as weak labels. The subset is split into a training partition (171 WSIs) representing 145 patients from 19 medical centres, a validation partition (84 WSIs), representing 70 patients from 21 medical centres, and a test partition (46 WSIs), representing 38 patients from 12 medical centres. The composition of the dataset is shown in Table 5. The number of medical centres implies that the visual appearance of the WSIs is highly heterogeneous, mostly regarding the stain colour and tissue structures, as shown in Figure 1.

2.2. Data analysis pipeline

2.2.1. Teacher/student paradigm

The training schema is based on the teacher/student paradigm, previously used by Lee (2013); Hady and Schwenker (2013); Yalniz et al. (2019). The paradigm involves two convolutional neural networks, named the teacher and the student model. An overview of the training schema is illustrated in Figure 2.

The teacher model is a high-capacity neural network, trained to annotate pseudo-labeled samples within weakly-annotated or unlabeled data. The pseudo-labels are created in the following way. For each of the samples without local annotations the teacher model computes a probability array. It labels the sample as the class with the highest probability. The pseudo-labels are used as if a human expert had performed the annotation (Lee,

[8] https://portal.gdc.cancer.gov/projects/TCGA-PRAD. Retrieved 20th of March, 2020

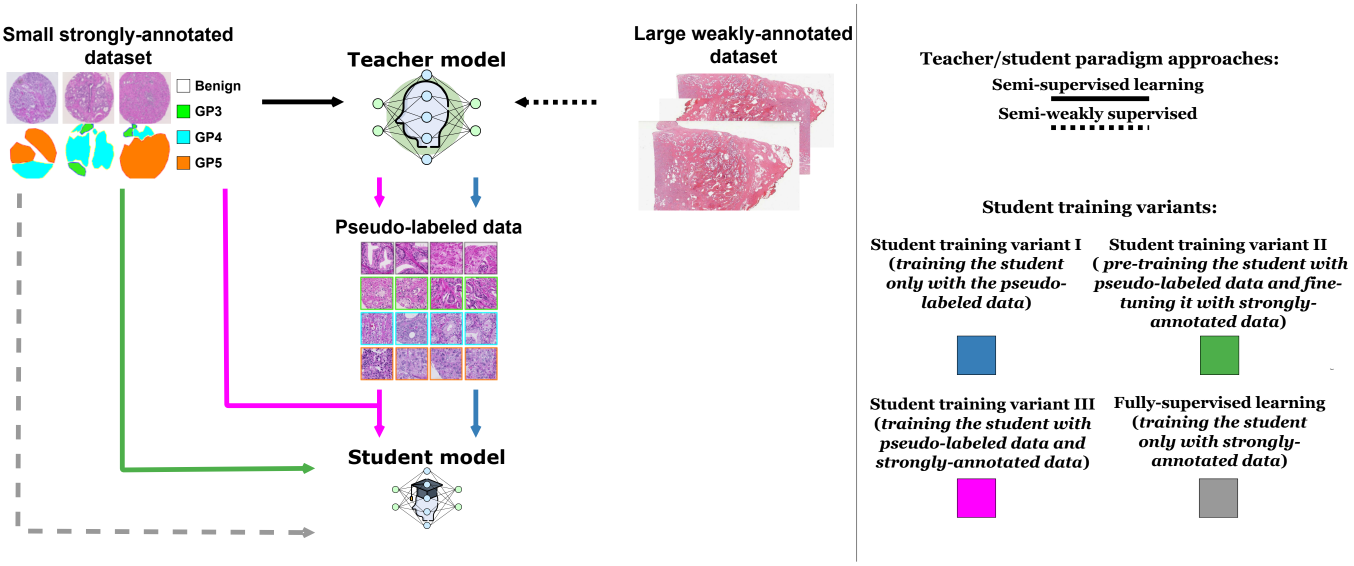

Small strongly-annotated dataset
Benign
GP3
GP4
GP5
Teacher model
Large weakly-annotated dataset
Pseudo-labeled data
Student model
Teacher/student paradigm approaches:
Semi-supervised learning
Semi-weakly supervised
Student training variants:
Student training variant I (training the student only with the pseudo-labeled data)
Student training variant II (pre-training the student with pseudo-labeled data and fine-tuning it with strongly-annotated data)
Student training variant III (training the student with pseudo-labeled data and strongly-annotated data)
Fully-supervised learning (training the student only with strongly-annotated data)

Fig. 2. Scheme for the training of the teacher (top) and the student (bottom) models. Two approaches of the paradigm are shown: the semi-supervised learning (solid line), the semi-weakly supervised learning (dotted line). They influence the training of the teacher model. For each of the two approaches, three student training variants are presented. The different colours represent the three student training variants described. They are compared with a fully-supervised learning approach of the student model (dashed grey line).

2013). A subset of the pseudo-labels includes labels where the ground truth matches the real class (relevant labels), while the other subset includes labels where the ground truth does not match the real class (noisy labels) (Yalniz et al., 2019; Han et al., 2018). The latter labels influence the learning process negatively (Natarajan et al., 2013; Karimi et al., 2019). With high-capacity models, it is possible to better separate relevant labels from the noisy ones (Han et al., 2018). High-capacity models can be trained from large image datasets with weak labels, such as hashtags from vast social media datasets. Yalniz et al. (2019) et al. show how very large capacity CNN models trained with a large amount of data outperform the low-capacity CNN models trained with standard datasets. The student model is a small neural network in terms of number of parameters (compared to the teacher) and it is trained using pseudo–labeled and/or strongly-annotated data depending on the student training variant. The student model is designed to be efficient (fast in the evaluation of the inputs (Chen et al., 2019)) at testing time, with a performance comparable to that of the teacher (Guo et al., 2019).

2.2.2. *Teacher/student approaches*

Two approaches for the teacher/student paradigm are presented (see Figure 2): the semi-supervised learning and the semi-weakly supervised learning. The difference between the approaches concerns how the teacher model is trained. In the semi-supervised learning approach, the teacher model is trained only with strongly-annotated data (solid lines in Figure 2). In the semi-weakly supervised learning approach, the teacher model is pre-trained with weakly-annotated data (dotted line in Figure 2) and then it is fine-tuned with strongly-annotated data (dashed line in Figure 2).

2.2.3. *Student training variants*

Each teacher/student approach is evaluated using three variants to train the student model (Figure 2). The student training variants concern how the pseudo-labeled and strongly-annotated data are combined for training the student model. A fully-supervised learning approach is also evaluated as the baseline for the methods.

Student training variant I. In the student training variant I (blue line in Figure 2), the student model is trained using only the pseudo-labeled data.

Student training variant II. In the student training variant II (green line in Figure 2), the student model is pre-trained using the pseudo-labeled data and then it is fine-tuned using the strongly-annotated data.

Student training variant III. In the student training variant III (magenta line in Figure 2), the student model is trained using both the pseudo-labeled data and the strongly-annotated data in the same training phase.

Fully-supervised training approach. In the fully-supervised learning approach of the student training (grey dashed line in Figure 2), the student model is trained using only the strongly-annotated data. This approach provides a baseline to evaluate the student training variants.

2.3. *Experimental setup*

2.3.1. *Image Preprocessing*

The images of both datasets are pre-processed with the same strategy: they are tiled into patches, the background regions are removed and finally the patches are extracted and selected.

Tiling is required since modern GPU hardware cannot handle a very large image size due to limited memory. The images are initially split into patches of 750x750 pixels at 40x magnification. They are then down-sampled to 224x224 pixels as it is the required size for using the pre-trained networks, following an approach similar to the one proposed by Arvaniti and Claassen (2018). Several patch size configurations were tested: 224x224, 250x250, 500x500, 750x750 and 1'000x1'000 pixels. A size of 750x750 pixels was chosen by visual inspection as it provides enough context to capture the morphology of the glands while also keeping details of the cancer cells. The approach is applied to both datasets. The patches are tiled at 40x magnification because this is the only magnification available for the images from the TMAZ data. Patches corresponding to the background regions are not considered in the analysis since they are not informative. In the TMAZ images, the background is removed according to the pathologist annotations. In the TCGA-PRAD images, the background is removed using tissue masks generated with the HistoQC tool (Janowczyk et al., 2019). On TCGA images, HistoQC is also used to curate the WSIs, by detecting and removing pen-markings manually made by the pathologists. The patch extraction and selection processes are different, as the datasets have different characteristics in terms of image size, resolution and type of annotations. In the TMAZ dataset, 30 patches (with possible overlapping) are randomly extracted from each TMA core. The patches are selected so they contain at least 60% tissue. The number of patches extracted is chosen considering the size of the patches (750x750), the overlapping between them and the size of the cores (3100x3100). In the TCGA-PRAD dataset, the WSIs are divided into grid cells (without overlapping) and then the patches are densely extracted. The number of patches extracted varies between 400 and 12'000 per WSI, depending on the size of the WSIs and the amount of background. Two sets of TCGA-PRAD patches are used for training the models: the first set is used for training the teacher model, while the second set is used during the training of the student model. The first set is composed of patches used as weakly-annotated data that are used to pre-train the teacher in the semi-weakly supervised learning approach. The TCGA-PRAD dataset has no pixel-wise local annotations and therefore it is not possible to distinguish between healthy and cancer tissue. In this case, the patches are labeled with the primary Gleason pattern of the corresponding WSI. In order to reduce the noise due to the use of weak labels, only a subset of the patches is selected to train the teacher model. The subset of the patches is selected using the Blue-Ratio (BR) technique (Chang et al., 2012). BR ranks patches starting with dense nuclei regions first, avoiding patch extraction from areas without nuclei, such as those containing fat or connective tissue. The patches are sorted in decreasing order by their BR and only the ones with the highest and the lowest values are selected for each WSI. The second set is composed of patches used as unlabeled data and the teacher automatically annotates all of them. In order to reduce noise in the labeling, only the top-ranked samples are selected for training the student model. A probability array is created with the softmax probability of each sample. For each class, the probability array is sorted in descending order by the class probability and the first K top-ranked samples are selected. K denotes the number of patches from each class selected by the teacher model.

