WeGleNet: A Weakly-Supervised Convolutional Neural Network for the Semantic Segmentation of Gleason Grades in Prostate Histology Images

Julio Silva-Rodríguez[a], Adrián Colomer[b], Valery Naranjo[b]

[a] *Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain*
[b] *Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain*

Abstract

Background and Objective:

Prostate cancer is one of the main diseases affecting men worldwide. The Gleason scoring system is the primary diagnostic tool for prostate cancer. This is obtained via the visual analysis of cancerous patterns in prostate biopsies performed by expert pathologists, and the aggregation of the main Gleason grades in a combined score. Computer-aided diagnosis systems allow to reduce the workload of pathologists and increase the objectivity. Nevertheless, those require a large number of labeled samples, with pixel-level annotations performed by expert pathologists, to be developed. Recently, efforts have been made in the literature to develop algorithms aiming the direct estimation of the global Gleason score at biopsy/core level with global labels. However, these algorithms do not cover the accurate localization of the Gleason patterns into the tissue. These location maps are the basis to provide a reliable computer-aided diagnosis system to the experts to be used in clinical practice by pathologists.

In this work, we propose a deep-learning-based system able to detect local cancerous patterns in the prostate tissue using only the global-level Gleason score obtained from clinical records during training.

[*] This work was supported by the Spanish Ministry of Economy and Competitiveness through projects DPI2016-77869 and PID2019-105142RB-C21.

Email addresses: jjsilva@upv.es (Julio Silva-Rodríguez), adcogra@ui3b.upv.es (Adrián Colomer), vnaranjo@dcom.upv.es (Valery Naranjo)

Methods:
The methodological core of this work is the proposed weakly-supervised-trained convolutional neural network, WeGleNet, based on a multi-class segmentation layer after the feature extraction module, a global-aggregation, and the slicing of the background class for the model loss estimation during training.

Results:
Using a public dataset of prostate tissue-micro arrays, we obtained a Cohen's quadratic kappa (κ) of 0.67 for the pixel-level prediction of cancerous patterns in the validation cohort. We compared the model performance for semantic segmentation of Gleason grades with supervised state-of-the-art architectures in the test cohort. We obtained a pixel-level κ of 0.61 and a macro-averaged f1-score of 0.58, at the same level as fully-supervised methods. Regarding the estimation of the core-level Gleason score, we obtained a κ of 0.76 and 0.67 between the model and two different pathologists.

Conclusions:
WeGleNet is capable of performing the semantic segmentation of Gleason grades similarly to fully-supervised methods without requiring pixel-level annotations. Moreover, the model reached a performance at the same level as inter-pathologist agreement for the global Gleason scoring of the cores.

Keywords: Gleason grading, prostate cancer, semantic segmentation, tissue micro-arrays, weakly supervised.

1. Introduction

Prostate cancer is one of the most common diseases affecting men worldwide. It constitutes 14.5% of all cancers affecting men [1], and, according to the World Health Organization, the yearly number of new cases will increase by up to 1.8 million people in this decade [2]. The gold standard for prostate cancer diagnosis and prognosis prediction is the analysis of prostate biopsies under the Gleason grading system [3]. This system defines a series of cancerous patterns related to the morphology, distribution, and degree of differentiation of the glands in the tissue. Specifically, in histology slides, the observable Gleason grades (GG) range from 3 (GG3) to 5 (GG5). Examples of those patterns are presented in Figure 1.

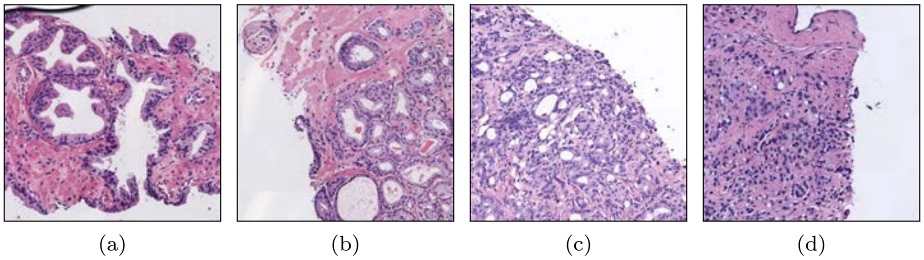

(a) (b) (c) (d)

Figure 1: Histology regions of prostate biopsies. (a): region containing benign glands, (b): region containing GG3 glandular structures, (c): region containing GG4 patterns, (d): region containing GG5 patterns. GG: Gleason grade.

In clinical practice, small portions of tissue are extracted, laminated, stained with Hematoxylin and Eosin, and finally analyzed under the microscope by expert pathologists using this system. Local cancerous regions of the sample are classified according to the Gleason grades, and finally, the two majority patterns are grouped to obtain a Gleason score as prognosis biomarker (e.g. the Gleason score $5 + 4 = 9$ would be assigned to a sample in which the main cancerous Gleason grade is GG5 and the second is GG4). Due to the large size of the biopsies augmented under a microscope, this process results in a high time-consuming and repetitive task, and presents a large intra and inter pathologist variability [4].

In the last decades, the development of digitization devices has allowed the storage of biopsies at microscopic magnifications as digital images. Due to this advance, the field of Computer-Aided Diagnosis (CAD) systems to support pathologists based on computer-vision techniques has experienced a great growth. However, the development of those applications is limited due to the high data-demanding character of deep learning algorithms, and the difficulty in obtaining pixel-level labeled histology images [5]. Normally, pathologists store in the clinical history the global-level diagnosis of the biopsy (e.g. the Gleason score per prostate biopsy). In order to train/build models or develop algorithms able to detect and grade local cancerous patterns, a laborious manual annotation process is required, which must be performed by expert pathologists due to the complexity of the task. In the case of prostate cancer, the different tumor patterns have to be accurately delimited at the pixel level to avoid noisy annotations. Even though multi-resolution graphical user interfaces are provided to clinicians for performing this task, it is a tedious process prone to error. These limitations encourage the development of weakly-supervised deep-learning techniques able to utilize global labels during the training process to accurately identify local cancerous patterns in the images. The main benefit of those methods is that they are not limited to the annotated samples. They can work using histology images labeled only in the global-level patient diagnosis. Recent advances in the literature have proposed the use of the global Gleason score (obtained from the clinical record) to develop CAD systems for biopsy scoring (these works are detailed in Section 2.2). Nevertheless, these methods focus on predicting only global biopsy-level markers, while the location of the cancerous structures in the tissue is qualitatively evaluated or not addressed. The classification of local Gleason grades in prostate biopsies is the basis of CAD systems during its use in clinical practice. Accurate heat-maps provide confidence to the pathologists in the daily use of the CAD system, and support the biopsy-level markers provided by the system.

In this work we propose a deep-learning architecture based on convolutional neural networks able to perform a semantic segmentation of the Gleason grades (i.e. non-cancerous tissue, GG3, GG4 or GG5 classes) in prostate histology images, trained via weak supervision using the diagnosed Gleason score of the sample. To the best of the authors' knowledge, this is the first time in the literature that weakly-supervised methods are explored and quantitatively assessed for the local segmentation of cancerous Gleason grades. The main contributions of this research are the following: (i): a weakly-supervised framework based on a convolutional neural network (CNN) architecture able to obtain complementary semantic segmentation maps based on a novel configuration of multi-class activation maps, aggregation layers and the slicing of the background class prediction during training; (ii) the validation of different aggregation layers and regularization techniques to optimize the model; and (iii) the comparison of the proposed weakly-supervised model with fully-supervised state-of-the-art methods.

