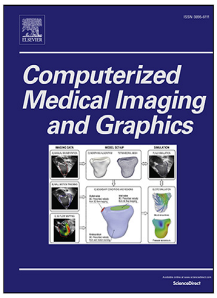
Journal Pre-proof



Labeling confidence for uncertainty-aware histology image classification

Rocío del Amor, Julio Silva-Rodríguez, Valery Naranjo

PII:	S0895-6111(23)00049-6
DOI:	https://doi.org/10.1016/j.compmedimag.2023.102231
Reference:	CMIG 102231

To appear in: *Computerized Medical Imaging and Graphics*

Received date :	14 December 2022
Revised date :	23 February 2023
Accepted date :	27 March 2023

Please cite this article as: R.d. Amor, J. Silva-Rodríguez and V. Naranjo, Labeling confidence for uncertainty-aware histology image classification. *Computerized Medical Imaging and Graphics* (2023), doi: https://doi.org/10.1016/j.compmedimag.2023.102231.

Labeling confidence for uncertainty-aware histology image classification

Rocío del Amor[a], Julio Silva-Rodríguez[b], Valery Naranjo[a]

[a] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano,
Universitat Politècnica de València, Valencia, Spain
[b] ÉTS Montréal,
Montréal, Québec, Canada

Abstract

Deep learning-based models applied to digital pathology require large, curated datasets with high-quality (HQ) annotations to perform correctly. In many cases, recruiting expert pathologists to annotate large databases is not feasible, and it is necessary to collect additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth, non-expert annotators). Learning from datasets with noisy labels is more challenging in medical applications since medical imaging datasets tend to have instance-dependent noise and suffer from high inter/intra-observer variability. In this paper, we design an uncertainty-driven labeling strategy with which we generate soft labels from 10 non-expert annotators for multi-class skin cancer classification. Based on this soft annotation, we propose an uncertainty estimation-based framework to handle these noisy labels. This framework is based on a novel formulation using a dual-branch min-max entropy calibration to penalize inexact labels during the training. Comprehensive experiments demonstrate the promising performance of our labeling strategy. Results show a consistent improvement by using soft labels with standard cross-entropy loss during training (~ 4.0% F1-score) and increases when calibrating the model with the proposed min-max entropy calibration (~ 6.6% F1-score). These improvements are produced at negligible cost, both in terms of annotation and calculation.

Keywords: Digital pathology, Non-expert annotators, Uncertainty estimation, Model calibration

1. Introduction

Digital pathology research has experienced significant growth in recent years thanks to the advent of novel computer vision techniques based on deep learning [1]. The deployment of convolutional neural networks (CNNs) has allowed the automatic identification of new biomarkers and innovative features in the whole slide images (WSIs) that support the diagnostic process. In particular, these techniques have shown promising results for computer-aided diagnosis on different applications such as prostate [2], breast [3] and skin cancer detection [4], tissue segmentation [5], or mitosis detection [6], among others. Nevertheless, deep learning models require large and curated datasets with high-quality (HQ) annotations to perform properly. In the case of digital pathology, a popular choice is the use of weakly supervised strategies with WSI-level annotations. In the multi-class scenario, an expert pathologist assigns a unique label to the whole biopsy based on diagnostic or prognostic features. Then, deep learning models are trained using multiple instance learning (MIL) to automatically solve the task at hand. However, this pipeline does not consider real-world limitations and noise sources inherent to the annotation process, which may hinder the performance of the model. These limitations are accentuated in some applications requiring a high level of expertise, such as several skin neoplasm diagnosis (i.e., cutaneous spindle cell neoplasms, one of the most challenging skin neoplasms not studied in previous studies [7]). In many cases, recruiting expert pathologists to annotate large databases is not feasible. Unfortunately, without sufficient labels, the data-hungry learning-based methods often struggle with overfitting, leading to inferior performance [8]. To alleviate this issue, collecting additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth,

Email addresses: madeam2@upv.es (Rocío del Amor), julio-jose.silva-rodriguez@etsmtl.ca (Julio Silva-Rodríguez), vnaranjo@dcom.upv.es (Valery Naranjo)

non-expert annotators) or using machine-generated labels is a common practice. However, directly introducing data with low-quality (LQ) noisy labels may confuse the network training, which easily leads to performance degradation [9, 10]. Therefore, how to effectively and robustly exploit the additional information in plentiful LQ noisy labeled data is crucial to the medical image analysis community.

Learning from noisy labels is a widely recognized challenge in classical image recognition. Several efforts have been made to mitigate the negative impact of LQ labels in medical image analysis [10–13]. However, this is still an under-explored area, as existing literature on learning with noisy labels lacks a clear distinction of applicable scenarios, leading to ambiguous benchmarks. Some approaches [12, 13] assumed mixed data from multiple sources, i.e., set-HQ and set-LQ labels are indiscriminate. In contrast, other techniques [10, 11] were developed for a scenario where experts label a small data set, making LQ and high-quality (HQ) labels separated. A main body of literature exploits multiple annotators in a crowdsourcing scenario, to extract the underlying noise-free label distribution. Nevertheless, gathering multiple annotators in the medical context may be unrealistic. The high level of expertise required, as well as the time-consuming nature of such annotation, is a barrier to the implementation of these methods in real-world applications. These findings highlight the need for developing uncertainty-aware pipelines to address the inherent uncertainty in the annotation process, which may not require from multiple label sources.

Based on these observations, we propose a novel uncertainty-driven labeling strategy for histology skin cancer classification. The key contributions of our work can be summarized as follows:

- A single-annotator uncertainty-aware labeling strategy with which we generate soft labels from 10 non-expert annotators for multi-class skin cancer classification that quantify uncertainty in the annotations.
- Based on these annotations, we present an extensive study for the use of soft label model calibration compared to the ground truth, labeled by an expert pathologist.
- In addition, we propose a novel formulation based on dual-branch entropy calibration (DBEC) to calibrate both, overconfident outputs and uncertain soft labels, during training.
- Comprehensive experiments demonstrate the promising performance of our labeling strategy. By incorporating uncertainty during labeling we found average improvements of nearly ∼ 4.0% in averaged F1-score using the baseline methods, which increases up to ∼ 6.6% using the proposed dual-branch calibration.