Table 2. Number of patches for each Gleason pattern in the TMAZ dataset.

Class/Set	Training	Validation	Test
Benign	2'010	1'350	127
GP3	5'992	1'352	1'602
GP4	4'472	831	2'121
GP5	2'766	457	387
Total	15'240	3'990	4'237

Table 3. Number of TMA cores for each Gleason score in the TMAZ dataset.

Class/Set	Training	Validation	Test
Benign	61	42	12
GS6	158	35	79
GS7 (3+4)	47	14	28
GS7 (4+3)	18	11	23
GS8	119	15	84
GS9 - 10	105	16	19
Total	508	133	245

Table 4. Number of patches for each Gleason pattern selected in the TCGA-PRAD dataset (used only for pre-training the teacher model).

Class/Set	Training	Validation	Test
Benign	1710	840	460
GP3	28'919	15'443	4'000
GP4	48'398	22'500	13'633
GP5	8'000	4'000	3'000
Total	87'027	42'783	23'093

2.3.2. Dataset composition

TMAZ patches. The patches selected from the TMAZ dataset are used as strongly-annotated data for training and testing both models. The dataset includes four classes: benign, Gleason pattern 3, Gleason pattern 4 and Gleason pattern 5. The detailed number of patches (divided per class) is reported in Table 2.

TCGA-PRAD patches. TCGA-PRAD patches are used to train the teacher model, to train the student model and to test the student model at the WSI-level, as shown in Figure 3. The patches used to train the teacher model include the 500 top-ranked and the 100 bottom-ranked patches per WSI regarding BR. These patches are used as weakly-annotated data. The label assigned to each patch is the corresponding primary Gleason pattern included in the medical report. However, TCGA-PRAD does not include WSIs with benign as primary Gleason score. Only Gleason pattern 3, Gleason pattern 4 and Gleason pattern 5 are represented, as prostatectomies are not performed for lower Gleason patterns. Several numbers of patches (100, 250, 500, 750, 1'000) selected with the BR were tested. Visual inspections allowed to verify that 500 is the number that guarantees patches with nuclei, avoiding the selection of patches with stroma. For each of the WSIs, 100 patches with

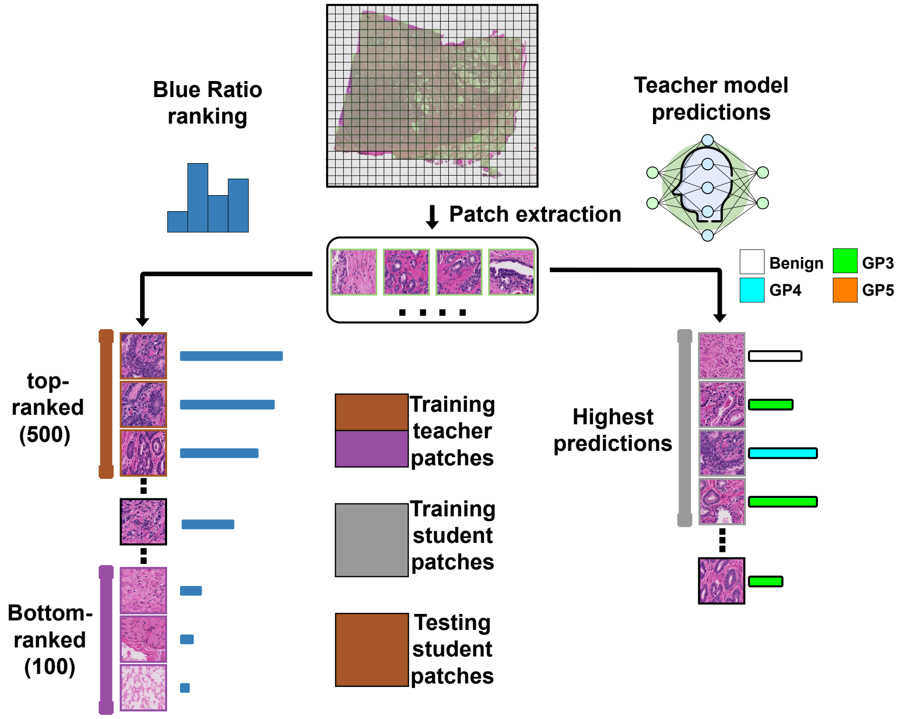


Fig. 3. Scheme for the selection of the patches from TCGA-PRAD WSIs. Patches from TCGA-PRAD are ranked according to the Blue Ratio (left part of the Figure) and according to the predictions made by the teacher model (right part of the Figure). For each WSI, the 500 top-ranked patches (brown line) and the 100 bottom-ranked patches (purple line) according to Blue Ratio are included in the training partition of the teacher model. The 500 top-ranked patches are used to test the performance of the student model at the WSI-level. The teacher model makes predictions on the patches from TCGA-PRAD. The patches with the highest predictions (grey line) are labeled as pseudo-labels, to train the student model, as described in 2.2.1.

Table 5. Number of WSIs for each Gleason score in the TCGA-PRAD dataset.

Class/Set	Training	Validation	Test
GS6	13	20	5
GS7 (3+4)	42	10	6
GS7 (4+3)	30	14	11
GS8	37	12	13
GS9 - 10	49	28	11
Total	171	84	46

very low BR were selected and labeled with the benign class so that the model could be trained with samples labeled in four classes, similar to the strongly-annotated data. The detailed number of patches (divided per class) is reported in Table 4. The patches used to train the student model include the patches annotated with the highest predictions by the teacher model. These patches are used as pseudo-labeled data. The set includes four classes: benign, Gleason pattern 3, Gleason pattern 4 and Gleason pattern 5. The training partition of the pseudo-labeled data includes different subsets, each one with a different number of samples for each class (K). The subsets are incrementally added to the partition. Different K values are tested in the paper. The first subset includes the top K-ranked samples, for each of the classes. Each of the following subsets added to the partitions includes the next top K-ranked samples per class. The K values tested vary between 1’000 and 10’000 pseudo-annotated patches for each class. The difference between two consecutive K values is 1’000, i.e. the first subset has 1’000 patches per class and the second one 2’000 patches per class etc.. The highest K value tested is 10’000, then the biggest subset has 10’000 samples pseudo-labeled for each class. The validation partition of the pseudo-labeled data includes 8’000 patches (2’000 for each class). The testing partition of the pseudo-labeled data includes 8’000 patches (2’000 for each class). The patches used to test the student model include the the 500 top-ranked patches per WSI regarding BR. The patches are used as unlabeled data. The student model predicts Gleason patterns for each patch and aggregates the predictions in order to have primary and secondary Gleason patterns for the WSI. TCGA-PRAD does not include benign images, therefore the analysis of the top-ranked patches allows to exclude patches including stroma or healthy tissue.

2.3.3. *Data analysis implementation*

The teacher and the student are implemented according to the teacher/student paradigm, with the same software and hardware solutions and trained with the same strategy.

The paradigm includes the architecture size of the CNN adopted as model. The teacher model is a Resnext50_32x4d (Xie et al., 2017) implemented by Yalniz et al. (2019). The choice of the teacher model is made considering the paradigm constraints (the teacher must be a high-capacity model) and the classification performance of the model. Therefore, the teacher model has up to 22 million parameters and it guarantees high-level classification performance. The network is initialized with weights of the model pre-trained with the YFCC100M dataset (Thomee et al., 2015) using 1 billion Instagram images and their corresponding hashtags. The choice of using pre-trained networks is made to speed up the model convergence during the training. The student model is a DenseNet121 (Huang et al., 2016) implemented within the PyTorch framework[9]. The choice of the student model is made considering the paradigm constraints (i.e. to have a model smaller than the teacher and fast in the evaluation of the inputs). The student model has up to 7 million parameters and it guarantees classification performance comparable with the one of the teacher. The network is initialized with weights of the model pre-trained with ImageNet (Deng et al., 2009) using 1 million natural images. Both architectures are modified using a classifier different from the original one. The original networks were pre-trained with datasets including samples from 1'000 classes. Therefore they had 1'000 nodes in the last dense layer (1'000 is the number of classes in the dataset). The new classifier used in both models has only four output nodes in the last layer, equal to the number of Gleason patterns.

