Semi-supervised learning with a teacher-student paradigm for histopathology classification: a resource to face data heterogeneity and lack of local annotations

Niccolò Marini*[*†], Sebastian Otálora*[*†], Henning Müller*[‡], Manfredo Atzori*
* Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais)
Technopôle 3, 3960 Sierre, Switzerland
{niccolo.marini, juan.otaloramontenegro, henning.mueller, manfredo.atzori}@hevs.ch
† Centre Universitaire d'Informatique, University of Geneva, 1227 Carouge, Switzerland
‡ Medical Faculty, University of Geneva, 1211 Geneva, Switzerland

Abstract*—Training classification models in the medical domain is often difficult due to data heterogeneity (related to acquisition devices and protocols) and due to the difficulty of getting sufficient amounts of annotations from specialized medical doctors. It is particularly true in digital pathology, where models do not generalize easily. This paper presents a novel approach for the generalization of models in conditions where heterogeneity is high and annotations are few, based on the application of a teacher/student approach to different datasets and annotations. The approach relies on a semi-supervised teacher/student paradigm. The paradigm combines a small amount of strongly-annotated data (tissue microarrays), with a large amount of unlabeled data from whole slide images, for training a Convolutional Neural Networks (CNN). Two CNNs are involved: the teacher and the student model. The teacher model is trained with strong labels and used to generate pseudo-labeled samples from the unlabeled data. The student model is trained with the pseudo-labeled samples and then fine-tuned with a small amount of strongly-annotated data. The paradigm is evaluated on the student model performance of Gleason pattern and Gleason score classification in prostate cancer images. The paradigm is compared with a fully-supervised learning approach for training the student model. In order to evaluate the capability of the approach to generalize, the datasets used for the evaluation are highly heterogeneous in visual characteristics and are collected from different medical institutions. The models, trained with the teacher/student paradigm, show an improvement in performance above the fully-supervised training. The models generalize better on both the datasets, despite the inter-datasets heterogeneity, alleviating the overfitting. The classification performance shows an improvement both in the classification of Gleason pattern at patch level ($\kappa = 0.6129 \pm 0.0127$ from $\kappa = 0.5608 \pm 0.0308$) and at in Gleason score classification, evaluated at WSI-level $\kappa = 0.4477 \pm 0.0460$ from $\kappa = 0.2814 \pm 0.1312$).

***Index Terms*—Digital Pathology, Deep Learning, Semi-Supervision, Prostate Cancer**

I. Introduction

The lack of large datasets with local annotations and the highly-heterogeneous data represent a critical challenge for developing machine learning algorithms that generalize well in the digital pathology domain [1], despite the increasing amount of datasets available with repositories such as TCGA (The Cancer Genome Atlas).

Machine learning algorithms, particularly Convolutional Neural Networks (CNNs), are the state-of-the-art for analyzing digital pathology images [2], [3] (such as, for instance, whole slide images, WSIs, or tissue-micro-arrays, TMAs). Convolutional neural network models usually require large datasets with local annotations to train robust models [4] that generalize well to unseen data [5]. The annotation of the digital pathology images is a time-consuming and expensive process that requires medical experts, such as the pathologists. Therefore, only a small amount among the publicly available datasets is locally annotated, e.g. the Camelyon dataset [6].

Despite the small number of datasets that are locally annotated, an increasing number of datasets with histopathological images is available, e.g. The Cancer Genome Atlas (TCGA)[1]. Most of these datasets come without local annotations (strong annotations) of the region of interest for the diagnosis. Some of these datasets are released with medical reports and some are unlabeled. The reports include the final diagnostic, among other information, that can instead be used as weak annotations for digital pathology images.

The amount of strongly-annotated data is much smaller than the unlabeled and the weakly-annotated data. This fact constitutes a challenge for training supervised CNN models in a fully-supervised fashion.

Furthermore, histopathological images that come from different sources are highly-heterogeneous. The staining procedure applied to the samples and the variability in the tissue structures cause the heterogeneity. Hematoxylin and eosin (H&E) represent the golden standard for staining the samples within a WSI [7]. Although H&E is a standard, their preparation procedures are not fully standardized, often leading to inter-dataset heterogeneity [8], [9]. This heterogeneity leads

[1]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Retrieved 9th of March, 2020

to models that are more prone to overfitting, compared with models trained in conditions where there is no heterogeneity between different datasets. Therefore, many CNN models, trained to analyze histopathological images, face a decrease in their performance when they are tested on data originated from a different source, as shown in previous works [10], [11].

