Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Proportion constrained weakly supervised histopathology image classification

Julio Silva-Rodríguez [a,*], Arne Schmidt [b], María A. Sales [c], Rafael Molina [b], Valery Naranjo [d]

[a] Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain
[b] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[c] Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain
[d] Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Keywords:
Multiple instance learning
Histology
Proportion
Inequality constraints
Extended log-barrier

ABSTRACT

Multiple instance learning (MIL) deals with data grouped into bags of instances, of which only the global information is known. In recent years, this weakly supervised learning paradigm has become very popular in histological image analysis because it alleviates the burden of labeling all cancerous regions of large Whole Slide Images (WSIs) in detail. However, these methods require large datasets to perform properly, and many approaches only focus on simple binary classification. This often does not match the real-world problems where multi-label settings are frequent and possible constraints must be taken into account. In this work, we propose a novel multi-label MIL formulation based on inequality constraints that is able to incorporate prior knowledge about instance proportions. Our method has a theoretical foundation in optimization with log-barrier extensions, applied to bag-level class proportions. This encourages the model to respect the proportion ordering during training. Extensive experiments on a new public dataset of prostate cancer WSIs analysis, SICAP-MIL, demonstrate that using the prior proportion information we can achieve instance-level results similar to supervised methods on datasets of similar size. In comparison with prior MIL settings, our method allows for $\sim 13\%$ improvements in instance-level accuracy, and $\sim 3\%$ in the multi-label mean area under the ROC curve at the bag-level.

1. Introduction

In the supervised learning paradigm, deep learning methods have shown promising performance in a wide range of medical imaging applications. Nevertheless, these methods usually require large amount of data for training, which must be labeled by expert clinicians. Obtaining these labeled datasets is a time-consuming process and is susceptible to inter-annotator variability, which complicates the use of these models in practice. This is the case for histology image analysis, whose large size of tissue images magnified on whole slide images (WSIs), patterns heterogeneity, and the high level of expertise required to annotate the data make this learning paradigm unfeasible. Considering these limitations, the most popular choice in this field has become the use of weakly supervised learning strategies under the multiple instance learning (MIL) paradigm. In particular, typically the training dataset is composed of bags (WSIs) that are known to have cancer or not. Each bag consists of instances (tissue tiles), of which the label is not accessible during training. Under this setting, different works have demonstrated outstanding results for both WSI-level cancer detection [1] and instance-level cancer localization [2]. Nevertheless, these methods require very large datasets (i.e. thousands of biopsies) to compensate for the absence of greater supervision. One common limitation is that these methods tend to focus on only a limited number of instances of each bag during training. Very recent literature has resort to instance-dropout [3] during training to alleviate this issue. Despite the improvement it produces, this solution does not involve classifying more positive instances systematically, but depends on the samples randomly discarded in the dropout, without prior knowledge. To improve the performance of MIL models with the help of prior knowledge, constraint deep learning has been proposed using previously estimated tumor size [4] to guide the weakly supervised optimization. Although this method shows a promising performance, in this case the tumor size estimation is a tedious task, which can be as costly as performing instance-level annotations. All these limitations are accentuated in the multi-label scenario, where it is desired to differentiate between different types of tissues, which may coincide in the same bag. In

[*] Corresponding author.
E-mail addresses: jjsilva@upv.es (J. Silva-Rodríguez), arne@decsai.ugr.es (A. Schmidt), salesman@gva.es (M.A. Sales), rms@decsai.ugr.es (R. Molina), vnaranjo@dcom.upv.es (V. Naranjo).

https://doi.org/10.1016/j.compbiomed.2022.105714
Received 7 February 2022; Received in revised form 14 May 2022; Accepted 4 June 2022
Available online 10 June 2022

contrast to the binary scenario classification, multi-label MIL literature still remains scarce in histology image analysis [5].

Based on these observations, we propose a novel formulation for MIL in the multi-label scenario, applied to histology prostate cancer grading in WSIs. The key contributions of our work can be summarized as follows:

- A novel constrained formulation for instance-level MIL, which integrates an auxiliary term that forces to increase the number of instances classified on positive classes.
- In addition, our formulation leverages prior knowledge in terms of relative tissue proportions (i.e. primary cancerous grade in the WSI) by imposing inequality constraints on bag(WSI)-level class proportions.
- We benchmark the proposed model against a relevant body of literature on SICAP-MIL, a new publicly available dataset containing 350 prostate WSIs with global labels, as well as instance-level labels to test weakly-supervised methods on tumor localization.
- Comprehensive experiments demonstrate the superior performance of our model. By simply incorporating relative proportion information during training (easily accessible from medical records in many cancer types) we found improvements of nearly $\sim$ 3% in mean AUC for bag-level classification and $\sim$ 13% for instance-level cancer grading accuracy compared to prior MIL methods.

2. Related work

2.1. Multiple instance learning

In computer vision, multiple instance learning (MIL) is a learning paradigm that works with independent images (instances) that form groups (bags), and only bag-level information is known. In the multi-label scenario, each instance belongs to one class, but different classes could coincide at bag level [6]. Modern MIL methods using convolutional neural networks (CNNs) for feature extraction usually process each instance independently, and then combine the instance-level information into one bag-level output. Methods that combine instance-level features are known as embedding-based, which require a subsequent classification layer. In contrast, instance-based architectures combine directly instance-level predictions into the bag classification. Beyond the basic mean and maximum aggregation functions, recent methods have proposed the use of weighted-averaged embeddings, using instance-specific attention weights learned via a multi-layered perceptron projection [7] or recurrent neural networks [1]. It is noteworthy to mention that, although embedding-based approaches have yielded slightly better bag-level results in previous literature, they do not provide instance-level probability outputs. In this work, we are interested in both: instance and bag-level classification. Since we aim to include prior knowledge referred to class-wise proportions, our proposed method follows the instance-based learning paradigm.

