Annotation Protocol and Crowdsourcing Multiple Instance Learning Classification of Skin Histological Images: the CR-AI4SkIN Dataset

Rocío del Amor*[a], Jose Pérez-Cano*[b], Miguel López-Pérez*[b], Liria Terradez[c], Jose Aneiros-Fernandez[d], Sandra Morales[a], Javier Mateos[b], Rafael Molina[b], Valery Naranjo[a]

[a] *Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, Valencia, Spain*

[b] *Department of Computer Science and Artificial Intelligence, University of Granada, 18010 Granada, Spain.*

[c] *Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Spain.*

[d] *Pathology Department. Hospital San Cecilio de Granada, Granada, Spain.*

Abstract

Digital Pathology (DP) has experienced a significant growth in recent years and has become an essential tool for diagnosing and prognosis of tumors. The availability of Whole Slide Images (WSIs) and the implementation of Deep Learning (DL) algorithms have paved the way for the appearance of Artificial Intelligence (AI) systems that support the diagnosis process. These systems require extensive and varied data for their training to be successful. However, creating labeled datasets in histopathology is laborious and time-consuming. We develop a crowdsourcing-multiple instance labeling/learning protocol that is applied to the creation and use of the CR-AI4SkIN dataset[a]. CR-AI4SkIN contains 271 WSIs of 7 Cutaneous Spindle Cell (CSC) neoplasms with expert and non-expert labels at region and WSI levels. It is the first dataset of these types of neoplasms made available. The regions selected by the experts are used to learn an automatic extractor of Regions of Interest (ROIs) from WSIs. To produce the embedding of each WSI, the representations of patches within the ROIs are obtained using a contrastive learning method, and then combined. Finally, they are fed to a Gaussian process-based crowdsourcing classifier, which utilizes the noisy non-expert WSI labels. We validate our crowdsourcing-multiple instance learning method in the CR-AI4SkIN dataset, addressing a binary classification problem (malign vs. benign). The proposed methodobtains an F1 score of 0.7911 on the test set, outperforming three widely used aggregation methodsfor crowdsourcing tasks. Furthermore, our crowdsourcing method also outperforms the supervised model with expert labels on the test set (F1-score = 0.6035). The

*These authors contributed equally.

[a] The dataset will be publicly available upon acceptance of the paper.

promising results support the proposed crowdsourcing multiple instance learning annotation protocol. It also validates the automatic extraction of interest regions and the use of contrastive embedding and Gaussian process classification to perform crowdsourcing classification tasks.

Keywords: Histopathology, Skin cancer, Gaussian Processes, Multiple Instance Learning, Crowdsourcing

1. Introduction

Digital Pathology (DP) has experienced a significant growth in recent years, becoming essential for the diagnosis and prognosis of tumors. DP involves capturing, storing, and analyzing high-resolution digital images of tissues, known as Whole Slide Images (WSIs). WSIs are vital in the pathological diagnosis process because they allow easy data sharing, storing, and analysis on the computer [1]. WSI analysis provides pathologists with a comprehensive understanding of the data, leading to more accurate diagnoses of tumors and various cancer subtypes. Furthermore, the availability of WSIs has facilitated the implementation of novel computer vision techniques based on deep learning, which allow the automatic identification of new biomarkers and innovative features in the images to enhance the diagnostic process [2]. Unfortunately, for these deep learning techniques to perform effectively, they require large and diverse datasets [3].

Generating large-scale labeled histology datasets is a time-consuming and error-prone task. Recently, crowdsourcing has emerged as an appealing proceduretо labeling histopathological datasets. Crowdsourcing distributes the effort among a large number of annotators who may have varying degrees of expertise. In medical image-based diagnostic studies, crowdsourcing has produced accurate results in microtasks, e.g. nuclei detection [4] or identification of cancer cells [5]. In more specialized tasks, e.g. tissue classification, annotators with less expertise may introduce noisy labels [6, 7]. To address this issue, a common strategy is to aggregate the different labels to generate a more accurate label set [8]. Then, a regular classification method can be applied to this noise-free set of labels. However, recent research suggests that this strategymay not be optimal, as it typically leads to poorer performance compared to models that consider each annotator's confusion as part of the training process [9]. To address this issue, several methods have been developed to learn from noisy non-expert crowdsourcing labels, being their performance comparable to that of supervised methods that use expert labels in histopathological tissue classification [10, 11, 12].

Obtaining fine-grained WSI annotations through crowdsourcing remains a challenging task, as it requires a significant amount of effort and time from the annotators. They have to delineate and annotate the structures present in the WSIs. Multiple Instance Learning (MIL) is a promising solution to tackle the problem of detailed labeling. MIL considers samples to be grouped in bags. Then, the labels are collected per bag, and there is no need for (fine-grained) individual labels. In the case of MIL in WSI analysis, a bag represents a WSI, and the instances represent smaller regions within that slide. Therefore, the labels are only collected at WSI level, and there is no need for pixel or region labels, which streamlines the labeling process [13]. MIL methods have been applied to histopathological images with promising results [14, 15].

Based on these observations, in this work, we design and develop a crowdsourcing-MIL protocol to alleviate and distribute the burden of WSI labeling. This protocol combines two methods, (i) MIL: the global annotations are collected at the WSI level to alleviate the burden of detailed labeling, and (ii) crowdsourcing: the annotation effort is distributed among several non-expert annotators. We apply this protocol to create the *CRowdsourcing - Artificial Intelligence for cutaneouS spindle cell neoplasm hIstopathological diagNosis* (CR-AI4SkIN) dataset. CR-AI4SkIN comprises 271 WSIs of Cutaneous Spindle Cell (CSC) neoplasms with expert and non-expert WSI annotations. To the best of our knowledge, this is the first available dataset of CSC neoplasms and the first dataset of medical images with crowdsourcing-MIL annotations. Secondly, in this work, we also propose a new crowdsourced classifierfor histological image analysis that utilizes these noisy global labels. The method is named *Contrastive Learning Representations - Gaussian Processes for CRowdsourcing* (CLR-GPCR). It combines embedded-based MIL and crowdsourcing classification. We first learn to automatically select Regions of Interest (ROIs) from a reduced set of expertly annotated WSIs. Using a contrastive learning paradigm, we extract features from the patches of those ROIs. Then, we aggregate the embeddings to obtain a global latent representation of each WSI. Finally, we utilize Gaussian Processes for crowdsourcing classification with this global embedding and the noisy global WSI labels provided by the non-experts for each WSI. The overview of our proposed method is depicted in Figure 1. Specifically, the main contributions of our work are:

- A new procedurefor high-quality crowdsourcing dataset creation and a publicly-available histological dataset of 271 WSIs. They correspond to seven different types of spindle cell neoplasms diagnosed by two expert pathologists and 10 in-training pathologists.

- We propose CLR-GPCR, a novel formulation based on self-supervised learning and MIL combined with crowdsourcing Gaussian processes for the tumor classification task. We use the noisy WSI labels provided by non-expert pathologists. To the best of our knowledge, this is the first methodthat formulates and tackles crowdsourcing with global labels.

- Comprehensive experiments demonstrate the promising performance of our crowdsourcing method. With this method, we found averaged improvements of nearly $\sim 9.0\%$ in averaged F1-score compared to majority voting and expert labeling.