2. Related work

2.1. Skin WSIs

According to the World Health Organization, nearly one in three diagnosed cancers worldwide is a skin cancer [14]. Different techniques, such as dermatoscopy, wood lamp, CT scan and histopathology, are utilized for the diagnosis of skin diseases. However, the gold standard for skin cancer detection is histological image analysis. Traditionally, histological slides would be viewed with a light microscope. However, digitization has created opportunities for automated analysis using WSI. Applying deep-learning models to computer vision problems shows excellent potential in skin cancer detection. Most research was based on the analysis of dermoscopic images [15–21] and few studies have focused on the analysis of WSI [3, 4, 22–25]. In this vein, MIL approaches have been successfully applied to Basal carcinoma (BCC) [3] or melanoma [4], reducing the time required to perform precise annotations. However, many types of skin cancer have not yet been explored. These include cutaneous spindle cell neoplasms (CSC), predominantly composed of spindle-shaped neoplastic cells arranged in sheets and fascicles [26]. These lesions are relatively common. For example, cutaneous squamous cell carcinoma is the second most common epidermal cancer representing 20 % to 50% of skin cancers [27] and spindle cell melanoma contributes 3% to 14% of all melanoma cases [28]. CSC neoplasms are challenging to diagnose due to the considerable morphological overlap between the different tumor types that make up this group [7], which poses a particular problem for less experienced pathologists. This hampers an accurate diagnosis and the application of effective clinical treatment [29] in neoplasms in which early detection and appropriate treatment are essential for a good prognosis in malignant cases. Despite the complexity of these neoplasms, they had not been previously studied in the literature. Therefore, the main objective of this paper is to classify, under a MIL-based approach, the seven types of fusocellular skin neoplasms identified by expert pathologists as the most challenging: leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (mfc), fibroxanthomas (fxa) and squamous cell carcinoma (cef).

2.2. *Uncertainty estimation*

Uncertainty estimation methods are expected to improve the understanding and quality of deep learning models to enhance their generalization during inference. These methods have an outstanding interest in medical applications due to the high expertise required to obtain quality labels, the variability in acquisition systems and noise present in many databases [30], and the known inter-annotator variability in different medical applications [31, 32]. For these reasons, training uncertainty-aware models is key to the success of diagnostic support systems in medical applications. An uncertainty-aware deep learning model training usually covers two steps: uncertainty quantification and model calibration. Uncertainty quantification aims to assess the prior probability of error for certain samples during training. From the perspective of noisy labels, a main core of previous literature use multiple annotators in crowd-sourcing scenarios to quantify inter-observer agreement for each sample [33–36]. Thus, crowd-sourcing methods aim to predict the underlying noise-free label distribution by simultaneously training annotator-specific projections over the feature space [33–37]. Other solutions focus on prior task-specific knowledge such as avoiding overconfident outputs on neural networks [38] or leveraging high confidence on non-informative regions [39]. Other uncertainty quantification approaches focus on sample noise estimation, which may raise from image quality, feature extraction, or out-of-distribution domains. Previous literature in this regard use a trained student model to study the confidence of the model via Monte Carlo dropout with image augmentations [24, 36, 40], curriculum learning [41], or co-teaching [42, 43]. After uncertainty estimation, deep learning models are calibrated to overcome the limitations detected in the training samples. Some approaches include sample weighting based on divergence observed by the Student-based methods [36], or calibrating the output of the network based on label smoothing [44] and entropy regularization [38, 45, 46].

In this paper, we focus on label-noise calibration, and we study the feasibility of estimating uncertainty from single annotator labels. Contrary to much of the previous literature, we study the case in which multiple annotators are not available. To this end, we define a soft label-based annotation protocol. Then, we propose a dual-branch criterion for calibrating the trained neural network based on entropy regularization. The underlying idea is two-fold: (i) penalizing overconfident predictions on high-certain samples, and (ii) forcing the network to produce confident outputs on uncertain cases, to overcome the limitations of the noisy labels based on the features of each sample. Note that although we trained 10 models, one for each non-expert to validate the proposed methodology, these models are independent since only the labels of a single annotator are used to train the algorithm each time.

3. Methods

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

Problem Formulation. Under the paradigm of Multiple Instance Learning (MIL), instances are grouped in bags of instances $X = \{x_n\}_{n=1}^{N}$ that exhibit neither dependency nor ordering among them, and its number N is arbitrary for each bag. In the multi-class scenario, each bag is a member of one of K mutually exclusive classes, such that $Y_k \in \{0, 1\}$. Note that, in contrast to other MIL formulations, the individual instances do not have an associated label, but rather the label of the bag is determined by the combination of features of the different instances.

Embedding-based MIL. In this work, we aim to train a model 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[1]. Let us denote a neural network model, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by $\boldsymbol{\theta}$, which projects instances $x \in \mathcal{X}$ to a lower dimensional manifold $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^d$, with d the embedding dimension. Then, we define an aggregation, $f_a(\cdot)$, which is in charge of combining the instance-level projections into a global embedding, Z. In particular, we use a global-average pooling along instances, such that: $Z = \frac{1}{N}\sum_n \{f_{\boldsymbol{\theta}}(x_n)\}_{n=1}^{N}$. Finally, a neural network classifier, $f_{\boldsymbol{\phi}}(\cdot) : \mathcal{Z} \rightarrow \mathcal{S}$, is in charge of predicting softmax bag-level class scores, S_k, such that $S_k \in [0, 1]$. The optimization of the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is driven by the minimization of standard categorical cross-entropy loss between the reference labels and predicted scores such that:

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

[1] Based on the denomination proposed in [47]