2.2. Constrained classification

Constrained classification aims to guide the training of a CNNs towards a solution that satisfies a given condition, which takes advantage of additional knowledge to the main labels. This learning paradigm has gained popularity on weakly supervised scenarios (e.g. weakly supervised segmentation or MIL), since it allows to incorporate local information to the global annotations. In a usual constraint weakly supervised setting, an additional loss term enforces the sum of the instance-level predictions to match a given proportion using an L_2 penalty [8]. Similarly, it has been applied in unsupervised anomaly segmentation, to force attention maps to focus on all patterns of training images [9], or in semi-supervised learning, to match the predicted size distributions to the ones observed in the supervised subset using a KL-divergence term [10]. While the aforementioned equality-constrained formulations proposed in weakly supervised settings are very promising, they demand exact knowledge of the prior. For instance, in the case of histology tumor grading, this would require to know the cancerous tissue proportion extent. Therefore, recent works have preferred the use of inequality constraints to relax the prior assumptions, allowing more flexibility. This approach allows, for example, to set some tolerance margins on target size using L_2 penalties [11,12], or Lagrangian optimization [13]. Following the example above, these works would require approximate knowledge of tumor size, and a tolerance margin would be applied to smooth the constraint. Unlike these works on weakly supervised classification, our formulation does not require prior information on the absolute size of the target. In contrast, we seek to constrain the training to account for relative relationships between proportions within the same global image. In the case of histological whole slide image classification in a multi-label setting, this formulation incorporates information about which tumor grade is in the majority (primary) and which is in the minority (secondary), so that the proportion of the primary grade must be greater than that of the secondary grade. Thus, we use inequality constraints to (i) encourage classification of instances to positive classes at the bag level, and (ii) incorporate relative relationships between class proportions within bags.

3. Methods

An overview of our proposed method is depicted in Fig. 1. In the following, we describe the problem formulation, and each of the proposed components.

Problem formulation. In the paradigm of Multiple Instance Learning (MIL), instances are grouped in bags of instances $X = \{x_n\}_{n=1}^{N}$, that exhibit neither dependency nor ordering among them, and its number N is arbitrary for each bag. In the multi-label scenario, there are multiple labels per bag, $Y = (Y_1, \dots, Y_k, \dots, Y_K)$, where $k \in \{1, \dots, K\}$ denotes each one of the K categories. Also, individual labels, $y_{n,k} \in \{0, 1\}$, exist for each instance in the bags, but they remain unknown during training. In the standard MIL formulation, a bag label is considered positive if at least one instance in the bag is positive for that category. We can rewrite this assumption in the following forms:

$$Y_k = \begin{cases} 1, & \text{iff } \sum_n y_{n,k} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\equiv \quad Y_k = \max_n \{y_{n,k}\} \tag{2}$$

Instance-based MIL. In this work, we aim to training a model capable of extracting both: instance and bag-level labels, which falls into the instance-based MIL paradigm.[1] Let us denote a neural network model, $f_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{H}^{K+1}$, parameterized by θ, which processes instances $x \in \mathcal{X}$ to predict softmax instance-level class scores, $\{h_k\}_{k=0}^{K} \in \mathcal{H}$, such that $\mathcal{H} \in [0, 1]$. Note that $k = 0$ represents a category for instances negative at all classes. Also, we use a parameter free aggregation function, $f_a(\cdot)$, in charge of pooling the instance-level scores into one global score $H = (H_1, \dots, H_k, \dots, H_K)$ such that $H = f_a(\{f_\theta(x_n)\}_{n=1}^{N})$. Then, the optimization of θ is driven by the minimization of cross entropy loss between reference and predicted bag-level score.

$$\mathcal{L}_{ce} = -\frac{1}{K} \sum_{k=1}^{K} Y_k log(H_k) + (1 - Y_k) log(1 - H_k) \tag{3}$$

[1] Based on the denomination proposed in [7]

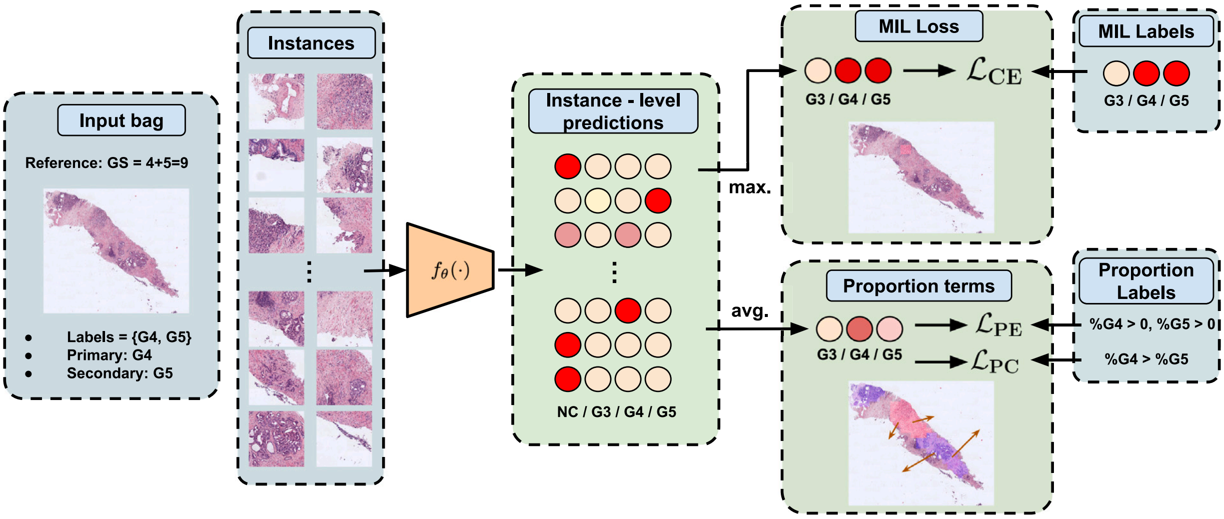


Fig. 1. Method overview. In this work, we face weakly supervised histology image classification under the Multiple Instance Learning (MIL) paradigm. Each biopsy is a bag, while its patches are the instances conforming it. In the case of prostate analysis, expert labels are conformed by the Gleason score, that are the sum of the two most predominant tumor grades (i.e. G3, G4 or G5). In order to extract both instance and bag-level labels, an standard instance-level MIL with max aggregation is trained via cross-entropy loss, $\mathcal{L}_{ce}$ (see Eq. (3)). Then, prior information is incorporated via inequality constraints that (i) force the classifier to predict instances that are present in the biopsy ($\mathcal{L}_{PE}$, see Eq. (5)), and (ii) ensure that the proportion of the primary grade is superior than the secondary grade ($\mathcal{L}_{PC}$, see Eq. (7)). Colored tissue indicates: blue: Gleason grade 4; red: Gleason grade 5. Circles in instance-level predictions indicate soft-max scores, $y_{n,k}$. The more intense the color, the higher the score.

3.1. Inequality constraints for MIL

Previous literature on instance-level MIL have proposed aggregation functions $f_a(\cdot)$ based on mean or maximum operator. The second solution is used based on the direct interpretation of maximum operation on MIL formulation (Eq. (2)). Nevertheless, training a neural network via this aggregation produces well-known problems such as gradient vanishing of non-maximum instances. This limitation produces the network to focus only on discriminative instances during training, which leads to poor generalization performance on unseen samples. To alleviate this issue, we focus on the MIL formulation in Eq. (1), which interpretates a positive bag via an inequality that forces the sum of instances scores to be greater than zero. In this line, we incorporate to the base instance-based MIL training a term that increases the proportion of positive instances classification for a given class k, $p_k = \frac{1}{N}\sum_n^N h_{n,k}$, by minimizing $-\lambda log(p_k)$. Nevertheless, this log-term is non-differentiable when $p_k \rightarrow 0$. To solve this limitation we resort to a smooth, duality-gap bound approximation. Concretely, we use the formulation proposed in [13] on constrained optimization that models inequality constraints using the approximation of log-barrier that is formally defined as:

$$\widetilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(z) & \text{if } z \geq \frac{1}{t^2} \\ -tz - \frac{1}{t}\log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise,} \end{cases} \tag{4}$$

where *t controls* the barrier during training, and z is the objective term.

This log barrier extension is applied on the proportion term p_k of the bags that are positive for the class k at bag level (i.e. $Y_k = 1$). It is noteworthy to mention that this proportion is the objective term z in Eq. (4). Hereafter, we refer to this term as positives expansion (PE) constraint.

$$\mathcal{L}_{PE} = \sum_{k:Y_k=1} \widetilde{\psi}_{t_{PE}}(p_k) \tag{5}$$

Thus, we propose a MIL loss that combines the maximum formulation in Eq. (2) via the aggregation function $f_a(\cdot) = \max_n\{y_{n,k}\}$, and the PE term as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{PE}\mathcal{L}_{PE} \tag{6}$$

where $\lambda_{PE} \in \mathbb{R}^+$ weights the importance of each term during training. Note that the positives expansion term, $\mathcal{L}_{PE}$, is only applied for those positive categories at bag-level.

3.2. Incorporating proportion information

In some applications, prior knowledge of the bags is known. In this work, we focus on an information usually recorded on medical domains: data regarding the proportion of categories in the image (i.e. primary or secondary tumor grades in the tissue). This information can be formulated as an inequality constraint between categories proportions such that: $p_{k'} > p_{k''}$, where k' denotes the larger proportion category, and k'' its respective counterpart. Note that this relation can be established between any pair of positive categories in the bag for which we have this information available. Thus, we contemplate an arbitrary number of conditions I for each bag, which could give complete or partial information (i.e. the formulation could be applied for only few known inequalities). For each condition i, both major (k') and minor (k'') categories should be indicated. Again, we make use of extended log-barrier (see Eq. (4)) to solve this inequality constraint, which has demonstrated good performance when multiple constraints are used [13]. In this case, the objective term z in Eq. (4) is the different between major and minor proportions in a given bag: $(p_{k'_i} - p_{k''_i})$. Hereafter, we refer to this additional term as proportion constraint (PC).

$$\mathcal{L}_{PC} = \sum_{i}^{I_b} \widetilde{\psi}_{t_{PC}}(p_{b,k'_i} - p_{b,k''_i}) \tag{7}$$

where b indicates the bag index over the complete dataset, $\lambda_{PC} \in \mathbb{R}^+$ weights the relative importance of the proportion term during training, t_{PC} controls the barrier slope over time. It is noteworthy to mention that the proportion term is not taken into account for bags with only one positive category, or which the proportion information is unknown.

Taking into account the different terms previously detailed, θ is trained to solve the multi-label MIL formulation using the following optimization criteria via standard Gradient Descent:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{PE}\mathcal{L}_{PE} + \lambda_{PC}\mathcal{L}_{PC} \tag{8}$$

Table 1
SICAL-MIL dataset. Whole slide images partition and Gleason scores (GS) distribution. NC: non-cancerous.

Partition	NC	GS6	GS7	GS8	GS9	GS10	Total
Train	77	10	61	7	25	8	188
Validation	19	2	26	5	10	2	64
Test	17	9	28	13	27	4	98
Total	111	21	115	25	62	14	350

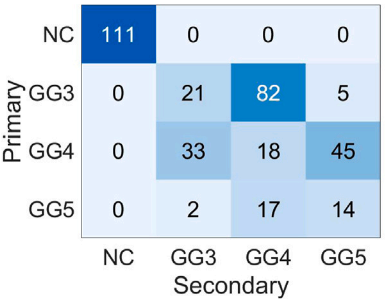


Fig. 2. SICAP-MIL dataset description. The confusion matrix shows the distribution of global labels in terms of primary and secondary Gleason grades per Whole Slide Image. GG: Gleason grade. NC: non-cancerous.

4. Experiments and results

4.1. Experimental setting

Datasets. In this work, we present a new dataset for prostate histological image analysis: SICAP-MIL.[2] This dataset is an extension of the previously published SICAP versions [14,15], which is expanded with 168 new WSIs. The dataset introduced is composed of 350 WSIs from 271 patients. The samples were digitized using the Ventana iScan Coreo scanner at $40x$ magnification. The slides were analyzed by a group of expert urogenital pathologists at Hospital Clínico of Valencia, and a combined Gleason score (GS) was assigned per biopsy. The Gleason score is the sum of the two main (primary and secondary) Gleason grades (GG) in the biopsy regarding its extent and severity. The clinical report specifies both the score and the primary and secondary grades that constitute the score. SICAP-MIL is specially design to serve as a benchmark for MIL methods. Each WSI is considered as a bag, from which instances are obtained by tiling the images using non-overlapped moving-windows of 512^2 pixels at 10× of resolution level. Note that tiles with less than 20% of tissue were excluded. The dataset is divided into three class-wise balanced groups for training, validation and testing. A summary of the dataset in terms of the labeled Gleason scores and proposed partitions is presented in Table 1

From the WSI-level Gleason scores, bag-level labels referred to the presence of each Gleason grade in the WSI are inferred. Also, the relative-proportion information of the primary and secondary grades is obtained from this score. We show in Fig. 2 the information regarding the primary and secondary Gleason grades for each WSI. It is observed that most cases present at least two tumor types, and thus two proportion expansion (PE) constraints and one proportion constraint (PC) in the proposed formulation. Also, the difficulty of training a classifier capable of distinguishing between different Gleason grades in a weakly supervised manner is appreciated, since the biopsy rarely presents a single tumor type.

Table 2
Datasets with patch-level Gleason grade annotations used for testing. Distribution of the patches among non-cancerous (NC) and the different Gleason grades (GG).

Partition	NC	GG3	GG4	GG5
Test	448	289	632	132

In addition, SICAP-MIL includes instance-level annotations, which allow to test the capability of MIL methods to leverage instance classifications in a weakly-supervised manner. To do so, annotated WSIs are kept into the test subset. Note that instance-level labels are obtained from pixel-level annotations done by expert pathologist. Non-cancerous patches are obtained only from benign WSIs, while cancerous patch-level labels are obtained by majority voting of segmentation masks. The distribution of instance-level annotated subset from the test cohort is presented in Table 2.

Implementation details. The proposed methods were trained using the train subset from SICAP-MIL. The backbone $f_\theta(\cdot)$ used was a VGG16 [16] pre-trained on Imagenet [17], which takes as input instances resized to 224 × 224 images. First, the PE setting was trained by empirically fixing $\lambda_{PE} = 0.1$ and $t_{PE} = 15$. Training was carried out during 100 epochs using a batch size of 1 bag and the SGD optimizer with a learning rate $\eta = 1 \cdot 10^{-2}$. After 50 epochs, η is decreased in a factor to 10×. During training, bag-level mAUC is monitored in the validation set, and early stopping is applied if this figure of merit does not improve during 20 epochs. Then, the PC formulation is trained keeping constant the PE hyperparameters, and empirically setting $\lambda_{PC} = 1$ and $t_{PC} = 5$. The training is carried out using the same training conditions as the PE setting. Nevertheless, instead of using mAUC from validation subset as early stopping criterion, we use the average proportion constraint satisfaction, $z = p_{b,k'_i} - p_{b,k''_i}$ in Eq. (7) from the training set to determine the best model. The hyperparameters and early stopping criterion used are further justified by means of ablation experiments. The code and trained models are publicly available on https://github.com/jusiro/mil_histology.

Instance-level student. In this work, we complement the proposed models for instance-level prediction with a second model, Student, trained with instance-level hard pseudo-labels as described in [2]. This second stage has demonstrated to increase model performance without any modification of the architecture as described in [2]. Note that we use as Teacher any trained instance-level classifier $f_\theta(\cdot)$ under the MIL paradigm with the proposed methodology. A Student model with the same complexity as the Teacher is trained following the Noisy Student paradigm on semi-supervised learning [18]. Concretely, a dropout rate of 0.20 is applied over the instance embedding, and data augmentation is applied to all instances using random rotations, translations, Gaussian blur and color jittery. Student is trained during 60 epochs with mini-batches of 32 images using SGD optimizer and a learning rate of $\eta = 1 \cdot 10^{-2}$.

Baselines. With the aim of comparing our approach to state-of-the-art methods, we implemented and tested prior methodologies on MIL for both instance-level and bag-level classification on SICAP-MIL dataset. **Instance-based MIL**. First, we compare our method with other instance-based MIL aggregation. Concretely, we use basic mean and max operations over the instance-level predictions to obtain the bag-level prediction. **Embedding-based MIL**. Secondly, we included embedding-based methods, which aim to obtain a bag-level embedding, on which a classifier is trained to predict bag-level labels. Aggregation methods of instance-level features include mean, max, attention mechanism, and recurrent neural networks (RNN). AttentionMIL [7] aims to obtain a weighted feature representation, which highlights positive instances in the bag. The weights are obtained using a multi-layered perceptron as detailed in [7]. We implemented the gated attention mechanism with

[2] SICAP-MIL is available at https://cvblab.synology.me/PublicDatabases/SICAP_MIL.zip or on the GitHub repository of the project: https://github.com/jusiro/mil_histology

Table 3
Quantitative comparison to prior literature at instance level on SICAP-MIL dataset. Results derived from the proposed methods in gray. Best results in bold. NC: non-cancerous; GG: Gleason grade; κ: Cohen's quadratic kappa.

Method	Acc	F1-score					κ
		NC	GG3	GG4	GG5	Avg.	
mean	0.458	0.312	0.383	0.548	0.411	0.413	0.431
max	0.484	0.604	0.295	0.411	0.199	0.377	0.262
max (Student) [2]	0.573	0.716	0.398	0.529	0.320	0.490	0.454
max - w. PE	0.535	0.644	0.259	0.533	0.217	0.413	0.296
max - w. PE (Student)	0.610	0.748	0.302	0.616	0.341	0.502	0.481
max - w. PE w. PC	0.639	0.706	0.686	0.611	0.309	0.578	0.450
max - w. PE w. PC (Student)	**0.705**	**0.818**	**0.692**	**0.691**	**0.417**	**0.655**	**0.655**

Table 4
Quantitative comparison to prior literature at instance level. Results derived from the proposed methods in gray. TMAs: tissue micro arrays; WSIs: whole slide images; NC: non-cancerous; GG: Gleason grade; κ: Cohen's quadratic kappa.

Method	Paradigm	Training Dataset		Acc	F1-score					κ
		TMAs	WSIs		NC	GG3	GG4	GG5	Avg.	
Arvaniti et al. [19] (2018)*	supervised	508	-	-	-	-	-	-	-	0.67/0.55
Nir et al. [20] (2019)*	supervised	333	-	-	-	-	-	-	-	0.61
Silva-Rodrguez et al. [15] (2020)	supervised	-	160	0.67	0.86	0.59	0.54	0.61	0.65	0.77
Otálora et al. [21] (2020)*	semi-supervised	508	171	-	-	-	-	-	-	0.59/0.55
Silva-Rodrguez et al. [2] (2021)	MIL	-	10, 000	0.797	0.901	0.714	0.798	0.601	0.754	0.830
max - w. PE w. PC (Student)	MIL	-	188	0.705	0.818	0.692	0.691	0.417	0.655	0.655

* Results reported on different patch size and resolutions, on private datasets.

an intermediate layer with $D = 128$ neurons. Campanella et al. [1] proposed a RNN based aggregation over the top-k positive instances of each bag to produce bag-level classifications. We increased $k = 10$ to support the multi-label scenario, and a RNN with a hidden state of 128 neurons was trained. All methods are train under the same training setup (i.e. backbone, learning rate, scheduler, batch size, etc.) as our baseline. Only the learning rate of the methods based on attention mechanisms was changed to $\eta = 1 \cdot 10^{-3}$. Note that embedding-based method do not make instance-level predictions, and is therefore only used as a comparison of the results at the bag level. Although attention-based methods include instance-level importance weights, these are not true predictions at the instance level, as they are sensitive to the number of instances in the bag.

Evaluation metrics. We evaluate the different models in this work using standard metrics on MIL for both instance and bag-level performance on the test subset. Concretely, for instance-level validation we obtain accuracy (Acc), and f1-score per class and micro-averaged. Also, as the Gleason grades constitutes a set of ordered classes, we obtain Cohen's quadratic kappa (κ) as figure of merit. Regarding the bag-level predictions, we evaluate them using the area under ROC curve (AUC). In the multi-label scenario, AUC is obtained class-wise, and it is averaged (mAUC). In order to facilitate the comparison of our methods with previous literature at the bag level, we also obtained the AUC for binary cancer vs. non-cancer detection by combining each class prediction and target via max-aggregation. For each experiment, the metrics shown are the mean of three consecutive repetitions (with its respective standard deviation) of the model training, to account for the variability of the stochastic factors in the process.

4.2. Results

Comparison to the literature. The quantitative results obtained by the proposed model and baselines on the test cohort are presented at instance level in Table 3, and at bag level in Table 5 and Fig. 3. Also, we include results reported in a relevant body of literature for both tasks, using different datasets and experimental settings for instance level in Table 4, and at bag level in Table 6.

Table 5
Quantitative comparison to prior literature at bag level in SICAP-MIL dataset. The metric presented is the Area Under ROC curve (AUC). Results derived from the proposed methods in gray. Best results in bold.

Method	Cancer Detection	Multilabel
Embedding + mean	0.952(0.013)	0.844(0.009)
Embedding + max	0.951(0.019)	0.834(0.002)
Embedding + RNN [1]	0.967(0.014)	0.855(0.011)
Embedding + AttentionMIL [7]	0.961(0.006)	0.848(0.007)
Instance + mean	0.701(0.090)	0.769(0.071)
Instance + max	0.955(0.012)	0.867(0.005)
Instance + max w. PE	0.962(0.009)	0.873(0.019)
Instance + max w. PE w. PC	**0.979**(**0.005**)	**0.899**(**0.007**)

Instance-level results. The proposed constrained formulation using a positive expansion constraint term (PE) to enhance positive instances prediction outperforms in $\sim$ 5% the accuracy for instance-level classification of max-aggregation baseline. Adding the Student stage, the model reaches an accuracy of 0.610, which outperforms on SICAP-MIL the Teacher–Student strategy using only max aggregation in [2]. The observed improvement could be caused by the larger number of instances classified using the inequality constraint, which avoids over-fitting the model to focus only on very discriminative instances. Note that, although still the results reported in [2] in prior literature are better, the training dataset required to accomplish these results is too large: around 10,000 WSIs. Once we introduce the proportion information in terms of primary and secondary classes in the bag via the proportion inequality constraint (PC), results reach an accuracy of 0.705 and average F1-score of 0.655. It is noteworthy to mention that these results are similar to the ones obtained in prior literature under full supervision on similar sized datasets [15,19–21]. Under our proposed formulation, the model is capable of grading cancerous patches at the same performance of using pixel-level annotated datasets, by providing only WSI-level information about the most abundant grade.

Bag-level results. Regarding the MIL bag-level results obtained, our PE formulation improved around $\sim$ 0.7% the baseline instance-based

Table 6
Quantitative comparison to prior literature at bag level. Results reported on different datasets, patch size and resolutions. The metric presented is the Area Under ROC curve (AUC). Results derived from the proposed methods in gray. WSI: whole slide image [22,23].

Method	Training WSIs	Cancer Detection	Multilabel
Campanella et al. [1] (2019)	24, 859	0.994	–
Ström et al. [22] (2020)	6, 682	0.997	–
Bulten et al. [23] (2020)	5, 759	0.990	–
Li et al. [3] (2021)	9, 638	0.982	–
max - w. PE w. PC (Student)	188	0.979(0.005)	0.899(0.007)

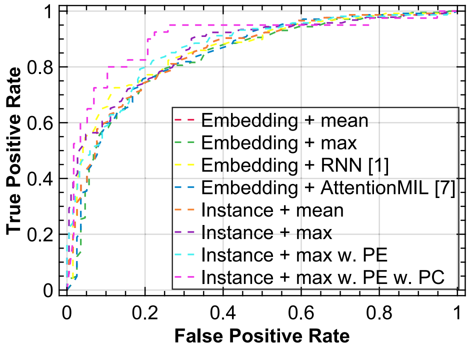
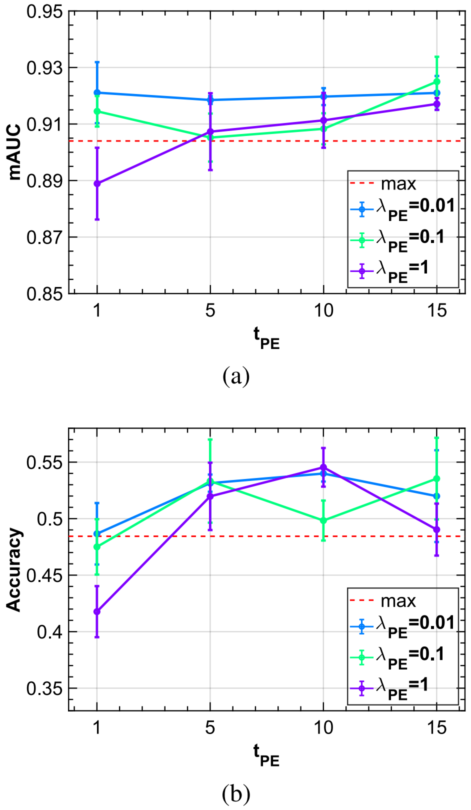


Fig. 3. Overall receiver operating characteristic (ROC) curves for the multilabel bag-level prediction of proposed methods and baselines on SICAP-MIL dataset.

maximum aggregation. This modest improvement may be due to the fact that, because of the maximum-based inference, it is only necessary to locate one positive sample to get the bag-level prediction right. These observations are in line with previous literature, which highlights that the best classifier at the bag level need not be the best classifier at the instance level [24]. Once we incorporate the proportion information during training, the proposed model increases the multilabel mAUC in $\sim 3.3\%$ from the baseline, and reaches mAUC of 0.899 in the multi-label scenario and 0.979 in the binary prediction (see Table 5). Note that this result almost reaches the ones reported in previous literature (see Table 6), which use thousands of WSIs during training. However, it is worth noting the limitations of this indirect comparison. The methods used in previous works may have different levels of supervision, and the datasets used are larger. Next, we perform a direct comparison of the weakly supervised methods in the database used in this work, SICAP-MIL (see Table 5). Specifically, we pay attention to embedding-based methods performance at bag level. The obtained results using mean and max aggregation are similar to the baseline instance-based max approach. However, in the multi-label scenario, these methods perform worse. Moreover, since they cannot provide instance-level labels, they cannot take advantage of the information referred to the proportion during training. It is notable that deep-learning based aggregation modules such as AttentionMIL or RNN do not perform properly in this training setting. This could be due to the complexity of having multiple classes in some bags, the over-fitting tendency of neural networks, and the incapacity of AttentionMIL to get class-specific attention weights. Finally, We would like to point out that a significant body of previous work validates multi-class methods at the bag level on the basis of Gleason scores. However, this score is beyond the scope of MIL. Its derivation involves a decision making according to the severity of the grades in the tissue by the clinical expert, which does not fit a proper formulation of MIL (see Eq. (1)), based on the presence of each class in the bags of instances.

Ablation studies. In the following, we provide comprehensive ablation experiments to validate several elements of our model, and motivate the choice of the values employed in our formulation, as well as our experimental setting.



(a)

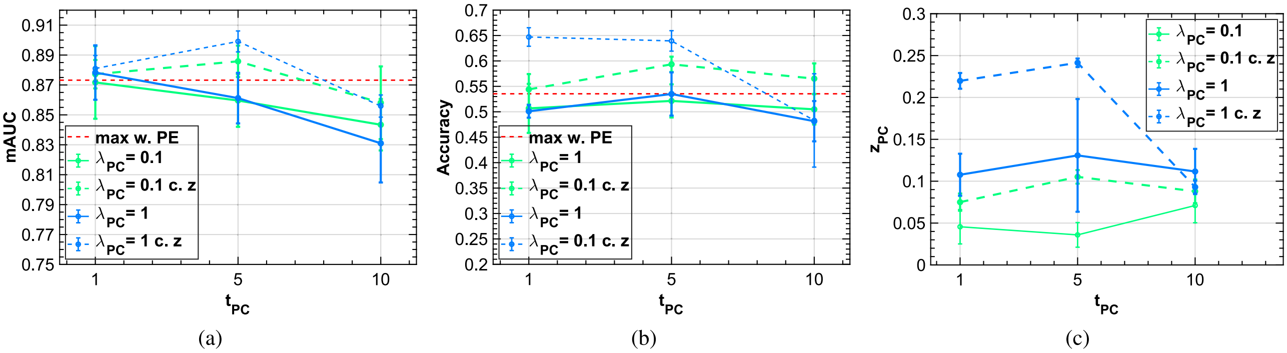
(b)

Fig. 4. Ablation studies on positive expansion (PE) MIL formulation. Hyperparameters study for λ_{PE} and t_{PE} are performed for bag-level mAUC on validation set (a), and instance-level accuracy (b).

First, we optimized the proposed formulation only with the inequality constraint term in Eq. (6). Using the training setting previously described, validated different values of $\lambda_{PE} = \{0.01, 0.1, 1\}$ and slopes of the log-barrier inequality $t_{PE} = \{1, 5, 10, 15\}$. Using the mAUC on validation subset as an early stopping criteria, we obtained bag-level mAUC from the validation subset and instance-level accuracy from the test cohort. Results are presented in Fig. 4. These show that the inclusion of the PE term improves both the performance at both bag-level and instance-level under most of the settings. Thus, we selected $t_{PE} = 15$ and $\lambda_{PE} = 0.1$, which led the best results at bag level in the validation cohort.

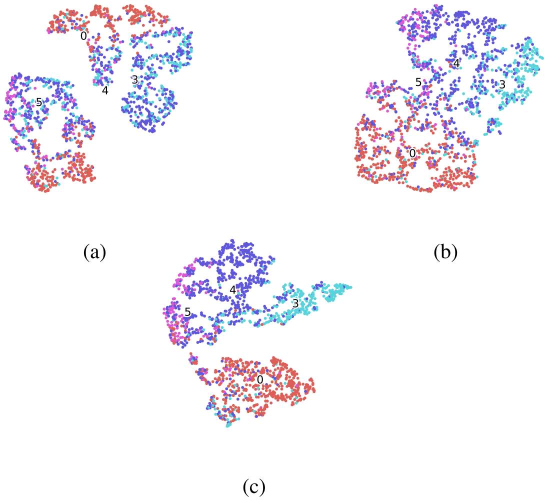
Then, using the best configuration reached for the PE term, we optimized the proportion constraint configuration (PC) in Eq. (8). During empirical experimentation, we appreciated that the instance-level model performance on the test subset did not always correlate with the bag-level performance on the validation or test cohort when applying early stopping based on mAUC metric. As the proposed PC loss term provides information about the correct prediction of proportions, we evaluated this term as an early stopping criterion. Thus, we also kept track of the epoch average of $z = \frac{1}{B} \sum_b p_{b,k_i'} - p_{b,k_i''}$. Among the full range of hyperparameter values, the ones that showed best stability during training were $\lambda_{PC} = \{0.1, 1\}$ and $t_{PC} = \{1, 5, 10\}$. We show the results obtained at bag-level and the instance-level accuracy on test cohort, as well as the proportion constraint satisfaction on the train subset for both early stopping criterion in Fig. 5.

The figures of merit indicate that the criterion based on constraint satisfaction (dashed lines) consistently outperforms the validation mAUC criteria (solid line) at both instance and bag level for all



(a) (b) (c)

Fig. 5. Ablation studies on proportion constraint (PC) MIL formulation. Hyperparameters study for λ_{PC} and t_{PC} are performed for bag-level mAUC on test set (a) and instance-level accuracy on test set (b). Also, two early stopping criterion are validated: mAUC on validation set (solid lines) and proportion constraint satisfaction z_{PC} (dashed lines), which values are illustrated in (c).



(a) (b)

(c)

Fig. 6. Visualization of the embedding space produced by baselines and the proposed method models on the labeled instances from SICAP-MIL test cohort. (a) instance-max; (b) instance-max w. PE; (c) instance-max w. PE w. PC. Red: non-cancerous; light blue: Gleason grade 3; dark blue: Gleason grade 4; purple: Gleason grade 5.

settings. This could be explained by the possibles bias introduced using the validation subset due to class imbalance. Likewise, maximizing the difference in proportion between the majority and minority classes can help to better distinguish between them. The results obtained are in line with these observations, since lower values of t_{PC} seem to obtain better results. Due to the formulation of the barrier extension (Eq. (4)), low values of t contribute not only to fulfill the constraint, but also to maximize it by using a slope proportional to $1/t$. Therefore, we selected the setting that gives the largest proportion of difference between the primary and secondary grade on the train cohort: $t_{PC} = 5$ and $\lambda_{PC} = 1$.

Qualitative evaluation. Finally, we want to get a more intuitive view of how the different terms of the proposed methodology are influencing the extraction of discriminative features. For that purpose, we depict the feature representation of the embedding space produced by the encoder networks on the instance-level labeled test cohort using the t-sne [25] in Fig. 6. Concretely, we obtained the two-dimensional t-sne embedding using a perplexity value of 40, and 300 iterations. The t-sne representation is obtained on the instance-max setting 6(a), instance-max with PE term 6(b) and instance-max with PE and PC terms 6(c) after Student model training.

Features obtained using the basic max aggregation are quite overlapped on the cancerous classes. Although the PE term slightly improves this condition, only once the PC term is included it is possible to distinguish class-wise clusters between Gleason grades 3 and 4. These grades tend to coincide in WSIs, with Gleason score 7 (whole slide images that include both tumor growth patterns of grade 3 and 4) being the most common in the database used (see Table 1 and Fig. 2). This fact produces noise during training, as many bags are positive for both classes simultaneously, making it difficult to distinguish between the two types of instances. However, when we introduce the relative proportion information of both classes during training, this facilitates the network to promote a distinction between them.

Also, we introduce in Fig. 7 visualizations of the obtained instance-level classifications, compared to pathologists annotations and baselines. Instance-level predictions are performed on the test subset biopsies using an overlap of 75% between instances, to gain spatial resolution. Then, the instance-level scores are assigned to each pixel of the patch, and they are averaged among the overlapped patches. From the selected representative examples, it is observed how once the different proportion constraints are introduced, the model is able to differentiate best between the different Gleason grades (first and second rows), and locates more cancerous regions (third row).

5. Conclusions

In this work, we have presented a novel constrained multi-label instance-based MIL formulation that encourages the network to focus on many positive instances, and allows to impose restrictions about relative proportions of class size within the bag. In particular, we combine a standard instance-based max aggregation with additional inequality constrains terms via a flexible log-barrier extension. We validate the proposed formulation on a new publicly available dataset of prostate histology cancer WSIs images, SICAP-MIL. In the experimental stage, our method shows that forcing the network to classify more positive instances, the results improve in ∼ 5% at instance level classification accuracy. By simply incorporating relative proportion information about the primary grade in the WSI, which is usually easily accessible from medical records, our method reports improvements of ∼ 9% accuracy at instance level, and ∼ 3.3% mAUC at bag level. In addition, the target relative proportion difference between primary and secondary classes in the bag has proven to be a good criterion when optimizing the model, obtaining more generalizable results than using the mAUC at the bag level. The obtained results are comparable to prior works using similarly-sized datasets under the supervised paradigm, which require tedious instance-level annotations.

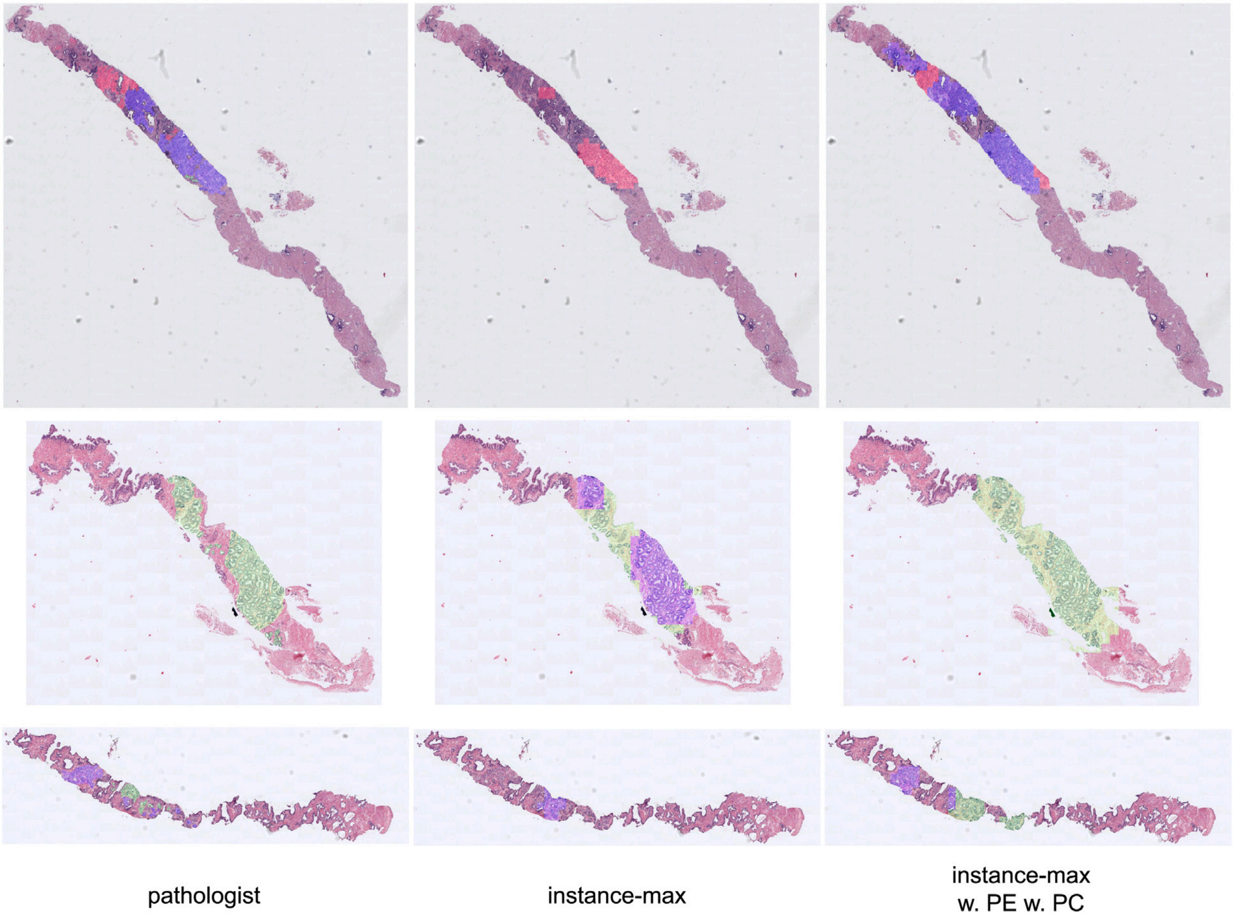


Fig. 7. Visual examples of the proposed model performance on instance-level prostate cancer grading. In particular, the pathologists annotations are depicted with the instance-based MIL baseline using max aggregation, and the results when we introduce the proportion priors. In green: Gleason grade 3; blue: Gleason grade 4; red: Gleason grade 5.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C2. The work of A. Schmidt was funded from the European Union's Horizon 2020 research and innovation programme under the Marie SkÅodowska Curie grant agreement No 860627 (CLARIFY Project). The hardware NVIDIA DGXA100 used in the experimental part of this work was funded by Generalitat Valenciana/European Union through the European Regional Development Fund (ERDF) of the Valencian Community (IDIFEDER/2020/030). Funding for open access charge: Universitat Politècnica de València.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Urology Service from Hospital Clínico de Valencia for their collaboration and support provided during the data collection.

References

[1] 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, Nat. Med. 25 (8) (2019) 1301–1309.

[2] J. Silva-Rodriguez, A. Colomer, J. Dolz, V. Naranjo, Self-learning for weakly supervised Gleason grading of local patterns, IEEE J. Biomed. Health Inf. 25 (8) (2021) 3094–3104.

[3] J. Li, W. Li, A. Sisk, H. Ye, W.D. Wallace, W. Speier, C.W. Arnold, A multi-resolution model for histopathology image classification and localization with multiple instance learning, Comput. Biol. Med. 131 (November 2020) (2021) 104253.

[4] Z. Jia, X. Huang, E.I. Chang, Y. Xu, Constrained deep weak supervision for histopathology image segmentation, IEEE Trans. Med. Imaging 36 (11) (2017) 2376–2388.

[5] C.L. Srinidhi, O. Ciga, A.L. Martel, Deep neural network models for computational histopathology: A survey, Med. Image Anal. 67 (2021) 101813.

[6] Z.H. Zhou, M.L. Zhang, Multi-instance multi-label learning with application to scene classification, Adv. Neural Inf. Process. Syst. (2007) 1609–1616.

[7] M. Ilse, J.M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: 35th International Conference on Machine Learning, Vol. 5, ICML 2018, 2018, pp. 3376–3391.

[8] Z. Jia, X. Huang, E.I. Chang, Y. Xu, Constrained deep weak supervision for histopathology image segmentation, IEEE Trans. Med. Imaging 36 (11) (2017) 2376–2388.

[9] S. Venkataramanan, K.C. Peng, R.V. Singh, A. Mahalanobis, Attention guided anomaly localization in images, in: ECCV 2020, 12362 LNCS, 2020, pp. 485–503.

[10] Y. Zhou, Z. Li, S. Bai, X. Chen, M. Han, C. Wang, E. Fishman, A. Yuille, Prior-aware neural network for partially-supervised multi-organ segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, Vol. 2019-Octob, 2019, pp. 10671–10680.

[11] H. Kervadec, J. Dolz, E. Granger, I. Ben Ayed, Curriculum semi-supervised segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11765 LNCS, 2019, pp. 568–576.

[12] M. Bateson, J. Dolz, H. Kervadec, H. Lombaert, I.B. Ayed, Constrained domain adaptation for image segmentation, IEEE Trans. Med. Imaging 40 (7) (2021) 1875–1887.

[13] H. Kervadec, J. Dolz, J. Yuan, C. Desrosiers, E. Granger, I.B. Ayed, Constrained deep networks: Lagrangian optimization via log-barrier extensions, 2019, pp. 1–23, URL http://arxiv.org/abs/1904.04205.

[14] A.E. Esteban, M. López-Pérez, A. Colomer, M.A. Sales, R. Molina, V. Naranjo, A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes, Comput. Methods Programs Biomed. 178 (2019) 303–317.

[15] J. Silva-rodríguez, A. Colomer, M.A. Sales, R. Molina, V. Naranjo, Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection, Comput. Methods Programs Biomed. 195 (2020).

[16] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, Vol. 1, 2014, pp. 1–14.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[18] Q. Xie, M.-T. Luong, E. Hovy, Q.V. Le, Self-training with noisy student improves ImageNet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10687–10698.

[19] 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, Sci. Rep. 8 (1) (2018) 1–11.

[20] 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 Netw. Open 2 (3) (2019).

[21] S. Otálora, N. Marini, H. Müller, M. Atzori, Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks, in: IMIMIC 2020, MIL3ID 2020, LABELS 2020: Interpretable and Annotation-Efficient Learning for Medical Image Computing, Vol. 12446 LNCS, no. September, 2020, pp. 193–203.

[22] 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, Lancet Oncol. 21 (2) (2020) 222–232.

[23] 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, Lancet Oncol. 21 (2) (2020) 233–241.

[24] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.

[25] L. Van Der Maaten, H. Geoffrey, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.