Despite the lack of large datasets that are locally annotated and the highly-heterogeneous data, new methods were proposed recently for training the models with small datasets of local annotations, showing partial success, such as semi-supervised learning learning [12]–[20], active learning [21]–[26] and weakly supervised learning [5], [27]–[34].

This paper represents a novelty in a domain where there is a lack of large datasets with local annotations and the data are highly heterogeneous. The semi-supervised teacher/student paradigm is applied to the digital pathology task of prostate cancer classification, using two datasets.

Prostate cancer (PCa) is one of the most common cancer for worldwide healthcare systems[2] and is diagnosed through the Gleason grading system[3]. Prostate cancer is the fourth most frequent cancer in the entire human population[4]. Prostate cancer is diagnosed using the Gleason grading system, which is based on two steps: first, the identification of Gleason patterns, second the computation of the Gleason Score. The identification of Gleason patterns is made to estimate the aggressiveness of cancer. The tissue structures in a sample are distinguished in different Gleason patterns, according to their cell abnormality and their gland deformation. The Gleason patterns range from 1 to 5. According to the guidelines described by the Union for International Cancer Control and the World Health Organization/International Society of Urological Pathology, the Gleason score is computed by evaluating the most diffused primary and secondary patterns. Typically, malignant prostate cancer has a Gleason score ranged from 6 to 10. The recent advancements in the digital pathology cancer prostate classification task are summarized in the Table I.

In this paper, two highly-heterogeneous datasets are used for training the models: a small strongly-labeled dataset with pixel-wise annotations and a large unlabeled dataset of whole slide images. The strongly-annotated dataset is the Tissue Micro-Arrays Zurich dataset (TMAZ). The non locally annotated dataset is a cohort of The Cancer Genome Atlas PRostate ADenocarcinoma (TCGA-PRAD).

The approach proposed follows the teacher/student paradigm and consists of two models: a high-capacity model, called *teacher model*, and a smaller model, called the *student model*. The teacher model generates pseudo-labeled examples from the unlabeled data. The student model is trained combining the pseudo-labeled examples and the strongly-annotated data.

The teacher and the student models are implemented using large pre-trained models and following the paradigm constraints. The teacher model must be a high-capacity model, while the student model must be efficient at test time. The teacher model is a high-capacity ResNexT based model (22 million of parameters), pre-trained with a dataset of one billion natural images retrieved from Instagram [19]. The model is trained with the strongly-annotated data and it creates the pseudo-labeled examples annotating the unlabeled data. The student model is a DenseNet121, pre-trained with ImageNet weights. The student architecture is a small model, compared with the model used for implementing the teacher. The model is trained first with the pseudo-labeled data and then fine-tuned with the strongly-annotated data.

The models' performance is compared with the fully-supervised learning of the student model, considered as the baseline. The teacher/student paradigm, as shown in the experimental results, performs better than the fully-supervised CNN (trained only with strongly-annotated data), both at the Gleason pattern level and at the Gleason score level. The approach allows leveraging large unlabeled datasets as a source of supervision for training CNN models in digital pathology.

II. Methods

A. Datasets

Two open-access datasets are adopted for the evaluation of the teacher/student paradigm. They are highly heterogeneous and they are pre-processed with the same approach.

The two datasets selected are heterogeneous, which makes them similar to real clinical classification problems. In both datasets, the images are pre-processed dividing them into patches and removing the background regions. The images are divided into tiles of 750x750 pixels, and then they are resized to 224x224 pixels to fit as input to the chosen networks. Only the patches extracted from tissue regions are selected (background regions are non-informative). The HistoQC tool [36] is used for generating tissue masks of the images so that only patches that include tissue are extracted.

The two datasets are the tissue microarray dataset (TMAZ) released by Arvanity et al. [35] and a cohort of the TCGA-PRAD dataset[5].