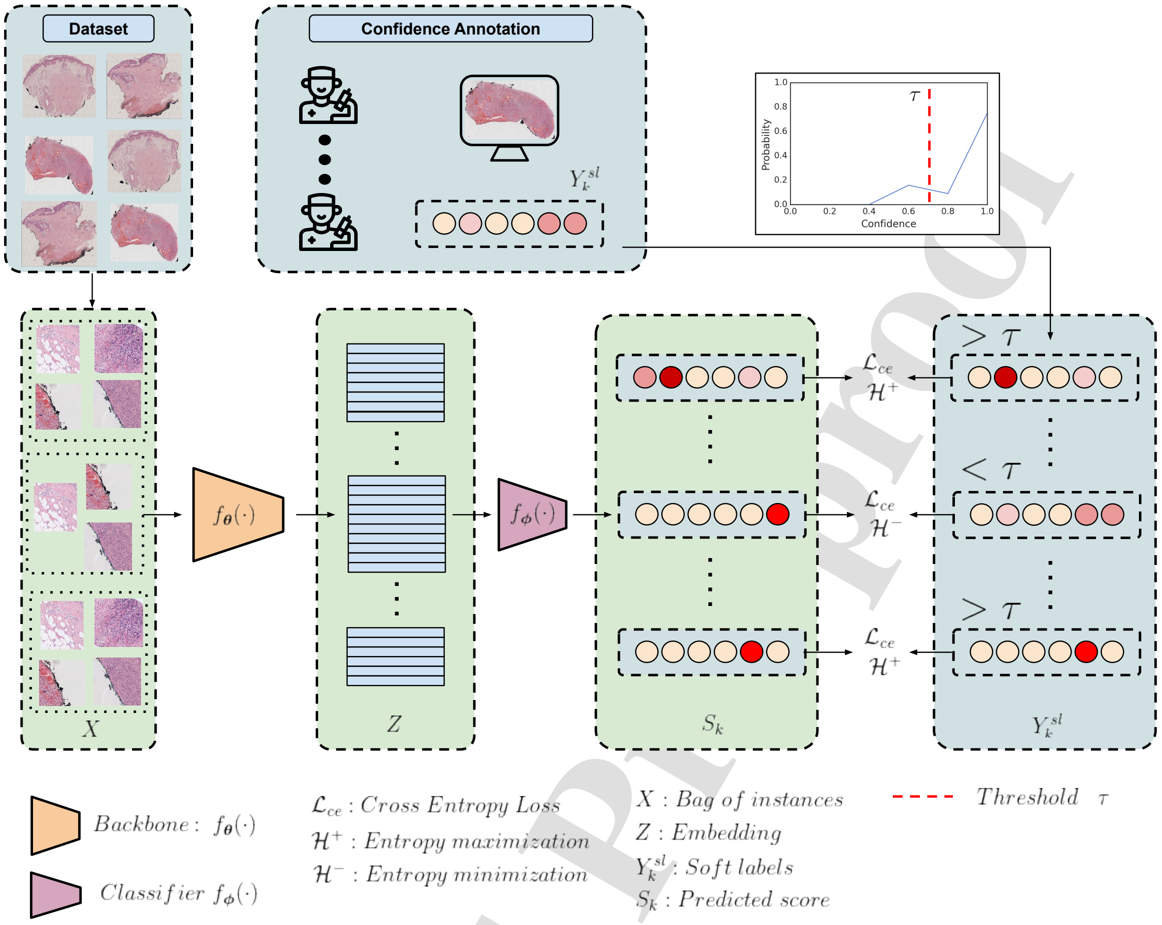


Figure 1: **Method overview**. In this work, we address weakly supervised histology image classification on skin WSIs by quantifying the uncertainty of the individual annotators during labeling. Concretely, we train an embedding-based Multiple Instance Learning (MIL) model to predict up to six different categories using standard cross-entropy loss. We propose to quantify annotator-specific uncertainty by following a soft labels annotations protocol, such that $Y_k^{sl} = [0, 1]$, and $\sum_k Y_k^{sl} = 1$. In this fashion, our model captures information regarding inter-category dependencies and avoids over-fitting to uncertain, noisy annotations. Then, we propose a dual-branch min-max uncertainty calibration (DBEC) based on the annotated soft labels. Based on uncertainty calibration using Shannon entropy regularization (see Eq. 3), we propose to (i) maximize the entropy on high-confidence labeled samples, by entropy maximization (H^+), and (ii) to minimize the entropy on samples labeled with low-confidence (H^-). Thus, entropy minimization encourages the network to produce confident outputs on uncertain cases, based on the features of the sample, and thus diminishing noise propagation. A threshold τ is empirically fixed to differentiate low and high certain labels, and the dual-branch min-max uncertainty is combined with cross-entropy loss (see Eq. 5). Circles in bag-level predictions and references indicate soft-max scores. The more intense the color, the higher the score.

3.1. Labeling uncertainty

Uncertainty estimation methods assume that different noise sources are present in the dataset, both in image noise and inter and intra- annotator variability. The objective is to calibrate the trained model to account for quantified uncertainties. Regarding inter-annotator variability, a large body of literature quantifies this uncertainty by obtaining labels from multiple annotators. However, obtaining multiple annotators may not be possible in specific scenarios requiring a high level of specialization or covering proprietary solutions, such as medical applications. To overcome this limitation, we propose an annotator-level uncertainty quantification by annotating the confidence associated with each sample in the form of soft labels. To this end, we differentiate between the labeled samples using hard labels (HL), Y_k^{hl}, and soft labels (SL), Y_k^{sl}. As previously described, hard labels assign a discrete value for each label such that $Y_k^{hl} \in \{0, 1\}$, where $Y_k = 1$ indicates that the corresponding sample belongs to the class k. It is worth mentioning that, in the multi-class scenario, categories are considered mutually excluded, and only one tag is given to each sample. Nevertheless, this labeling strategy fails to capture the certainty of the annotator for each sample. To gather this information, we propose to use soft labels, such that $Y_k^{sl} \in [0, 1]$. Note that in this case, Y_k is a continuous value that corresponds to the probability that the annotator assigns to each class, such that: $\sum_k Y_k^{sl} = 1$. For instance, in a case with high uncertainty, the annotator might assign the following labels: $Y^{sl} = [0, 0, 0, 0, 0.9, 0.1, 0]$, whereas in a uncertain case, the total probability might be more distributed among categories: $Y^{sl} = [0.2, 0.2, 0, 0, 0.6, 0, 0]$. Then, the MIL classification model previously described is trained using standard cross-entropy loss in Eq. 1 using soft annotation labels. We believe that, in this fashion, the model might capture information regarding inter-category dependencies and avoid over-fitting to uncertain cases, as supported in the experimental stage of the present work.

3.2. Dual-branch uncertainty calibration

The aforementioned soft-labeling strategy can differentiate between high-certain and uncertain labels provided by the annotator. Still, using standard cross-entropy might produce ill-calibrated models. These limitations include reaching trivial solutions by producing overconfident outputs from high-certain samples or trivial, uniform outputs on low-certainty samples. In addition, we want to consider that samples labeled with low confidence might belong to a class other than the one most likely to be noted. To this end, we propose calibrating the model during training to deal differently with both types of samples in a dual-branch fashion.