The teacher and the student models are implemented using the PyTorch framework [10] and using the Cartesius cluster as infrastructure, provided by the SURFsara High Performance Computing centre (HPC)[11]. PyTorch is a framework developed for deep learning research. The version adopted is 1.1.0. Pytorch is used to write our experiments in a high-level language that uses the computational power of the graphics processing unit (GPU) and its application programming interface. The Cartesius cluster contains more than 130 nodes, each equipped with two Tesla K40m GPUs. The models are trained using two GPUs for each CNN.

The strategy adopted for training the teacher and the student models concerns the hyperparameters of the networks and the solution adopted for facing the non-deterministic condition of the training.

The hyperparameters adopted are selected considering the convergence of the models (evaluated on the validation partition). The batch size is chosen considering the samples in the training partition. If the partition includes a large number of samples (i.e. more than 60'000 samples), a batch size of 128 samples is selected. If the partition includes a smaller number of samples (fewer than 60'000 samples), a batch size of 32 samples is chosen, in order to reduce the effort in terms of time to train the models. The number of epochs is chosen considering the performance on the validation partition. When the model is tested with a large dataset (e.g. weakly-annotated dataset), it is empirically observed that the model converges within 10 epochs and that it reaches the minimum value in the loss function. When the model is tested with a smaller dataset, it is empirically observed that the model converges within 15 epochs and reaches the minimum value in the loss function. A grid search algorithm (Chicco, 2017) allows to identify the optimal values (i.e. the values that allow the CNN to minimize the loss function on the validation partition) for the hyperparameters (i.e. optimizer algorithm, the learning rate and the decay rate). The optimizer selected is Adam (Adam and SGD were tested). The learning rate selected is 10^{-3} (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} were tested). The decay weight selected is 0 (10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 0 were tested).

The teacher can be trained with weakly-annotated or with strongly-annotated data. In the first case (pre-training of the teacher model), the model is trained for ten epochs, with a batch size of 128 samples (the partition includes more than 87'000 samples). In the second case, the model is trained for fifteen epochs, with a batch size of 32 samples. The student is trained for fifteen epochs for each of the training variants with a batch size of 32 samples. The loss function in each epoch minimizes the cross-entropy loss function between the predicted class and the ground truth label. At the end of each epoch, the loss function is evaluated on the validation partition of the corresponding dataset. The model weights are saved only when the loss function is lower than the loss function in previous epochs. When the teacher model is trained with weakly-annotated data, the validation partition includes the validation partition of the TCGA-PRAD dataset presented for the teacher model (Paragraph 2.3.2). When the teacher model is trained with strongly-annotated data, the validation partition includes the validation partition of the TMAZ dataset (Paragraph 2.3.2). In all the student training variants presented (Section 2.2.3), the validation partition includes the validation partition of the strongly-annotated dataset (i.e. TMAZ) (Paragraph 2.3.2) and the validation partition of the pseudo-labeled data (Paragraph 2.3.2). The validation partitions are evaluated separately. The choice of using both datasets during the evaluation is made to avoid that the model overfits on one of the two datasets. In this case, the model weights are saved only when the loss function is the lowest one for both partitions within an epoch, compared with the loss functions of the other epochs.

The training of the models is not deterministic because of the stochastic gradient descent optimizer (Adam optimizer) used to train the models and also partially because of the probabilistic rate applied to the data augmentation (presented in Section 2.3.4). In order to limit the non-deterministic effects introduced in the training, both models are trained ten times in each step of the pipeline, except for the training of the teacher model in the semi-weakly supervised learning approach (see Section 2.2.2). In the semi weakly-supervised learning approach, the teacher model is pre-trained with weakly-annotated data only once because of the large number of patches within the dataset. Considering the training and the validation partitions, the pre-training is made with up to 130'000 patches, while in the other steps of the pipeline, fewer than half of the patches are used for training the models. In both teacher/student approaches (Section 2.2.2), the teacher model selected to annotate the unlabeled data among the ten repetitions is the one that shows the best performance in κ, both in the TMAZ validation partition (Paragraph 2.3.2) and in TCGA-PRAD validation partition (Paragraph 2.3.2). This criterion is selected considering

[9] https://pytorch.org/hub/pytorch_vision_densenet. Retrieved 20th of March, 2020

[10] hhttps://pytorch.org/. Retrieved 20th of March, 2020

[11] https://userinfo.surfsara.nl/systems/hpc-cloud. Retrieved 7th of February, 2020

the inter-dataset heterogeneity so that the model can generalize on both datasets, avoiding a model overfitting on one of them. In TMAZ, the two approaches are tested at the patch level, while in TCGA-PRAD, they are tested at the WSI level, as described in Paragraph 2.3.5. In the student training variant II (Section 2.2.3), the student model selected to be fine-tuned with strongly-annotated data is the one that shows the best performance in κ, in the TMAZ validation partition (Paragraph 2.3.2).

2.3.4. Data augmentation

Class-wise data augmentation (CWDA) is applied during the training of the CNN models. The class-wise data augmentation is composed of three kinds of operations and it is applied for avoiding overfitting. The three operations are rotation, flipping and colour augmentation. Rotation augmentation is applied with randomly rigid rotations (90,180,270 degrees). Flipping augmentation is applied flipping the image vertically and/or horizontally. Colour augmentation is applied shifting the hue, saturation and brightness values of the original image. The colour augmentation parameters are selected according to the parameters suggested by Janowczyk[12]. The parameters for the colour augmentation are: the hue shift is limited to be between -9 and 9, the saturation shift is limited to be between -25 and 25 and the brightness shift is limited to be between -10 and 10. Each of these operations is applied to the training images, with a probability of 0.5. The pipeline is implemented using the Albumentations open-source library (Buslaev et al., 2020).

Overfitting can be caused by the unbalanced distribution of the class distribution within the datasets and by the small amount of data available (Tables 2, 3, 4, 5), particularly for the TMAZ dataset. Class-wise data augmentation (CWDA) is also used to solve this problem, by augmenting classes that are less frequently represented in a stronger way. The implementation is based on the open access repository of Ufoyn[13].

2.3.5. Evaluation

The models are evaluated on the classification of the Gleason score and the Gleason patterns of histopathological images given the image patches. The quadratic weighted Cohen Kappa score (κ) is used for assessing the model performance. The optimal Kappa is 1 and a random distribution would be 0, as kappa is normalized by agreement by chance. The average and the standard deviation of the κ of the models are reported. The Gleason pattern classification task is evaluated at the patch-level, while the Gleason score classification is evaluated at the core level (for TMAZ) and at the WSI level (for TCGA-PRAD). The Gleason score is evaluated aggregating the Gleason patterns classified at the patch level using a majority voting system. The most frequently predicted class is selected as primary Gleason score, while the second most frequently predicted as secondary Gleason pattern. The majority voting has two main drawbacks. The first drawback regards the fact that the majority voting does not work well for the TMAs or the WSIs where the primary and the secondary Gleason patterns coincide. In the TCGA-PRAD dataset, the drawback is handled in the following way: if the predominant pattern is represented in more than twice the amount of patches as the second pattern, it is considered to be both the primary and the secondary Gleason pattern. The second drawback regards noisy patches that can influence the voting. Not all the patches include tissue (e.g. stroma or healthy tissue) that belongs to one of the four classes analyzed in this paper (benign, Gleason pattern 3, Gleason pattern 4, Gleason pattern 5). The model is trained to classify one of these classes and the predictions can thus introduce noise. This drawback is limited selecting only patches from un-healthy tissue (500 patches) for each WSI, using the Blue-Ratio, as explained in Section 2.3.1. The majority voting follows the revised grading system (Pierorazio et al., 2013) proposed by the American Urology Association[14]. In the TMAZ dataset, in order to make results comparable, the majority voting approach used is the one adopted by Arvaniti et al. (2018).

The κ measures the agreement between raters. The quadratic weighted Cohen Kappa score is adopted because it penalizes predicted values far from their actual class stronger:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

The same formula is applied to both Gleason patterns and Gleason scores. i, j are the ordered patterns, $N = 4$ is the total number of Gleason patterns ($N = 5$ is the total number of Gleason scores). $O_{i,j}$, is the number of images that were classified with a pattern (score) i by the first rater and j by the second. $E_{i,j}$ denotes the expected number of images receiving rating i by the first expert and rating j by the second. The quadratic term $w_{i,j}$ penalizes the ratings that are not close to the right value. When the predicted Gleason pattern (score) is far from the ground-truth class, $w_{i,j}$ gets closer to 1. E.g. if the ground truth of a patch is Gleason pattern 5 and it is predicted as Gleason pattern 4, it is penalized less than if the predicted class is benign.

Cohen's κ is usually adopted in prostate classification to assess the level of agreement in multiclass problems (Berg et al., 2011; Arvaniti et al., 2018; Ström et al., 2019; del Toro et al., 2017; Otálora et al., 2020b), as shown in Table 1. The inter-pathologist agreement in prostate cancer classification can be used as an overall reference or baseline for evaluating the models. In Arvaniti et al. (2018), two pathologists pixel-wise annotated 245 tissue micro arrays. They reached $\kappa = 0.67$ in Gleason grading, while $\kappa = 0.71$ in Gleason scoring. In Tolkach et al. (2020), three pathologists pixel-wise annotated 453 large tumour images from a cohort of TCGA-PRAD dataset. They reached $\kappa = 0.70$ in Gleason grading.

