Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Efficient Cancer Classification by Coupling Semi Supervised and Multiple Instance Learning

ARNE SCHMIDT[1], JULIO SILVA-RODRÍGUEZ[2], RAFAEL MOLINA[2] (SENIOR MEMBER, IEEE), AND VALERY NARANJO.[4]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain (email: arne@decsai.ugr.es)
[2]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain (email: jjsilva@upv.es)
[3]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain (email: rms@decsai.ugr.es)
[4]Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain (email: vnaranjo@upv.es)

Corresponding author: Arne Schmidt (e-mail: arne@decsai.ugr.es).

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627 (CLARIFY Project) and from the Spanish Ministry of Science and Innovation under project PID2019-105142RB-C22.

ABSTRACT The annotation of large datasets is often the bottleneck in the successful application of artificial intelligence in computational pathology. For this reason recently Multiple Instance Learning (MIL) and Semi Supervised Learning (SSL) approaches are gaining popularity because they require fewer annotations. In this work we couple SSL and MIL to train a deep learning classifier that combines the advantages of both methods and overcomes their limitations. Our method is able to learn from the global WSI diagnosis and a combination of labeled and unlabeled patches. Furthermore, we propose and evaluate an efficient labeling paradigm that guarantees a strong classification performance when combined with our learning framework. We compare our method to SSL and MIL baselines, the state-of-the-art and completely supervised training. With only a small percentage of patch labels our proposed model achieves a competitive performance on SICAPv2 (Cohen's kappa of 0.801 with 450 patch labels), PANDA (Cohen's kappa of 0.794 with 22,023 patch labels) and Camelyon16 (ROC AUC of 0.913 with 433 patch labels). Our code is publicly available at https://github.com/arneschmidt/ssl_and_mil_cancer_classification.

INDEX TERMS Cancer Classification, Histopathology, Multiple Instance Learning, Semi-Supervised Learning, Whole Slide Images

I. INTRODUCTION

THE analysis of histopathological biopsies is the gold standard for the diagnosis of many different cancer types. In the last years, Computer-Aided Diagnosis (CAD) systems based on artificial intelligence have gained attention as a promising tool to reduce pathologists' workload, improve the repeatability and to avoid the variability of diagnostic processes. For the training of deep learning algorithms, initially many approaches relied on detailed local-level annotations of the digitized biopsies by pathologists [1]. Unfortunately, due to the large size of the WSIs, this process is a time-consuming task which makes it difficult to obtain large and heterogeneous annotated datasets. This recently led to the rise of approaches that do not need detailed local-level annotations. Instead, they utilize the MIL assumption where the image patches form the instances and the complete WSI forms the bag [2]. In this setting, no patch-level annotations are needed and only the diagnosis of the biopsies are used for training. Another strategy to learn with fewer patch-level annotations is SSL where only a subset of the image patches must be labeled. Still, existing methods have some common limitations: while SSL techniques do not incorporate the WSI diagnosis (global label) and therefore show a limited performance, MIL methods often can not make accurate patch-level predictions or have to be trained on very large datasets. For example, in [2] the authors conclude that at least 10,000 slides are necessary for a good performance. These limitations encourage the development of novel data-efficient methodologies which balance the amount of patch-level annotations and size of the required datasets and can

flexibly adapt to different scenarios.

A. CONTRIBUTIONS

We propose a new machine learning method based on MIL and SSL and an efficient labeling strategy to perform cancer classification with fewer annotations and reduced human workload. The contributions of this work are:

- A novel cancer classification method utilizing the global WSI diagnosis, unlabeled image patches and a limited number of labeled image patches for training. The proposed method exploits pseudolabeling techniques to combine both global labels in the MIL perspective and scarce patch-level annotations under the SSL setting. This combined approach overcomes current limitations of existing MIL and SSL methods and shows a significant improvement in comparison to the SSL and MIL baselines.
- An Efficient Labeling (EL) technique to achieve the best possible performance with a limited amount of annotations. Instead of annotating complete WSIs we propose to annotate only some cancerous patches per WSI for each cancer class.

We make an extensive quantitative validation of the performance on three different datasets and show that our deep learning framework achieves very competitive results without the need for detailed patch labels or an excessive amount of WSIs. With just a few patch labels per WSI we get a similar performance as in a supervised setting, even on relatively small datasets. The success of our algorithm supports the following labeling paradigm: A good performance of deep learning algorithms is already possible if pathologists only point out a few cancerous image patches per WSI instead of spending a lot of time with the detailed annotation.

B. RELATED WORK

To structure the related work into Multiple Instance Learning (MIL) and Semi Supervised Learning (SSL) approaches, we first clarify the definition of both, following the terminology of Cheplygina et al. [3]. Under the MIL assumption, instances (patches) are grouped into bags (WSIs), where only the label of the entire bag is known and the instance labels remain unobserved. In this paradigm, learning is driven by known global information (WSI diagnosis). SSL describes a learning scenario with two sets of samples: a labeled set and an unlabeled set. SSL methods use the unlabeled set (additionally to the labeled set) to find a better decision boundary and improve the classifier. In the given use-case, this means SSL methods use labeled and unlabeled patches for training, but not the WSI labels.

MIL approaches for histopathological images are becoming more and more popular because they do not require detailed local annotations, but only bag labels for training. [2] Usually, a bag-level representation is obtained by the aggregation of either the instance-level features (embedding-based) or their predictions (instance-based). Recently, the classical aggregation functions based on max or average pooling have been replaced by more advanced mechanisms, such as learnable attention methods [4]. Campanella et al. [2] showed promising results processing the top-ranked positive instance features with an RNN. In other works, the use of instance-based aggregations based on top and bottom ranked instances [5] or min-max aggregation [6] have been proposed. Further approaches use embedding-based MIL via multi-head attention mechanisms [7] or combine instance-level predictions with embeddings [8]. Hashimoto et al. [9] use multiple scales with attention mechanisms and domain adversarial training for malignant lymphoma subtype classification. Common limitations of existing approaches are the requirement of very large datasets [2] and the incapability to make class predictions at instance level [4] [2] [9]. Further, recent approaches often include complex multi-stage training procedures with multiple models [2] [9]. This motivates the development of well performing, but simpler approaches for an easy application in clinical practice.

SSL approaches use labeled and unlabeled patches for training. For histopathological images, most existing SSL approaches rely on pseudo-labeling techniques such as Pulido et al. [10] who apply MixMatch [11] and FixMatch [12] under a highly noisy and imbalanced data setting. Jaiswal et al. [13] combine pseudo-labeling techniques with a novel learning rate schedule (one cycle policy). The approaches of Shaw et al. [14] and Marini et al. [15] are based on teacher-student models, where the teacher model trains with the labeled set of images. The SSL component of our work is related to FixMatch and Unsupervised Data augmentation (UDA) [16]: UDA proposes to use unlabeled images for so-called consistency regularization. Fixmatch extends the idea of consistency regularization with pseudolabels: Based on weak image augmentations, pseudo labels are assigned to confident predictions while the network is trained with strong image augmentations. The common drawback of all the mentioned SSL methods is that they do not make use of global information (bag labels) and always require a certain amount of labeled instances.

SSL+MIL approaches were proposed very recently for histopathological images, but existing methods show some major differences to our work. Otalora et al. [17] propose an SSL+MIL method based on teacher-student networks, but it is specialized for prostate cancer and uses micro tissue arrays for pre-training. Although this approach is theoretically interesting, the performance gap to the supervised state-of-the-art models is quite large in practice (listed in Table 2). Li et al. [18] and Lu et al. [19] also propose hybrid models of SSL+MIL, but the applications are not comparable to our work: while the first approach is applied to binary semantic segmentation of WSIs, the latter is used for binary classification of histopathological images of 2048×1536 pixels that are much smaller than WSIs.

Our method takes advantage of both SSL and MIL learning strategies and is able to perform multi-class classification on WSIs for different cancer types. It incorporates the aug-

mentation strategy of FixMatch [12] and the consistency regularization of Unsupervised Data augmentation (UDA) [16] while the pseudo label assignment is driven by the MIL perspective. As a result, the proposed method inherits the advantages of SSL and MIL while overcoming their existing limitations: our method achieves competitive results on small datasets, provides multi-class instance-level predictions, only needs one training procedure, one stage and one model that performs the common mini-batch training but still has the capability to include the bag label information.

C. PAPER STRUCTURE

The rest of the paper is organized as follows: In section II we describe the problem in theoretical terms, the proposed efficient labeling strategy (II-A), the image augmentation strategy (II-B), the training framework (II-C) and the theoretical background of the proposed method (II-D). In the experiment section III we first outline the description of dataset (III-A) and implementation (III-C). In the ablation studies (III-D) we show experimentally the effect of the different loss components. Finally, we highlight the effect of the proposed efficient labeling strategy (III-E) and compare with state-of-the-art methods (III-G) before concluding our article (IV).

II. MODEL DESCRIPTION

Let us consider a WSI classification problem where images are assigned a single class Y or a primary and secondary class Y^1 and Y^2. We refer to these WSI labels as 'bag labels' in the context of MIL and to the image patches as 'instances'. Each patch can be either non-cancerous (NC) or contain one of the cancer classes. There are many problems that can be formulated this way. For the example of prostate cancer, the tissue is classified as non-cancerous (NC), Gleason grade 3 (GG3), Gleason grade 4 (GG4) or Gleason grade 5 (GG5). The primary Gleason grade Y^1 and the secondary Gleason grade Y^2 of a WSI are assigned based on the two most prominent grades. In other cancer classification tasks like the lymph node detection of the Camelyon16 challenge, just one global label is assigned. Our approach works in both cases.

To translate the problem into a mathematical notation, we denote the bag indices as $B = \{1, 2, .., N\}$ where N is the number of WSIs in the training set. Let further $I_b = \{1, 2, .., M_b\}$ be the index set for the image patches (instances) in bag b. The complete set of image patches and their true cancer class can now be defined as

$$\{x_{bi}, y_{bi}\} \;\; b \in B, i \in I_b \tag{1}$$

To describe the labels let us first define the subset of non-cancerous WSIs $B^- \subset B$ and cancerous WSIs B^+. Following the MIL assumption we know that for each negative bag $b \in B^-$:

$$y_{bi} = NC \; \forall i \in I_b \tag{2}$$

For all positive bags we know that some patches must contain the pattern of the present cancer class Y_b. For each bag $b \in B^+$:

$$\exists i \in I^b : y_{bi} = Y_b \tag{3}$$

which in the case of a primary and secondary label applies to both Y_b^1 and Y_b^2.

Note that the targets y are represented as a C-dimensional probability vector with each dimension representing one class probability and the class labels are described as one-hot vectors.

A. EFFICIENT LABELING

We propose a data setting that we name Efficient Labeling (EL): For each cancerous WSIs the pathologist only points out a few cancerous patches instead of annotating the whole WSI. For each global label Y_b some corresponding patch labels $y_{bi} = Y_b$ are assigned. We consider the annotation of a few cancerous patches per WSI a realistic and time-efficient strategy for the annotation of a new dataset from scratch or the data collection in already deployed CAD systems. In the latter case, the pathologist provides labels during the diagnostic process (human-in-the-loop, see f.e. [20]).

In our experiments, this data setting is simulated by picking randomly a certain amount of patch labels and hide the others during model training. This allows us to systematically study the effect of a varying amount of patch labels.
We divide the indices of each positive bag into the set of labeled ($L \subset I_b$) and the set of unlabeled ($U \subset I_b$) instances such that all labels $\{y_{bi} | i \in L_b\}$ are available due to pathologists annotation, while the labels $\{y_{bi}, | i \in U_b\}$ remain unknown.

B. IMAGE AUGMENTATION

Our image augmentation strategy is related to FixMatch [12] and Unsupervised Data augmentation (UDA) [16]: UDA proposes to use unlabeled images for so-called consistency regularization: for two versions of a randomly augmented image the network is trained to predict the same class probabilities. The FixMatch algorithm combines consistency regularization with pseudolabeling. Here, a weak image augmentation is applied to the unlabeled images, the class is estimated by a CNN and pseudo labels are assigned to the images with confident class predictions. Then the network is trained to predict these pseudo labels given a strongly augmented version of the unlabeled images. Both approaches have in common that random image augmentation is a key component.

Similar to [12] the weak and strong image augmentation for the image patches play an important role in our approach. The strong image augmentation in our implementation uses a very strong random brightness shift that leads to substantially darker and brighter versions of the original image. The weak augmentation only applies a mild version of the brightness shift, leading to images similar to the original. Applying only

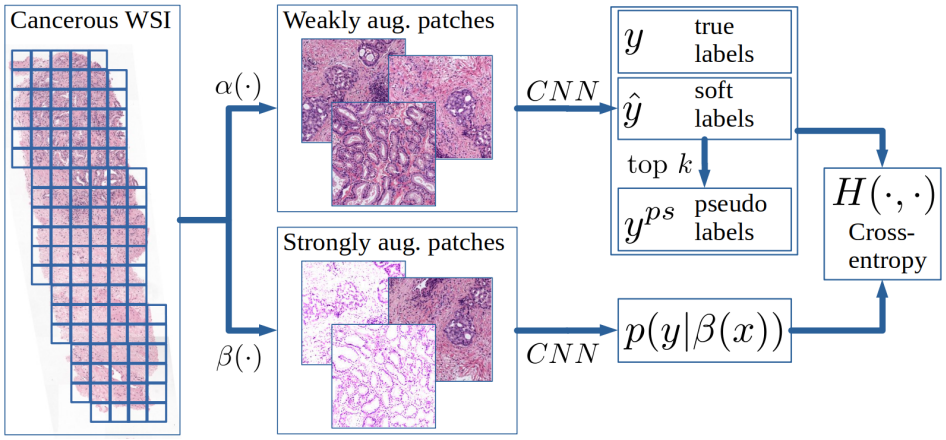

FIGURE 1. Proposed training framework for cancerous WSI, combining MIL and SSL. We take all patches of the WSI and apply a weak augmentation to obtain soft labels and pseudo labels by the CNN predictions. Based on these labels, we train the same CNN with the strongly augmented patches.

a weak augmentation makes it easier for the network to obtain a correct prediction and is therefore used to estimate pseudo labels and soft labels. The strongly augmented images are more challenging to predict and are therefore used to train the network. We denote $\alpha(\cdot)$ as the operator of weak random image augmentation and $\beta(\cdot)$ as a strong random image augmentation. For more details, we refer to the theoretical background (II-D) and the implementation details (III-C).

C. PROPOSED TRAINING FRAMEWORK

The goal is to train a patch classifier $p_\theta(y|x)$ which predicts class probabilities y for a given patch x and is parametrized by the model weights θ (following the notation of [12]). The training procedure (Figure 1) can be applied to any classification model and is divided into three steps that are repeated for each training epoch:

Step 1 Obtain the CNN predictions of the weakly augmented image patches in the positive bags. For a given image patch x_{bi} of the positive bag $b \in B^+$, we apply the weak image augmentation α. The weakly augmented image patch $\alpha(x_{bi})$ is used to predict the CNN output probability vector $p_\theta(y|\alpha(x_{bi}))$ which we define as $\hat{y}_{bi}$:

$$\hat{y}_{bi} := p_\theta(y|\alpha(x_{bi})) \; \forall b \in B^+ \tag{4}$$

As some of these vectors of probabilities will serve later as training targets, we will call $\hat{y}$ soft labels.

Step 2 Calculate pseudo labels for each positive bag $b \in B^+$. We know from equation (3) that some patches have the same class as the WSI. Given the global label Y_b of the bag, we assign this pseudo label to the k patches whose class probabilities of class Y_b are the largest of all instances in the bag. Concretely, this is done by the following steps:

(i) Create a list of probability vectors $\hat{y}_{bi}$ ordered with respect to class Y_b.

(ii) Select the k first items of this list to define the index set $P_b \subset I_b$.

(iii) Assign the one-hot class label Y_b to the patches indexed by P_b as a pseudo label y^{ps}:

$$y^{ps}_{bi} = Y_b \; \forall i \in P_b, b \in B^+ \tag{5}$$

In the case of two or more global labels, this pseudo label assignment is performed for each of them.

Step 3 Use the strongly augmented image patches $\beta(x)$ and a combination of groundtruth labels, pseudo labels and soft labels to train the CNN. Mathematically, the loss function is described as:

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{b\in B^-} \underbrace{\sum_{i\in I_b} H(y_{bi}, p_\theta(y|\beta(x_{bi})))}_{A} \\ &+ \sum_{b\in B^+} \Bigg(\underbrace{\sum_{i\in L_b} \lambda H(y_{bi}, p_\theta(y|\beta(x_{bi})))}_{B} \\ &\quad + \underbrace{\sum_{i\in P_b} H(y^{ps}_{bi}, p_\theta(y|\beta(x_{bi})))}_{C} \\ &\quad + \underbrace{\sum_{i\in U_b\setminus P_b} H(\hat{y}_{bi}, p_\theta(y|\beta(x_{bi})))}_{D} \Bigg)\end{aligned} \tag{6}$$

Here, $H(\cdot,\cdot)$ denotes the cross-entropy loss for classification and λ is a hyperparameter to assign a higher weight to the groundtruth labels of the cancer classes. Note that all terms of the loss function (A, B, C, D) split into sums over the instances. Training can therefore be performed in minibatches via stochastic gradient descent. In comparison to semi-supervised methods, our algorithm is still able to train without any patch labels (MIL setting): In this case, the loss term of positive instance labels (B) can be simply omitted, and the training can be performed based only on negative, pseudo and soft labels (terms A, C and D).

The proposed training framework is summarized in Algorithm 1. For notational coherence, we describe the algorithm for an instance-wise optimization. In practice, the prediction of step 1 and the gradient update of step 3 can be performed in common mini-batches for efficient computational parallelization.

D. BACKGROUND

In the following subsection, we want to explain the derivation and theoretical background of the different loss components and the image augmentation.

The MIL component of our method enables the model to incorporate information from the global WSI labels during training and constitutes the loss terms A and C of equation 6. The loss term A uses the MIL property of equation 2: all instances in a negative bag must be negative. With these negative instances, the model can perform supervised training. Further, the pseudo labels of term C are derived by the MIL perspective: From equation (3) we know that some instance labels are equal to the bag label Y_k. A natural assumption is that instances with the highest class probabilities (of class Y_k) are the best candidates for the assignment of label Y_k. When no positive instance labels are available, this label assignment enables the model to learn the positive classes at instance level through the bag labels. The proposed MIL component can be seen as an extension of the max-pooling which is used for example in [2] in the first training phase. Instead of assigning the global label just to one instance with the highest probability, we assign it to multiple instances (k in total) with the highest probabilities. Further, we extend the binary case [2] to multiple classes using the class probabilities of the given global label, as described in step 2. The empirical improvement of our algorithm over max-pooling is discussed in section III-D.

The SSL component of our method ensures that the labeled (term A and B of equation 6) and unlabeled (term C and D of equation 6) patches are used to improve the classifier. From a theoretical point of view, it has been shown that pseudo labels (loss term C) can be interpreted as a form of entropy minimization [21].

As the conditional entropy of class probabilities is a measure of class overlap, the optimization will favor putting the class decision boundary in a low density area and leads to a better separation of classes [22]. The loss term D with soft labels serves as an additional consistency regularization: for two randomly augmented versions of the image, the classifier is trained to predict the same output. This technique has been proven to lead to better generalization and stability of the classifier [23]. Further, we want to discuss the role of weak and strong image augmentations for label propagation. The basic assumption of semi-supervised learning algorithms is that the data distribution of unlabeled datapoints can help a model to find a better decision boundary between the classes. One strategy is to propagate label information from one datapoint to nearby unlabeled datapoints during the model training (so called 'label propagation', see f.e. [24] or [16]).

Algorithm 1 Proposed Training Procedure

Input: For each bag $b = 1, .., N$: Image patches $\{x_{bi}\}_{i=1,..,M_b}$, a reduced number of patch labels $\{y_{bi}\}_{i \in L_b}$, WSI labels $\{Y_b\}$, number of epochs E, learning rate η

Output: Optimal model parameters θ

```
for e = 1 to E do
  # Step 1
  for b = 1 to N do
    for i = 1 to M_b do
      estimate ŷ_bi ← p_θ(y|x_bi) (eq. 4)
    end for
    # Step 2
    Order {ŷ_bi} regarding class Y_b (Step 2 (i))
    Define P_b as the k max. probabilities (Step 2 (ii))
    Assign y_bi^ps ← Y_b for i ∈ P_b (Step 2 (iii))
  end for
  # Step 3
  for b = 1 to N do
    for i = 1 to M_b do
      θ ← θ − η δL(θ)/δθ (eq. 6, using {y_bi}, {y_bi^ps}, {ŷ_bi})
    end for
  end for
end for
return θ
```

The final goal is to assign a consistent label in high density areas provided by some labeled datapoints. Label propagation with loss terms C and D in combination with weak and strong image augmentation (α and β) can be explained in the following way: Let $V_\alpha(x)$ and $V_\beta(x)$ be the space of all possible image augmentations with α and β, respectively, for a given image patch x. As the strong image augmentation β leads to a higher distortion of the image, the image space $V_\beta(x)$ is larger than $V_\alpha(x)$ and we assume $V_\alpha(x) \subset V_\beta(x)$ when the same random augmentations are applied for α and β. As the pseudo and soft labels are predicted on $\alpha(x)$ and the network is trained on $\beta(x)$, label information is propagated from $V_\alpha(x)$ to $V_\beta(x)$ during training, as shown in Figure 2. Other unlabeled datapoints that are in or close to $V_\beta(x)$ are more likely to be assigned the same class as x in the next iteration. Therefore, the available patch labels are propagated to unlabeled patches in areas of high data density. As a result, the model is encouraged to assign a similar label to all instances in a data cluster and to define the decision boundaries between those data clusters.

The SSL component of our work is inspired by Fixmatch [12] and UDA [16] and in the following, we briefly discuss similarities and differences. Fixmatch has a similar augmentation strategy as the proposed method, but we extend it with soft labels and a MIL-driven pseudo label assignment instead of using a probability threshold as in the original work. This enables the model to incorporate bag labels during training while maintaining the benefits of SSL. The soft label assignment is inspired by UDA, such that loss terms

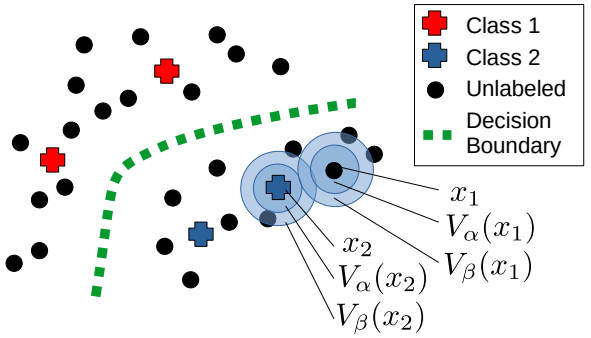

FIGURE 2. Simplified illustration of label propagation with weak and strong image augmentation. The datapoints correspond to image patches in our case. Shown are two example points x_1 and x_2 with the corresponding regions of weak and strong image augmentation ($V_\alpha(x)$ and $V_\beta(x)$ respectively).

B and D are similar to the UDA training. The idea of label propagation by consistency regularization was presented in the context of UDA, together with a theoretical proof based on graph theory. Apart from the soft-label assignment, UDA is lacking the other components of our proposed method (weak/strong augmentation, pseudo label assignment, bag label incorporation).

III. EXPERIMENTS AND DISCUSSION

We performed extensive experiments to evaluate our proposed training framework as well as the proposed efficient labeling (EL) strategy.

A. DATASETS

The experiments were conducted using three different public WSIs datasets of prostate and breast cancer. The gigapixel WSIs are sliced into smaller patches that form the instances of the MIL problem, while the WSI diagnosis is the bag label. The used datasets have both biopsy-level and patch-level annotations available, which made them particularly suitable for the validation of the proposed method. The SICAPv2[1] [25] dataset was used to validate the proposed method on prostate cancer for the multiclass Gleason grading scenario. This dataset contains 155 prostate WSIs which are sliced into 512x512 overlapping patches. The primary and secondary Gleason grade for all WSIs as well as patch-level labels are included for a large number of instances in the dataset. In our work, we maintained the proposed partitions of the original dataset for training, validation and testing.
Additionally, we use the PANDA dataset [2] for prostate cancer classification, which is substantially larger than SICAPv2, to test our method on a dataset with a different size. It consists of 10,415 WSIs and was presented at the MICCAI 2020 conference as a challenge. As the test set of the PANDA dataset is not public, we use the available WSIs to generate a train/validation/test split of 8469, 353 and 1794 WSIs, respectively. We extract 512x512 patches with a 50% overlap from the WSIs. The data was collected from two datacenters ('Radboud' and 'Karolinska') but only the WSIs from Radboud have local annotations of the Gleason grade while the annotations from Karolinska distinguish non-cancerous and cancerous. The exact classes of these cancerous patches remains unknown, so they can only be used as unlabeled data for training. We disregard all patches with less than 50% tissue and assign the label 'non-cancerous' to all patches that have at least 95% pixels annotated as non-cancerous or background. To the cancerous patches of Radboud we assign the Gleason grade which has the highest amount of pixels in comparison to the other cancer classes. The breast cancer experiments were conducted on Camelyon16[3] which contains 130 WSIs for testing and 270 WSIs for training/validation. We split them into 80% for training and 20% for validation. For the training/validation set, detailed annotations are available while the testing is done only at the WSI level in a binary manner (cancer vs no-cancer). For our experiments, we sliced the WSIs into non-overlapping 512x512 patches at 20x magnification and filtered out the patches that contain less than 5% tissue.

B. METRICS

To compare our method, we use the metrics that are reported for other state-of-the-art methods for the different datasets. For prostate cancer (SICAPv2 and PANDA), the common metric for comparison is Cohen's quadratic kappa, which measures the inter-rater reliability between the pathologist's annotations and the model's predictions. It is calculated based on the confusion matrix, and a kappa value of 0.0 indicates agreement by chance while 1.0 means complete agreement. This metrics takes into account that, in a set of ordered classes, error between consecutive classes should be less penalized and therefore it is especially suitable for Gleason grading. Further, we report the average F1 score, as in [25], [26]. The F1 score is based on the recall and precision per class and then averaged over the classes. For the breast cancer dataset Camelyon16, the commonly reported metric is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings and measures the diagnostic ability of a binary classifier. The AUC of the ROC is 0.5 for a random classifier and 1.0 for a perfect classifier. As our model uses pseudo labels, it is especially important to prove the reliability and robustness of our model. We perform multiple independent runs on the independent test sets to assure a reliable high performance of our method. The results are therefore reported as mean and standard deviation of the above described metrics.

[1] Available at: https://data.mendeley.com/datasets/9xxm58dvs3/1
[2] Available at: https://www.kaggle.com/c/prostate-cancer-grade-assessment
[3] Available at: https://camelyon16.grand-challenge.org/Data/

C. IMPLEMENTATION DETAILS

We implemented our model in tensorflow 2.3 and used one TITAN X (Pascal) GPU with 12 Gb for training with mini-batches of 16 patches. The training until convergence with the EfficientNet-B5 backbone took approximately 7 hours (30 epochs) for SICAPv2, 7 days (10 epochs) for PANDA and 6 days (15 epochs) for Camelyon16. The time to perform predictions for inference is negligible for applications in clinical practice and lies below 2 seconds for a complete WSI on average for all used datasets. The model selection and hyperparameter tuning was performed on the four-fold cross validation set of SICAPv2. For the classification backbone, this work utilized the state-of-the-art image classification model EfficientNet [27] which was pre-trained on ImageNet and can scale with 8 different levels of complexity (B0-B7). We used the four-fold cross validation of SICAPv2 for the model selection and observed that an increasing complexity of the model led indeed to a better performance until EfficientNet-B5. Models B6 and B7 did not show any further improvements, so we chose EfficientNet-B5 as our backbone. The hyperparameters of the model were set to $k = 5$ (tested: $k = 1, 3, 5, 10$, used in Step 2 in II-C) and $\lambda = 3$ (tested $\lambda = 1, 2, 3, 5$, used in equation 6 D) which showed the best results. The network was fine-tuned with stochastic gradient descent and the learning rate 0.01. For the relatively small dataset SICAPv2, class balanced loss was used (based on true y and estimated labels $\hat{y}$) to stabilize the training as we sometimes observed the convergence to 'bad' local minima (for the experiments $P = 0$ and $P = 1$ in Figure 3). We resized the image patches to 250x250 which is the input resolution for our model.

For image augmentation brightness shift, random flip and rotation were used. The difference between weak and strong augmentation in our experiment was the intensity of the brightness shift (multiplication of the alpha channel with a factor) which led to a darker or brighter version of the original image. While the weak augmentation α uses random brightness shift factors between 0.9 and 1.1, the strong augmentation β uses a range from 0.5 to 1.5. The stronger the brightness shift, the harder it gets to visually recognize the pattern in the images.

D. ABLATION STUDIES

To study the effect of the different loss components and the improvement over the SSL and MIL baselines, we performed an ablation study for the SICAPv2 dataset with efficient labeling (see section II-A) and 5 patch labels per WSI and global label (equal to the experiment P=5 of section III-E). In Table 1 we first compare 4 different label settings: only the available ground truth labels (GT), ground truth and soft labels (GT + SL), ground truth and pseudo labels (GT + PL) and ground truth, soft and pseudo labels (GT + SL + PL) which is our proposed setting. In the loss equation, terms A and B represent the ground truth, term C the pseudo labels and term D the soft labels. The model trained only with ground truth labels can be seen as a baseline because it simply uses all available labels in a supervised fashion. We observe that pseudo label as well as soft labels improved this baseline in both metrics. The best result was obtained using ground truth, pseudo and soft labels, and we therefore proved that all loss terms are relevant in practice.

TABLE 1. Ablation studies on SICAPv2 with 5 patch labels per cancerous WSI and global label. The results are reported as the mean and standard deviation of 5 independent runs.

Model	Cohen's quadr. kappa	avg. F1 Score
GT	0.768 ± 0.009	0.688 ± 0.012
GT + PL	0.774 ± 0.012	0.697 ± 0.007
GT + SL	0.780 ± 0.012	0.698 ± 0.012
GT + SL + PL	$\mathbf{0.801 \pm 0.013}$	$\mathbf{0.700 \pm 0.011}$
MIL (Max-pooling)	0.545 ± 0.038	0.492 ± 0.026
SSL (Fixmatch)	0.774 ± 0.031	0.676 ± 0.009

We also compared our method to the SSL and MIL baselines to highlight the improvement of our combined solution. The chosen baseline implementations Max-pooling and Fixmatch are the algorithms that are the most related approaches in the fields of SSL and MIL (for details, see section II-D). For the MIL baseline, we disregarded the available patch labels for training. Max-pooling inspired our pseudo-label assignment and is commonly used, f.e. in [2]: the global label was assigned to one patch with the highest class probability. The poor results of only 0.545 (Cohen's kappa) and 0.492 (F1 Score) highlight that the dataset is too small for this MIL-baseline method. Including some patch labels with our proposed method performs much better. To compare with the SSL baseline, we implemented the Fixmatch algorithm [12], which uses the available patch-labels but can not integrate the global WSI labels for training. For a fair comparison, we assigned negative patch labels to all patches of a negative WSI, although this is already beyond SSL in a strict sense. As proposed in the original paper, pseudo labels were assigned for cancer class predictions higher than 0.95. In this setup, the Fixmatch baseline showed a comparable performance to our proposed pseudo-label assignment (GT+PL) in terms of Cohen's kappa, but the F1 score was significantly lower. In comparison to our proposed final model (GT+SL+PL), the SSL baseline performed approximately 2.5 percentage points worse in both Cohen's kappa and F1 score. Overall, we see that utilizing a reduced number of patch labels and WSI labels with our approach achieved a substantial improvement over the SSL and MIL baselines.

E. EFFICIENT LABELING VS. COMPLETE ANNOTATION

In the next experiment, we compared two different data settings for the prostate cancer dataset SICAPv2: We wanted to study whether with limited resources it is better to use Efficient Labeling (EL, see section II) with a few patch labels from all available WSI or a few WSI with the Complete Annotation (CA). In the first case (EL) we randomly sampled a certain amount P of patch labels for the primary and secondary Gleason grade of each WSI. For the second approach

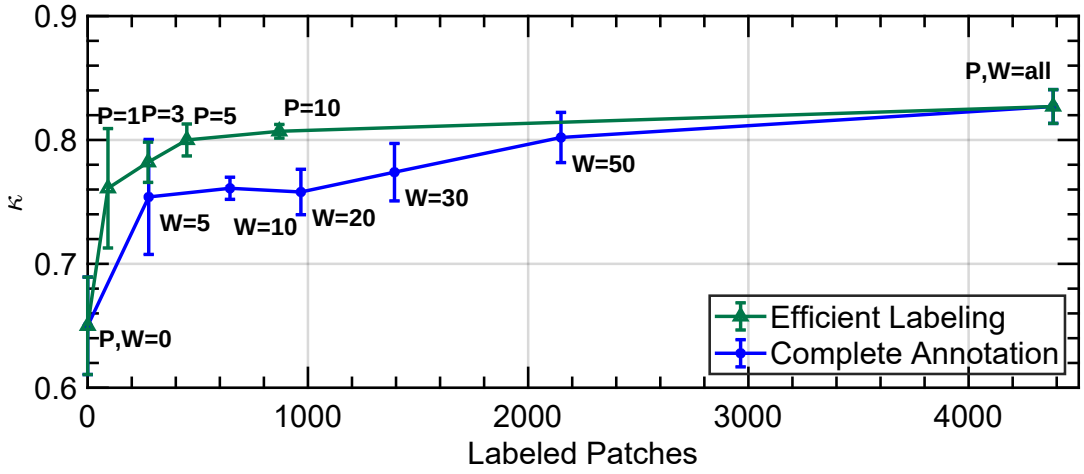

FIGURE 3. Comparison of data label settings. Efficient Labeling (EL) with P annotated patches per primary and secondary Gleason grade of all WSIs and Complete Annotation (CA) of W WSIs with all the patch labels. We plot the mean and standard deviation of Cohen's quadratic kappa (patch level) of five runs against the total amount of annotated training patches of SICAPv2.

(CA) we randomly selected W WSIs and used all patch labels of this selection for training. In Figure 3 we observe that the EL setting required substantially fewer labels than CA to obtain good results. We explain this by the higher variability of the annotated patches with EL that allows the network to learn from more diverse examples. The annotated patches of CA have a higher co-similarity and therefore contribute less information to the model training. The steep ascent of the performance from $P = 0$ to $P = 5$ proofs the efficiency of our learning approach and EL. To estimate the saved time and resources (to annotate a dataset) for the model training with our approach, we use the total amount of local annotations. Concretely, we count the total number of labeled patches used for training. We compare settings with a reduced number of patches with supervised training using all available patch labels. In the case of SICAPv2, our model with $P = 5$ and EL showed a performance close to the supervised one with only using 450 of the 4384 available patch labels. This means that approximately 10 times less labeled patches were needed for training.

For PANDA, the ratio of saved labeling effort is comparable: the model trained with $P = 5$ uses approximately 10 times less patch annotations for training than the supervised setup, but with much higher absolute numbers: while the supervised model is trained with 205,111 labeled patches, the model with $P = 5$ obtained 22,023 patch labels. For Camelyon16 (Table 3), the advantage is even bigger: the model with P=5 and EL used 433 patch labels while the supervised model trained with 21437 patch labels. This means, that only 2% of the complete training data was needed for the proposed approach, while the result remains close to the supervised performance, as reported in the next section.

F. QUALITATIVE EVALUATION

We qualitatively assessed the WSI predictions for SICAPv2 and show visual examples in Figure 4. For comparison, the predictions of our proposed model trained without any patch labels ($P = 0$), with some patch labels ($P = 5$) and all patch labels ($P = all$) are depicted as well as the ground truth annotations. We observe that the model trained without any patch labels in a MIL setting correctly marks the cancerous areas but has problems to assign the right classes to the tissue. This highlights the limitations of MIL models trained on relatively small datasets for complex multi-class scenarios. The model with some patch annotations ($P = 5$) shows a robust performance which is close to the prediction of the supervised model ($P = all$). This confirms the reliability of the proposed method, which uses pseudo labels to complement a small amount of patch labels. Note that both models, $P = 5$ and $P = all$, highlight some areas as Gleason Grade 3 that are annotated as non-cancerous. This can be explained by the interpolation in between patches to produce the graphic and the ambiguity in the Gleason grading task: even between pathologists, a complete agreement on the exact cancerous regions is rare, as reported in Table 2.

G. COMPARISON WITH STATE OF THE ART

In this section we report the results for the three datasets: SICAPv2, PANDA and Camelyon16, and compare our proposed method with efficient labeling (EL, see section II-A) to other state-of-the-art approaches. In Table 2 we show the performance of patch level classifiers of Gleason grades. We observe that our model is able to achieve competitive results with only 5 patch labels per WSI and global label. For the relatively small dataset SICAPv2, our model with $P = 5$ achieves a remarkable result of 0.807 Cohen's kappa, outperforming the existing supervised state-of-the-art [25] for this dataset. In this setting, the model only required a total of 433 labeled patches. Our model in the completely supervised setting reached a slightly better result, but using approximately 10x more patch labels. For a larger prostate cancer dataset, PANDA, we observe similar results. The

TABLE 2. Comparison with previous works of prostate cancer patch-level Gleason grading. We report the average result of 5 independent runs.

Method	Learning	Dataset	Cohen's quadr. kappa	avg. F1 Score
Arvaniti et al. [26] (2018)	Supervised	(other)*	0.55/0.49	–
Nir et al. [28] (2018)	Supervised	(other)*	0.60	–
Otálora et al. [17] (2020)	MIL + SSL	(other)*	0.59/0.55	–
Silva-Rodríguez et al. [25] (2020)	Supervised	SICAPv2	0.77	0.66
Ours (P=5; 450 positive patch labels)	MIL + SSL	SICAPv2	0.801	0.700
Ours (P=10; 870 positive patch labels)	MIL + SSL	SICAPv2	0.807	0.710
Ours (P=all; 4,384 pos. patch labels)	Supervised	SICAPv2	0.827	0.718
Ours (P=5; 22,023 positive patch labels)	MIL + SSL	PANDA	0.794	0.739
Ours (P=10; 41,910 positive patch labels)	MIL + SSL	PANDA	0.830	0.735
Ours (P=all; 205,111 pos. patch labels)	Supervised	PANDA	**0.891**	**0.812**
Inter-Pathologists [26]*			0.65	–

* Results reported on different datasets, patch size and resolutions, see [26], [28] and [17] for details.

TABLE 3. Comparison with previous works of metastasis detection in sentinel lymph nodes of breast cancer patients (Camelyon16). Our reported results are the average of 3 independent train and test runs.

Method	Learning	ROC AUC
Camelyon16 Winner [29]	Supervised	0.923
Camelyon16 Best on Leaderboard [29], [30]	Supervised	**0.994**
Campanella et al. [2] **	MIL	0.899
Campanella et al. [2] ***	MIL	0.965
Ours (P=5; 433 positive patch labels)	MIL + SSL	0.913
Ours (P=all; 21,437 pos. patch labels)	Supervised	0.933
Pathologists with time constraints [30]		0.810
Pathologists without time constr. [30]		0.966

** Tested on Camelyon16, trained on MSK breast dataset (total 9894 WSIs, see [2] for details)
*** Trained and tested on MSK breast dataset (total 9894 WSIs, see [2] for details)

model with $P = 10$ achieved a remarkable Cohen's kappa value of 0.830 and an average F1 score of 0.735. Note that the gap in comparison to the supervised model (with a Cohen's kappa of 0.891 and an average F1 Score of 0.812) is slightly larger than for the SICAPv2 experiment. This can be explained by the much higher absolute number of labeled patches for the supervised setting (205,111 patch labels). In this case, the model learns to mimic the pathologist's annotation very accurately. It is noteworthy to mention that, as the inter-pathologist agreement for this task lies round 0.65 [26], all Cohen's kappa values above 0.8 indicate a very high agreement with the given annotation. The proposed SSL+MIL approach with $P = 10$ shows a very good performance, while 163,201 less patch labels were used than in the supervised approach (P=10: 41,910 patch labels; supervised: 205,111 patch labels).

Table 3 shows the results for the detection of lymph node metastasis of breast cancer with the dataset Camelyon16. As Camelyon16 allows only the evaluation of the global WSI labels, we derived the cancer probability simply from the highest patch probability per WSI. Although our model's primary strength is the instance (patch-level) classification, we obtained a competitive Camelyon16 result with $P = 5$ (ROC AUC $= 0.913$) close to the supervised performance $P = all$ (ROC AUC $= 0.933$) while using approximately 50 times less patch labels during training. Further, the results with $P = 5$ are still more than 10 percentage points above pathologists with realistic time constraints (ROC AUC $= 0.810$). The strong performance proves the model's good generalization to different cancer types and the high accuracy of the instance predictions: bag labels can reliably be derived from them by a simple heuristic. Note that the MIL approach of Campanella et al. [2] had strong results but has some limitations: the method trained on a 20 times larger dataset and only predicted binary labels on the bag level. Our model, trained only on the Camelyon16 training set, is able to provide patch-level predictions and extendable to multiclass-settings.

H. ADVANTAGES AND LIMITATIONS OF THE PROPOSED METHOD

The proposed method has several advantages in comparison to other existing approaches. First of all, it showed a high performance while training with limited resources: A total of 450 labeled patches of 155 WSIs for prostate cancer and 433 labeled patches of 400 WSIs for breast cancer were sufficient to obtain competitive results. This confirms the effectiveness of the proposed combination of MIL and SSL techniques. Furthermore, it can adapt flexibly to any amount of available patch labels, as shown in the experiments of Figure 3. Depending on the available annotations, even unlabeled WSIs or completely annotated WSIs can be easily integrated in the training procedure. Regarding the best labeling strategy, the proposed efficient labeling strategy showed very good results with limited annotations, as highlighted in subsection III-E. It can be recommended for the future annotation of datasets. Still, there are some limitations of our method. When no patch labels are available, the proposed method can still be used for training, but the performance was not comparable to the supervised training result, as shown in Figure 3.

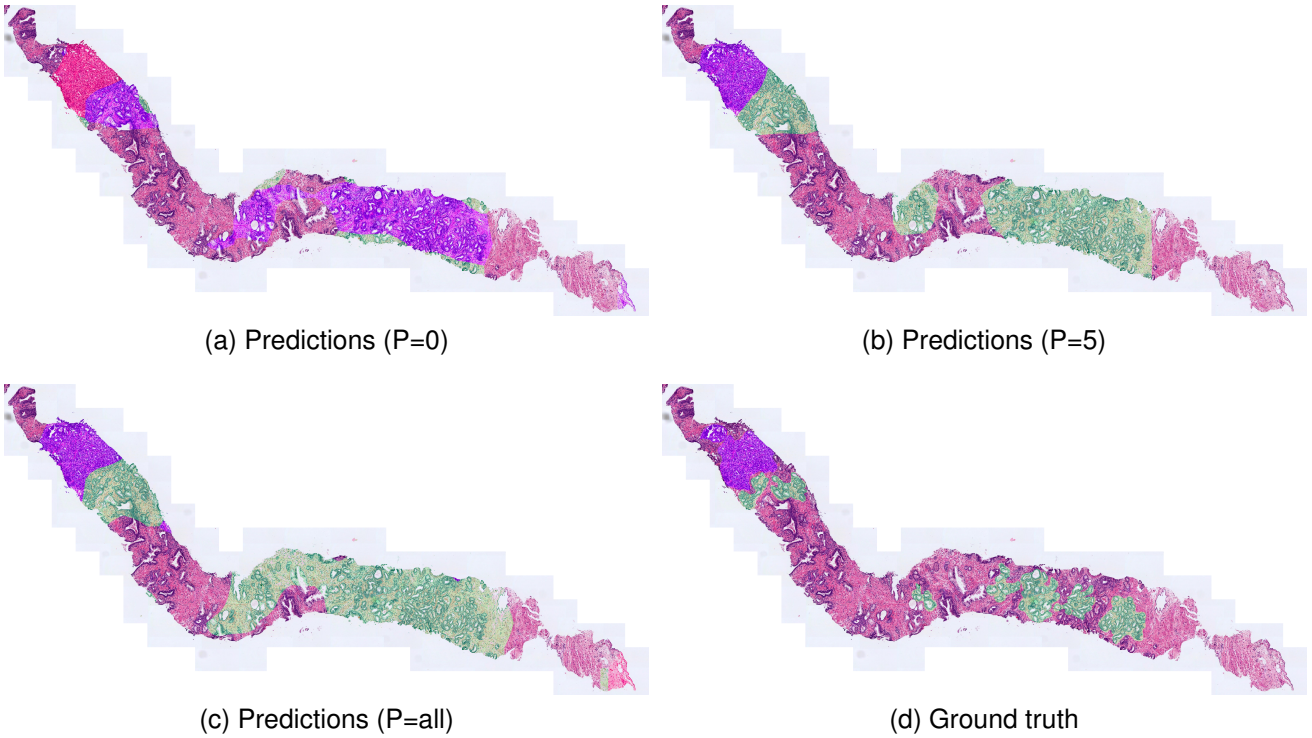

FIGURE 4. Visual example of model predictions for a test WSI of SICAPv2. The cancerous areas are marked in green (Gleason Grade 3), blue (Gleason Grade 4) and red (Gleason Grade 5). We compare the model predictions trained with $P = 0$ (MIL), $P = 5$ (some patch labels), $P = all$ (supervised), corresponding to the patch labels per class and WSI available during training. The marked areas of predictions are interpolated from patch-level predictions and therefore not as fine-grained as the ground-truth annotation. While the model trained in a MIL setting (a) correctly identifies the cancerous areas, the predicted classes are incorrect. The model predictions with the setting $P = 5$ depicted in (b) are very similar to those of the supervised model (c) and the ground truth (d).

In the default MIL setting, other specialized MIL methods might provide a better performance [2], [9]. Furthermore, our method assumes that the label classes on instance and bag level are the same. For problems where the local cancer class differs from the overall WSI labels, the proposed algorithm needs to be adjusted.

IV. CONCLUSIONS

We have presented a flexible deep learning framework for cancer classification which is able to make very accurate local as well as global predictions while requiring significantly fewer annotations than supervised approaches. The success of this approach can be attributed to the combination of semi-supervised and multiple instance learning as well as the proposed efficient labeling strategy, which was experimentally quantified. The work of the pathologist in our setting reduces to the annotation of some cancerous patches in each WSI and the final diagnosis. With this work, we hope to significantly contribute to the efforts of improving cancer diagnosis with the help of deep learning. By reducing the dependency on large, completely annotated datasets, we lower the threshold for new applications of artificial intelligence. With our approach, researchers and engineers can train deep learning models for cancer classification problems for which deep learning was not yet applied because of data limitations. To further improve our approach, we propose two future research directions: (i) active learning algorithms to choose the most discriminative patches for labeling and (ii) the use of an additional bag-level classifier based on the models feature maps to obtain even better WSI-level results.

REFERENCES

[1] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, p. 264, 2019.
[2] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
[3] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.
[4] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," pp. 3376–3391, 2018.
[5] M. Lerousseau, M. Vakalopoulou, M. Classe, J. Adam, E. Battistella, A. Carré, T. Estienne, T. Henry, E. Deutsch, and N. Paragios, "Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation," 2020, pp. 470–479.
[6] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation," *International Conference for Commputer Vision (ICCV)*, no. cMIL, 2019.
[7] Y. Huang and A. C. Chung, "Evidence localization for pathology images using weakly supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 613–621.
[8] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classifica-

tion," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020, pp. 519–528.

[9] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3851–3860.

[10] J. V. Pulido, S. Guleria, L. Ehsan, M. Fasullo, R. Lippman, P. Mutha, T. Shah, S. Syed, and D. E. Brown, "Semi-Supervised Classification of Noisy, Gigapixel Histology Images," in *International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 563–568.

[11] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[12] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 596–608.

[13] A. K. Jaiswal, I. Panshin, D. Shulkin, N. Aneja, and S. Abramov, "Semi-supervised learning for cancer detection of lymph node metastases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] S. Shaw, M. Pajak, A. Lisowska, S. Tsaftaris, and A. O'Neil, "Teacher-student chain for efficient semi-supervised histology image classification," in *International Conference on Learning Representations (ICLR)*, 2020.

[15] N. Marini, S. Otálora, H. Müller, and M. Atzori, "Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification," vol. 73, p. 102165, 2021.

[16] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *International Conference on Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 6256–6268.

[17] S. Otálora, N. Marini, H. Müller, and M. Atzori, "Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks," *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, vol. 12446 LNCS, pp. 193–203, 2020.

[18] J. Li, W. Chen, X. Huang, S. Yang, Z. Hu, Q. Duan, D. Metaxas, H. Li, and S. Zhang, "Hybrid supervision learning for pathology whole slide image classification," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021, pp. 309–318.

[19] M. Y. Lu, R. J. Chen, and F. Mahmood, "Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding," in *Medical Imaging 2020: Digital Pathology*, vol. 11320, 2020.

[20] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[21] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2005, p. 17.

[22] D. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," in *International Conference on Machine Learning (ICML)*, 2013.

[23] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2016, pp. 1171–1179.

[24] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5065–5074.

[25] J. Silva-rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the Gleason scoring scale : An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, 2020.

[26] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports*, vol. 8, no. 1, pp. 1–11, 2018.

[27] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.

[28] 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, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," pp. 167–180, 2018.

[29] "Camelyon16 challenge results," https://camelyon16.grand-challenge.org/Results/, accessed: 2021-12-21.

[30] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. V. van Dijk, P. Bult, F. Beca, A. Beck, D. yong Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. Lin, P. Heng, C. Hass, E. Bruni, Q. J. J. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Z. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernández-Carrobles, I. Serrano, Ó. Déniz, D. Racoceanu, and R. Venâncio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Journal of the American Medical Association*, vol. 318, p. 2199–2210, 2017.

ARNE SCHMIDT studied Mathematics and received his Bachelor degree at the Freie Universität Berlin in 2015 and his master degree at the Technische Universität Berlin in 2018.

After working for Astrofein GmbH, Fraunhofer Heinrich Hertz Institut and TomTom as a programmer specialized in deep learning, he started his PhD studies in 2020 under the supervision of Prof. Rafael Molina at the University of Granada. His research interests are probabilistic deep learning, Gaussian processes and crowdsourcing with applications to medical images. The PhD is part of the CLARIFY project which focuses on digital pathology for cancer classification.

JULIO SILVA-RODRÍGUEZ received Bach. and M.Sc. degree in biomedical engineering from Universitat Politècnica de València, Spain, in 2017 and 2018, respectively.

He is currently pursuing his Ph.D. studies at Universitat Politécnica de València (UPV). He has co-authored an awarded Best Paper at 21st International Conference on Intelligent Data Engineering and Automated Learning in 2020. His research interests focuses mainly on deep learning applied to medical image analysis, concretely under weakly and unsupervised scenarios.

VALERY NARANJO (received the Ph.D. degree in telecommunications in 2002. She is a Professor at the Universitat Politècnica de València, Spain. She lectures Digital Signal Processing, Digital Image Processing and Molecular Imaging in the Telecommunication and biomedical degree and in the Master's degree in biomedical engineering of the UPV. Since 2008 she has been carrying out her research work at the Institute for Research and Innovation in Bioengineering (I3B). Since 2016 she is the leader of the CVBLab (Computer Vision and Behaviour Analysis Laboratory) research group from the I3B where leads numerous national and European projects basically focused on image analysis and machine learning. Since 2021 she is also the I3B's director.

One of its main lines of research is the development of computer-aided systems applied to fields such as histopathology images, fundus images, magnetic resonance imaging, computerized axial tomography (CT), OCT, spectroscopy, etc. In the recent years she has specialized in Artificial Intelligence techniques applied to different industrial sectors.

RAFAEL MOLINA (SM 2013) received the M.Sc. degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from the University of Granada, Granada, Spain, in 1979 and 1983, respectively. He was the Dean of the Computer Engineering School, University of Granada, from 1992 to 2002, where he became a Professor of computer science and artificial intelligence in 2000. He was the Head of the Computer Science and Artificial Intelligence Department, University of Granada, from 2005 to 2007. He has coauthored an article that received the runner-up prize at reception for early stage researchers at the House of Commons in 2007. He has coauthored an awarded Best Student Paper at the IEEE International Conference on Image Processing in 2007, the ISPA Best Paper in 2009, and the EUSIPCO 2013 Best Student Paper. His research interest focuses mainly on using Bayesian modeling and inference in image restoration (applications to astronomy and medicine), super-resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, supervised learning, and crowdsourcing.

Dr. Molina has served as an Associate Editor for Applied Signal Processing from 2005 to 2007 and the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He has been serving as an Area Editor for Digital Signal Processing since 2011.