Shannon entropy for confidence regularization. One of the main approaches to calibrating neural networks is using an auxiliary term to regulate the output probabilities. Originally developed to reduce overconfident predictions, which are produced by training models using cross-entropy and hard labels, one of the main approaches lies in forcing the output distribution to approximate a uniform distribution [38, 44]. To this end, the neural network is trained to minimize the Kullback − Leibler (KL) distance, $D_{KL}(p\|u) = H(p, u) - H(p)$ between an output distribution, p and an uniform distribution, u. Note that $H(p, u)$ indicates the cross-entropy between both distributions, and $H(p) = H(p, p)$ is the Shannon entropy or self-entropy, such that $H(p) = -\frac{1}{K}\sum_k p_k \cdot log(p_k)$. It is straightforward to see that, in the case of a target uniform distribution, minimizing the KL distance is equivalent to maximizing the Shannon entropy of the output distribution.

$$D_{KL}(p\|q) = H(p, q) - H(p) \stackrel{c}{=} -H(p) \tag{2}$$

where $\stackrel{c}{=}$ indicates equality up to an additive constant.

Thus, standard model calibration using Shannon entropy includes a regularization term to the standard cross-entropy loss weighted by an hyper-parameter $\beta > 0$, such that:

$$\mathcal{L} = \mathcal{L}_{ce} - \beta H(p) \tag{3}$$

Dual-branch min-max entropy calibration. Inspired by previous literature on model calibration, we propose to use the Shannon entropy regularization in a dual-branch fashion. First, we want the model to avoid overconfident outputs on high-certainty labeled samples, similarly to Eq. 3. Secondly, we aim to calibrate the model to assign a confident category to each sample, even though the annotator might have high uncertainty in the label. For the latter, we draw on Shannon entropy minimization, which encourages the output scores to differ from the uniform distribution (see Eq. 2). It is worth mentioning that, in the case of minimum entropy, the output scores tend to produce hard labels. Thus, we hypothesize that the model may be able to overcome the potential noise from the uncertain labels, and produce more accurate predictions based on the features of the sample. This formulation is inspired by the semi-supervised learning literature, in which entropy maximization is used as a proxy to learn from unlabeled samples [48]. From now on, and for simplicity in the context of loss functions, we refer to the entropy-maximization criteria $-H(p)$ as H^{+}, and the opposite minimization term as $H^{-} = H(p)$.

Thus, we propose a dual-branch optimization criterion to independently calibrate low and high-certainty labeled samples, using the bag-level predicted scores, S_k, such that:

$$\mathcal{L}_H = \begin{cases} H^{+}(S_k), & \text{if } \max_k Y_k^{sl} > \tau \\ H^{-}(S_k), & \text{otherwise} \end{cases} \tag{4}$$

where τ is an empirically-fixed threshold that divides the input samples based on its certainty, quantified by the confidence of the predominant category per sample, $\max_k Y_k^{sl}$.

Since using entropy calibration alone may yield trivial results [49], the MIL model is trained with annotated soft labels, Y_k^{sl}, and the dual-branch entropy calibration, using the overall following loss function:

$$\mathcal{L} = \alpha^{+/-}\mathcal{L}_{ce} + \beta^{+/-}\mathcal{L}_H \tag{5}$$

Note that $\mathcal{L}_H$ is the cross entropy loss at bag level in Eq. 1, and $\mathcal{L}_H$ refers to the dual-branch calibration presented in Eq. 4, and $\alpha^{+/-}$ and $\beta+/-$ are disentangled in two terms, one for high-certainty labeled samples (α^+, β^+), and other for the opposite case (α^-, β^-). It is worth mentioning that the values of threshold value τ in Eq. 4 as well as the relative weight of the min-max entropy duality, β^+ and $\beta-$, and cross-entropy loss, α^+ and α^-, are hyperparameters empirically optimized during the experimental stage. Hereafter, we refer to this dual-branch min-max entropy calibration term as DBEC.

4. Experimental setting

4.1. Dataset

To validate the proposed approach, we use the *AI4SKINV1* database. This database comprises two private databases (DSV and DSG) from the University Clinic Hospital of Valencia (Spain) and San Cecilio University Hospital in Granada (Spain). DSV and DSG are composed of histopathological skin images from different body areas that contain cutaneous spindle cell (CSC) neoplasms, i.e, leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (mfc), fibroxanthomas (fxa) and squamous cell carcinoma (cef). Each database (DSV and DSG) comprises 180 and 91 different patients who signed the pertinent informed consent. Two expert pathologists established the WSI-level label of the whole database, 271 images. A summary of the database description is presented in Table 1.

Table 1: Database distribution. DSV: database from Valencia; DSG: database from Granada. Lm:leiomyomas; lms: leiomyosarcomas; df:dermatofibromas; dfs: dermatofibrosarcomas; fxa: fibroxanthomas; spindle cell melanomas; cef: squamous cell carcinoma.

	lm	lms	df	dfs	mfc	fxa	cef	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

Regarding the non-experts labeling, an annotation protocol was designed to ensure that 106 WSIs were annotated by all non-expert annotators (dense set). In contrast, the rest were only annotated by some non-expert pathologists (non-dense set). It is worth mentioning that the use of a dense set allows us to establish data-balanced comparisons between annotators, without requiring everyone to annotate the entire data set, with the burden that this process entails. Table 2 shows images used by each non-expert annotator for training, validation and testing of the models. To establish fair comparisons the validation and test images belonged to the dense set. Note that the images were annotated following the soft strategy proposed in Sec. 3.1 [2].

To process the large WSIs, these were downsampled to 10x resolution and divided into patches of size 512x512x3 with a 50% overlap. Aiming at preprocessing 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 database.

Table 2: Number of images used for training, validation and testing the models of each non-expert annotator (ten in total). 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
Tain	148	142	151	143	154	145	155	149	152	150
Val	26									
Test	54									

4.2. ROI extraction

To select the instances with tumor from the WSI to train and validate the proposed approach, we extend the model proposed in [50] for the six neoplasms under study. This method was based on a teacher-model paradigm to increase the annotated database while avoiding manual annotations. In this vein, this approach enhances the detection of tumor regions in WSI using pseudolabels from non-labeled data. As the output of this section, we obtain the patches with tumor lesions used as input for the MIL-based model.