[12] http://www.andrewjanowczyk.com/employing-the-albumentation-library-in-pytorch-workflows-bonus-helper-for-selecting-appropriate-values/. Retrieved 8th of January, 2020

[13] https://github.com/ufoym/imbalanced-dataset-sampler. Retrieved 6th of February, 2020

[14] https://www.auanet.org/education/auauniversity/education-products-and-resources/pathology-for-urologists/prostate/adenocarcinoma/prostatic-adenocarcinoma-gleason-grading-(modified-grading-by-isup. Retrieved 5th of February, 2020

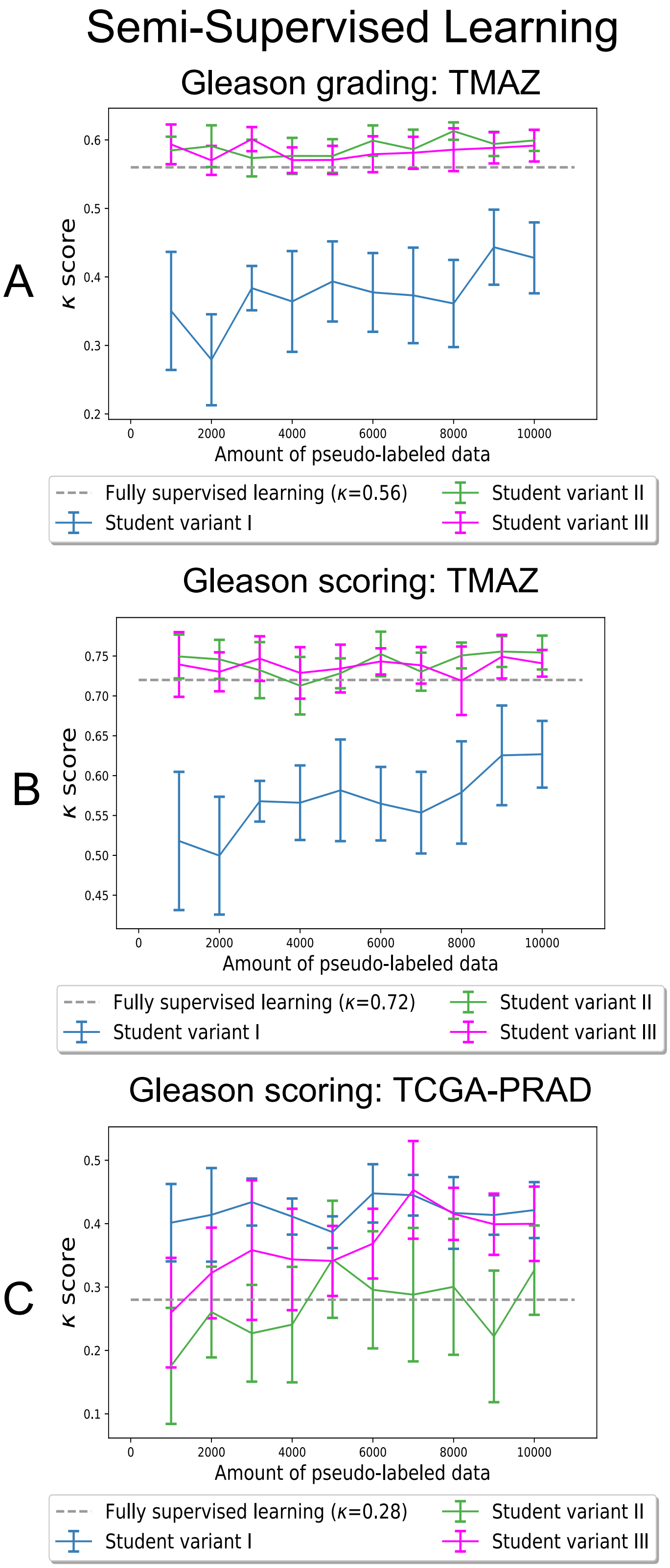


Fig. 4. Results for the average performance of the student training, trained with the semi-supervised learning approach (the teacher model is trained with strongly-annotated data). They are measured by the κ as a function of the amount of pseudo-labeled data used to train the student model. The sub-figure A includes the performance evaluated in Gleason grading at patch level using the TMAZ test data, the sub-figure B includes the performance evaluated in Gleason scoring at TMA core level using the TMAZ test data, the sub-figure C includes the performance evaluated in Gleason scoring at the WSI level, using the TCGA-PRAD test data.

3. Results

The tested semi-supervised learning and semi-weakly supervised learning approaches improve the student model performance both in Gleason grading and Gleason scoring compared with the fully supervised learning approach. The improvement is statistically significant, as the results are tested also using the Wilcoxon Rank-Sum test (Wilcoxon, 1992). The Wilcoxon Rank-Sum test is used for determining if two probabilistic populations have the same distribution (null hypothesis). If the test confirms the null hypothesis (p-value > 0.05), it means that the populations have the same distribution. If it rejects the null hypothesis (p-value < 0.05) the populations are assumed to have different distributions.

In both classification tasks the best performance is reached with the semi-supervised learning approach. Figures 4 and 5 show the performance of the training/student approaches presented. In each figure, the performance is presented for the three student variants and the fully-supervised baseline described in Section 2.2.3. The performance is measured as a function of the amount of pseudo-labeled data used to train the student model, as described in Section 2.3.5. Each figure includes three curves and a constant line. The curves represent the student training variants, as presented in Section 2.2.3. The blue curve represents the performance obtained with the student training variant I (training the student only with pseudo-labeled data). The green curve represents the performance obtained with the student training variant II (pre-training the student with pseudo-labeled data and fine-tuning it with strongly-annotated data). The magenta curve represents the performance obtained with the student training variant III (training the student with both pseudo-labeled data and strongly-annotated data). The constant line (grey dashed line) represents the fully-supervised learning approach (training the student only with strongly-annotated data).

Figure 4 and Table 6 shows the performance of the semi-supervised learning approach. Figure 4 includes three sub-figures: 4.A, 4.B and 4.C. In sub-figure 4.A, the models are evaluated in Gleason grading on the TMAZ test partition, at the patch level. Two student training variants exceed the baseline: student training variant II and student training variant III. In sub-figure 4.B, the models are evaluated for Gleason scoring on the TMAZ test partition, at TMA core level. Two student training variants exceed the baseline: student training variant II and student training variant III. In sub-figure 4.C, the models are evaluated in Gleason scoring on the TCGA-PRAD test partition, at WSI level. All the student training variants exceed the baseline. Table 6 summarizes the peak results for the semi-supervised learning approach, highlighting the ones that are statistically significant.

Figure 5 and Table 7 show the performance of the semi-supervised learning approach. Figure 5 includes three sub-figures: 5.A, 5.B and 5.C. In Figure 5.A, the models are evaluated for Gleason grading on the TMAZ test partition at the patch level. Two student training variants exceed the baseline in κ: student training variant II and student training variant III. In Figure 5.B, the models are evaluated for Gleason scoring on the TMAZ test partition, at TMA core level. Two student train-