The TMAZ includes 886 prostate TMA core images with pixel-wise annotations, made by pathologists. Each TMA core has a size of 3100^2 pixels, scanned at 40x resolution (0.23 microns per pixel). The arrays are scanned at the same medical center, the University Hospital of Zurich (NanoZoomer-XR Digital slide scanner, Hamamatsu). The TMAZ dataset includes four classes: benign, Gleason pattern 3, Gleason pattern 4, Gleason pattern 5. It is split into three partitions: the training partition is composed of 508 cores, the validation partition is

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

[3]https://www.pcf.org/about-prostate-cancer/diagnosis-staging-prostate-cancer/gleason-score-isup-grade/, Retrieved 16th of March, 2020

[4]https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data. Retrieved 16th of March, 2020

[5]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Retrieved 9th of March, 2020

TABLE I
STATE-OF-THE-ART WORKS FOR GLEASON PATTERNS AND GLEASON SCORING DEEP LEARNING MODELS.

Reference	Classes	Results	Dataset	Annotations
Arvaniti [35]	Benign,GP3,GP4,GP5	$\kappa = 0.53$	886 TMAs	Strong
Ström [34]	GP1, GP2, GP3, GP4, GP5	$\kappa = 0.67$	6682 WSIs	Strong
Ström [34]	Benign vs malignant cancer	AOC = 0.997	6682 WSIs	Strong
This work	Benign,GP3,GP4,GP5	$\kappa = 0.61$	886 TMAs + 341 WSIs	Strong + Weak
Arvaniti [35]	GS6,GS7=3+4,GS7=4+3,GS=8,GS=9-10	$\kappa = 0.75$	886 TMAs	Strong
Arvaniti [29]	GS6,GS7,GS8,GS9,GS10	AUC = 0.882	886 TMAs + 447 WSIs	Strong + Weak
Jimenez-del-Toro [28]	[GS6,GS7] vs [GS8,GS9,GS10]	ACC = 0.78	235 WSIs	Weak
Otálora [33]	GS6,GS7=3+4,GS7=4+3,GS=8,GS=9-10	$\kappa = 0.44$	341 WSIs	Weak
Bulten [27]	GS6,GS7=3+4,GS7=4+3,GS=8,GS=9-10	$\kappa = 0.72$	1243 WSIs	Strong + Weak
Campanella [5]	Benign vs Cancer	AUC = 0.986	24859 WSIs	Weak
This work	GS6,GS7=3+4,GS7=4+3,GS=8,GS=9-10	$\kappa = 0.44$	886 TMAs + 341 WSIs	Strong + Weak