4.3. Implementation details

The proposed methods were trained using the different train subsets for each non-expert annotator (10 in total), see Table 2. The backbone $f_{\theta}(\cdot)$ used was a VGG16 [51] pre-trained on Imagenet [52], using patches resized to 224×224 images. Models were trained during 120 epochs with a batch size of 1 whole slide image, using a learning rate of $\eta = 1 \cdot 10^{-3}$ with SGD optimizer. The model performance was continuously monitored on the validation subset, and early stopping was applied to keep the model with the best accuracy on this subset. The proposed uncertainty calibration DBEC in Eq. 5 was trained similarly, but the learning rate was exponentially decreased in the last 20 epochs to ensure

[2] The soft labels will be available on request.

stability. In this case, early stopping was not applied since the calibration moved predictions away from the domain of the training labels. Hyperparameters were fixed empirically such that: $\alpha^+ = 1, \beta^+ = 0.1, \alpha^- = 0.1$, $\beta^- = 1$, and $\tau = 0.7$. For the motivation of these values, we refer the reader to the ablation experiments. All the validated experiments were implemented using Pytorch version 1.9.1 and Python 3.7. Experiments were conducted on the NVIDIA DGXA100 system. The code is publicly available on `https://github.com/cvblab/Labeling_Uncertainty`.

4.4. Evaluation metrics

In order o evaluate the performance of the proposed approaches regarding previous literature, we use standard metrics for multi-class classification. In particular, we obtain accuracy (ACC) and macro-averaged F1-score. It is worth mentioning that, although explicitly mentioned, metrics are obtained using as reference the ground truth, labeled by the expert pathologists, Y_k.

5. Results

5.1. Comparison to the literature

In this subsection, we study the obtained results by the proposed methods, concerning previous literature. We also carried out a detailed study of the success cases and limitations encountered, by means of a detailed study of the annotations made by the in-training pathologists.

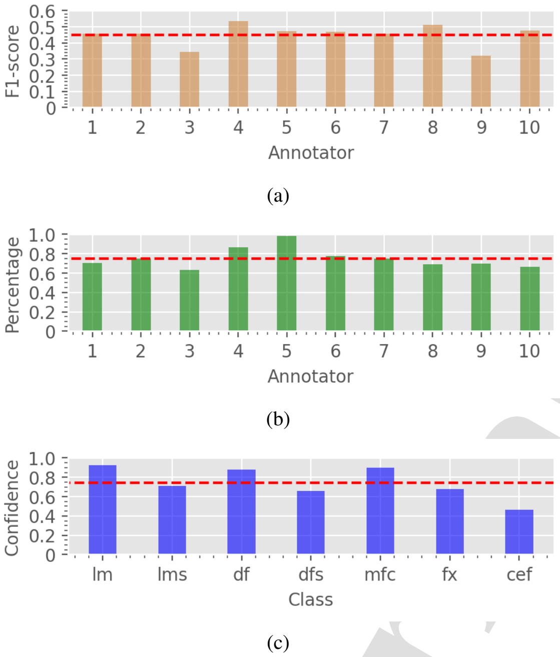
Quantitative evaluation. The quantitative results obtained training the model using expert labels, and non-expert labels using hard labels (HL), annotated soft labels (SL), and the proposed dual-branch entropy calibration (DBEC) on the respective test subset of each non-expert annotator are depicted in Table 3. Results obtained using annotated soft labels from non-expert pathologists reach an average F1-score of 0.364, which shows an improvement of ∼ 4.0% compared to hard labels by simply training the model using standard cross-entropy loss. This fact demonstrates that the annotation protocol developed in the paper is optimal for model training when expert labels are not available. Once our proposed dual-branch entropy calibration (DBEC, see Eq. 4) is incorporated during training, results achieve an average F1-score of 0.389. In addition, some noteworthy improvements can be observed for some non-expert annotators. For example, annotators 1, 2, and 8 show improvements of ∼ 13.1%, ∼ 21.5% and ∼ 13.2%, respectively. Although the results obtained are still far from those obtained using the ground truth from the expert pathologists, the models obtained bridge the gap, going from a difference of ∼ 25% to ∼ 18% regarding F1-score. Furthermore, this paper is the first study to address the multi-class problem of spindle cell neoplasms. While previous studies focus on binary problems to identify benignity or malignity of neoplasms [50], in this study we try to identify the distinct neoplasms that have considerable morphological overlap between them. Therefore, the results obtained in this paper establish a benchmark for the comparison of further models.

In-depth results analysis. Although, as discussed above, the methodology based on confidence annotation offers promising results, the variability in the results observed among different annotators calls for an in-depth analysis of the annotated labels, their advantages, and limitations. To this end, we proceed to study the accuracy of the annotations made by non-expert pathologists in the training subset, the number of samples labeled with low confidence, and their distribution in relation to the classes, in Figure 2. Likewise, we display the confusion matrices obtained by the non-expert annotators concerning the expert annotations, as well as those obtained using the model trained with hard labels and the proposed dual-branch entropy calibration, in Figure 3.

Regarding the gap observed between models trained using the ground truth or non-expert labels, this is due to the quality of the latter labels, which shows an average F1-score of 0.4510 (see Figure 2 (a)), which sets an upper limit on the results that the model can extract using pathologist-in-training labels. As observed in the corresponding confusion matrix (see Figure 3 (a)), this problem accentuates in certain classes such as lms and cef, which show lower prevalence concerning other classes in the used dataset (see Table 1). In addition, it can be observed how non-expert pathologists show lower confidence when labeling a sample corresponding to those categories (see Figure 2 (b)). This make sense since, for example, in the case of lms the pathologists-in-training are often confused with lm as they have the same morphological features. These limitations produce the drop in results between both types of labels observed in the quantitative metrics, which can be observed in the corresponding confusion matrix (see Figure 3 (b)). Interestingly, once the proposed dual-branch calibration is used, obtained results for those low-confidence classes improve (see Figure 3 (c)). Concretely, promising improvements for the classes lms, dfs, and fx are observed, which coincide with those categories that pathologists show the least confidence (see Figure 2 (c)). This may

Table 3: Quantitative comparison to prior literature. The metrics presented are the accuracy and micro-averaged F1-score (ACC/F1-score). The model trained with expert labels (second column) is used as the upper bound of the non-expert-based models. Colored values indicate the relative improvement of each method concerning the baseline using hard labels from non-expert in terms of the F1-score. Green indicates improvement and red a worsening lack. HL: hard labels; SL: soft labels; H^+: entropy maximization. Gray background highlights the averaged results.