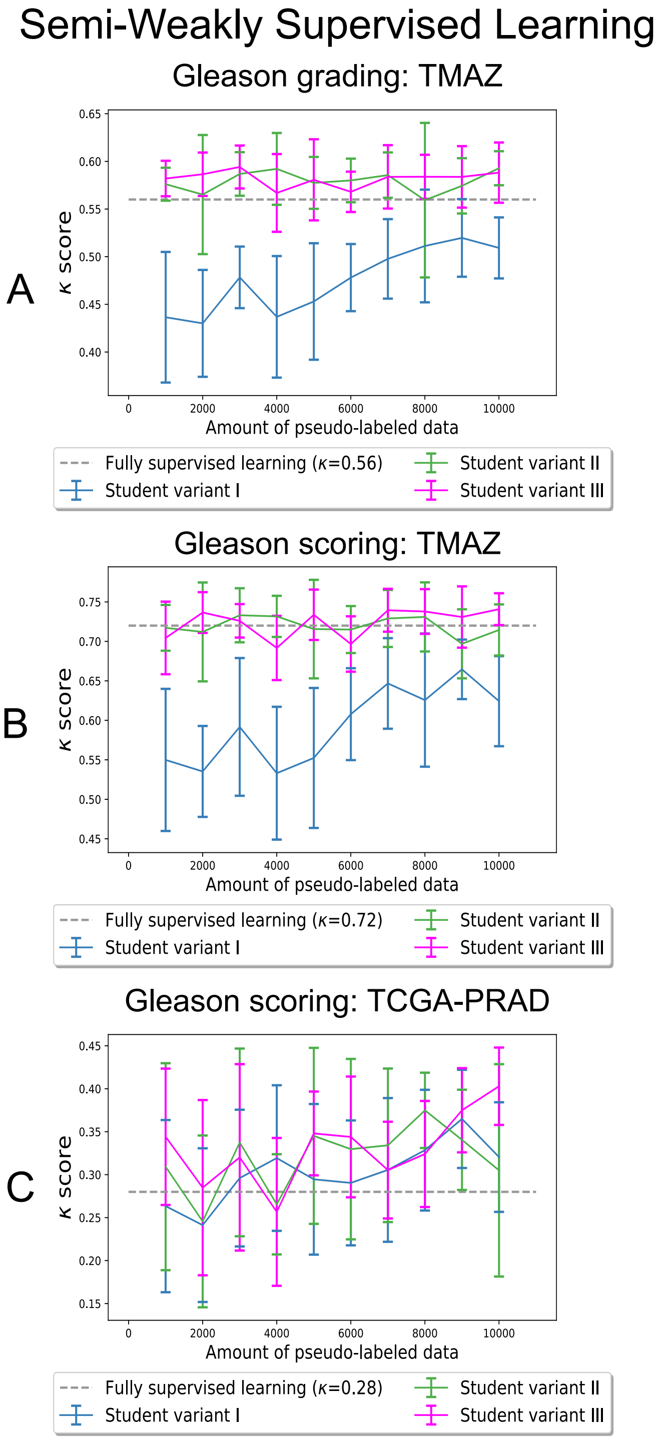


Fig. 5. Results for the average performance of the student models with the semi-weakly supervised learning approach (the teacher model is pre-trained with weakly-annotated data and it is fine-tuned with strongly-annotated data). They are measured by the κ as a function of the amount of pseudo-labeled data used to train the student model. Sub-figure A includes the performance evaluated in Gleason grading at the patch level using the TMAZ test data, the sub-figure B includes the performance evaluated in Gleason scoring at TMA core level using the TMAZ test data, sub-figure C includes the performance evaluated in Gleason scoring at the WSI level, using the TCGA-PRAD test data.

Table 6. Performance measured for the semi-supervised learning approach evaluated in κ. For each of the student training variants, the peak value and the corresponding number of pseudo-labeled patches per class are reported. The results that are statistically significant (compared with the baseline) are reported with an asterisk (*).

Training approach	**κ-score**	**# pseudo-labels**
Gleason grading: TMAZ dataset		
Fully-supervised	0.5667 ± 0.0285	-
Training variant I	0.4643 ± 0.0385	9’000
Training variant II	**0.6127** ± **0.0133***	10’000
Training variant III	0.6086 ± 0.0176*	3’000
Gleason scoring: TMAZ dataset		
Fully-supervised	0.7186 ± 0.0306	-
Training variant I	0.6284 ± 0.0492	9’000
Training variant II	**0.7645** ± **0.0231***	9’000
Training variant III	0.7562 ± 0.0293*	9’000
Gleason scoring: TCGA-PRAD dataset		
Fully-supervised	0.2293 ± 0.1350	-
Training variant I	**0.4529** ± **0.0512***	7’000
Training variant II	0.3981 ± 0.1085	1’000
Training variant III	0.4353 ± 0.0483*	8’000

ing variants exceed the baseline in κ: student training variant II and student training variant III. In Figure 5.C, the models are evaluated for Gleason scoring on the TCGA-PRAD test partition, at WSI level. All the student training variants exceed the baseline. Table 7 summarizes and the peak results for the semi-weakly supervised learning approach, highlighting the ones that are statistically significant.

Figure 6 shows confusion matrices of the CNNs (among all the repetitions) that reach the highest performance in Gleason grading and Gleason scoring tasks. They are all obtained using the semi-supervised learning approach. For each task, the matrix with the raw values (on the left) and the normalized matrix are shown (on the right). Figure 6.A shows confusion matrices of the best CNN in Gleason grading, evaluated in TMAZ. Figure 6.B shows confusion matrices of the best CNN in Gleason scoring, evaluated on TMAZ. Figure 6.C shows confusion matrices of the best CNN in Gleason grading, evaluated on TCGA-PRAD.

4. Discussion

Although deep learning is the state-of-the-art technique in classification tasks, it is not easy to train models that generalize well on varying datasets, particularly when data are highly-heterogeneous and few data with local annotations are available. The teacher/student approaches presented in this paper allow training models that reach high performance in prostate cancer classification, improving the performance of a CNN (compared with a CNN trained only with strongly-annotated data), generalize well on different heterogeneous datasets, facing the inter-dataset heterogeneity and overcoming the lack of large datasets with local annotations.

Prostate cancer classification is still an open challenge in digital pathology despite much work on the topic. The analysis usually involves two tasks: Gleason grading and Gleason

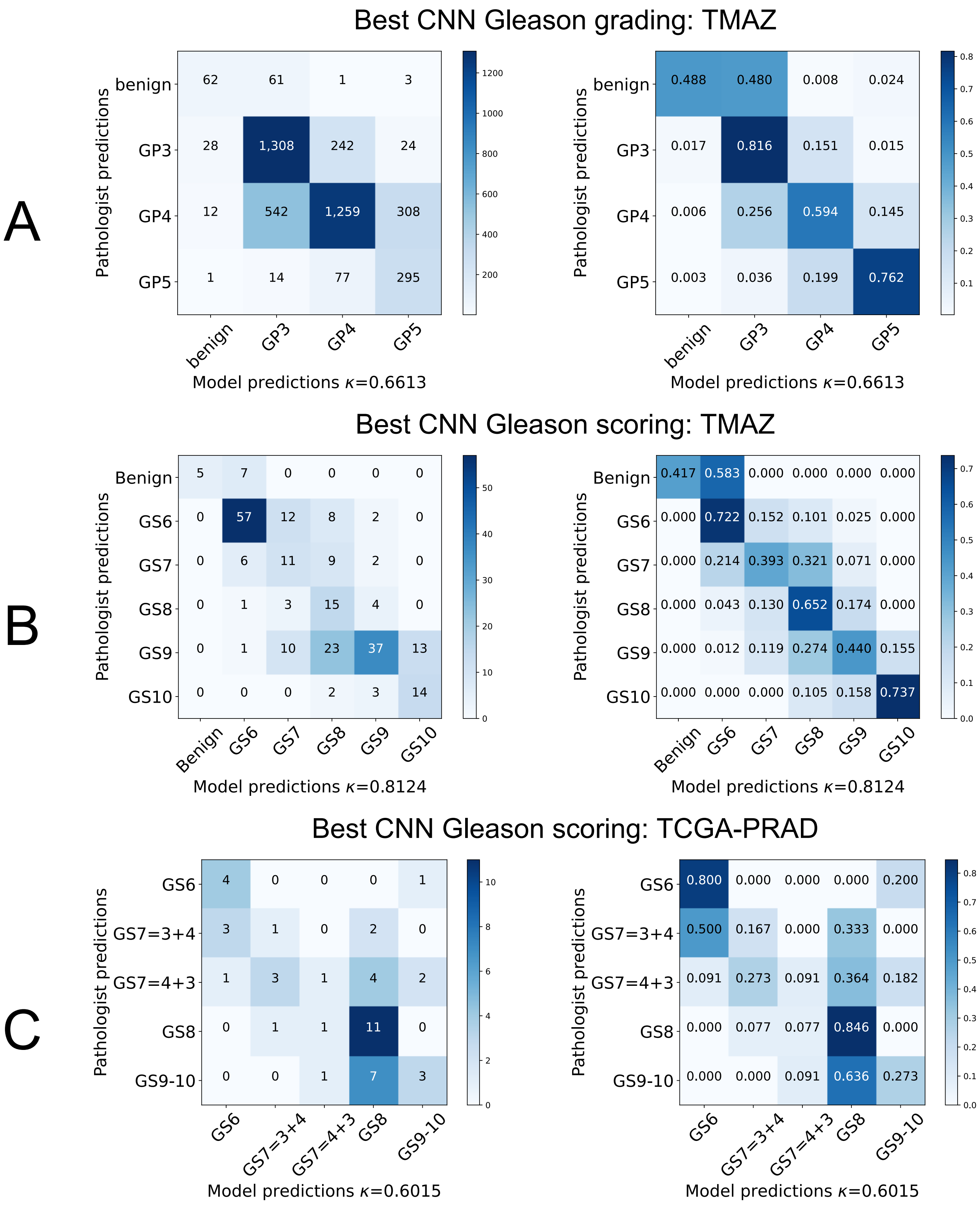


Fig. 6. Confusion matrices of the CNN models that show the highest performance in the tasks proposed in the paper. Sub-figure (A) includes the best CNN on Gleason grading in TMAZ, sub-figure (B) includes the best CNN on Gleason scoring in TMAZ, sub-figure (C) includes the best CNN on Gleason scoring in TCGA-PRAD. For each sub-figure, two confusion matrices are shown: on the left, the confusion matrix with the raw predictions of the CNN, on the right the confusion matrix with the normalized predictions.

Table 7. Performance measured for the semi-weakly supervised learning approach evaluated in κ. For each of the student training variants, the peak value and the corresponding number of pseudo-labeled patches per class are reported. The results that are statistically significant (compared with the baseline) are reported with an asterisk (*).