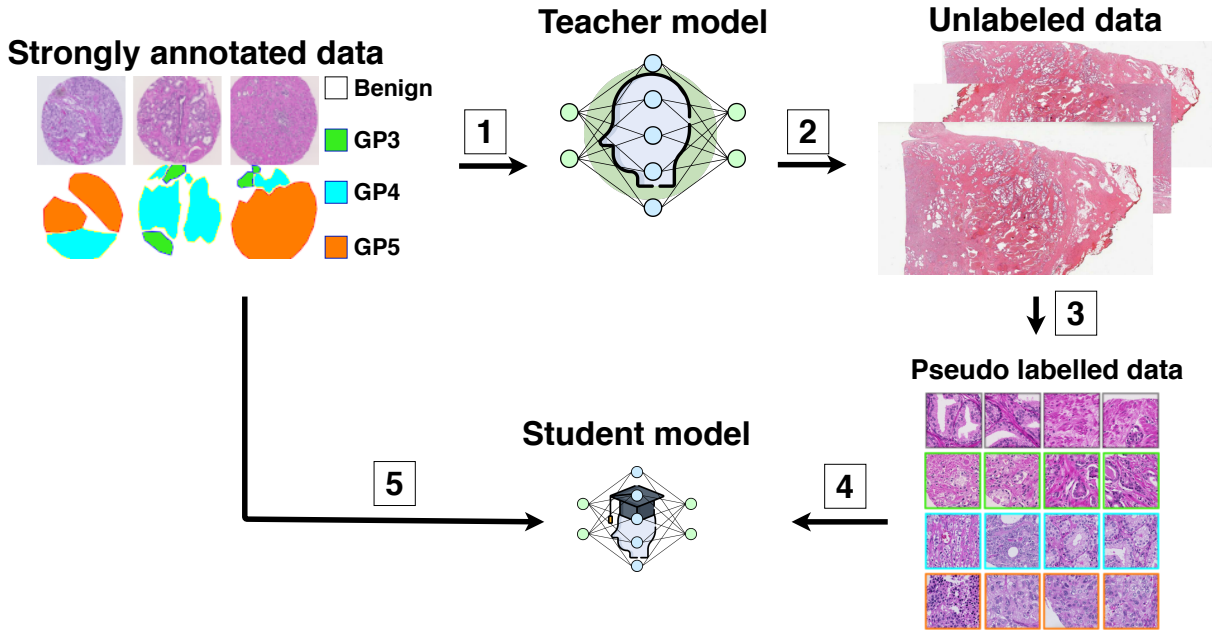

Fig. 1. Overview of the teacher/student training model. In the step one, the teacher is trained with strongly-annotated data. In the step two, the teacher predicts the class probabilities for the unlabeled data. In the step three, the samples with the highest probabilities are selected (pseudo-labeled data). In the step four, the student model is trained using the pseudo-labeled data. In the step five, the student model is trained using the strongly-annotated data.

composed of 133 cores, and the test partition of 245 cores. The partitions of the dataset are shown in table II. From each TMAZ core, 30 patches are randomly extracted. The number of patches to extract is chosen considering the trade-off between the patch size and the whole tissue covered within the TMA. The number of patches for each class is summarized in Table III.

TCGA-PRAD[6] is a data repository of digitized radical prostactectomies (made up of $100'000^2$ pixels) with no pixel-wise annotations. The cohort of the TCGA-PRAD dataset includes 301 WSIs, paired with their primary and secondary Gleason pattern within the corresponding pathology report. The WSIs in the cohort are collected from 20 medical centers. This large number of medical centers leads to a highly heterogeneous visual content of the WSIs. The dataset is split into three partitions (as shown in Table IV): the training set is composed of 171 WSIs, the validation set composed of 84 WSI, and the test set composed of 46 WSIs. In this paper, the TCGA-PRAD patches are annotated with pseudo-labels

[6]https://portal.gdc.cancer.gov/projects/TCGA-PRAD. Retrieved March 1, 2020

TABLE II
NUMBER OF TMA CORES FOR EACH GLEASON SCORE IN THE TMAZ DATASET.

Class/Partition	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 III
NUMBER OF PATCHES FOR EACH GLEASON PATTERN IN THE TMAZ DATASET.

Class/Partition	Training	Validation	Test
Benign	1830	1260	127
GP3	5992	1352	1602
GP4	4472	831	2121
GP5	2766	457	387
Total	15060	3900	4237

TABLE IV
NUMBER OF WSIS FOR EACH GLEASON SCORE IN THE TCGA-PRAD DATASET.

Class/Partition	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

by the teacher model. It predicts a probability vector for each of the patches within the WSIs. The probability vectors are sorted in descending order by the class probabilities and the top-ranked P patches are selected. Different values of P are tested for the training partitions of pseudo-labeled data. They vary between 1000 and 10'000 patches per class. They are explored increasing the value of 1000 patches per class, between two consecutive P values. Therefore 1000 patches per class are included in the first subset and 2000 per class in the second one. The validation and test partition include both 8000 patches (2000 samples for each class).

B. Teacher/Student paradigm

The presented semi-supervised learning approach is a pipeline based on teacher/student paradigm [20], [37]. Figure 1 shows an overview of the training schema. The paradigm includes two distinct CNNs, called respectively the teacher model and the student model.

The teacher model is a high-capacity neural network, trained to annotate pseudo-labeled examples from the unlabeled data. The pseudo-labels are the labels predicted by a model, in this case, the teacher model [20]. They are assigned considering the prediction vector and selecting the class with the maximum predicted probability. They are used as they were labels made by an expert [20]. For a subset of these labels, the assigned ground truth matches with the correct class (relevant label), while for the other subset, the ground truth does not match with the correct class (noisy labels) [17], [19]. Noisy labels can compromise the learning process [17]. The choice to use high-capacity models permits to better separate noisy labels from correct labels [17]. Furthermore, high-capacity models can leverage the large amount of data better [19]. The teacher model annotates unlabeled data with pseudo-labels that are used for training the student model. The annotation process is made predicting the class probabilities of unlabeled data [20]. The relevant samples are labeled with the highest probabilities for separating them from noisy examples.

The student model is a smaller (compared to the teacher) neural network, trained using a combination of pseudo-labeled and strongly-annotated data. The choice to use a smaller network is made so that the model can be highly efficient at test time, but guaranteeing performance comparable to the teacher [38].