Annotator	Expert	Non-Expert				
	HL+H$^+$	HL	SL		DBEC	
1	0.653/0.620	0.408/0.277	0.428/0.295	↑ 1.8%	0.530/0.408	↑ 13.1%
2	0.571/0.467	0.408/0.288	0.530/0.424	↑ 13.6%	0.571/0.503	↑ 21.5%
3	0.612/0.584	0.448/0.386	0.489/0.401	↑ 1.5%	0.428/0.330	↓ 5.6%
4	0.551/0.520	0.448/0.309	0.428/0.355	↑ 4.6%	0.489/0.364	↑ 5.5%
5	0.673/0.601	0.551/0.448	0.571/0.460	↑ 1.2%	0.530/0.442	↓ 0.0%
6	0.591/0.555	0.428/0.298	0.428/0.304	↑ 0.6%	0.428/0.315	↑ 1.7%
7	0.673/0.602	0.469/0.348	0.551/0.427	↑ 7.9%	0.530/0.444	↑ 9.6%
8	0.693/0.655	0.367/0.259	0.408/0.270	↑ 1.1%	0.469/0.391	↑ 13.2%
9	0.653/0.614	0.387/0.280	0.469/0.323	↑ 4.3%	0.387/0.299	↑ 1.9%
10	0.632/0.525	0.469/0.353	0.530/0.390	↑ 3.7%	0.530/0.398	↑ 4.5%
Avg.	0.630/0.574	0.438/0.324	0.473/0.364	↑ 4.0%	0.489/0.389	↑ 6.6%



(a)

(b)

(c)

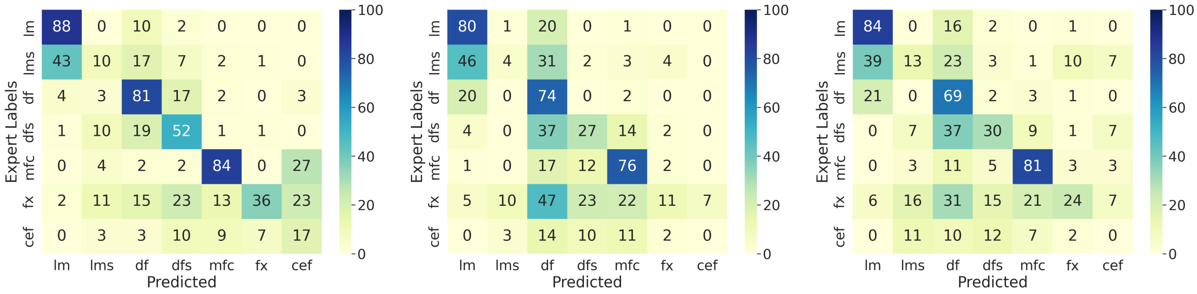
Figure 2: In-depth study of the soft labels annotated by in-training pathologists. (a) Quality of the labels, in terms of F1-score, in the training subset. Reference labels are the expert ground truth. (b) Percentage of samples with maximum confidence above the threshold $\tau = 0.7$. (c) Average confidence per each class, on positive samples. Dashed, red lines indicate average values.

be produced by the lower-confidence entropy minimization, which encourages the model to produce confident predictions in those cases in which confidence falls below the fixed threshold τ. In this fashion, predicted labels move away from the annotator bias, based on the inherent features of each sample, and show the best generalization compared to expert annotations. Although the proposed approach offers consistent improvements among most annotators, still some limitations can be observed. For instance, it shows the least effect when noise increases. Annotators 3 and 9, which show low accuracy on the training dataset (see Figure 2 (a)), also offer worse results regarding the proposed approach. Also, if no use is made of soft labels (see 2 (b), annotator 5), the results remain the same as using hard labels.

5.2. Ablation studies

The following experiments aim to demonstrate the convenience of the proposed approaches in an empirical fashion. First, we compare the benefits of labeling uncertainty instead of using a direct calibration of hard labels. Then, we motivate the choice of the components and hyper-parameters used for the proposed dual-branch uncertainty calibration setting in Eq. 5.

Artificial vs. annotated soft labels. As previously discussed, we propose in this work to calibrate the model training to the inherent uncertainty of non-expert labeling by annotating the confidence for each independent class per sample. The benefit of calibrating CNNs to avoid overconfident predictions has already been demonstrated in previous literature [38]. We follow two main artificial methods used in this regard: label smoothing (LS) [44] and entropy regularization (H) [38]. Concretely, LS modifies the hard labels to assign a uniform distribution over non-positive categories such that: $Y_k^{LSR} = (1-\epsilon)Y_k + \frac{\epsilon}{K}$. Entropy calibration is based on Shannon entropy maximization (H^+), as described



(a) (b) (c)

Figure 3: Normalized confusion matrices, averaged among non-expert annotators, obtained using (a) raw hard labels, (b) the model trained using hard labels, and (c) the model trained using the dual-branch entropy calibration proposed in Eq. 5. Reference labels are the expert ground truth.

in the method section (see Eq. 3). In our experiments, we empirically optimized the hyper-parameters for both $\epsilon = 0.2$ and $\beta = 0.2$. We depict in Figure 4 the results using hard labels (HL), both artificial regularization approaches (LS and H^{+}), and the model trained using the proposed annotated soft labels (SL).

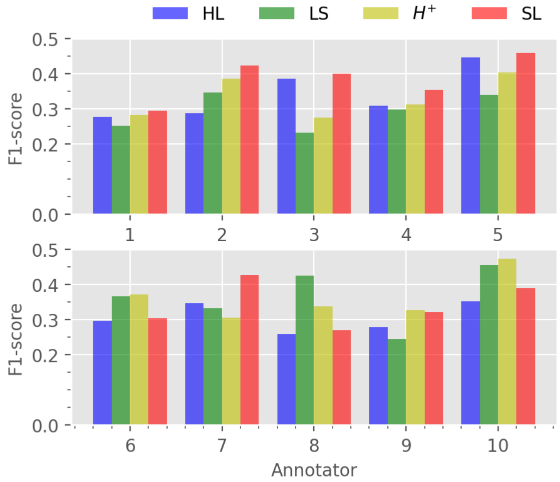


Figure 4: Ablation study on the use of artificial model calibration of hard labels (HL) or annotated soft labels (SL). For the first approach, label smoothing (LS) and entropy maximization (H^{+}) are used. F1-score is presented for each method and non-expert annotator.

The obtained results show that regularizing neural network outputs improves the model performance. In particular, entropy-based regularization outperforms label smoothing, as indicated by previous literature [38, 46]. Concretely, average improvements of F1-score of $\sim$ 0.6% and $\sim$ 2.4% are obtained, respectively. The proposed labeling confidence approach outperforms the artificial entropy-based calibration across most annotators (see Figure 4 annotators 2, 3, 4, 8 $-$ 10). Concretely, an average improvement of $\sim$ 4% is observed, as already depicted in Table 3. This indicates that labelling the confidence of the annotator for the different classes for each sample offers benefits beyond preventing the model from producing overconfident outputs. It is worth mentioning that this improvement is produced at a negligible cost, both in terms of annotation time and computational level. This may be because it introduces a sample-dependent distribution over labels, as opposed to these artificial methods.