Training approach	κ-score	# pseudo-labels
Gleason grading: TMAZ dataset		
Fully-supervised	0.5667 ± 0.0285	-
Training variant I	0.5062 ± 0.0487	9’000
Training variant II	0.6067 ± 0.0152*	8’000
Training variant III	**0.6104 ± 0.0158***	5’000
Gleason scoring: TMAZ dataset		
Fully-supervised	0.7186 ± 0.0306	-
Training variant I	0.6517 ± 0.0437	9’000
Training variant II	0.7516 ± 0.0274*	7’000
Training variant III	**0.7588 ± 0.0192***	7’000
Gleason scoring: TCGA-PRAD dataset		
Fully-supervised	0.2293 ± 0.1350	-
Training variant I	0.3593 ± 0.0496*	8’000
Training variant II	**0.4121 ± 0.0963***	5’000
Training variant III	0.4065 ± 0.0725*	9’000

scoring. Several approaches were developed: fully-supervised learning approaches (Arvaniti et al., 2018; Ström et al., 2019; Nagpal et al., 2019), weakly supervised (del Toro et al., 2017; Arvaniti and Claassen, 2018; Otálora et al., 2020b; Campanella et al., 2019) and semi-supervised approaches (Bulten et al., 2020; Tolkach et al., 2020). Fully-supervised learning shows that CNNs need large datasets with local annotations to be trained and to guarantee high performance. Unfortunately, large datasets usually come without local annotations, since producing them requires a large amount of work. Furthermore, it is hard to collect locally-annotated heterogeneous data, as datasets often originate from a single medical source (e.g. Arvaniti et al. (2018); Ström et al. (2019); Nagpal et al. (2019)). Weakly-supervised learning approaches aim to reduce the dependency from local annotations, using global annotations as weak labels. However, the weak labels introduce noise in training, and therefore a large amount of data is needed to reach high performance, as shown in Campanella et al. (2019). In del Toro et al. (2017); Arvaniti and Claassen (2018); Otálora et al. (2020b) the authors applied weakly-supervised learning using subsets of a publicly available dataset (TCGA-PRAD). In order to achieve comparable performance, the experiments required a number of pixels that are 500 times larger than the locally annotated dataset (TMAZ). Semi-supervised learning approaches aim to alleviate the lack of large datasets with local annotations and train models that generalize better to heterogeneous datasets. The published literature that builds on top of the semi-supervised techniques for histopathology (particularly in prostate cancer) using the teacher/student paradigm shows an appreciable level of technical sophistication, such as the work by Cheng et al. (2020); Shaw et al. (2020); Bulten et al. (2020); Tolkach et al. (2020). There is literature that builds on top of the semi-supervised techniques for histopathology using the teacher/student paradigm, namely the work of Cheng et al. (2020); Shaw et al. (2020); Bulten et al. (2020); Tolkach et al. (2020). Here, we want to discuss significant similarities and missing points from these papers, as well as in our proposed method. In Cheng et al. (2020), the authors train a teacher/student model for a segmentation task, exploiting partially labeled WSIs using spatially related patches to filter out noisy model predictions. The authors show that their strategy improves the fully-supervised baseline. The CNN is evaluated on a homogeneous external prostate dataset, showing an improvement in segmentation performance. A homogeneous dataset does not account for the realistic spectra of variations in clinical scenarios, where WSIs can be scanned with different scanners and have heterogeneous visual appearance (as we have previously shown on TCGA-PRAD). In the work of Shaw et al. (2020) the authors train a chain of teacher-student models to annotate image regions of colorectal cancer. Their model showed an impressive sample complexity by using only a small fraction (less than 1%) of the annotated data to achieve comparable performance to the fully-supervised model with 100% of the labels. Nevertheless, their experimental setup lacks validation on unseen centers. An interesting possible validation for future work is to perform teacher/student model training using only one center and comparing it with the reported inter-center performance. In Bulten et al. (2020), the authors train a U-net based network to generate and to pseudo-label prostate cancer regions. They use pure Gleason Pattern WSIs (where the primary and the secondary Gleason patterns coincide) to reduce the noisy labels and avoid annotations with the wrong patterns. The model reaches very high performance in the internal test set and pathologist level performance on the TMAZ dataset (used as external test partition). However, this work suffers of the same experimental problem shown in the other work, since TMAZ originates from one single medical source: the model is not validated on a heterogeneous external dataset. In Tolkach et al. (2020), the authors train a model to classify Gleason patterns using a large dataset with pixel-wise annotated patches. The model is used to annotate unseen data from complex patterns (where primary Gleason pattern differs from secondary Gleason pattern) and to fine-tune the model. The model is tested on both an internal test partition and on a TCGA-PRAD cohort. Even though the model is tested on a heterogeneous external dataset it does not exploit data heterogeneity during the training, since it is used to pseudo-annotate images from the same medical center. Despite the noticeable differences in the individual technical components of each presented teacher/student approach, it is noteworthy that in Cheng et al. (2020), Shaw et al. (2020)), the student models always outperform the fully-supervised baseline. Furthermore, in the work of Shaw et al. (2020), the authors show that when using 100% of the labeled data, the student's performance is inferior to the one obtained with fewer labels. This result suggests that there can be a substantial amount of noise in the full dataset and that the student can distil a right amount of labels to reduce its inclusion in the model.

The semi-supervised and semi-weakly supervised approaches presented in this paper show better performance than a fully-supervised learning approach and other published ap-

Table 8. **Comparison between Arvaniti et al. (2018), Bulten et al. (2020) and this work on TMAZ dataset, for both Gleason grading and Gleason scoring. The comparison involves the performance of best CNN trained and the average of ten CNNs (if reported).**

Work	Best result	Average result
Gleason grading: (inter-pathologist agreement κ = 0.67)		
Arvaniti et al. (2018)	$\kappa = 0.55$	not reported
This work	κ **= 0.6613**	κ **= 0.6127 ± 0.0133**
Gleason scoring: (inter-pathologist agreement κ = 0.71)		
Arvaniti et al. (2018)	$\kappa = 0.75$	not reported
Bulten et al. (2020)	$\kappa = 0.72$	not reported
This work	κ **= 0.8124**	κ **= 0.7645 ± 0.0231**

proaches (Arvaniti et al., 2018; Bulten et al., 2020) tested on the TMAZ dataset. In Gleason grading, the fully-supervised approach obtains κ= 0.5667 ± 0.0285, while the semi-supervised approach obtains κ=0.6127 ± 0.0133 and the semi-weakly supervised approach κ=0.6104 ± 0.0158. In the Gleason scoring task, evaluated on TMAZ, the fully-supervised approach obtains κ= 0.7186 ± 0.0306, while the semi-supervised approach obtains κ=0.7645 ± 0.0231 and the semi-weakly supervised approach κ=0.7588 ± 0.0192. In the Gleason scoring task, evaluated on the TCGA-PRAD, the fully-supervised approach obtains κ= 0.2293 ± 0.1350 semi-supervised approach obtains κ=0.4529 ± 0.0512 and the semi-weakly supervised approach κ=0.4121 ± 0.0963.

The models trained with the semi-supervised learning approach outperform the models proposed by Arvaniti et al. (2018) and Bulten et al. (2020) in both Gleason grading and Gleason scoring, evaluated on the TMAZ test partition, as reported in Table 8.

The performance obtained with the semi-supervised approach are slightly higher than the one of the semi-weakly supervised approach, considering the peak values in Gleason grading and scoring reported in Table 6 and 7. The difference can be explained with the pseudo-labeled patches annotated by the teacher model. In the semi-supervised learning approach, the teacher model provides pseudo-labeled patches that are less noisy than the patches of the other approach. This is particularly true considering the performance of the models trained with the training variant I (only pseudo-labeled patches) in Gleason scoring, evaluated on TCGA-PRAD: in the semi-supervised approach, the models reach κ=0.4529 ± 0.0512 as peak value, while in semi-weakly supervised approach, the models reach κ=0.3593 ± 0.0496. The patches provided by the teacher are a consequence of the training schema used. The teacher is trained only with strongly-annotated data in the semi-supervised learning approach, while otherwise it is first pre-trained with weakly-annotated data and then it is fine-tuned with strongly-annotated data. Therefore, the weakly-annotated dataset used for pre-training the teacher model in the semi-weakly supervised learning approach can lead to noisy patches. The classification performance of the models depends on the amount of pseudo-labeled data used for training. A trade-off between the two is identified. As shown in Section 2.3.2, one of the most critical parameters for the paradigm is the number of pseudo-labeled patches (K parameter) used for training the student model. As expected, including the pseudo-labeled data improves the performance, since the dataset is larger and it includes data from different sources. However, the performance is not monotonically increasing, considering the amount of pseudo-labeled data used for training the student model. Since pseudo-labeled data can be noisy, the more pseudo-labeled data are collected the higher is the noise in the training data. When data are too noisy, the performance can also decrease. This is the case for the work presented: for each of the student variants tested a peak value in κ is identified. For each of the models, the peak value corresponds to the subset of pseudo-labeled data where the noisy labels are less represented than in the other subsets.

The teacher/student approaches allow alleviating the inter-dataset heterogeneity, mitigating the overfitting and allowing the models to generalize on datasets from several sources. The inter-dataset heterogeneity leads to models that adapt their weights on the data used for training them. The overfitting lowers the model performance when it is trained on a dataset and is evaluated on another one (from a different source). It happens for both the TMAZ dataset (fully-supervised learning approach of the student) and the TCGA-PRAD dataset (student training variant I). The overfitting influences less strongly the student training variant II and it is limited adopting the student training variant III. The student model trained only with TMAZ patches (fully-supervised learning approach) obtains good results when it is tested on the test partition of the same dataset (Figure 4.A,B and Figure 5.A,B, dashed line) but it fails to generalize on the TCGA-PRAD dataset (Figure 4.C and Figure 5.C, dashed line). The student model trained only with TMAZ patches and tested on the TCGA-PRAD dataset reaches close to the lowest performance. The student model trained only with pseudo-labeled data (student training variant I) obtains excellent performance when it is tested on an internal test partition (Figures 4.C and 5.C, blue curve) but it fails to generalize on the TMAZ dataset (Figures 4.A,B and 5.A,B, blue curve). The same thing happens for both the teacher/student paradigm approaches. The student model trained only with pseudo-labeled patches and tested on the TMAZ dataset has the lowest performance. The inter-dataset heterogeneity leads to overfitting also for the student training variant II but in this case it is less strong than the previous examples. In this student training variant, the student model is trained in two steps. First, it is pre-trained with pseudo-labeled data and then it is fine-tuned with strongly-annotated data. The model first adapts its weights on the pseudo-labeled data and then it adapts its weights on strongly-annotated data. For the TMAZ dataset the student training variant II obtains among the best performance when it is tested on its test partition (Figures 4.A,B and 5.A, green curve) but it fails to generalize well on the TCGA-PRAD dataset (Figures 4.C, 5.C, green curve). Considering both approaches, the model trained with this student training variant and tested on the TCGA-PRAD dataset has a better performance than the fully-supervised learning approach. However, in the semi-supervised learning approach, it is exceeded by 0.096 in κ, compared with the same model trained using the student training variant I (only with pseudo-labeled data). The overfitting is limited using the student training variant III. In this student training variant, the model is trained combining both the pseudo-labeled and the strongly-annotated data. In this case, the performance is good both for the Gleason score and Gleason pattern classification (Figures 4.A,B,C, 5.A,B,C, magenta curves). In Gleason pattern classification, this student training variant reaches the best result in the semi-weakly supervised learning approach (Figure 5.A, magenta curve) and a performance similar to the best in the semi-supervised learning approach (Figure 4.A, magenta curve). The same models generalize well in the TCGA-PRAD dataset (Figures 4.C and 5.C, magenta curves). In both approaches, the results are similar to the ones obtained with the student training variant II regarding the evaluation at patch level, but much better regarding the evaluation at the WSI level.

The teacher/student approaches allow overcoming the lack of large datasets with locally-annotated data. Locally-annotated data are needed to train CNNs but only small datasets with local annotations are available. The approaches exploit large amounts of unlabeled or weakly-annotated data that are publicly available, automatically annotating them with pseudo-labels. Figure 7 shows a set of pseudo-labeled patches from the teacher model, trained with the semi-supervised learning approach. For each class, the upper row includes some of the top-ranked patches, while the bottom row includes a few of the patches with the lowest probabilities. The top-ranked patches match the corresponding tissue morphology.

The methods shown in the paper can be still improved, even though they reach high performance and outperform models from other approaches evaluated on the same datasets. Figure 6 shows the best CNN trained for Gleason grading and scoring (both on TMAZ and TCGA-PRAD). The matrices show that the models have poorer performance in distinguishing between benign and Gleason pattern 3 when evaluated on TMAZ at patch level (Figure 6.A), in distinguishing between benign and Gleason score 6 when evaluated in TMAZ at core level (Figure 6.B) and in predicting high risk Gleason scores on TCGA-PRAD (Figure 6.C). Figures 4 and Figure 5 show that the models sometimes have large confidence intervals and unstable performance, especially when evaluated in TCGA-PRAD. The large confidence intervals can occur due to the aggregation method used to evaluate the primary and the secondary Gleason patterns and to the heterogeneous datasets used to train and test the models. The aggregation is made with a majority voting method and it involves patches that may include large portions of stroma (since the patches are the top-ranked within a WSI according BR). Despite stroma and healthy tissue representing most of the tissue within an image, they are not very important for the diagnosis process, since the diagnosis is made considering cancer tissue that occupies only a smaller portion of the images. Therefore, the aggregation algorithm considers also portions of stroma in the diagnosis that a pathologist would discard. The models that show the largest confidence intervals and most unstable performance are the ones trained with data from a dataset and evaluated on another dataset. It can be explained considering that the datasets used are heterogeneous and that the models tend to adapt their weights on the training data. In Figure 4.A,B and Figure 5.A,B the models trained with variant I (only pseudo-labeled patches from TCGA-PRAD) have the largest confidence interval, since they are evaluated on TMAZ. The models trained with variant II and III include patches from TMAZ in the training partitions, therefore the performance are more stable and have smaller confidence intervals when the models are evaluated in TMAZ. In Figure 4.C and Figure 5.C the models trained with variant II (pre-training with pseudo-labeled patches from TCGA-PRAD and fine-tuning with strongly-annotated patches from TMAZ) have the largest confidence interval and unstable performance, since they are evaluated on TCGA-PRAD.

5. Conclusion

Training classification models that generalize on several datasets raises questions when data are highly-heterogeneous and only limited amounts of locally-annotated data are available. In this paper, two teacher/student approaches (semi-supervised learning and semi-weakly supervised learning) allow to train CNN models that generalize on datasets from varied sources. The approach allows to improve the performance of the CNN, to face inter-dataset heterogeneity and to overcome the lack of large datasets with local annotations. The approaches are evaluated for Gleason pattern and Gleason score classification. They are compared with a fully-supervised learning approach and other approaches. In both cases, the models trained with the teacher/student paradigm improve their performance compared with a fully-supervised learning used to train the models. In particular, the models trained using the semi-supervised approach show the best performance for both Gleason pattern and Gleason score classification. The teacher/student paradigm allows to face the inter-dataset heterogeneity and to limit the overfitting during the training of the models. The models trained with the approach generalize on both datasets used in this paper, the TMAZ and the TCGA-PRAD. The approach allows to overcome the lack of large locally-annotated datasets to train the models, exploiting a large amount of unlabeled and/or weakly-annotated data to automatically generate pseudo-labeled data. In future work, the teacher/student paradigm will be applied to different types of tissues and different K values will be tested. K values, larger than the ones presented in this paper, will be tested in the future to make it possible to learn with a large amount of weakly-annotated or unlabeled data, building more robust and more accurate CNNs. The code and all the pre-trained models will be made publicly available on Github (https://github.com/ilmaro8/Semi_Supervised_Learning) upon paper publication. The pseudo-labeled data are available from the corresponding author on request.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825292 (ExaMode, `htttp://www.examode.eu/`). Infrastructure from the SURFsara HPC centre was used

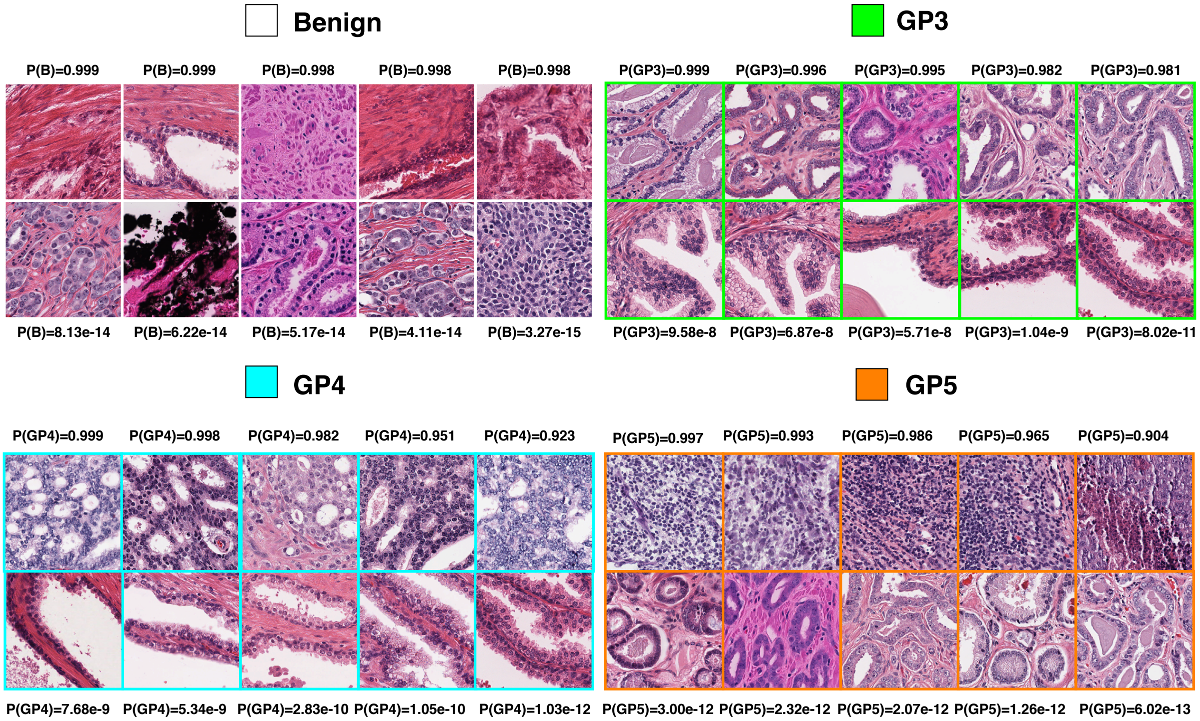


Fig. 7. A few examples of the pseudo-labeled patches, annotated by the teacher model. For each of the classes, the upper row includes the top-ranked patches, while the bottom row includes the lowest probabilities. For each of the patches the probability of belonging to the class is shown. The Xe-Y is an abbreviation for for $X \times 10^{-Y}$.

to train the CNN models in parallel. The authors also thank Nvidia for the Titan Xp GPUs used for some of the weakly supervised experiments. Sebastian Otálora thanks to the Colombian science ministry Minciencias for partially funding his Ph.D. studies through the call ”756 - Doctorados en el exterior”.

References

Arvaniti, E., Claassen, M., 2018. Coupling weak and strong supervision for classification of prostate cancer histopathology images. Medical Imaging meets NIPS Workshop, NIPS 2018 .

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports 8.

Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning, in: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 17–36.

Berg, K.D., Toft, B.G., Roder, M.A., Brasso, K., Vainer, B., Iversen, P., 2011. Prostate needle biopsies: interobserver variation and clinical consequences of histopathological re-evaluation. Apmis 119, 239–246.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology .

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. Information 11, 125.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine 25, 1301–1309.

Chang, H., Loss, L.A., Parvin, B., 2012. Nuclear segmentation in h&e sections via multi-reference graph cut (mrgc), in: International symposium biomedical imaging.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q., 2019. Data-free learning of student networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3514–3522.

Chen, N., Zhou, Q., 2016. The evolving gleason grading system. Chinese Journal of Cancer Research 28, 58.

Cheng, H.T., Yeh, C.F., Kuo, P.C., Wei, A., Liu, K.C., Ko, M.C., Chao, K.H., Peng, Y.C., Liu, T.L., 2020. Self-similarity student for partial label histopathology image segmentation, in: European Conference on Computer Vision, Springer. pp. 117–132.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54, 280–296.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData mining 10, 35.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A., 2016. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. The American journal of surgical pathology 40, 244–252.

Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R., 2008. Hematoxylin and eosin staining of tissue and cell sections. Cold spring harbor protocols 2008, pdb–prot4986.

Foucart, A., Debeir, O., Decaestecker, C., 2019. Snow: Semi-supervised, noisy and/or weak data for deep learning in digital pathology, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 1869–1872.

Grönberg, H., Adolfsson, J., Aly, M., Nordström, T., Wiklund, P., Brandberg, Y., Thompson, J., Wiklund, F., Lindberg, J., Clements, M., et al., 2015. Prostate cancer screening in men aged 50–69 years (sthlm3): a prospective population-based diagnostic study. The lancet oncology 16, 1667–1676.

Guo, T., Xu, C., He, S., Shi, B., Xu, C., Tao, D., 2019. Robust student network learning. IEEE transactions on neural networks and learning systems .

Hady, M.F.A., Schwenker, F., 2013. Semi-supervised learning, in: Handbook on Neural Information Processing. Springer, pp. 215–239.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M., 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels, in: Advances in neural information processing systems, pp. 8527–8537.

Huang, G., Liu, Z., Weinberger, K.Q., 2016. Densely connected convolutional networks. CoRR abs/1608.06993. URL: `http://arxiv.org/abs/1608.06993`, `arXiv:1608.06993`.

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A., 2019. Histoqc: an open-source quality control tool for digital pathology slides. JCO clinical cancer informatics 3, 1–7.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2019. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. arXiv preprint arXiv:1912.02911 .

Katharopoulos, A., Fleuret, F., 2019. Processing megapixel images with deep attention-sampling models, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA. pp. 3282–3291. URL: `http://proceedings.mlr.press/v97/katharopoulos19a.html`.

Komura, D., Ishikawa, S., 2018. Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal 16, 34–42.

van der Laak, J., Ciompi, F., Litjens, G., 2019. No pixel-level annotations needed. Nature Biomedical Engineering , 1–2.

Lee, D.H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on challenges in representation learning, ICML, p. 2.

Li, J., Li, W., Gertych, A., Knudsen, B.S., Speier, W., Arnold, C.W., 2019. An attention-based multi-resolution model for prostate whole slide image classification and localization. Medical Computer Vision Workshop - CVPR 2019-32 .

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al., 2018. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 7, giy065.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Madabhushi, A., Feldman, M.D., Leo, P., 2020. Deep-learning approaches for gleason grading of prostate biopsies. The Lancet Oncology 21, 187–189.

Marini, N., Otálora, S., Müller, H., Atzori, M., 2020. Semi-supervised learning with a teacher-student paradigm for histopathology classification: A resource to face data heterogeneity and lack of local annotations., in: ICPR Workshops (1), pp. 105–119.

Montironi, R., Mazzuccheli, R., Scarpelli, M., Lopez-Beltran, A., Fellegara, G., Algaba, F., 2005. Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. BJU international 95, 1146–1152.

Mormont, R., Geurts, P., Marée, R., 2018. Comparison of deep transfer learning strategies for digital pathology, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2262–2271.

Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al., 2019. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. NPJ digital medicine 2, 1–10.

Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A., 2013. Learning with noisy labels, in: Advances in neural information processing systems, pp. 1196–1204.

Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., Müller, H., 2019. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. Frontiers in Bioengineering and Biotechnology 7, 198.

Otálora, S., Marini, N., Müller, H., Atzori, M., 2020a. Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks, in: Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer, pp. 193–203.

Otálora, S., Marini, N., Müller, H., Atzori, M., 2021. Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification. BMC Medical Imaging 21, 1–14.

Otálora, S., Perdomo, O., González, F., Müller, H., 2017. Training deep convolutional neural networks with active learning for exudate classification in eye fundus images, in: Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 146–154.

Otálora, S., del Toro, O.J., Atzori, M., Khan, A., Andrearczyk, V., Müller, H., 2020b. A systematic comparison of deep learning strategies for weakly supervised gleason grading, in: Medical Imaging 2020: Digital Pathology.To appear, International Society for Optics and Photonics.

Pierorazio, P.M., Walsh, P.C., Partin, A.W., Epstein, J.I., 2013. Prognostic g leason grade grouping: data based on the modified g leason scoring system. BJU international 111, 753–760.

Prior, F.W., Clark, K., Commean, P., Freymann, J., Jaffe, C., Kirby, J., Moore, S., Smith, K., Tarbox, L., Vendt, B., et al., 2013. Tcia: an information resource to enable open science, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 1282–1285.

Schulte, E., 1991. Standardization of biological dyes and stains: pitfalls and possibilities. Histochemistry 95, 319–328.

Settles, B., 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Settles, B., 2011. From theories to queries: Active learning in practice, in: Active Learning and Experimental Design workshop In conjunction with AISTATS 2010, pp. 1–18.

Shaw, S., Pajak, M., Lisowska, A., Tsaftaris, S., ONeil, A., 2020. Teacher-student chain for efficient semi-supervised histology image classification, in: Proceedings of the ICLR 2020 conference. Workshop on AI for Affordable Healthcare, pp. 7340–7351.

Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al., 2019. Pathologist-level grading of prostate biopsies with artificial intelligence. arXiv preprint arXiv:1907.01368 .

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis 58, 101544.

Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L., 2015. The new data and new challenges in multimedia research. CoRR abs/1503.01817. URL: `http://arxiv.org/abs/1503.01817`, `arXiv:1503.01817`.

Titford, M., 2005. The long history of hematoxylin. Biotechnic & histochemistry 80, 73–78.

Tolkach, Y., Dohmgörgen, T., Toma, M., Kristiansen, G., 2020. High-accuracy prostate cancer pathology using deep learning. Nature Machine Intelligence 2, 411–418.

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary oncology 19, A68.

Jimenez-del Toro, O., Otálora, S., Atzori, M., Müller, H., 2017. Deep multimodal case–based retrieval for large histopathology datasets, in: International Workshop on Patch-based Techniques in Medical Imaging, Springer. pp. 149–157.

del Toro, O.J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., Müller, H., 2017. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score, in: Medical Imaging 2017: Digital Pathology, International Society for Optics and Photonics. p. 101400O.

Veeraraghavan, H., Miller, J.V., 2011. Active learning guided interactions for consistent image segmentation with reduced user interactions, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE. pp. 1645–1648.

Wilcoxon, F., 1992. Individual comparisons by ranking methods, in: Breakthroughs in statistics. Springer, pp. 196–202.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.

Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D., 2019. Billionscale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 .

Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J., 2017. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7340–7351.