The remainder of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces and details the new publicly available

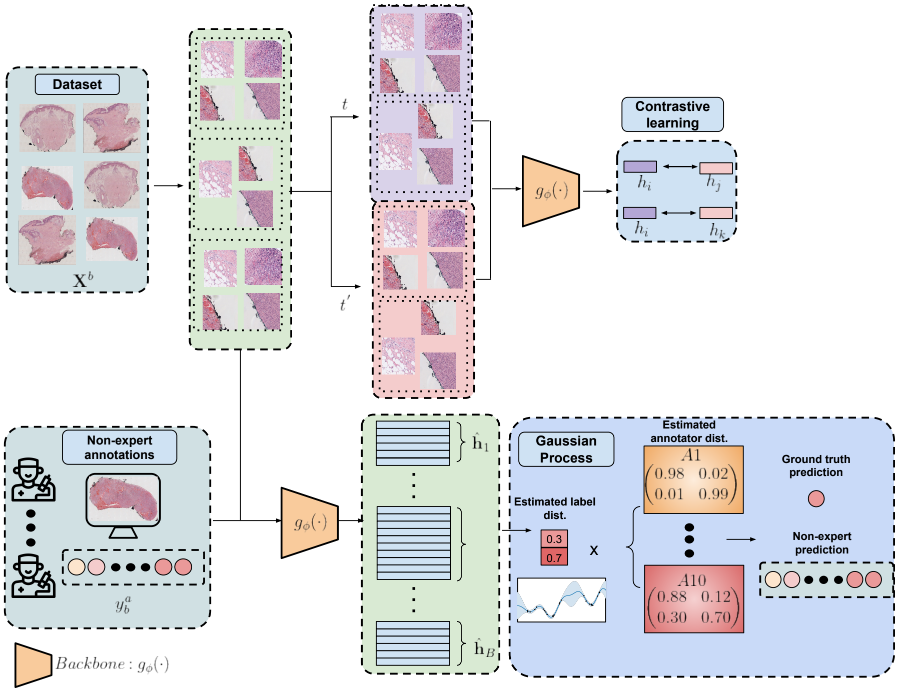


Figure 1: **Overview of the proposed CLR-GPCR model.** In this work, we address (weakly supervised) histology image classification on skin WSIs using noisy global labels provided by several non-expert annotators. Our method consists of two steps: 1) self-supervised feature extraction based on Contrastive Learning detailed in subsection 4.1; 2) Crowdsourcing classification using Gaussian processes, explained in subsection 4.2.

CR-AI4SkIN dataset. Section 4 details the proposed method for crowdsourcing-MIL, named (CLR-GPCR). This section is composed of Section 4.1, where the WSI embedding construction using self-supervised learning is explained, and Section 4.2, where we present the Gaussian Process-based method for crowdsourcing classification. The experiments and results, as well as a discussion of them, are described in the Section 5. Section 6 includes conclusions and future work. Finally, Appendix A includes further details on the CSC neoplasms that compose the CR-AI4SkIN dataset.

2. Related work

2.1. Digital pathology for skin cancer

According to the World Health Organization, nearly one in three diagnosed cancers worldwide is a skin cancer [16]. Different techniques, such as dermatoscopy, wood lamp, CT scan and histopathology, are utilized to diagnose this disease. However, the gold standard for skin cancer detection is histological image analysis. Traditionally, histological slides were viewed with a light microscope which is a highly time-consuming task. The digitization of biopsies has created opportunities for automated analysis of WSIs using machine learning-based methods. Applying deep-learning models to computer vision problems shows excellent potential in skin cancer detection. However, most research is based on the analysis of dermoscopic images [17, 18, 19, 20, 21, 22, 23] and few studies have been focused on the analysis of WSIs [24, 25, 26, 27, 28, 29]. Hekler et al. [24] used transfer learning on a pre-trained ResNet50 convolutional neural network (CNN) to differentiate between two classes, benign and melanoma tissues. The main limitation of this work is that they cannot analyze entire WSIs but only a characteristic tumor sub-region. In De Logu et al. [25], a pre-trained Inception-ResNet-v2 network was then used to distinguish cutaneous melanoma areas from healthy tissues. In [26], the authors developed a deep learning system to automatically detect malignant melanoma in the eyelid from histopathological sections. The authors used the VGG16 model to assign patch-level classification. The patches were embedded back into each WSI using the malignant probability from the CNN to generate a visualization heatmap. They utilized a random forest model to establish a WSI-level diagnosis between malignant and benign samples. Current methods based on MIL have been successfully applied to basal carcinoma (BCC) [29] or melanoma [27, 28], reducing the time required to perform precise annotations. However, many types of challenging skin cancer have not yet been explored. These include CSC neoplasms, predominantly composed of spindle-shaped neoplastic cells arranged in sheets and fascicles [30]. CSC neoplasms are challenging to diagnose due to the considerable morphological overlap between the different tumor types that make up this group [31], which poses a particular problem for less experienced pathologists. To the best of our knowledge, there is only one study in which deep learning techniques have been applied to automatically analyze this type of lesion [32]. In this case, techniques based on self-training were used to detect tumor regions. In line with that study, in this paper, we develop the first deep learning-based classifier to identify the malignancy or benignity of different types of CSC lesions.

2.2. Multiple Instance Learning in digital pathology

In the MIL framework, instances are grouped in bags and the only available labels are at the bag level. The standard MIL assumption considers that a bag is labeled as positive if at least one instance belongs to the positive class. Among other tasks, it has been used to detect breast cancer [33] and grade local patterns in prostate cancer [34]. This assumption makes sense when the labels at the pixel/region label directly affect the WSI label. However, there are cases where this is not true. Regarding CSC neoplasms, the WSI outcome is a combination of features among the different patches [27], which is called the bag-embedding MIL [15]. The most common technique is to obtain the bag-level representation by instance-level aggregation of features extracted from each instance by a CNN backbone. The feature extraction is frequently performed with pretrained networks [35], transfer learning [36], and more interestingly, following a contrastive learning strategy[37]. The contrastive learning representations do not need instance features, being useful to obtain bag descriptors. It is a self-supervised method, imposing similarity among similar patches (from the same WSI) by means of a contrastive loss [38]. Then, the aggregation of the patch features results in the bag embedding. The most straightforward and non-trainable aggregation techniques are batch global average (BGAP) [39] and batch global max pooling (BGMP) [33]. Other aggregation techniques include trainable parameters, such as weighted embeddings based on attention [40] or recurrent neural networks (RNN) [14]. Recently, authors in [41] proposed a Transformed based correlated MIL (TransMIL) that considers the morphological and spatial correlation between instances.

2.3. Crowdsourcing in digital pathology

The concept of learning from crowds was first introduced in the biomedical domain in 2016 [42]. The authors involved non-experts in an online system to label mitosis in breast cancer histological images. They adapted a CNN to learn from noisy observations. Since then, more sophisticated crowdsourcing labeling strategies have been developed for more complex tasks. Amgad et al. [6] labeled a triple negative breast cancer dataset with a panel composed of twenty medical students, three junior pathologists, and two senior pathologists. They designed a structured protocol where medical students segmented most of the WSIs. Junior and senior pathologists annotated the most challenging ones. In this protocol, medical students and junior pathologists obtained feedback and reviews from senior pathologists. Later, the same authors in [7] further annotated these images for nuclei classification. The protocol was similar and employed a collaborative effort of the different participants. They also utilized non-supervised segmentation methods and region labels to suggest delineations with associated classes.

Other recent works avoided label curation and directly utilized the crowdsourcing labels. The most straightforward way to utilize these noisy labels is to aggregate them by Majority Voting.More elaborated methods considered the biases of the different annotators, yielding a better-calibrated set of training labels, see [8, 43, 44, 45, 46]. The quality of the labels can be further improved by considering correlations with related samples, i.e., their nearest neighbors [47, 48].However, a recent work found that when labels from multiple annotators are available, methods that model observer confusion as part of the training process generally perform better than methods that aggregate the labels in a separate step prior to training [9]. In this vein, Nir et al. [10] used the Gleason2019 challenge[1] data to exploit the multiple opinions for Gleason grading. The authors jointly estimated a latent classifier (logistic regression) and the reliability of each participant during the learning process. They obtained an overall agreement with the pathologists consistent with the agreement levels reported in the literature. Following a similar strategy, López-Pérez et al. [11]

[1] https://gleason2019.grand-challenge.org/

applied crowdsourcing Gaussian Processes (GP) to breast cancer images labeled by several medical students. Crowdsourcing GPs trained with noisy labels were competitive with the ones trained with expert labels. They automatically estimated the reliability of each participant and a latent GP classifier for predicting the actual class.

These results suggested the feasibility of using data labeled by non-experts without the need for expert review to feed machine learning systems. Following this strategy, we propose a crowdsourcing methodfor CSC neoplasm classification under a multiple instance learning paradigm using WSI labels. To the best of our knowledge, this work is the first on crowdsourcing classification with (global) WSI labels in general and specifically applied and studied in CSC neoplasm classification.

3. CR-AI4SkIN: Data acquisition and annotation protocol

3.1. Dataset description

For this study, we used a scaled-down version of the CR-AI4SkIN dataset consisting of WSIs provided by the University Clinic Hospital of Valencia (HCUV) and San Cecilio University Hospital in Granada (HUSC) of skin tissue biopsies containing CSC neoplasms. A complete description of the CR-AI4SkIN dataset is included in Appendix A. For the purpose of this study, 227 images were used. In total, 123 were diagnosed as benign and 104 as malignant.

3.2. Methodology Part 1: Extraction of expert and automatically selected ROIs

Since WSIs are high-resolution images, not all the information is relevant for the final diagnosis. Aiming at pre-processing the biopsies and reducing the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the dataset. Additionally, expert pathologists annotated the regions of interest (ROIs), i.e. the tumoral areas that have a clinical impact on the outcome, in 15% of the slides using in-house software based on the OpenSeadragon library [49]. These expert annotations were then used to train an automatic ROI prediction algorithm, designed as a teacher-student network and trained on a few WSIs labeled by the expert pathologists. For further details of the methodology employed, see [32]. Note that in this case 10% was used to train the ROI extraction method and the rest to validate the automatic ROI extraction process. Finally, all WSIs in the database were automatically analyzed by the ROI prediction algorithm, obtaining the ROIs of each WSI. The final dataset is composed of 512 × 512 patches (with 50% of overlap) from the ROIs at a magnification of 10×.

The purpose of the ROI extraction was twofold. First, it aimed at guiding non-expert participants through the labeling process. Second, since the images did not fit in computer memory, only the ROIs were passed to the classification method.

Table 1: Number of images used for training, validating, and testing the models of each non-expert annotator. Note that for the validation and test set, the same samples labeled by all non-experts were used.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Train	148	142	151	143	154	145	155	149	152	150
Val	22									
Test	44									

3.3. Methodology Part 2: Non-expert annotation protocol

Ten non-expert pathologists participated in the annotation of the whole dataset: four resident pathologists from HCUV and six from HUSC. 70% of non-expert pathologists were in their third and fourth year of residency, while the remaining 30% were first and second-year medical residents. We designed the following annotation protocol for the annotation process. First, to simplify the task of analyzing WSIs for non-expert participants, we provided them with the ROI generated by the automatic algorithm trained on expert annotations, see subsection 3.2. During the examination of the WSIs, the non-experts performed two tasks: i) first, they indicated whether they agreed with the ROI proposals or wished to manually annotate additional regions. If they disagreed with every ROI proposal, they were compelled to manually delineate at least one ROI. An ROI proposed by the model is correct if at least 50% of the area contains tumoral tissue; ii) secondly, they also assigned a global label to the WSI from the set of the seven classes.

To mitigate individual biases and test the concordance among annotators, from the 271 WSIs that compose the CR-AI4SkIN dataset, we gathered 106 of them, referred to as the “dense set”, that were annotated by all non-expert participants. The remaining images were only annotated by a subset of non-expert pathologists. Table 1 displays the images annotated by each non-expert pathologist for training, validation, and testing of the models. To ensure fair comparisons, the validation and test images were chosen from the dense set.

4. Classification framework for Multiple Instance Learning with crowdsourcing labels

In our medical problem, we observe the training data $\mathcal{D} = \{(\mathbf{X}_b, \mathbf{y}_b^a) : b = 1, ..., B; a \in A_b\}$, where $\mathbf{X}_b$ is the b-th WSI and $\mathbf{y}_b^a$ is the label provided by the a-th non-expert annotator for the b-th WSI. There are A different annotators and we note by A_b the set of annotators who provided a label to the b-th WSI. In this work, we address a binary classification problem. For this purpose, we group the malignant and the benign classes. We left out from this study the afx lesions as the distinction between benign and malignant is unclear. The WSI labels for each annotator are $\mathbf{y}_b^a \in \{0, 1\}$, denoting 1 the positive (malign) class. Under the MIL paradigm, each WSI is a bag composed of patches, i.e., $\mathbf{X}_b = \{\mathbf{x}_i | i \in \text{bag } b\}$. These patches do not have an associated label that could determine the WSI label, but rather the combination of features of the different patches determines the label of the WSI. In this case, we aim to train a model

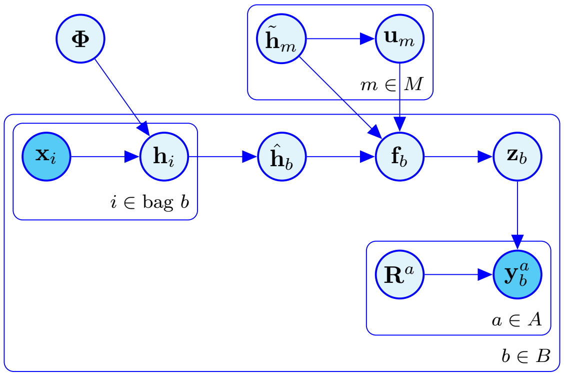


Figure 2: **Probabilistic graphical model of our CLR-GPCR model.** Dark circles stand for observed variables, while light circles stand for latent variables. We project the patches $\mathbf{x}_i$ onto a latent embedding space to obtain their low-dimensional representations $\mathbf{h}_i$, using a neural network parametrized by the set of weights $\mathbf{\Phi}$ with contrastive learning. Then, the low-dimensional representations are aggregated using the average to obtain a global embedding of the whole ROI, $\hat{\mathbf{h}}_b$. Gaussian Processes for Crowdsourcing model the latent real label of the ROIs, $\mathbf{z}_b$. Then, the observed (noisy) labels $\mathbf{y}_b^a$ are obtained using the latent real $\mathbf{z}_b$ and the reliability matrix $\mathbf{R}^a$ for each annotator. We introduce the sparse approximation for inference, i.e., we summarize the GP with a set of M inducing points $\mathbf{u}_m$ with inducing locations $\tilde{\mathbf{h}}_m$.

capable of predicting bag-level labels using a combination of features extracted at the instance level. This learning strategy falls under the embedding-based MIL paradigm[2].The following sections explain the feature extraction process to obtain bag embeddings and the crowdsourcing classifier.

4.1. Self-Supervised feature extraction: SimCLR

To obtain a low-dimensional representation for each instance from the predicted ROIs (see subsection 3.1), we use a self-supervised method. Typically, self-supervised methods are trained so that similar images have embeddings that are close and dissimilar images have embeddings far away from each other. More concretely, we have chosen to use the SimCLR method [38]. Under this framework, the notion of similarity is built around a set of predefined transformations $\mathcal{T}$. These are commonly augmentation transformations, e.g., rotations, translations, etc. Then, two images are considered similar if there exists a transformation $t \in \mathcal{T}$ that converts one into another. To learn the weights, SimCLR utilizesthe normalized temperature-scaled cross-entropy loss (NT-Xent) [50, 51, 52] defined as follows:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k=1}^{2S} \mathbb{1}_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)} \tag{1}$$

where $(\mathbf{h}_i, \mathbf{h}_j)$ are the embeddings (i.e., the output of the encoder network)of two augmented patches $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i')$ from the same patch $\mathbf{x}_i \in \mathcal{X}_b$. Each augmented

[2] Based on the denomination proposed in [40].

patch is obtained by applying a different augmentation $t, t' \in \mathcal{T}$ so that $\mathbf{h}_i = g_{\mathbf{\Phi}}(\tilde{\mathbf{x}}_i) = g_{\mathbf{\Phi}}(t(\mathbf{x}))$ and $\mathbf{h}_j = g_{\mathbf{\Phi}}(\tilde{\mathbf{x}}'_i) = g_{\mathbf{\Phi}}(t'(\mathbf{x}))$. In total, S different patches are selected, being S the batch size. τ is the temperature that controls how smooth the function is and sim : $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is any similarity function between vectors. We use the cosine similarity function, which is normally used [38, 37]. The other embeddings $\mathbf{h}_k$ correspond to augmentations of other images in the batch, that is, they are dissimilarsamples. This loss function $(l_{i,j})$ is low if the similarpair has resembling embeddings, however, the resemblance of $\mathbf{h}_i$ with respect to other dissimilarsamples in the batch is taken into account in the denominator. This way $\mathbf{h}_i$ not only must be close to $\mathbf{h}_j$ but also far from the other $\mathbf{h}_k$.

Once we have obtained a feature descriptor for each patch of the region of interest, we aggregate them using the average. If we have a bag of patches $\{\mathbf{x}_i\}_{i=1}^{N_b}$, then the embedding of the bag is $\hat{\mathbf{h}}_b = \frac{1}{N_b}\sum_{i=1}^{N_b} g_{\mathbf{\Phi}}(\mathbf{x}_i)$. Where $g_{\mathbf{\Phi}}$ is the nonlinear projection learned by minimizing the loss defined in eq. (1).

4.2. Crowdsourcing classification: Sparse Gaussian Processes

SVGPCR predicts the observed noisy WSI label $\mathbf{y}_b^a$ for each WSI embedding $\hat{\mathbf{h}}_b$ using a (latent) Gaussian Process (GP) $\mathbf{f}$ and a global reliability matrix $\{\mathbf{R}^a\}_{a=1}^{A}$ for each non-expert annotator. The reliability matrix of each annotator is governed by two parameters $\{\alpha_a, \beta_a\}$. This matrix is expressed as follows:

$$\mathbf{R}^a = \begin{pmatrix} \alpha_a & 1-\beta_a \\ 1-\alpha_a & \beta_a \end{pmatrix} \tag{2}$$

where the parameters are the specificity and sensitivity of each annotator, i.e., $\mathrm{p}(y_b^a = 0 | z_b = 0) = \alpha_a$ and $\mathrm{p}(y_b^a = 1 | z_b = 1) = \beta_a$. Then, the probability of the observed (noisy) label for each non-expert annotator is given by the following Bernoulli distribution:

$$\mathrm{p}(y_b^a | z_b, \mathbf{R}^a) = \prod_{a=1}^{A} \left[(\alpha_a^{y_b^a} (1-\alpha_a)^{1-y_b^a} \right]^{z_b} * \left[\beta_a^{1-y_b^a} (1-\beta_a)^{y_b^a} \right]^{1-z_b} \tag{3}$$

We assume that the annotators label the different samples independently,

$$\mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{R}) = \prod_{b=1}^{B} \prod_{a \in A_n} \mathrm{p}(y_b^a | z_b, \mathbf{R}^a). \tag{4}$$

The prior distribution for the annotator behavior is modeled with a Beta distribution on the parameters α and β which is conjugated with the Bernoulli distribution of Eq. (3). This prior distribution introduces the prior beliefs on the annotators' expertise. In this work, we use a non-informative prior since this information is unavailable. Finally, the latent (real) label $\mathbf{z}$ is estimated using a latent variable, $\mathbf{f}$. The likelihood $\mathrm{p}(\mathbf{z}|\mathbf{f})$model for this problem is a Bernoulli distribution, whose parameter is obtained by applying a sigmoid function to $\mathbf{f}$.

SVGPCR imposes a GP prior on $\mathbf{f}$, i.e., $\mathbf{f}|\hat{\mathbf{H}}$ follows a multivariate Gaussian distribution.

SVGPCR also introduces M inducing points for scalability $\{\mathbf{u}_m\}_{m=1}^{M}$, which summarizes the information of the GP $\mathbf{f}$. These inducing points are the realizations of the GP in the inducing locations $\{\tilde{\mathbf{h}}_m\}_{m=1}^{M}$, i.e., $\mathbf{u} = \mathbf{f}(\tilde{\mathbf{h}})$. The probabilistic model is given by

$$\mathrm{p}(\mathbf{y}, \mathbf{z}, \mathbf{f}, \mathbf{u}, \mathbf{R}|\boldsymbol{\Theta}) = \underbrace{\mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{R})\mathrm{p}(\mathbf{R})}_{\text{CR modeling}} \underbrace{\mathrm{p}(\mathbf{z}|\mathbf{f})}_{\text{GP likelihood}} \underbrace{\mathrm{p}(\mathbf{f}|\mathbf{u}, \boldsymbol{\Theta})\mathrm{p}(\mathbf{u}|\boldsymbol{\Theta})}_{\text{GP prior}}. \tag{5}$$

An overview of the graphical probabilistic model is depicted in Figure 2. The goal here is to obtain the posterior $\mathrm{p}(\mathbf{z}, \mathbf{f}, \mathbf{u}, \mathbf{R}|\mathbf{y}, \boldsymbol{\Theta})$. To fulfill this, we perform stochastic variational inference [53]. We aim to find an approximate posterior $\mathrm{q}(\mathbf{z}, \mathbf{f}, \mathbf{u}, \mathbf{R})$ by maximizing the Evidence Lower BOund (ELBO), see [54] for further details on the inference process.

Once we learn the posterior distribution q, we predict new WSIs with the following process. First, we extract features using the fine-tuned neural network (detailed in subsection 4.1) and aggregate them using the average pooling to obtain a global vector for this new WSI, $\hat{\mathbf{h}}_*$. Then, we use the trained SVGPCR model on these features, $\hat{\mathbf{h}}_*$, to predict $\mathrm{p}(z_*)$ via Monte Carlo sampling.he whole training and inference procedures are summarized in Algorithm 1 and 2.

Algorithm 1: CLR-GPCR High-level description of the training loop

```
Data: A set of WSIs with crowdsourced labels
Result: A trained encoder network $\boldsymbol{\Phi}$ and SVGP for WSI classification
Initialize encoder network $\boldsymbol{\Phi}$
Initialize temperature $\tau$
dataset ← All the images (without labels)
for epoch in range(num_epochs) do
    for batch in DataLoader(dataset) do
        images ← augment(batch)
        embeddings ← $\boldsymbol{\Phi}$(images)
        loss ← NT-XENT(embeddings, $\tau$)
        update_parameters(loss, $\boldsymbol{\Phi}$)
embeddings ← $\boldsymbol{\Phi}$(dataset)
dataset ← aggregate_by_patient(embeddings) ∪ crowdsourced labels
Initialize SVGP
for epoch in range(num_epochs) do
    likelihood ← compute_likelihood(dataset, SVGP)
    update_parameters(likelihood, SVGP)
```

```
Algorithm 2: CLR-GPCR High-level description of one inference step
Data: A batch of WSIs, a trained encoder network $\mathbf{\Phi}$ and a trained
      SVGP
Result: The probabilities of each WSI of being malignant
predictions $\leftarrow \emptyset$
for WSI in batch do
    patches $\leftarrow$ extract_patches(WSI)
    embeddings $\leftarrow \mathbf{\Phi}$(patches)
    global_embedding $\leftarrow$ average(embeddings)
    mean, variance $\leftarrow$ SVGP(global_embedding)
    predictions $\leftarrow$ predictions $\cup$ mean
```

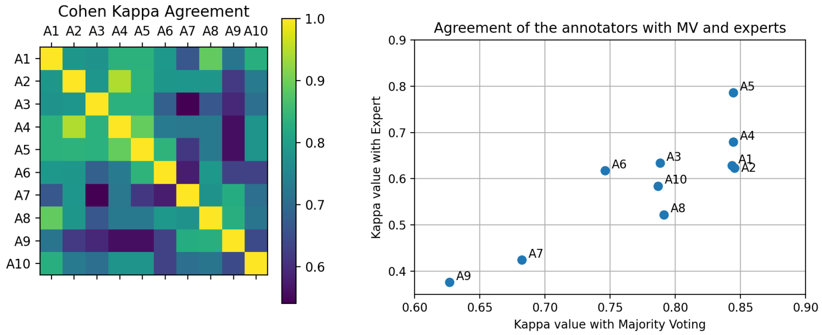
Table 2: Number of images used for training, validating and testing the models. B: benign lesions; M: malignant lesion; MV: majority voting of labels provided by non-experts; % Dis: % discrepancy between expert pathologists and majority voting of non-experts.

	Experts		**MV**		**% Dis**
	B	**M**	**B**	**M**	
Train	95	55	107	44	12%
Val	9	13	16	6	31.81%
Test	19	25	26	18	15.90%

5. Experiments and Results

5.1. Implementation

1) **Dataset (CR-AI4SKIN)**: We used the binary version of theCR-AI4SKIN databasedescribed in section 3 to assess the efficacy of our proposed crowdsourcing-MIL method (CLR-GPCR). The dataset was divided into three splits (training, validation and testing) containing 70%, 10% and 20% of the global dataset, respectively. Several slides can belong to the same patient. We ensured a strict separation of patients among the three sets. Table 2 displays the difference in labeling provided by experts and the majority vote (MV) of non-experts for the training, validation and test sets in terms of accuracy. Additionally, we report the kappa agreement value for the non-expert annotators. The kappa value ranges between -1 and 1, where -1 represents total disagreement, 0 represents randomness, and 1 represents total agreement. The kappa is an insightful metric to assess agreement among annotators as it considers the possibility of concurrence due to randomness. We calculate this metric for the test images, which all participants annotated, and these values are displayed in Figure 3. The non-expert annotators show a low-moderate level of agreement among them, as shown in Figure 3a. While some pairs exhibit strong agreement with values over 0.8, others show weak agreement with values near 0.5. Furthermore, the annotators tend to agree more frequently with the MV labeling than with the expert labeling, recall Figure 3b. While some have a strong agreement with the expert labeling, others display a weak agreement. From this figure, we can conclude that the annotators have varying levels of expertise.



(a)

(b)

Figure 3: Kappa values for agreement among annotators. (a) The kappa values for each pair of non-expert annotations; (b) the kappa value for each non-expert annotator and either the expert annotation or the MV.

2) **Model hyperparameters**: CLR-GPCR consists of two modules: a self-supervised feature extractor pretrained with all the images (SimCLR) and a Gaussian process-based classifier using crowdsourcing labels (SVGPCR). We used RESNET18 with an output dimension of 256 features as the backbone for the feature extractor. Note that the inputs for the feature extractor are the ROIs extracted from each WSI automatically by the ROI prediction algorithm [32]. We optimized the feature extractor with Adam, a learning rate of 10^{-5} and weight decay of 10^{-6}, which remained fixed for all experiments. Additionally, we used the cosine similarity with a temperature value of 0.5 in the loss function. In the case of the Gaussian process classifier, the only fixed hyperparameter was the Adam optimizer. The remaining hyperparameters were tuned using the validation set, i.e., the learning rate and the number of inducing points. The number of epochs was adapted to the experiment performed. Regarding the labels used for the training, we consider three different:

- Expert labels. They are the labels provided by expert pathologists. They are considered the ground truth for this task. These labels may also contain some noise due to the high inherent subjectivity in this field.

- Aggregated labels. They are the result of combining non-expert annotations. In this work, we employ three distinct methods: Majority Voting, Dawid-Skene [43] and MACE [44]. Majority Voting offers a simple aggregation method that does not consider annotators' biases. In contrast, Dawid-Skene explicitly computes and integrates annotators' biases using confusion matrices, similar to our method. MACE adopts a distinctive approach by utilizing annotators' biases selectively, particularly when detecting potential spamming behavior. Furthermore, MACE assumes that annotators' biases remain consistent across instances, regardless of the true underlying labels.

- Crowdsourcing labels. They are the noisy labels the non-expert annotators provide without any aggregation or processing. Our proposed CLR-GPCR uses these labels.

3) **Evaluation**: The quantitative comparison of the different methodologies was handled using different figures of merit for binary classification, such as accuracy (ACC), F1-score (F1S), and area under ROC Curve (AUC). Note that in the different experiments carried out, these metrics represent the average of 10 executions to obtain more precise estimations. Results for each model were presented based on the type of label used during training, and expert labels were always used for evaluation.

5.2. Ablation experiments

1) **The role of SimCLR**: This section examines the self-supervised feature extraction process. The great advantage of this procedure is that it does not need labels. Three configurations for feature extraction using a RESNET18 [55] as the backbone were analyzed, and their performance was evaluated based on how well the subsequent classification model performed. The first configuration utilized the pre-trained weights reported by Bin Li et al. [37] for WSI lung tumor detection. The other two configurations obtained the weights through fine-tuning the SimCLR framework [38] on our data. Since SimCLR is known to be sensitive to batch size [38, 37], two models were trained with batch sizes of 256 and 512. To compare the performance of the different feature extractors, we used SVGPCR classifier first proposed in [54]. Figure 4 depicts the results of the three configurations. Note that SVGPCR was trained using noisy annotations from each non-expert annotator.

The model that was trained using a batch size of 256 consistently performed better. Due to these findings, we used a batch size of 256 for further experiments. Additionally, these results demonstrate the effectiveness of retraining the feature extractor on our dataset compared to simply using a network pre-trained on an external dataset. From this section, we evaluate the global descriptor from the WSIs (in this case, from their ROIs). This global descriptor does not depend on labels but only visual features.

2) **Crowdsourcing study**: The performance of the Sparse Gaussian Processes model is largely influenced by two hyperparameters: the learning rate and the number of inducing points. To study the effect of the learning rate, we performed a grid search using 4 different learning rates (10^{-4}, 10^{-3}, 10^{-2} and 10^{-1}) with 4 different numbers of inducing points (16, 32, 64, and 128). We evaluated each combination using the same 3 binary classification metrics as in the previous section (F1 Score, Accuracy and ROC AUC). The results are displayed in Figure 5 with the learning rate on the X-axis, separated by the number of inducing points. It can be seen that a low learning rate results in poorer performance compared to higher values. If the learning rate is too low, the model can fail to converge. This is why using a learning rate of 10^{-4} or 10^{-3} yields worse results. For larger values, the model does converge. It was observed that a learning rate of 10^{-2} allows the model to generalize better when

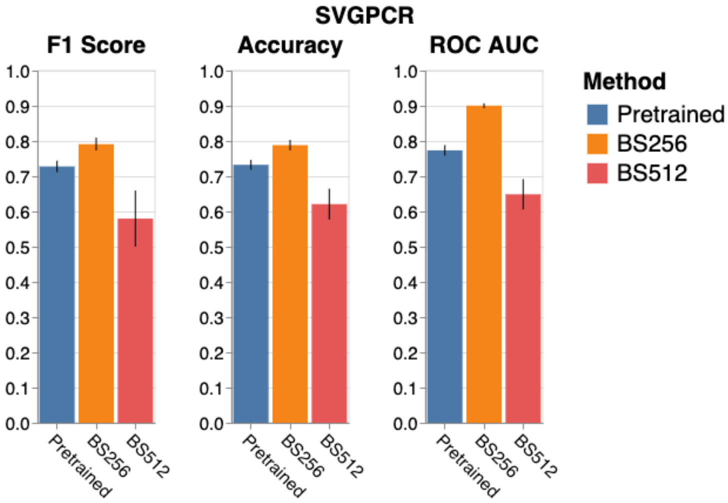


Figure 4: Influence of the SimCLR feature extractor on the classification stage. Metrics are averaged across ten trials. Each chart represents the result given a different backbone, where BS refers to the batch size employed when fine-tuning. The pre-trained model is obtained from [38]. The values correspond to the combination of hyperparameters that performed best in the validation set.

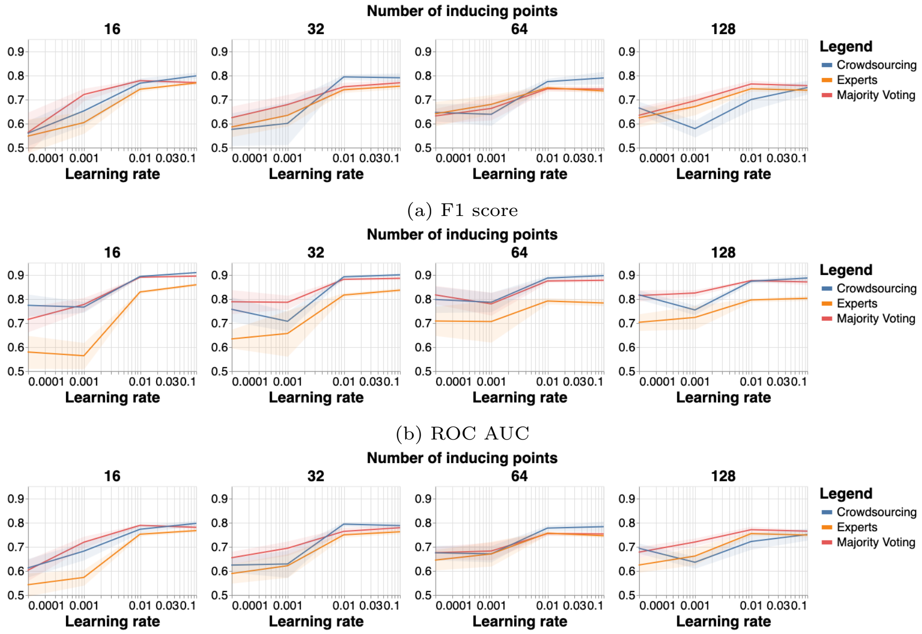
different inducing points are established. For this reason, we set the value of the learning rate to 10^{-2} for the rest of the experiments.

In Figure 6, we examine the influence of inducing points while keeping the learning rate fixed at 10^{-2}. Since we have a limited number of samples, we only use up to 128 inducing points. Although this hyperparameter affects the model's performance, its impact is not as significant as that of the learning rate. As shown in Figure 6, increasing the number of inducing points appears to worsen the results of our method. A small number of inducing points seem to improve the performance because it summarizes well the information and generalize better. Based on this information, we conclude that the optimal number of inducing points is 32.

5.3. Results

After performing the ablation study, we concluded that the best result was obtained by retraining the SimCLR method with a batch size of 256. For our classification method, the optimal hyperparameter values were a learning rate of 0.01 and a number of inducing points of 32. We compare our crowdsourcing method with the same classifier trained with aggregated labels. We use three different aggregation methods: Majority Voting, Dawid-Skene[43] and MACE [44]. We also train the classifier with expert labels. Results are presented in Table 3.

Our proposed CLR-GPCR outperforms all the aggregation methods, justifying the use of all the labels during classifier learning and not only an aggregation



(a) F1 score

(b) ROC AUC

(c) Accuracy

Figure 5: Each line chart represents the performance of the models trained with different labels and different learning rates but fixing the number of inducing points. Each experiment was repeated 10 times to estimate the mean value and the confidence interval.

of them. Furthermore, the remarkable performance of our method and Dawid-Skene highlight the importance of explicitly modeling the bias of each annotator (as highlights in Table 2).

Furthermore, our method also improves over the expert labels by a relative improvement of 31%. Since histopathological analysis is an error-prone task, we can benefit from a larger number of annotators. A sufficiently large number of annotators may reduce the variance in the labels. In this work, we built the dataset with 10 in-training pathologists, we can suppose that the majority might be right in most cases. For example, when one is wrong or is prone to make more mistakes, we can rely on the rest. A single mistake should not ruin or introduce a source of sensible noise to the classifier. Although the expert labels are considered the ground truth for this data, our method leverages crowdsourcing labels provided by non-experts.

5.4. Computational cost

Our method CLR-GPCR is scalable on the WSI size. The complexity of the encoder network (i.e., a convolutional network) scales linear, and SVGP yields a computational cost of $\mathcal{O}(M^2N_s)$ where M is the number of inducing points and N_s is the batch size [53]. Notice that the number of inducing points and batch size remain constant regardless of the volume of data.

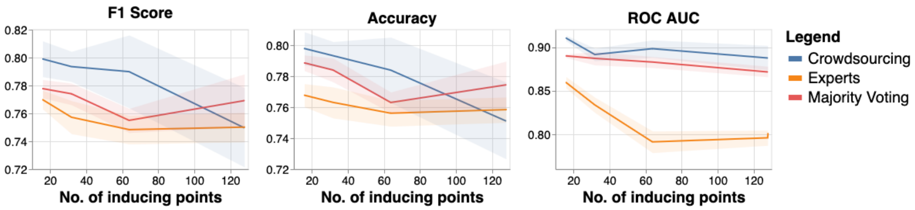


Figure 6: Comparison of the performance for different values of inducing points fixing the value of the learning rate.

Table 3: Different metrics on the test set of the same model trained with five different types of labels: expert, majority voting, David Skene and MACE aggregations of the non-expert annotators, and the crowdsourcing method, which uses all the labels of the non-expert annotators (without previous aggregation). Metrics are averaged across ten trials.

	F1 Score	Accuracy	ROC AUC
Expert labels	0.6035 ± 0.0451	0.5721 ± 0.0326	0.5629 ± 0.0544
Majority Voting	0.7200 ± 0.0254	0.7186 ± 0.0240	0.7764 ± 0.0286
David Skene [43]	0.7734 ± 0.0141	0.7767 ± 0.0114	0.8873 ± 0.0072
MACE [44]	0.6797 ± 0.0667	0.6837 ± 0.0532	0.7222 ± 0.0704
CLR-GPCR	**0.7911 ± 0.0177**	**0.7884 ± 0.0138**	**0.9007 ± 0.0063**

The most time-consuming aspect of the training process was the SimCLR method. The model trained with a batch size of 256 took 1.73 days to complete using two GPUs (NVIDIA TITAN X and NVIDIA GeForce RTX 2080 Ti). The model with a batch size of 512 required more memory and time, taking 5.13 days to be trained using two NVIDIA GeForce RTX 3090 GPUs.

In comparison, training the Gaussian processes was significantly faster. Each model only took a few minutes to fit to the data. To conduct the hyperparameter study, we ran each model several times, consuming a total of approximately 4 hours and 30 minutes of computation on the CPU and 1 hour and 30 minutes on the GPU. CLR-GPCR was trained on a single GPU, NVIDIA GeForce GTX 980 Ti.

The inference time is notably short. The neural network employed to extract embeddings has 11,560,896 parameters, and it takes 5.62 ms ± 1.07 ms to process one patch with a MacBook Pro M1 Max. The Gaussian process contains just 51,402 parameters and is even more efficient, requiring only 223 µs ± 1.5 µs to classify one WSI with a MacBook Pro M1 Max.

6. Conclusion and Future Work

This paper introduces CR-AI4SkIN, a new dataset of 271 WSIs containing seven different types of spindle cell neoplasms. We design an annotation protocol, combining expert and non-expert annotations with global WSI labels. We develop an automatic ROI extraction algorithm with a few expert annotations to discard irrelevant tissue and non-tumoral areas. These ROIs were shown to the non-expert participants to facilitate the labeling, and we only passed the ROIs to the classification method (to avoid memory issues). Our protocol scales and speeds up the labeling process by distributing and alleviating the annotation burden. To the best of our knowledge, this is the first dataset combining non-expert annotations (crowdsourcing) with global WSI labels (Multiple Instance Learning).

We propose and validate a Crowdsourcing-MIL method called CLR-GPCR for WSI classification. We perform the classification in two stages. First, we use self-supervised learning to obtain WSI representations. Second, we use crowdsourcing Gaussian Processes to classify the WSIs with crowdsourcing labels. Our approach leverages the developed annotation protocol and only requires global WSI labels provided by multiple noisy annotators.

Experimental results show the effectiveness of the proposed method and annotation protocol in distinguishing between malign and benign. Our method obtains satisfying results and outperforms the widely used Majority Voting in crowdsourcing. Our method, with non-expert labels, also outperforms the supervised method trained with single expert labels.

Future work will address the multiclass classification of the seven types of CSC neoplasms. We further improve the modeling for the annotators' behavior in this task, e.g., by considering the years of experience. Since we perform the classification in two independent stages, we will further investigate how to optimize the crowdsourcing and MIL parts end-to-end.

Funding

This work has received funding the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C21 (AI4SKIN), from Horizon 2020, the European Union's Framework Programme for Research and Innovation, under the grant agreement No. 860627 (CLARIFY), grant B-TIC-324-UGR20 funded by Consejería de Universidad, Investigación e Innovación (Junta de Andalucía) and by "ERDF A way of making Europe", and GVA through the project INNEST/ 2021/321 (SAMUEL). The work of Rocío del Amor has been supported by the Spanish Ministry of Universities (FPU20/05263). The work of Miguel López Pérez has been supported by the University of Granada postdoctoral program "Contrato Puente". The work of Sandra Morales has been co-funded by the Universitat Politècnica de València through the program PAID-10-20.

References

[1] D. R. Snead, Y.-W. Tsang, A. Meskiri, P. K. Kimani, R. Crossman, N. M. Rajpoot, E. Blessing, K. Chen, K. Gopalakrishnan, P. Matthews, et al., Validation of digital pathology imaging for primary histopathological diagnosis, Histopathology 68 (7) (2016) 1063–1072.

[2] C. L. Srinidhi, O. Ciga, A. L. Martel, Deep neural network models for computational histopathology: A survey, Medical Image Analysis 67 (2021).

[3] J. Van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic, Nature medicine 27 (5) (2021) 775–784.

[4] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, A. H. Beck, Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd, in: Pacific symposium on biocomputing Co-chairs, World Scientific, 2014, pp. 294–305.

[5] J. Lawson, R. J. Robinson-Vyas, J. P. McQuillan, A. Paterson, S. Christie, M. Kidza-Griffiths, L.-A. McDuffus, K. A. Moutasim, E. C. Shaw, A. E. Kiltie, et al., Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays, British journal of cancer 116 (2) (2017) 237–245.

[6] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, et al., Structured crowdsourcing enables convolutional segmentation of histology images, Bioinformatics 35 (18) (2019) 3461–3467.

[7] M. Amgad, L. A. Atteya, H. Hussein, K. H. Mohammed, E. Hafiz, M. A. Elsebaie, A. M. Alhusseiny, M. A. AlMoslemany, A. M. Elmatboly, P. A. Pappalardo, et al., Nucls: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer, GigaScience 11 (2022).

[8] A. Grote, N. S. Schaadt, G. Forestier, C. Wemmert, F. Feuerhake, Crowdsourcing of histological image labeling and object delineation by medical students, IEEE Transactions on Medical Imaging 38 (5) (2018) 1284–1294.

[9] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Medical Image Analysis 65 (2020) 101759.

[10] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, et al., Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts, Medical image analysis 50 (2018) 167–180.

[11] M. López-Pérez, M. Amgad, P. Morales-Álvarez, P. Ruiz, L. A. Cooper, R. Molina, A. K. Katsaggelos, Learning from crowds in digital pathology using scalable variational gaussian processes, Scientific reports 11 (1) (2021) 1–9.

[12] M. López-Pérez, P. Morales-Álvarez, L. A. D. Cooper, R. Molina, A. K. Katsaggelos, Deep gaussian processes for classification with multiple noisy

annotators. application to breast cancer tissue classification, IEEE Access 11 (2023) 6922–6934.

[13] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, P. Honeine, Multiple instance learning for histopathological breast cancer image classification, Expert Systems with Applications 117 (2019) 103–111.

[14] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Medicine 25 (8) (2019) 1301–1309.

[15] Chapter 22 - deep multiple instance learning for digital histopathology, in: S. K. Zhou, D. Rueckert, G. Fichtinger (Eds.), Handbook of Medical Image Computing and Computer Assisted Intervention, The Elsevier and MICCAI Society Book Series, Academic Press, 2020, pp. 521–546.

[16] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, D. Ioannides, Epidemiological trends in skin cancer, Dermatology practical & conceptual 7 (2) (2017) 1–6.

[17] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.

[18] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, nature 542 (7639) (2017) 115–118.

[19] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, et al., Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task, European Journal of Cancer 113 (2019) 47–54.

[20] S. H. Kassani, P. H. Kassani, A comparative study of deep learning architectures on melanoma detection, Tissue and Cell 58 (2019) 76–83.

[21] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, et al., A deep learning system for differential diagnosis of skin diseases, Nature Medicine 26 (6) (2020) 900–908.

[22] A. Astorino, A. Fuduli, P. Veltri, E. Vocaturo, Melanoma detection by means of multiple instance learning, Interdisciplinary Sciences: Computational Life Sciences 12 (1) (2020) 24–31.

[23] C. Yu, S. Yang, W. Kim, J. Jung, K.-Y. Chung, S. W. Lee, B. Oh, Acral melanoma detection using a convolutional neural network for dermoscopy images, PloS one 13 (3) (2018) 1–14.

[24] A. Hekler, J. S. Utikal, A. H. Enk, C. Berking, J. Klode, D. Schadendorf, P. Jansen, C. Franklin, T. Holland-Letz, D. Krahl, et al., Pathologist-level classification of histopathological melanoma images with deep neural networks, European Journal of Cancer 115 (2019) 79–83.

[25] F. De Logu, F. Ugolini, V. Maio, S. Simi, A. Cossu, D. Massi, et al., Recognition of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm, Frontiers in oncology 10 (2020) 1559.

[26] L. Wang, L. Ding, Z. Liu, L. Sun, L. Chen, R. Jia, X. Dai, J. Cao, J. Ye, Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning, British Journal of Ophthalmology 104 (3) (2020) 318–323.

[27] R. Del Amor, L. Launet, A. Colomer, A. Moscardó, A. Mosquera-Zamudio, C. Monteagudo, V. Naranjo, An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images, Artificial intelligence in medicine 121 (2021) 102197.

[28] L. Launet, A. Colomer, A. Mosquera-Zamudio, A. Moscardó, C. Monteagudo, V. Naranjo, A self-training weakly-supervised framework for pathologist-like histopathological image analysis, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022, pp. 3401–3405.

[29] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Medicine 25 (8) (2019) 1301–1309.

[30] V. Winnepenninckx, R. De Vos, M. Stas, J. J. van den Oord, New phenotypical and ultrastructural findings in spindle cell (desmoplastic/neurotropic) melanoma, Applied Immunohistochemistry & Molecular Morphology 11 (4) (2003) 319–325.

[31] J. H. Choi, J. Y. Ro, Cutaneous spindle cell neoplasms: pattern-based diagnostic approach, Archives of pathology & laboratory medicine 142 (8) (2018) 958–972.

[32] R. del Amor, A. Colomer, S. Morales, C. Pulgarín-Ospina, L. Terradez, J. Aneiros-Fernandez, V. Naranjo, A self-contrastive learning framework for skin cancer detection using histological images, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022, pp. 2291–2295.

[33] K. Das, S. Conjeti, J. Chatterjee, D. Sheet, Detection of breast cancer from whole slide histopathological images using deep multiple instance cnn, IEEE Access 8 (2020) 213502–213511.

[34] J. Silva-Rodriguez, A. Colomer, J. Dolz, V. Naranjo, Self-learning for weakly supervised gleason grading of local patterns, IEEE journal of biomedical and health informatics 25 (8) (2021) 3094–3104.

[35] M. E. Tschuchnig, P. Grubmüller, L. M. Stangassinger, C. Kreutzer, S. Couillard-Després, G. J. Oostingh, A. Hittmair, M. Gadermayr, Evaluation of multi-scale multiple instance learning to improve thyroid cancer classification, in: 2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2022, pp. 1–6.

[36] Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan, et al., Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4837–4846.

[37] B. Li, Y. Li, K. W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14318–14328.

[38] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[39] R. Tennakoon, G. Bortsova, S. Ørting, A. K. Gostar, M. M. Wille, Z. Saghir, R. Hoseinnezhad, M. de Bruijne, A. Bab-Hadiashar, Classification of volumetric images using multi-instance learning and extreme value theorem, IEEE Transactions on Medical Imaging 39 (4) (2019) 854–865.

[40] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: 35th International Conference on Machine Learning (ICML), 2018.

[41] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Advances in neural information processing systems 34 (2021) 2136–2147.

[42] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images, IEEE Transactions on Medical Imaging 35 (5) (2016) 1313–1321.

[43] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1) (1979) 20–28.

[44] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy, Learning whom to trust with MACE, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1120–1130.

[45] Z. Chen, L. Jiang, C. Li, Label augmented and weighted majority voting for crowdsourcing, Information Sciences 606 (2022) 397–409.

[46] Y. Zhang, L. Jiang, C. Li, Attribute augmentation-based label integration for crowdsourcing, Frontiers of Computer Science 17 (5) (2023) 175331.

[47] H. Li, L. Jiang, S. Xue, Neighborhood weighted voting-based noise correction for crowdsourcing, ACM Transactions on Knowledge Discovery from Data 17 (7) (2023) 1–18.

[48] L. Jiang, H. Zhang, F. Tao, C. Li, Learning from crowds with multiple noisy label distribution propagation, IEEE Transactions on Neural Networks and Learning Systems 33 (11) (2022) 6558–6568.

[49] Openseadragon, Archivo situacionista hispano, url http://openseadragon.github.io/ (1999).

[50] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, Advances in neural information processing systems 29 (2016).

[51] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[52] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[53] J. Hensman, A. G. de G. Matthews, Z. Ghahramani, Scalable variational Gaussian process classification, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015, 2015.

[54] P. Morales-Alvarez, P. Ruiz, S. Coughlin, R. Molina, A. K. Katsaggelos, Scalable variational gaussian processes for crowdsourcing: Glitch detection in ligo, IEEE Transactions on Pattern Analysis & Machine Intelligence 44 (03) (2022) 1534–1551.

[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[56] Z. Xu, P. Shi, F. Yibulayin, L. Feng, H. Zhang, A. Wushou, Spindle cell melanoma: Incidence and survival, 1973-2017, Oncology letters 16 (4) (2018) 5091–5099.

Appendix A. CR-AI4SKIN database description

Appendix A.1. Medical background

The CSC neoplasms are predominantly composed of spindle-shaped neoplastic cells arranged in sheets and fascicles [30]. Cutaneous spindle cell neoplasms are relatively common. For example, cutaneous squamous cell carcinoma is the second most common epidermal cancer representing 20 % to 50% of skin cancers and spindle cell melanoma contributes 3% to 14% of all melanoma cases [56]. The most common neoplasms in this group are: leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (scm), atypical fibroxanthomas (afx) and squamous cell carcinoma (scc), see Figure A.7 for some visual examples.

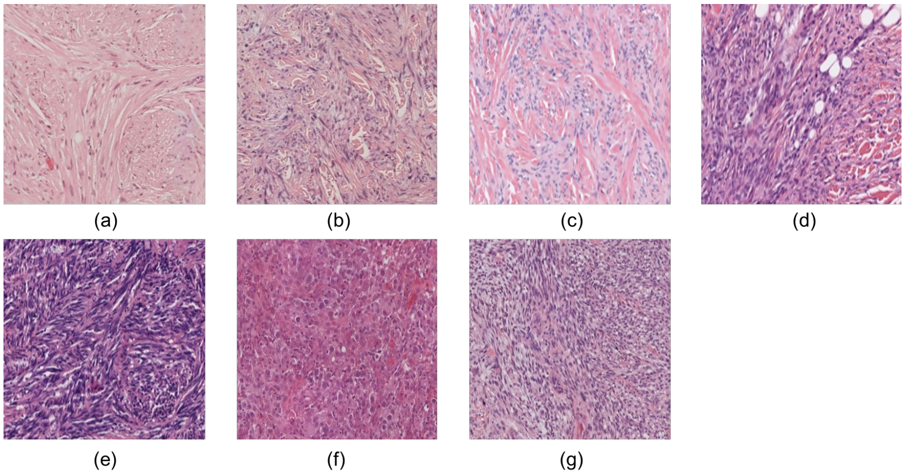


Figure A.7: Histological features of CSC neoplasms. (a) Leiomyoma, (b) leiomyosarcoma, (c) dermatofibroma, (d) dermatofibrosarcoma, (e) spindle cell melanoma, (f) atypical fibroxanthoma and (g) squamous cell carcinoma.

CSC neoplasms are challenging to diagnose due to the considerable morphological overlap between the different tumor types that fall under this category [31], which constitutes a particularly challenging problem for less experienced pathologists. Table A.4 displays the patterns and features that distinguish the different types of spindle cell neoplasms from one another. It is important to evaluate the following histological features to accurately classify these neoplasms: (1) architectural (growth) pattern of the tumor, (2) overall cellularity, (3) appearance of cells, (4) amount and type of matrix formation, (5) tumor and adjacent tissue interfaces, (6) vascularity, (7) tumor necrosis, and (8) mitotic activity [31]. At low resolution, preserving normal architecture, zonation, lesional symmetry, and overall cellularity aid to differentiate between benign and malignant tumors. At high magnification, atypical mitoses and nuclear atypia are more often associated with malignancy. In this difficult scenario, a specific immunohistochemical panel is often needed to evaluate and classify cutaneous spindle cell neoplasms. However, immunohistochemical analysis is expensive, and the observations must be carefully interpreted in context with other findings.

Table A.4: Histological features of spindle cell neoplasms in the dataset.

Tumor type		**Benignity**		**Malignancy**	
Origin tumor	***Significant patterns***	***Name***	***Features***	***Name***	***Features***
Smooth muscle cells	Spindle cells with eosinophilic cytoplasm Elongated nuclei (pure form)	**Leiomyoma**	No mitosis (exceptional) No frequent atypia	**Leiomyosarcoma**	Mitoses always present Nuclear atypia
Connective tissue cells	Spindle cells with a swirling or storiform pattern	**Dermatofibroma**	May have mitosis Multinucleated cells Epidermal ridges	**Dermatofibrosarcoma**	Few but present mitoses No multinucleated cells Epidermis more flattened
Melanocytic cells	Spindle cell fascicles	-	-	**Spindle cell melanoma**	Mitoses $\geq 2/\text{mm2}$ Significant atypia
Squamous cells	Spindle cell fascicles with variable cohesiveness	-	-	**squamous cell carcinoma**	Mitosis Nuclear atypia
Fibroblasts	Spindle-shaped, histiocytoid and multinucleated cells	-	-	**Atypical fibroxanthoma**[3]	Solar elastosis Multinucleated cells Atypical mitoses

Appendix A.2. Data summary

CR-AI4SkIN consists of two datasets from the University Clinic Hospital of Valencia (HCUV) and San Cecilio University Hospital in Granada (HUSC) of skin tissue biopsies. Each dataset, DSV for HCUV and DSG for HUSC, comprises, respectively, 180 and 91 different patients who signed the pertinent informed consent. The tissue samples from the different patients were sliced, H&E stained, and digitized at 40x magnification, obtaining WSIs. Two expert pathologists provided the WSI label for the dataset, consisting of 271 images. Specifically, one pathologist analyzed 91 images from HSUC, and the other pathologist analyzed the 180 images belonging to HCUV. Each WSI belongs to one of the seven types of CSC neoplasms previously described. A summary of the dataset is presented in Table A.5.

Table A.5: Dataset distribution. DSV: dataset from Valencia; DSG: dataset from Granada. Lm:leiomyomas; lms: leiomyosarcomas; df:dermatofibromas; dfs: dermatofibrosarcomas; scm: spindle cell melanomas; afx: fibroxanthomas; scc: squamous cell carcinoma.

	lm	**lms**	**df**	**dfs**	**scm**	**afx**	**scc**	**Total**
DSV	28	19	52	11	32	28	10	180
DSG	27	9	16	7	6	26	-	91
Total	55	28	68	28	38	44	10	271

Note that the number of WSIs is different from the one presented in subsection 3.1 due to the afx lesions were eliminated from the study as the distinction between benign and malignant is unclear.

[3] For this type, the borderline between benignity and malignancy is not obvious.