Uncertainty calibration optimization. The following experiments aim to demonstrate the convenience of the different components of the dual-branch entropy calibration (DBEC) for uncertainty assessment proposed in Eq. 4 when trained using soft labels (SL). Concretely, we fix the used threshold $\tau = 0.7$, then train and modify the relative weight of both branches to emulate the absence of each term. First, each term is trained individually, by using $\beta^{-} = 0$ and $\alpha^{-} = 0$, (DBEC (H^{+}) configuration), and $\beta^{+} = 0$ and $\alpha^{+} = 0$, (DBEC (H^{-}) configuration), respectively. Then, both terms are included as indicated in the implementation details. Average results among the 10 in-training pathologists are presented in Table4.

Table 4: Ablation experiment on the components of the proposed calibration formulation.

	Target Criteria			
	SL	DBEC (H^{+})	DBEC (H^{-})	DBEC ($H^{+/-}$)
ACC	0.438	0.461	0.386	0.489
F1-score	0.324	0.334	0.281	0.389

The results show that using only the positive entropy term, which calibrates the network by penalizing confident predictions, improves around $\sim$ 2% in terms of the

F1-score. In contrast, using only low-confidence samples during training does not show good results. However, by incorporating this term into the general formulation, the figures of merit reach the improvements discussed earlier in the article. These results show the usefulness of including both terms in the proposed double-branch formulation.

In the following, we perform a study regarding the threshold used to compute the positive or negative entropy calibration, τ. Concretely, we sample homogeneously τ values between $[0, 1]$. The obtained results for representative annotators are depicted in Figure 5.

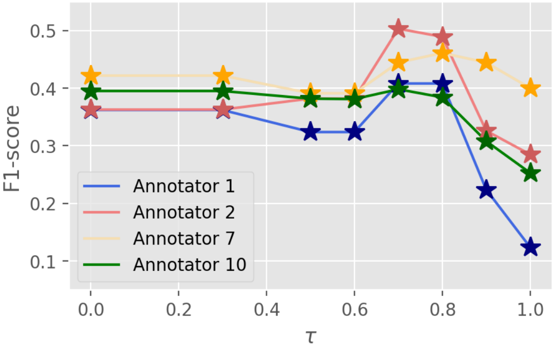


Figure 5: Ablation study on the effect of the confidence threshold τ on the proposed dual-branch entropy calibration (DBEC) based on annotated soft labels.

The performance of the DBEC proposed in relation to the τ value shows a characteristic shape. The non-expert annotators that show an improvement in the model performance using the proposed term first drop the obtained results when increasing τ. Then, an absolute maxima is reached around τ values of 0.7 and 0.8. Finally, increasing the hyper-parameter from this value worsen the performance, since entropy minimization is applied to all samples, even when high confidence is annotated. Based on these observations, we fixed $\tau = 0.7$ for the implementation of the dual-branch calibration.

6. Conclusions

A relevant body of literature on uncertainty estimation requires multiple annotators to quantify individual sample noise and inter-annotator variability. Nevertheless, acquiring multiple rater views is a limiting factor in a wide range of applications, such as medical imaging. In particular, in the case of digital pathology imaging, a high level of expertise is required to perform image labeling, which may make it unfeasible to recruit multiple annotators. To address this limitation, in this work we have proposed to capture individual uncertainties by annotating soft labels instead of unique categories. In addition, and inspired by previous literature on model calibration using Shannon entropy, we have proposed a dual-branch min-max entropy calibration (DBEC) criteria that optimize the model training to (i) avoid overconfident outputs by entropy maximization, and (ii) produce confident outputs on samples labeled with high uncertainty by Shannon entropy minimization, which focuses on inherent features of each sample.

The proposed uncertainty estimation method is validated in the challenging context of skin whole slide image (WSI) multi-class image classification, under the multiple instance learning (MIL) paradigm. It is worth highlighting the scarce literature on this field since, to the best of our knowledge, this is the first work that aims to distinguish among 6 different relevant pathological categories. Over the AI4SKIN dataset, we have generated new uncertainty-driven soft labels from 10 in-training pathologists, so-called non-expert annotators. Uncertainty-aware MIL models have been trained using soft labels, and the novel dual-branch min-max entropy calibration, and they have been evaluated using a ground truth annotated by expert pathologists. Results show a consistent improvement by using soft labels with standard cross-entropy loss during training (∼ 4.0% F1-score), and increases when calibrating the model with the proposed min-max entropy calibration DBCE (∼ 6.6% F1-score). In addition, we have observed that improvements using the DBCE appear in categories that non-expert annotators presented high uncertainty, which supports our claim that the entropy minimization term in this case helps the model to move away from the annotator bias. These improvements are produced at a negligible cost, both at the level of annotation and calculation.

Still, during the experimental stage, we found some limitations in our study. First, the proposed formulations are still highly dependent on the quality of the produced labels. In the context of non-expert annotators, this may produce limitations when labels are too noisy. Likewise, the annotation of soft labels depends on the commitment of the experts recruited and does not bring improvements when performed in a very low proportion. We believe that the framework developed in this work opens the door to different interesting lines of further research. Learning how to combine certain expert labels with uncertain non-expert labels might be of great interest, such as crowd-sourcing methods able to obtain the underlying label distribution using the least number of annotators, among others.

Funding

This work has received funding from Horizon 2020, the European Union's Framework Programme for Research and Innovation, under the grant agreement No. 860627 (CLARIFY), the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C21 (AI4SKIN) 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 J. Silva-Rodríguez was carried out during his previous position at Universitat Politècnica de València.

Acknowledgements

We gratefully acknowledge the support from the Generalitat Valenciana (GVA) with the donation of the DGX A100 used for this work, action co-financed by the European Union through the Operational Program of the European Regional Development Fund of the Comunitat Valenciana 2014-2020 (IDIFEDER/2020/030).

References

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

[2] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, K. A. Iczkowski, J. G. Kench, G. Kristiansen, T. H. van der Kwast, K. R. Leite, J. K. McKenney, J. Oxley, C. C. Pan, H. Samaratunga, J. R. Srigley, H. Takahashi, T. Tsuzuki, M. Varma, M. Zhou, J. Lindberg, C. Lindskog, P. Ruusuvuori, C. Wählby, H. Grönberg, M. Rantalainen, L. Egevad, M. Eklund, Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, The Lancet Oncology 21 (2) (2020) 222–232.