The remainder of this paper is divided into five sections. First, Section 2 presents an overview of the literature related to this research. Concretely, Section 2.1 describes the paradigm of weakly-supervised segmentation and the main related computer-vision techniques, Section 2.2 describes its applications in histology images and Section 2.3 details the state-of-the-art methods for Gleason grading of local cancerous patterns. Secondly, Section 3 describes the database used in the experimental stage. Then, Section 4 details the methodological core of this work. In particular, Section 4.1 and 4.2 describe WeGleNet, our proposed weakly-supervised CNN architecture (contribution (i)), and Section 4.3 presents the fully-supervised state-of-the-art architectures used as a benchmark to compare our model. Then, Section 4.4 describes the Gleason scoring method using the class-wise segmentation maps. Section 5 describes the experiments carried out in this work. The strategy and figures of merit used in this process are specified in Section 5.1. Section 5.2 presents different ablation experiments related to the optimization process of the proposed architecture (contribution (ii)), and Sections 5.3 and 5.4 expose an in-depth validation and comparison of our proposed model against fully-supervised models and previous literature (contribution (iii)). Finally, Section 6 summarizes the main conclusions extracted from our work.

2. Related Work

2.1. Weakly-Supervised Semantic Segmentation

Weakly-supervised learning deals with the challenge of using incomplete, scarce, inexact, inaccurate, or noisy information. The problem addressed in this work, image segmentation using just global labels during training, is covered within the Multiple Instance Learning (MIL) scope. MIL works with data clustered on bags of instances, under the assumption that bags labeled as a certain class present, at least, one instance belonging to that class. For one image X composed by the instances (pixels) x_{ij}, the bag-level label (Y) for a class (c) could be interpreted as:

$$Y_c = \begin{cases} 1, & \text{if } \exists\, x_{ij} : y_c = 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where y_c is the instance-level label for certain class c.

In this topic, two different kinds of classification problems are defined: the prediction of bag-level (global) labels, or the classification of individual instances. In this work, both problems are addressed. A recent extensive review of MIL and its characteristics can be found in [6]. Regarding MIL in image classification, convolutional neural networks (CNNs) are the most used technique, since they have demonstrated promising properties for locating objects while performing image-level classification tasks [7, 8].

The approaches to obtain segmentation maps from global-level image classification using CNNs can be divided into aggregation and gradient-based methodologies. Aggregation methods build segmentation maps into the CNN architecture. They are composed of three main blocks: a feature-extraction stage (or base model), an adaptation layer that constructs segmentation maps per class, and a global aggregation layer that resumes each map to one representative value. Then, a multi-label loss function is used to optimize the network weights. The main proposed architectures in this field are WILDCAT [9] and Attention-MIL [10]. WILDCAT constructs the adaptation layer by pooling activation maps after the last convolutional block of the base model and then applies a global-pooling operator to obtain the bag-level probabilities. Attention-MIL joins the adaptation and global aggregation layer by using an attention mechanism that combines all the features obtained in each instance by fully-connected layers. Regarding the gradient methods, the segmentation maps are obtained by post-processing the network output. In this line, the most relevant technique in the literature is the gradient-based class activation maps (Grad-CAMs) [11]. In this technique, the activation maps of the last convolutional block are linearly combined. Each map is weighted by back-propagating gradients in the network from the classification layer, and a ReLU activation is applied to the weights to keep just the features with a positive influence on the classification. Recently, the efforts on weakly supervised semantic segmentation have focused on self-supervised learning. In this methodology, CAMs obtained from gradient-based methods are used as pseudo labels to feed a pixel-level semantic segmentation network. Although these methods have reached promising results, they are still limited by the CAMs used, and the propensity of CNNs to look only at specific and discriminatory patterns. In this line, Ficklenet [12] and IRNet [13] have proposed the use of center-fixed spatial dropout and class propagation respectively to alleviate this limitation. In all the strategies, the aggregation of the different class-level maps (or CAMs) in a semantic segmentation mask is not straightforward. This process is usually carried out by hand-crafted post-processing. Some methods are based on simply assigning to each pixel the label with the highest probability and let as background those instances with probabilities below certain threshold [12]. Other works apply complex fully-connected conditional random fields (CRF) to combine the different class-level maps into one combined mask [9, 14, 15, 16]. In our work, we take steps forward in order to solve this limitation, and propose a CNN architecture that obtains complementary multi-class semantic segmentation maps without requiring any post-processing (see Section 4.1 for further explanation). An extensive survey regarding the application of weakly-supervised learning across different image domains and its current limitations was recently presented in [17].

2.2. Weakly-Supervised Segmentation in Histology Images

Weakly-Supervised learning is a field of increasing interest for histology images, due to the difficulty of preparing large datasets labeled by expert pathologists. While some works just focus on the prediction of bag-level labels in biopsy slides [18, 19, 20, 21] carrying out a qualitative evaluation of instance-level (local) classifications, others quantitatively evaluate their proposed models for the local-level classification task [22, 23, 24, 25]. Nevertheless, most of the works only focus on binary classification cancer/no cancer. Early work in [24] proposes a MIL model based on hand-crafted feature extraction (SIFT, color histogram, Local Binary Patterns, etc.), machine learning classifiers and aggregation of the instance-level probabilities for colon cancer detection. Lately, semi-supervised CNNs were used for gland segmentation in prostate images in [23]. However, the proposed UNet required to incorporate some instance-level annotations during training to perform properly. Finally, recent work in [22] included previous knowledge by applying constraints in the training stage of a weakly-supervised CNN to control the size of positive instances in the image for colon cancer detection. Recent works have used weakly-supervised CNNs approaches for multi-class semantic segmentation. Concretely, HistoSegNet, introduced in [25], performs a weakly-supervised segmentation of different tissue types in histology images based on CNNs and Grad-CAM gradient method. Then, a complex hand-crafted post-processing is proposed to join the class-level segmentation maps and to include the background class.

2.3. Prostate Gleason Grading

In the analysis of prostate histology samples, as mentioned previously, there are two main tasks: the grading of local structures using the Gleason system, and the global scoring.

First works in this field focused on fine-tuning well-known CNN architectures in a supervised patch-level classification, with the requirement of pixel-wise expert annotations. In this line, Nir et al. [26, 27] obtained a patch-level Cohen's quadratic kappa (κ) of 0.60 in the validation set, while 0.55 and 0.49 was reached by Arvaniti et al. in [28] in the test cohort referenced to two different pathologists. Then, the percentage of each cancerous tissue in the sample was calculated from the patch-level probabilities to predict the Gleason score of the sample. Arvaniti et al. [28] obtained with this method a κ of 0.76

and 0.71 against the annotations of two different pathologist, at the level of the inter-pathologist agreement ($\kappa = 0.71$).

Latest works in the literature have started to develop weakly-supervised techniques to avoid the tedious process of pixel-level labeling of Gleason grades. These techniques are based on assigning the global labels (i.e. the primary and secondary grades obtained from the Gleason score) to patch-level regions of interest (i.e. glandular or nuclei structures). Then, convolutional neural networks are trained to perform a patch-level classification with the obtained pseudo-ground truth. The selection of regions of interest in the tissue are based on different approaches, detailed in the following lines. The work in [29] developed a semi-supervised pipeline detecting the glandular tissue via a UNet trained with manual annotations. A few works works focus on selecting these regions with larger amounts of nuclei, based on color [30, 31] or Laplacian filters [32]. Finally, the work in [18] directly assigns the global label (cancerous against non cancerous) to all the patches in the tissue. All previous methods train patch-level convolutional neural networks with the obtained pseudo-ground truth, and finally they combine the patch-level predictions to obtain the global score. The first works aggregate the predictions using the percentage of each Gleason grade in the sample and then they train different machine learning models to predict the global Gleason score [29, 30, 32]. Also, novel approaches combine the patches using the features extracted by the CNN through recurrent neural networks [18]. Although the aforementioned methods provide promising results for Gleason scoring of prostate biopsies, the assumptions made to develop their weakly-supervised pipeline could be affecting the local grading of cancerous patterns. To the best of the authors' knowledge, none of previous works in the literature focus on locating the Gleason grades in the tissue using weakly-supervised learning. They only perform a qualitative evaluation of the heat-maps obtained by their models.

3. Dataset

The experiments described in this work were carried out using the public dataset presented by Arvaniti et al. in [28][1]. This dataset consists of 886 prostate Tissue Micro-Arrays (TMAs, samples of representative regions of cancerous biopsies known as cores), digitized at 40× magnification in images of size 3100^2 pixels. The cores include pixel-level annotations of Gleason grades and benign structures, and global labels of Gleason scores (primary and secondary Gleason grades in the sample). The distribution of the Gleason grades (GG) in the cores is distributed as follows: 421, 387 and 148 cores with GG3, GG4 and GG5, respectively. Regarding the pixel-level annotations, the dataset includes five different classes: benign tissue, GG3, GG4, GG5, and background. In order to evaluate our proposed methodology, the benign and background classes are joined in the non-cancerous (NC) class. To establish fair comparisons with previous literature, the partition of the dataset proposed by Arvaniti et al. was used for training, validating, and testing. Note that the test cohort contains pixel-level annotations made by two different expert pathologists.

4. Methods

4.1. WeGleNet: Weak-Supervised Gleason Grading Network

The methodological core of this work consists of a convolutional neural network able to predict semantic segmentation maps of non-cancerous, Gleason grade 3 (GG3), GG4, and GG5 tissue in prostate histology images, trained using global labels of the grades present in the tissue during training. The proposed weak-supervised Gleason grading network (WeGleNet) is presented in Figure 2.

The architecture is composed of three main components: the base model, the segmentation (also called adaptation) layer, and the global-aggregation operation, and it takes as input the prostate core image, which is resized to 750^2

[1] We contacted the corresponding authors to obtain the dataset.

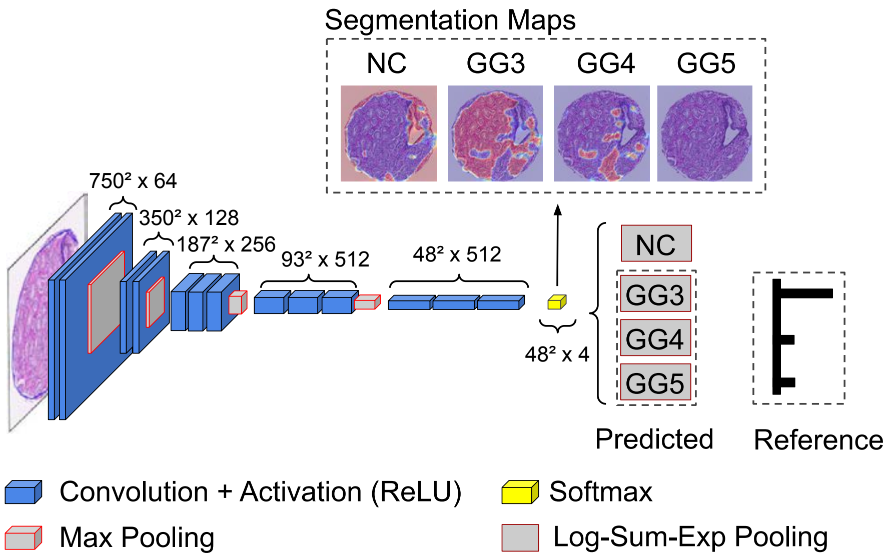


Figure 2: WeGleNet, weakly-supervised framework for semantic segmentation of local cancerous patterns via Gleason grading using the Gleason score of the global sample during the training stage. NC: non cancerous; GG3: Gleason grade 3; GG4: Gleason grade 4; GG5: Gleason grade 5.

pixels due to computational limitations. First, the base model is in charge of extracting automatic-learned features from the input image. Concretely, the VGG19 architecture [33] is used. This is based on convolutional blocks with an increasing number of filters with 3×3 kernels with ReLU activation and dimensional reduction via max pooling of size 2×2. In order to reduce the over-fitting during the training stage, weights are initialized using the VGG19 model pretrained in the ImageNet dataset [34]. Secondly, the segmentation layer applies to the output convolutional feature volume of the base model as many convolutional filters of size 1×1 as classes to be predicted. This layer also computes a softmax activation along the class dimension generating a multi-class segmentation volume of activation maps, in which each value represents the probability of that pixel of belonging to a class. During the inference stage, this layer will be the model output, and each segmentation map will be resized to the original core dimensions (3100^2 pixels). During the training stage, the pixel-level probabilities in the activation maps are aggregated in order to output one global probability per class ranging between 0 and 1. This operation is performed by a global-aggregation layer, which is detailed in Section 4.2. This aggregation of instance-level predictions embedded in the training stage of the model avoids previous assumptions in the literature to locate the regions of interest in the tissue. Then, binary cross-entropy is used as a loss function. As all cores contain non-cancerous regions, the loss function is only calculated using the Gleason grade classes (i.e. GG3, GG4, and GG5). Thus, the NC class segmentation map gathers those patterns not related to cancer but does not contribute to the calculation of the loss function. This strategy allows obtaining complementary segmentation maps including the background class (in our case non-cancerous class). This is a step forward compared to previous methods, which were based on the individual prediction of segmentation maps per class, and complex post-processing to join them including the background class (see Section 2.2 for a more detailed explanation of these methods).

During the training stage, two techniques are carried out to regularize the model and avoid over-fitting: data augmentation and hide-and-seek [35]. Data augmentation is performed by transforming the input images with random translations, rotations and mirroring in each iteration. Hide-and-seek (HS) is a method that regularizes weakly-supervised-trained architectures by replacing random patches of the images with the average intensity level of the input. In each iteration, the hidden patches vary, and thus the network is forced to focus on heterogeneous patterns during training. The input image is divided into patches of 75^2 pixels, which have a 25% probability of being hidden in each iteration.

4.2. Global-Aggregation Layers

Global-aggregation layers summarize information from all spatial locations in the activation maps (x_{ij}) to one representative value (p). For this task, we propose the use of the log-sum-exponential (LSE) layer [36] in WeGleNet, which is defined as:

$$p_{LSE} = \frac{1}{r} \cdot log \left[\frac{1}{S} \cdot \sum_{(i,j) \in S} exp(r \cdot x_{ij}) \right] \tag{2}$$

where S constitutes the number of pixels in the activation map x_{ij} and r is a parameter to be optimized.

The LSE operation permits us to obtain a domain-specific representation of the activation map via the parameter r, with large values of r ($r \rightarrow \infty$) similar to a global-max pooling operation (GMP) [8] and small values ($r \rightarrow 0$) equivalent to a global-average operation (GAP) [37]. The r parameter is empirically fixed by optimizing the model performance in the validation cohort (see Section 5.2). By this procedure, the training stage overcomes the limitations of the other global-aggregation layers (i.e. GAP assumes that the pattern is uniformly distributed across with the activation map, and GMP could produce over-fitting to small, specific patterns).

4.3. Fully-Supervised CNNs

To compare our proposed weakly-supervised framework, two state-of-the-art supervised architectures for semantic segmentation of Gleason grades are implemented. To take advantage of the pixel-level annotations, patches are extracted from the cores with a size of 750^2 pixels and a step of 350. Due to hardware limitations during training, patches are resized to 224^2 pixels. Then, a UNet architecture and a classifier based on a patch-level VGG19 fine-tuned network (VGG19Sup) are selected as supervised architectures to be compared to the WeGleNet model. It is important to highlight that these methods require an accurate pixel-level labeling of the images. The implementation of the models is detailed in the following lines.

VGG19Sup is based on training a patch-level multi-class classifier and then modifying the architecture to obtain segmentation maps. VGG19Sup is composed of a feature-extraction stage using VGG19 backbone pre-trained in Imagenet dataset, a global-average pooling (GAP) to aggregate the activation maps, and a fully-connected layer with as many neurons as classes to predict and soft-max activation as output. In this method, each patch is labeled as the majority grade annotated. If none Gleason grade is annotated, the patch is labeled as non-cancerous. Training is performed by optimizing the categorical cross-entropy as loss function. For the inference of segmentation maps, the output fully-connected layer is converted in a convolutional layer with kernel 1×1, which is applied over the activation volume previous to the GAP layer to obtain a segmentation map per class. This approach is equivalent to using a class activation map (CAM) post-processing, but the segmentation maps are obtained directly from the CNN in an end-to-end manner. This method was previously used by Arvaniti et al. in [28] to obtain the probability maps in prostate samples.

Regarding the UNet architecture [38], it is based on a symmetric encoder-decoder path. In the encoder, feature extraction is carried out based on convolutional blocks and dimensional reduction through max-pooling layers. Each convolutional block increases the number of filters by a factor of 2×, starting from 64 filters up to 1024. After each block, the max-pooling operation reduces the dimensions of the activation maps in a factor of 2×. Then, the decoder path builds the segmentation maps, recovering the original dimensions of the image. The reconstruction process is based on deconvolutional layers with filters of size 3×3 and ReLU activation. These layers increase the spatial dimensions of the activation volume in a factor of 2× while reducing the number of filters by a half. Then, the encoder features from a specific level are joined with the resulting activation maps of the same decoder level by a concatenation operation, feeding a convolutional block that combines them. The convolutional block used during both encoder and decoder paths includes residual connections [39] to improve the model optimization. This residual UNet configuration was proposed in [40], and showed to outperform other configurations for Gleason grading in [41]. It consists of three convolutional layers with 3×3 kernels and ReLU activation. The output of the last convolutional layer of the block in connected via a shortcut residual operation with the output of the first layer. Finally, after the decoder, a 1×1 convolutional layer creates the segmentation probability maps. The loss function used during the training process is the categorical *Dice* used in [41].

During the inference stage, the supervised models are used to predict the entire core instead of local patches. Cores are resized to match the resolution used during training, and then the output segmentation maps are resized to the original dimensions of the cores (3100^2 pixels).

4.4. Global Gleason Scoring

Once the probability maps per class are obtained, the Gleason score of the sample is inferred from the percentage of each class k in the tissue, w^k. In [28], the Gleason score is obtained assigning the majority and secondary grades in terms of percentage, considering only the classes above certain threshold c. In this work, we introduce another term, d, which models the tendency of pathologists to focus on the majority cancerous pattern if it is widespread in the tissue. Thus, the final percentage weights are assigned to each class such that:

$$w^k = \begin{cases} 0, & \text{if} \quad \max_{k'} \; w^{k'} > d \quad \text{and} \quad k \neq argmax_{k'} \; w^{k'} \\ w^k, & \text{otherwise} \end{cases} \tag{3}$$

where k denotes the different classes, i.e. Gleason grade 3, 4 and 5.

The operator d adapts the weakly-supervised framework to the global scoring procedure in clinical practice. Pathologists annotate regions focusing on primary patterns, while the weakly supervised model performs a more fine-grained segmentation, that increases the percentage of secondary patterns. Thus, d allows to suppress the system's confidence on these patterns for the global scoring task. The values of the parameters c and d are empirically fixed in the validation set to optimize the results.

5. Experiments and Results

5.1. Experimental Strategy and Metrics

In order to validate the proposed WeGleNet model, two types of figures of merit are extracted from the model output: global-level (bag-level in the MIL framework) and local-level (instance-level) metrics. Figure 3 illustrates the evaluation strategy.

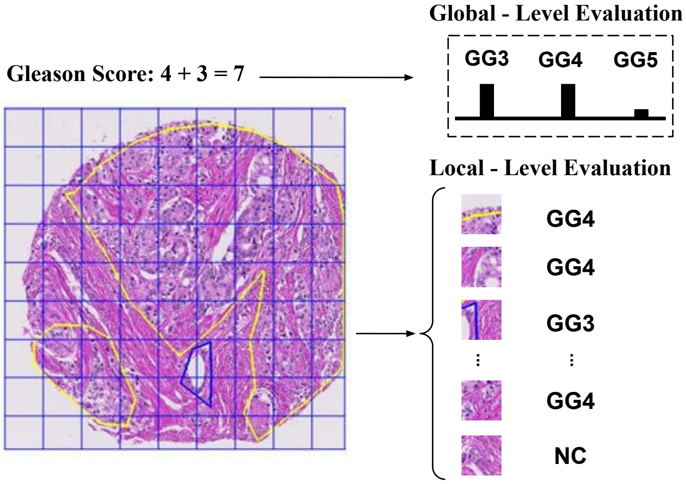


Figure 3: Strategies for the evaluation of the model performance. NC: non cancerous; GG: Gleason grade. The core-level (global) predictions are evaluated using the Gleason score. The local-level predictions are evaluated at pixel-level or using small patches extracted from the core.

Global-level metrics are obtained comparing the multi-label prediction of the WeGleNet in the global-aggregation layer and the Gleason grades observed in the core using the reference Gleason score. This evaluation is used to optimize the weakly-supervised model using the Area Under ROC curve (AUC) as a figure of merit. The decision of using this metric during the optimization stage is related to being closer to the output probabilities of the model. Finally, during the comparison of the model performance with previous literature (Section 5.3) the Cohen's quadratic kappa (κ) [42] is obtained for the Gleason score prediction.

This agreement statistic takes into account that in a set of ordered classes, errors between adjacent classes should be less penalized.

Regarding the local-level evaluation, this is performed to analyze the capability of the trained model for segmenting the Gleason grades in the tissue. During WeGleNet optimization and its comparison with fully-supervised methods for semantic segmentation, metrics are obtained at pixel level. The obtained figures of merit are the accuracy (ACC), f1-score per class, the macro-average (F1), mean intersection over union (mIoU) and Cohen's quadratic kappa (κ). Usually, in the Gleason grading literature, the local grading of cancerous patterns is evaluated at patch level to avoid underestimation of the model performance due to an inaccurate pixel-level annotation in the ground truth. Therefore, WegleNet is evaluated at patch level for the comparison of its performance with previous state-of-the-art works in this field. In order to establish fair comparisons with previous results reported in the literature in the used dataset, patch-level labels are obtained as proposed by Arvaniti et al. [28]. Concretely, patches are extracted using a moving-window of size 750^2 and a step of 350 pixels. Patches with multiple or no annotations were discarded, and the remaining were labeled by majority voting according to the annotations in the central region of the patch (i.e. benign, GG3, GG4 or GG5).

The remainder part of the experimental section describes the experiments carried out to optimize the WeGleNet architecture (Section 5.2), and its comparison on the local-level segmentation of Gleason grades with supervised methods (Section 5.3) and with previous works using the same dataset (Section 5.4).

5.2. Model Optimization

In the first experiments, the objective was to optimize the WeGleNet architecture for semantic segmentation using global-level labels (i.e. the presence of certain Gleason grade in the core). The model performance was studied under the different regularization techniques and global-aggregation layers. WeGleNet model was trained using the proposed log-sum-exponential (LSE), global-max (GMP) and global-average (GAP) pooling. In LSE layer, different values of the

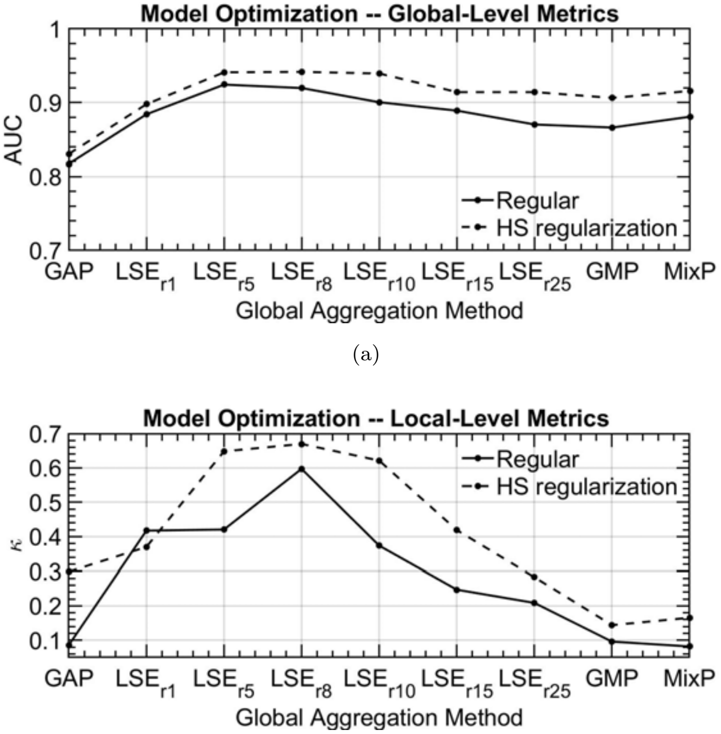
r parameter, $r = \{1, 5, 8, 10, 15, 25\}$, were used. In addition, to compare the performance of the LSE with respect to an automatic-learned combination of GMP and GAP, a mixed-pooling (MixP) aggregation layer is implemented such that:

$$p_{MixP} = \alpha \cdot p_{GMP} + (1 - \alpha) \cdot p_{GAP} \tag{4}$$

where α is a parameter learned during training.

The use of hide-and-seek (HS) regularization was validated by training the models with and without it. The training was performed in mini-batches of 8 images, and Stochastic Gradient Descent (SGD) was used as the optimizer with a learning rate of $1 \cdot 10^{-3}$. Exponential decay in the learning rate was applied in the last 20 epochs to stabilize the model weights such that: $\eta = 1 \cdot 10^{-3} \cdot e^{-0.1 \cdot t}$, where η is the applied learning rate and t is the epoch. The training was carried out during 120 epochs, which were increased to 400 when applying HS regularization. WeGleNet was trained using the training cohort, and early stopping was applied by keeping the weights of the model obtaining the best performance in the validation set (in terms of the obtained losses). After each experiment, segmentation maps were obtained from the segmentation layer, and core-level predictions were obtained from the global-aggregation layer using the images of the validation cohort. The scripts to reproduce the experiments reported in this work are publicly available on (`https://github.com/cvblab/prostate_wsss_weglenet`). Figures of merit related to global-level predictions and pixel-level segmentation are presented in Figure 4 (a) and (b) respectively.

Regarding the obtained results, LSE pooling showed superior performance compared to other global-aggregation techniques. In particular, the best results were obtained using $r = 8$ (LSE_{r8}), with an AUC of 0.9243 for the core-level detection of Gleason grades and a κ of 0.5973 for the pixel-level segmentation. Hide-and-Seek regularization (HS) showed to improve the results in all the experiments, forcing the model to focus on all the patterns of the images. Thus, results improved to an AUC of 0.9416 and a κ of 0.6699 in the best-performing



(a)

(b)

Figure 4: Model performance using different global-aggregation methods and regularization techniques. (a): global prediction performance; (b): pixel-level segmentation performance. HS: hide-and-seek; GAP: global-average pooling; LSE: log-sum-exponential pooling; GMP: global-max pooling; MixP: mixed pooling.

model, WeGleNet - LSE_{r8}. Finally, a high correlation was observed between the global-level and the local-level performance of the model. A Pearson correlation coefficient of 0.5462 was obtained between κ and AUC when using HS regularization. Then, improvements in the global-level predictions produced a better segmentation of the Gleason grades. This promising behavior indicates that the model can be optimized without any pixel-level annotations.

5.3. Weak Supervision vs. Strong Supervision

Once WeGleNet was optimized using the validation cohort, the best performing configuration, WeGleNet - LSE_{r8} with HS regularization, was used to predict the segmentation maps from the images of the test cohort. Representative examples of the obtained results are presented in Figure 5. This figure is organized as follows: each row is a different core and each column represents the ground truth of the Pathologist 1, and the predicted heatmaps for GG3, GG4 and GG5 classes, respectively. Finally, the last column presents the discrete-valued semantic segmentation maps, assigning to each pixel the class with the highest probability. In this figure, green, blue and red color indicate GG3, GG4 and GG5 patterns, respectively.

Then, we carried out experiments to compare our proposed weakly-supervised model with respect to the state-of-the-art supervised methods. UNet model was trained using Nadam as optimizer, with a learning rate of $1 \cdot 10^{-4}$ during 60 epochs. In each iteration, a mini-batch of 16 images was used to update the models weights. Regarding the VGG19Sup model, a learning rate of $1 \cdot 10^{-3}$ with SGD as optimizer was used. Training was performed in mini-batches of 64 images during 120 epochs. For both models, early stopping was applied to keep the best-performing model in the validation cohort (in terms of the obtained loss). From the trained models, the segmentation maps of the images in the test cohort were predicted. The figures of merit obtained by our proposed WeGleNet - LSE_{r8} and the supervised models are presented in Table 1, using as reference the annotations carried out by the pathologist 1 (the same pathologist that annotated the training and validation images). In order to perform a detailed comparison, accuracy (ACC), class-level f1-score (F1), average intersection over union (mIoU) and quadratic Cohen's kappa (κ) were obtained as detailed in Section 5.1.

WeGleNet - LSE_{r8} model reached a κ value of 0.6105, a $mIoU$ f 0.4368 and an average F1 of 0.5798 in the semantic segmentation of Gleason grades in the test cohort. Our proposed model outperformed the supervised SupVGG19 model segmentation ($\kappa = 0.2630$, $mIoU = 0.3497$ and $F1 = 0.4613$), and

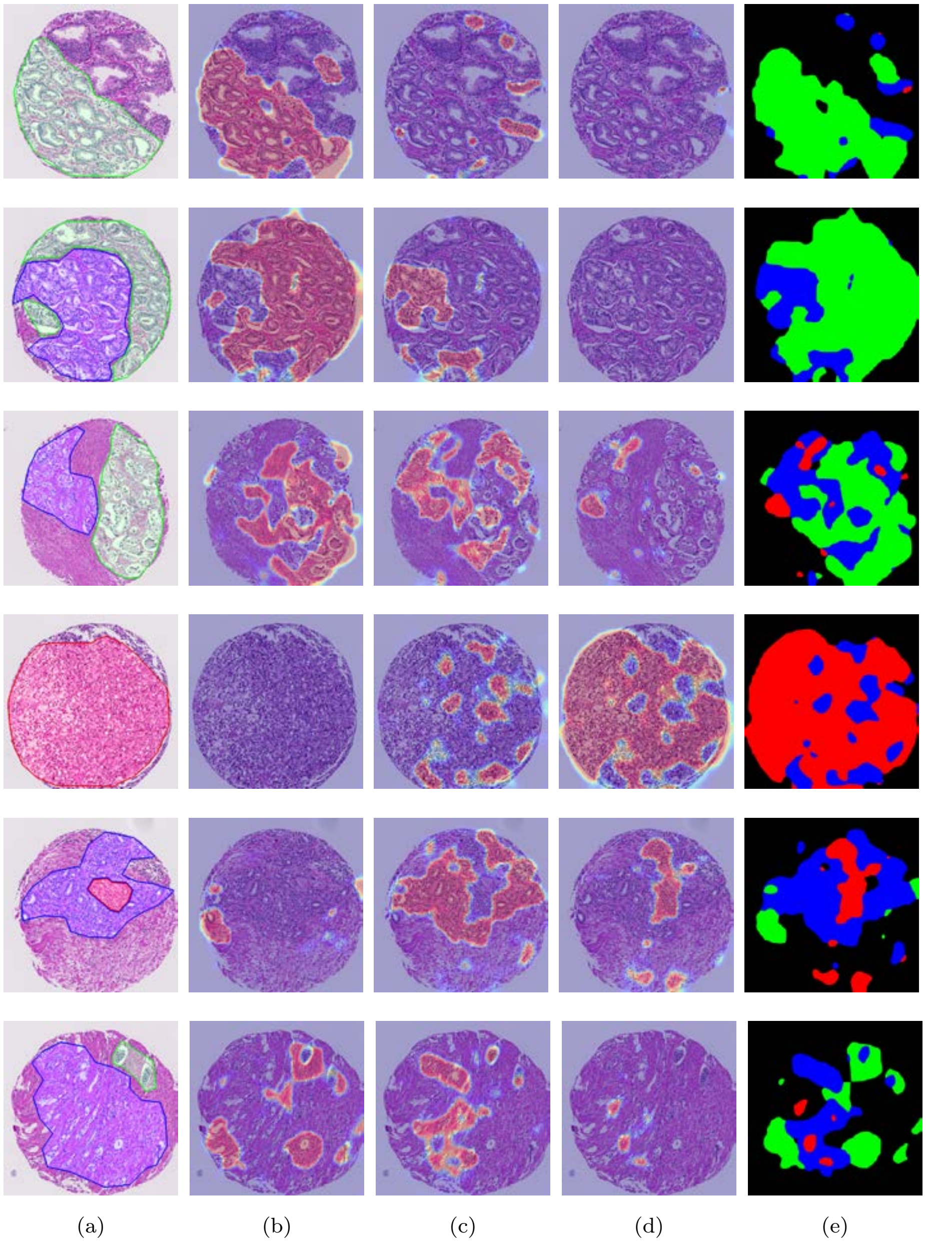


Figure 5: Examples of the proposed weakly-supervised model, WeGleNet, segmentation performance in the test set. The reference annotations are obtained from Pathologist 1. In green: Gleason grade 3; blue: Gleason grade 4 and red: Gleason grade 5. (a): Reference; (b): Gleason grade 3; (c): Gleason grade 4; (d): Gleason grade 5; (e): Semantic segmentation mask. The reference and predicted Gleason scores (Pathologist 1 - Pathologist 2 - Predicted), from top to bottom, are: (6 - 6 - 6); (6 - 7 - 7); (7 - 7 - 8); (10 - 10 - 10); (8 - 9 - 8) and (7 - 7 - 7).

Table 1: Results of the Gleason grades semantic segmentation using the proposed weakly-supervised model, WeGleNet, and two supervised approaches, SupVGG19 and UNet. The metrics presented are the accuracy (ACC), the F1-Score (F1), computed per class and its average, the mean intersection over union ($mIoU$) and the Cohen's quadratic kappa (κ).

Experiment	**ACC**	**F1**					**mIoU**	κ
		NC	GG3	GG4	GG5	Avg.		
WeGleNet - LSE_{r8}	0.6859	0.8155	0.5883	0.5622	**0.3531**	**0.5798**	**0.4368**	0.6105
SupVGG19	0.5426	0.6747	0.5195	0.4959	0.1551	0.4613	0.3497	0.2630
UNet	**0.6968**	**0.8383**	**0.5932**	**0.5737**	0.2419	0.5618	0.4178	**0.6387**

it performs similarly to the UNet model ($\kappa = 0.6387$, $mIoU = 0.4178$ and $F1 = 0.5618$). Although the UNet model reached better results in the non-cancerous class ($F1 = 0.8383$), WeGleNet - LSE_{r8} differentiated better the Gleason grades, reaching an F1 of 0.3531 for the GG5 class, a challenging task due to the low prevalence of these patterns. Thus, our proposed WeGleNet model performed at a level equivalent to supervised methods in the segmentation of Gleason grades, without requiring pixel-level annotations.

5.4. State-of-the-Art Comparison

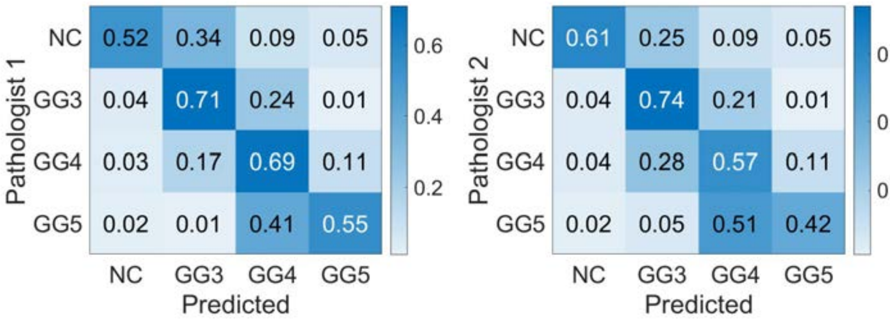
Finally, predictions were obtained at patch-level (which extraction is specified in Section 5.1) to compare WeGleNet against previous works in the used dataset. In the test cohort, patch-level classifications were obtained by majority voting of pixel-level predictions. Only fully non-cancerous patches were predicted as benign. The Cohen's quadratic kappa (κ) was obtained using the annotations of both pathologists. The figures of merit are presented in Table 2 and confusion matrices are presented in Figure 6.

Then, the global Gleason scoring of the cores was performed as described in Section 4.4. The parameters $c = 0.03$ and $d = 0.70$ were empirically fixed using the validation set. The κ and confusion matrices were obtained using as reference both pathologists, and the results are reported in Table 2 and Figure 7, respectively. Moreover, the obtained Gleason Score and references of representative cores are indicated in Figure 5.

In order to compare the obtained figures of merit with previous literature, the reported results for the patch-level grading and global scoring obtained using fully-supervised models with pixel-level annotations by Arvaniti et al. [28] are indicated in Table 2. Also, the results obtained in this test set by Bulten et al. [29] using semi-supervised models trained in a large set of biopsies (see Section 2.3 for a more detailed description) are pointed out in that table.

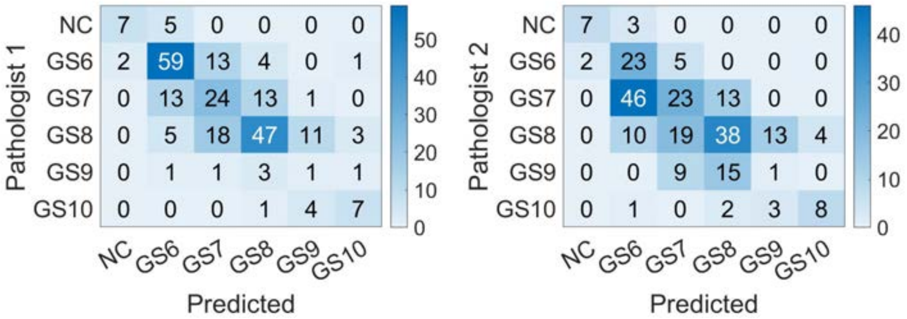
Table 2: Results of the patch-level Gleason grading and core-level scoring of the proposed model and comparison with previous literature. The metric presented is the Cohen's quadratic kappa (κ).

Approach	κ	
	Pathologist 1	**Pathologist 2**
Patch-Level Grading		
WeGleNet	0.59	0.50
Arvaniti et. al (2018) [28]	0.55	0.49
Pathologist 2	0.65	–
Core-Level Scoring		
WeGleNet	0.76	0.67
Arvaniti et. al (2018) [28]	0.75	0.71
Bulten et al. (2020) [29]	0.72	0.70
Pathologist 2	0.71	–



(a) (b)

Figure 6: Confusion Matrix of the patch-level Gleason grades prediction done by WeGleNet - LSE_{r8} network in the test subset. The reference labels in each matrix are obtained from: (a) pathologist 1, and (b) pathologist 2. GG: Gleason grade; NC: non cancerous.



(a) (b)

Figure 7: Confusion Matrix of the global-level Gleason scores prediction done by WeGleNet network in the test subset. The reference labels in each matrix are obtained from: (a) pathologist 1, and (b) pathologist 2.

The obtained results are in line with our previous experiments, and WeGleNet performed comparably to the fully-supervised approach used by Arvaniti et. al [28]. We reached a better κ value ($\kappa = 0.59$ against $\kappa = 0.53$) with the first pathologist, and similar performance was observed using the annotations from the second pathologist ($\kappa = 0.50$ against $\kappa = 0.49$). In addition, Figure 6 showed that most of the errors were conducted between adjacent classes.

Regarding the core-level Gleason scoring, the performance was also similar to previous works in the test set. A κ of 0.76 and 0.67 was obtained with each pathologist, respectively. In average, the obtained κ (0.715) is similar to the one obtained by Arvaniti et al. (0.730) and Bulten et al. (0.719). These results are at the same level of inter-pathologist agreement ($k = 0.710$). In addition, our approach obtained accurate localization heat-maps validated in Section 5.3 without using pixel-level annotations during training.

6. Conclusions

In this work, we have presented WeGleNet, a weakly-supervised trained architecture able to obtain semantic segmentation maps of Gleason grades in prostate histology images. The model is trained using just global-level labels, the Gleason score obtained from medical history, and it is capable of locating the local cancerous patterns in the tissue according to its grade.

Our proposed architecture makes use of multi-class segmentation layers after the feature-extraction stage, and a global-aggregation of the pixel-level probabilities into one representative value per class. Then, the output of the non-cancerous class (background) was sliced to obtain the loss of the model during training. This strategy allows us to obtain complementary maps in the architecture, without requiring complex post-processing of the output. In the experimental stage, we compared different global-aggregation layers and regularization techniques to optimize the model performance in the validation cohort. The log-sum-exponential pooling (LSE) showed superior performance than other layers, thanks to its ability to adapt the model to the specific domain via the adjustable parameter r. Thus, we have achieved a Cohen's quadratic kappa (κ) of 0.67 for the Gleason grading of local patterns in the validation cohort at the pixel level. During this optimization stage, we have observed a high correlation between global and local-level figures of merit. Thus, optimizing the proposed architecture using just global-labels involves improving the local-level localization of cancerous patterns. Additionally, we have compared the model performance with state-of-the-art supervised methods for semantic segmentation of Gleason grades in the test cohort. The proposed WeGleNet architecture performed similarly to supervised methods, without requiring any kind of pixel-level annotations during the training stage, reaching a pixel-level k of 0.61 and an average f1-score of 0.58. The performance for the core-level Gleason scoring was similar to previous works, and comparable to inter-pathologist agreement in the test cohort, reaching an average κ of 0.715. These promising results constitute a step forward in the literature of the analysis of prostate histology images and could avoid the tedious process of pixel-level generation of ground truth by expert pathologists.

Further research will focus on generalizing the proposed method to be trained using entire slices of biopsies digitized as whole slide images, whose larger size presents an added challenge in developing weakly-supervised methods for locating local cancerous patterns.

References

[1] World Cancer Research Foundation, Prostate cancer statistics (2019).
URL http://www.wcrf.org

[2] World Health Organization, Global cancer observatory (2019).
URL http://gco.iarc.fr

[3] D. F. Gleason, Histologic grading of prostate cancer: A perspective, human pathology (1992).

[4] M. Burchardt, R. Engers, M. Müller, T. Burchardt, R. Willers, J. I. Epstein, R. Ackermann, H. E. Gabbert, A. De La Taille, M. A. Rubin, Interobserver reproducibility of Gleason grading: Evaluation using prostate cancer tissue microarrays, Journal of Cancer Research and Clinical Oncology 134 (10) (2008) 1071–1078. doi:10.1007/s00432-008-0388-0.

[5] D. Komura, S. Ishikawa, Machine Learning Methods for Histopathological Image Analysis, Computational and Structural Biotechnology Journal 16 (2018) 34–42. doi:10.1016/j.csbj.2018.01.001.
URL https://doi.org/10.1016/j.csbj.2018.01.001

[6] M. A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognition 77 (2018) 329–353. doi:10.1016/j.patcog.2017.10.009.

[7] M. Oquab, Bottou, Is object localization for free?, Openaccess.Thecvf.Com (iii).

[8] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014) 1717–1724doi:10.1109/CVPR.2014.222.

[9] T. Durand, T. Mordan, N. Thome, M. Cord, WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua (1) (2017) 5957–5966. doi:10.1109/CVPR.2017.631.

[10] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, 35th International Conference on Machine Learning, ICML 2018 5 (Mil) (2018) 3376–3391.

[11] J. Li, W. Li, A. Gertych, B. S. Knudsen, W. Speier, C. W. Arnold, An attention-based multi-resolution model for prostate whole slide imageclassification and localization, ArXiv.
URL http://arxiv.org/abs/1905.13208

[12] J. Lee, E. Kim, S. Lee, J. Lee, S. Yoon, Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference, Pro-

ceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June (2019) 5262–5271. doi:10.1109/CVPR.2019.00541.

[13] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June (2019) 2204–2213. arXiv:1904.05044, doi:10.1109/CVPR.2019.00231.

[14] G. Papandreou, L. C. Chen, K. P. Murphy, A. L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, Proceedings of the IEEE International Conference on Computer Vision 2015 Inter (2015) 1742–1750. doi:10.1109/ICCV.2015.203.

[15] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, B. S. Manjunath, Weakly supervised localization using deep feature maps, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9905 LNCS (2016) 714–731. doi:10.1007/978-3-319-46448-0{_}43.

[16] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with Gaussian edge potentials, Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011 (2011) 1–9.

[17] L. Chan, M. S. Hosseini, K. N. Plataniotis, A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains, International Journal of Computer Vision 1. arXiv:1912.11186, doi:10.1007/s11263-020-01373-4.
URL https://doi.org/10.1007/s11263-020-01373-4

[18] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised

deep learning on whole slide images, Nature Medicine 25 (8) (2019) 1301–1309. doi:10.1038/s41591-019-0508-1.
URL http://dx.doi.org/10.1038/s41591-019-0508-1

[19] P. Courtiol, E. W. Tramel, M. Sanselme, G. Wainrib, Classification and Disease Localization in histopathology Using Only Global Labels: a Weakly Supervised Approach, ArXiv (2017) 1–13.

[20] E. Arvaniti, M. Claassen, Coupling weak and strong supervision for classification of prostate cancer histopathology images, ArXiv (Nips).
URL http://arxiv.org/abs/1811.07013

[21] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, P.-A. Heng, Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis, IEEE Transactions on Cybernetics (2019) 1–13doi:10.1109/tcyb.2019.2935141.

[22] Z. Jia, X. Huang, E. I. Chang, Y. Xu, Constrained Deep Weak Supervision for Histopathology Image Segmentation, IEEE Transactions on Medical Imaging 36 (11) (2017) 2376–2388. doi:10.1109/TMI.2017.2724070.

[23] J. Li, W. Speier, K. C. Ho, K. V. Sarma, An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies, Computerized Medical Imaging and Graphics 69 (2016) 125–133. doi:https://doi.org/10.1016/j.compmedimag.2018.08.003.

[24] Y. Xu, J. Y. Zhu, E. I. Chang, M. Lai, Z. Tu, Weakly supervised histopathology cancer image segmentation and classification (2014). doi:10.1016/j.media.2014.01.010.

[25] L. Chan, M. S. Hosseini, C. Rowsell, K. N. Plataniotis, S. Damaskinos, HistoSegNet : Semantic Segmentation of Histological Tissue Type in Whole Slide Images, IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 10662–10671doi:10.1109/ICCV.2019.01076.

[26] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, S. E. Salcudean, Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts (2018). doi:10.1016/j.media.2018.09.005.

[27] G. Nir, D. Karimi, S. L. Goldenberg, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, D. J. Thompson, P. C. Black, S. E. Salcudean, Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images, JAMA network open 2 (3) (2019) e190442. doi:10.1001/jamanetworkopen.2019.0442.

[28] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, M. Claassen, Automated Gleason grading of prostate cancer tissue microarrays via deep learning, Scientific Reports 8 (1) (2018) 1–11. doi:10.1038/s41598-018-30535-1.

[29] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, G. Litjens, Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study, The Lancet Oncology 21 (2) (2020) 233–241. doi:10.1016/S1470-2045(19)30739-9.
URL http://dx.doi.org/10.1016/S1470-2045(19)30739-9

[30] O. Jiménez del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, H. Müller, Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score, Medical Imaging 2017: Digital Pathology 10140 (2017) 101400O. doi:10.1117/12.2255710.

[31] S. Otálora, M. Atzori, V. Andrearczyk, A. Khan, H. Müller, Staining invariant features for improving generalization of deep convolutional neu-

ral networks in computational pathology, Frontiers in Bioengineering and Biotechnology 7 (AUG) (2019) 1–13. doi:10.3389/fbioe.2019.00198.

[32] 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, K. A. Iczkowski, J. G. Kench, G. Kristiansen, T. H. van der Kwast, K. R. Leite, J. K. McKenney, J. Oxley, C. C. Pan, H. Samaratunga, J. R. Srigley, H. Takahashi, T. Tsuzuki, M. Varma, M. Zhou, J. Lindberg, C. Lindskog, P. Ruusuvuori, C. Wählby, H. Grönberg, M. Rantalainen, L. Egevad, M. Eklund, Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, The Lancet Oncology 21 (2) (2020) 222–232. doi:10.1016/S1470-2045(19)30738-7.

[33] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations 1 (2014) 1–14.
URL http://arxiv.org/abs/1409.1556

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, 2009 IEEE Conference on Computer Vision and Pattern Recognitiondoi:10.1109/CVPR.2009.5206848.
URL http://arxiv.org/abs/1409.1556

[35] K. K. Singh, Y. J. Lee, Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization, Proceedings of the IEEE International Conference on Computer Vision 2017-Octob (2017) 3544–3553. doi:10.1109/ICCV.2017.381.

[36] P. O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with Convolutional Networks, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June (2015) 1713–1721. doi:10.1109/CVPR.2015.7298780.

[37] M. Lin, Q. Chen, S. Yan, Network In Network, International Conference of

Learning Representations (2014) 1–10.
URL http://arxiv.org/abs/1312.4400

[38] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351 (2015) 234–241. doi:10.1007/978-3-319-24574-4{_}28.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem (2016) 770–778. doi:10.1109/CVPR.2016.90.

[40] Z. Zhang, Q. Liu, Y. Wang, Road Extraction by Deep Residual U-Net, IEEE Geoscience and Remote Sensing Letters 15 (5) (2018) 749–753. doi:10.1109/LGRS.2018.2802944.

[41] A. Kalapahar, J. Silva-Rodríguez, A. Colomer, F. López-Mir, V. Naranjo, Gleason grading of histology prostate images through semantic segmentation via residual u-net, 2020 IEEE International Conference on Image Processing (ICIP) (2020) 2501–2505doi:10.1109/ICIP40778.2020.9191250.

[42] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, Psychological Bulletin 70 (4) (1968) 213–220. doi:10.1037/h0026256.