The training schema is composed of a pipeline of operations that are summarized here:

1) train the teacher with strongly-annotated data;
2) annotate pseudo-labeled data;
3) select pseudo-labeled data;
4) train the student with pseudo-labeled data;
5) fine-tune the student with strongly-annotated data.

In the first step of the training schema, the teacher model is trained with strongly-annotated data. Thus, it learns how to select relevant examples from the unlabeled data. In the second step, the teacher annotates unseen data, generating a prediction vector of the class probabilities from a softmax layer. In the third step, the teacher selects the pseudo-labeled samples to present to the student model. The samples selected are the ones with the highest probability of belonging to a class. The vectors are sorted in descending order by the class probability. P samples per class are selected from the highest-ranked ones [19]. In this step, it is essential to minimize the number of noisy samples selected [17]. Therefore, the right P value must be selected. However, this value is not possible to be identified a priori. In the fourth step, the student model is trained using the pseudo-labeled data. In this step, it is possible to explore different P values. Therefore, the model is trained with different subsets of pseudo-labeled data, each one including a different number of pseudo-labels per class. Among these models, the one that shows the best performance is the one trained with the subset with fewer noisy labels. Indeed, this subset includes the smallest number of noisy labels, compared with the others. In the fifth step, the student model is fine-tuned using the strongly-annotated data.

The learning paradigm is tested on the student model. The model is tested in two different steps of the pipeline and it is compared with fully-supervised learning approach. Firstly, it is tested after the training with only the pseudo-labeled data (Figure 1, step 4). Secondly, it is tested after the training with the pseudo-labeled and the fine-tuning with the strongly annotated data (Figure 1, step 5). In the fully-supervised learning approach, the student model is trained only with strongly-annotated data.

C. Implementation

The teacher model is Resnext50_32x4d, while the student model is DenseNet121 [39]. Both networks are implemented in PyTorch (version 1.1.0) and trained on the Cartesius cluster infrastructure, provided by the SURFsara HPC (High

Performance Computing) centre[7], using Tesla K40m GPUs. Both the architectures are trained with the same strategy to set the hyperparameters. In order to avoid overfitting, class-wise data augmentation is applied during the training, with a probabilistic rate.

The strategy for training the models regards the hyperparameters of the network, the weights used for initializing the models and the replacement of the last layer. Both models are trained ten different times, in order to avoid the non-deterministic effects caused by the stochastic gradient descent and the data augmentation pipeline. The average and standard deviation of the models are reported. The teacher model used for annotating the unlabeled data is the one that shows the best performance in the TMAZ validation set among the ten repetitions. The student model, selected to be fine-tuned with strongly annotated data, is the one that shows the best performance on the TMAZ validation set among the ten repetitions. Each of these training repetitions is trained for 15 epochs with a batch size of 32 samples. The hyperparameters adopted are the same for both models: they are optimized using Adam optimizer with a learning rate of 0.001 and a decay rate of 10^{-6}. Both the models are initialized with pre-trained weights. The teacher model has the initialized weights pre-trained with the YFCC100M dataset [40], which includes almost 1 billion Instagram images [19]. The student model has the initialized weights pre-trained with ImageNet images [41]. In both models, the architecture is changed for adapting the problem to the number of classes. The last layer of the original network architecture (1000 nodes) is changed with a new dense layer of four nodes (the number of classes in this classification problem).

A class-wise data augmentation (CWDA) solution is applied during the training phase of the CNNs. The class-wise data augmentation consists of three operations, applied in order to avoid overfitting. The operations of the pipeline are rotation, flipping and color augmentation, implemented with the Albumentations open-source library [42]. They are applied to the training images with a probability of 0.5 on each batch. The unbalanced distribution of the classes, combined with the small amount of data, can lead to overfitting. Class-wise data augmentation (CWDA) is applied to reduce the effect of unbalanced classes on training. It is implemented by the GitHub open access repository of Ufoyn[8].

III. Results

The models trained with the teacher/student paradigm perform better than the one trained with the fully-supervised training. The performance is evaluated with the weighted Cohen κ-score. The models are trained to classify the Gleason score and the Gleason patterns of histopathological image patches. The performance is evaluated on the student model and compared with a fully-supervised learning approach.

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

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

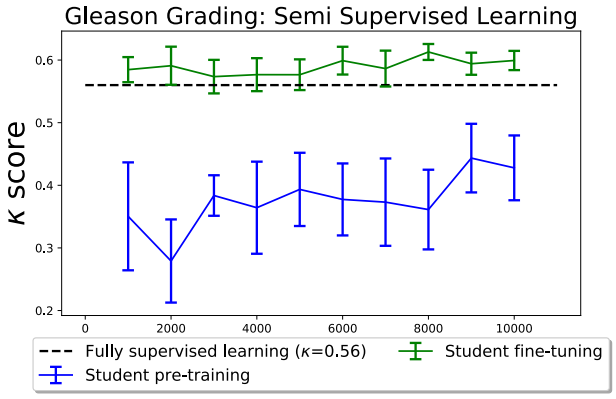

Fig. 2. Results of the student model average performance, trained with the semi-supervised approach, evaluated at the patch level, using the TMAZ test set. They are measured by the κ-score as a function of the amount of pseudo-labeled data used to train the student model.

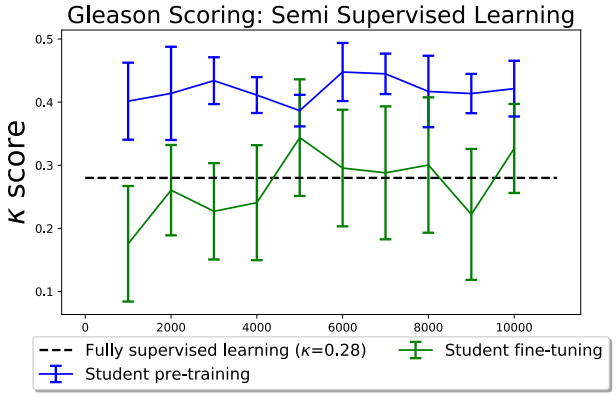

Fig. 3. Results of the student model average performance, trained with the semi-supervised approach, evaluated at the WSI level, using the TCGA-PRAD test set. They are measured by the κ-score as a function of the amount of pseudo-labeled data used to train the student model.

The performance is measured by the weighted Cohen κ-score as a function of the amount of pseudo-labeled examples (per class) used for training the student model. The weighted Cohen κ-score is a metric for measuring agreement between raters. The quadratically weighted κ is adopted for penalizing stronger predictions far from their real class. The Gleason score classification is evaluated at the WSI level, while Gleason pattern classification is evaluated at the patch-level. The Gleason score is measured by the aggregation of Gleason patterns at the patch level, using a majority voting system and the rules of the American Urology Association[9]. In this paper, the majority voting system is applied only on 1000 patches per WSI, selected with the Blue-ratio technique [43]. Blue-ratio permits to avoid the extraction of patches with a small number of nuclei, such as the ones that contain stroma or fat.

[9]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

Figures 2 and 3 show the performance of the training/student semi-supervised paradigm. In both figures, three curves are present. The blue curve represents the performance measured after training the student model with pseudo-labeled data. The green curve represents the performance measured after training the model with pseudo-labeled data and then fine-tuning it with strongly-annotated data. The dashed black line represents the performance of the fully-supervised training of the student model. The classification performance of Gleason patterns in the TMAZ dataset is presented in Figure 2, while the classification performance of Gleason scores in TCGA-PRAD is presented in Figure 3.

In Figure 2, the performance is measured on the TMAZ test set at the patch level. The baseline models (student model trained only with strongly–annotated data) reached a κ=0.5608 $\pm$ 0.0308. Each curve has a peak value since the curves are not monotonically increasing. The performance of the student model trained only with pseudo-labeled data (blue curve) is below the baseline, for each one of the amounts of samples per class tested. The peak value is κ=0.4434 $\pm$ 0.0547, reached with the pseudo-labeled training partition with 9000 patches pseudo-labeled per class. The performance of the student model trained with pseudo-labeled and fine-tuned with strongly-annotated data (green curve) exceeds the baseline, for each one of the amounts of pseudo-labeled data tested. The peak value is κ=0.6129 $\pm$ 0.0127, reached with the pseudo-labeled training partition with 8000 patches pseudo-labeled per class. Therefore, the model trained with pseudo-labeled and fine-tuned with strongly-annotated data exceeds the baseline by 0.052 in κ.

In Figure 3, the performance is measured on the TCGA-PRAD test set at the WSI level. The baseline models (student model trained only with strongly–annotated data) reached a κ=0.2814 $\pm$ 0.1312. Each curve has a peak value since the curves are not monotonically increasing. The performance of the student model trained only with pseudo-labeled data (blue curve) exceeds the baseline, for each one of the amounts of pseudo-labeled data tested. The peak value is κ=0.4478 $\pm$ 0.0460, reached with the pseudo-labeled training partition with 6000 patches pseudo-labeled per class. The lowest performance exceeds the baseline by 0.09 in κ, where the model is trained with 5000 pseudo-labeled samples per class. The performance of the student model trained with pseudo-labeled and fine-tuned with strongly-annotated data (green curve) exceeds the baseline, only for a range (from 5000 to 8000) of pseudo-labeled samples per class tested. The peak value is κ=0.3438 $\pm$ 0.0924, reached with the pseudo-labeled training partition with 5000 patches pseudo-labeled per class. Therefore, the baseline is exceeded by 0.062 in κ using the semi-supervised learning. The student model trained with the semi-supervised approach, in both the steps of the pipeline tested, exceed the baseline. The student model trained with pseudo-labeled data exceeds the baseline by 0.166 in κ. The student model trained with pseudo-labeled and fine-tuned with strongly-annotated data exceeds the baseline by 0.062 in κ. The results are summarized in Table V.

TABLE V
PERFORMANCE MEASURED FOR THE SEMI-SUPERVISED APPROACH, EVALUATED IN $\kappa-$SCORE

Fully-supervised	**Student pre-training**	**Student fine-tuning**
TMAZ dataset		
0.5608 $\pm$ 0.0308	0.4434 $\pm$ 0.0547	**0.6129 $\pm$ 0.0127**
TCGA-PRAD dataset		
0.2814 $\pm$ 0.1312	**0.4477 $\pm$ 0.0460**	0.3437 $\pm$ 0.0923

IV. DISCUSSION

The teacher/student paradigm permits to leverage on a large amount of the unlabeled data for training a more robust CNN model and improving its performance. The performance classification of the models trained with the paradigm is improved compared to a fully-supervised training schema. A trade-off is identified between the number of pseudo-labeled samples used for training and the model's classification performance. The paradigm permits to face the heterogeneity between datasets, limiting the overfitting.

As expected, in both the Gleason grading and the Gleason scoring, the models trained combining pseudo-labels and strongly-annotated data improve the performance, compared with the fully-supervised schema. This is explainable considering that the amount of data used (combining pseudo-labels and strongly-annotated) is increased. However, the metric curves are not monotonically increasing. A peak value in $kappa$ is identified for each of the approaches tested. This peak value allows to explore the best P parameter for the paradigm. P represents the amount of pseudo-labeled samples per class in a subset. The subset that reaches the peak value has less noisy pseudo-labels, compared with the other subsets. The higher the peak value, the fewer noisy labels are included in pseudo-label samples. Therefore, the higher the peak value, the higher is the performance.

The paradigm can alleviate overfitting caused by heterogeneity between datasets, although models tend to adapt their weights to the data with which they are trained (as it was expected). The results show that a model, trained on a dataset, does not generalize well for a different dataset. It is a consequence of the inter-dataset heterogeneity. This effect happens for both the datasets. The student model trained with the TMAZ patches reaches good results in its own set, but it fails to generalize in the TCGA-PRAD test partition, where it obtains some of the worst results (dashed line on Figure 3). The student model, trained with the pseudo-labeled samples, reaches the best results in TCGA-PRAD test set, but it fails to generalize in the TMAZ test partition, where it reaches the worst results (blue curve in Figure 2). The inter-dataset heterogeneity is the reason why the student model, trained only with pseudo-labeled data, performs better on TCGA-PRAD dataset, compared with the same model trained combining pseudo-labeled and strongly-annotated data. However, training the model combining the different data sources alleviates the overfitting. On the TMAZ dataset, the model trained with both

the dataset obtains the best performance (κ=0.6129 $\pm$ 0.0127), but it does not generalize well for the TCGA-PRAD dataset. The model's performance is better than the fully-supervised training of the student. However, the same model, trained only with pseudo-labeled data, exceeds this performance by 0.096 in κ.

V. Conclusion

In this paper, the classification of prostate cancer tissue is tackled with a novel approach, based on the semi-supervised teacher/student paradigm for training CNNs. It permits face data heterogeneity and alleviates the difficulty of obtaining a sufficient amount of locally annotated data for training the models. The approach is compared with a fully-supervised CNN learning approach. The teacher/student paradigm improves the performance of a CNN prostate cancer classification at the patch level and the WSI level. Therefore, it is possible to adopt it to leverage on a large amount of unlabeled data and then improve the fully supervised classification performance of CNNs. Furthermore, the teacher/student paradigm permits to face the heterogeneity of the datasets used for training the models. It permits to generalize better in datasets that come from different medical sources, reducing the effects caused by the overfitting. In the future works, the teacher/student paradigm will be tested on different types of biopsy tissues, with larger values of P parameter and testing more training steps and within the pipeline.

References

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

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

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

[4] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and structural biotechnology journal*, vol. 16, pp. 34–42, 2018.

[5] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[6] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels *et al.*, "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset," *GigaScience*, vol. 7, no. 6, p. giy065, 2018.

[7] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, "Hematoxylin and eosin staining of tissue and cell sections," *Cold spring harbor protocols*, vol. 2008, no. 5, pp. pdb–prot4986, 2008.

[8] K. Larson, H. H. Ho, P. L. Anumolu, and T. M. Chen, "Hematoxylin and eosin tissue stain in mohs micrographic surgery: a review," *Dermatologic surgery*, vol. 37, no. 8, pp. 1089–1099, 2011.

[9] T. Tsujikawa, G. Thibault, V. Azimi, S. Sivagnanam, G. Banik, C. Means, R. Kawashima, D. R. Clayburgh, J. W. Gray, L. M. Coussens *et al.*, "Robust cell detection and segmentation for image cytometry reveal th17 cell heterogeneity," *Cytometry Part A*, vol. 95, no. 4, pp. 389–398, 2019.

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

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

[12] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "A cluster-then-label semi-supervised learning approach for pathology image classification," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[13] A. Foucart, O. Debeir, and C. Decaestecker, "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, 2019, pp. 1869–1872.

[14] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An em-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Computerized Medical Imaging and Graphics*, vol. 69, pp. 125–133, 2018.

[15] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, and F. Mahmood, "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," *arXiv preprint arXiv:1910.10825*, 2019.

[16] S. Shaw, M. Pajak, A. Lisowska, S. A. Tsaftaris, and A. Q. O'Neil, "Teacher-student chain for efficient semi-supervised histology image classification," *arXiv preprint arXiv:2003.08797*, 2020.

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

[18] J. Bagherzadeh and H. Asil, "A review of various semi-supervised learning models with a deep learning and memory approach," *Iran Journal of Computer Science*, vol. 2, no. 2, pp. 65–80, 2019.

[19] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.

[20] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 2.

[21] S. Otálora, O. Perdomo, F. González, and H. Müller, "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, 2017, pp. 146–154.

[22] W. Shao, L. Sun, and D. Zhang, "Deep active learning for nucleus classification in pathology images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 199–202.

[23] L. Raczkowski, M. Mozejko, J. Zambonelli, and E. Szczurek, "Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

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

[25] H. Veeraraghavan and J. V. Miller, "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, 2011, pp. 1645–1648.

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

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

[28] O. J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, and H. Müller, "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score," in *Medical Imaging 2017: Digital Pathology*, vol. 10140. International Society for Optics and Photonics, 2017, p. 101400O.

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

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

[31] A. Katharopoulos and F. Fleuret, "Processing megapixel images with deep attention-sampling models," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 3282–3291. [Online]. Available: http://proceedings.mlr.press/v97/katharopoulos19a.html

[32] J. van der Laak, F. Ciompi, and G. Litjens, "No pixel-level annotations needed," *Nature Biomedical Engineering*, pp. 1–2, 2019.

[33] S. Otálora, M. Atzori, A. Khan, O. Jimenez-del Toro, V. Andrearczyk, and H. Müller, "A systematic comparison of deep learning strategies for weakly supervised gleason grading," in *Medical Imaging 2020: Digital Pathology*, vol. 11320. International Society for Optics and Photonics, 2020, p. 113200L.

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

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

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

[37] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," in *Handbook on Neural Information Processing*. Springer, 2013, pp. 215–239.

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

[39] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[40] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "The new data and new challenges in multimedia research," *CoRR*, vol. abs/1503.01817, 2015. [Online]. Available: http://arxiv.org/abs/1503.01817

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

[42] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018.

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