[3] 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) 13011309.

[4] 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.

[5] J. Silva-Rodríguez, A. Colomer, V. Naranjo, WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images, Computerized Medical Imaging and Graphics 88 (2021) (2021).

[6] H. Lei, S. Liu, A. Elazab, X. Gong, B. Lei, Attention-Guided Multi-Branch Convolutional Neural Network for Mitosis Detection from Histopathological Images, IEEE Journal of Biomedical and Health Informatics 25 (2) (2021) 358–370.

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

[8] Z. Xu, D. Lu, J. Luo, Y. Wang, J. Yan, K. Ma, Y. Zheng, R. K.-y. Tong, Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation, IEEE Transactions on Medical Imaging (2022) 1–13.

[9] Z. Xu, D. Lu, Y. Wang, J. Luo, J. Jayender, K. Ma, Y. Zheng, X. Li, Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021.

[10] W. Luo, M. Yang, Semi-supervised semantic segmentation via strong-weak dual-branch network, in: European Conference on Computer Vision (ECCV), 2020.

[11] J. Dolz, C. Desrosiers, I. B. Ayed, Teach me to segment with mixed supervision: Confident students become masters, in: International Conference on Information Processing in Medical Imaging (IPMI), 2021.

[12] H. Zhu, J. Shi, J. Wu, Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019.

[13] T. Zhang, L. Yu, N. Hu, S. Lv, S. Gu, Robust medical image segmentation from non-expert annotations with tri-network, in: International Conference on medical image computing and computer-assisted intervention (MICCAI), 2020.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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 algo-

rithm, Frontiers in oncology 10 (2020) 1559.

[24] L. Wang, L. Ju, D. Zhang, X. Wang, W. He, Y. Huang, Z. Yang, X. Yao, X. Zhao, X. Ye, Z. Ge, Medical Matting: A New Perspective on Medical Segmentation with Uncertainty, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021.

[25] C. Devalland, Spitzoid lesions diagnosis based on smote-ga and stacking methods, Advanced Intelligent Systems for Sustainable Development (AI2SD2019): Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health 1103 (2020) 348.

[26] 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.

[27] V. Lai, W. Cranwell, R. Sinclair, Epidemiology of skin cancer in the mature patient, Clinics in dermatology 36 (2) (2018) 167–176.

[28] 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.

[29] T. T. Ha Lan, S. J. Chen, D. P. Arps, D. R. Fullen, R. M. Patel, J. Siddiqui, S. Carskadon, N. Palanisamy, P. W. Harms, Expression of the p40 isoform of p63 has high specificity for cutaneous sarcomatoid squamous cell carcinoma, Journal of cutaneous pathology 41 (11) (2014) 831–838.

[30] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, E. Gibson, R. Vishwanath, A. Balachandran, J. M. Balter, Y. Cao, R. Singh, et al., Quantifying and leveraging predictive uncertainty for medical image assessment, Medical Image Analysis 68 (2021).

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

[32] A. Galdran, J. Dolz, H. Chakor, H. Lombaert, I. Ben Ayed, Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020.

[33] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, Y. Zheng, Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling, in: IEEE Conference on Computer Vision and Pattern Recoginition (CVPR), 2021.

[34] M. H. J. B, D. R. Jørgensen, R. Jalaboi, Improving Convolutional Neural Networks Using Inter-rater Agreement, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2019.

[35] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, M. Reyes, On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.

[36] L. Ju, X. Wang, L. Wang, G. S. Member, D. Mahapatra, Improving medical images classification with label noise using dual-uncertainty estimation, IEEE transactions on medical imaging 41 (6) (2022) 1533–1546.

[37] L. Zhang, R. Tanno, M. C. Xu, C. Jin, J. Jacob, O. Ciccarelli, F. Barkhof, D. C. Alexander, Disentangling human error from the ground truth in segmentation of medical images, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.

[38] G. Pereyra, G. Tucker, J. Chorowski, Å. Kaiser, G. Hinton, Regularizing neural networks by penalizing confident output distributions, in: International Conference on Learning Representations (ICLR), 2019.

[39] S. Belharbi, J. Rony, J. Dolz, I. B. Ayed, L. McCaffrey, E. Granger, Deep Interpretable Classification and Weakly-Supervised Segmentation of Histology Images via Max-Min Uncertainty, IEEE Transactions on Medical Imaging 41 (3) (2022) 702–714.

[40] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, P. Berens, Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection, Medical Image Analysis 64 (2020).

[41] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, D. Huang, CurriculumNet: Weakly supervised learning from large-scale web images, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.

[42] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, M. Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[43] L. Jiang, Z. Zhou, T. Leung, L. J. Li, L. Fei-Fei, Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, in: International Conference on Learning Representations (ICLR), 2018.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[45] B. Liu, I. B. Ayed, A. Galdran, J. Dolz, The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[46] C. Meister, E. Salesky, R. Cotterell, Generalized Entropy Regularization or: Theres Nothing Special about Label Smoothing, in: Annual Meeting of the Association for Computational Linguistics, 2020.

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

[48] Y. Grandvalet, Y. Bengio, Semi-supervised Learning by Entropy Minimization, 2004.

[49] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. B. Ayed, J. Dolz, Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?, 2021.

[50] 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), 2022.

[51] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: International Conference on Learning Representations (ICLR), 2014.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

HIGHLIGHTS

- Skin histological images are used for the first time to develop an end-to-end automatic system able to distinguish between seven different types of spindle cell neoplasms.
- We propose an uncertainty-aware labeling strategy with which generate soft labels from 10 non-expert annotators for multi-class skin cancer classification that quantify uncertainty in the annotations.
- Based on these annotations, we present an extensive study for the use of soft label model calibration compared to the ground truth, labeled by an expert pathologist.
- A novel formulation based on dual-branch entropy calibration (DBEC) is proposed to penalize both, overconfident outputs and uncertain soft labels, during training.
- Comprehensive experiments demonstrate the promising performance of our labeling strategy. By incorporating uncertainty during labeling, we found average improvements of nearly ~4.0% in averaged F1-score using the baseline methods, which increases up to ~ 6.6% using the proposed dual-branch calibration.

Credit Author statement:

Rocío del Amor: Software, Data Curation, Methodology, Writing - Original Draft

Julio Silva-Rodríguez: Software, Methodology, Writing - Original Draft

Valery Naranjo: Supervision, Writing - Review & Editing, funding acquisition, Project administration

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: