The CrowdGleason dataset: learning the Gleason grade from crowds and experts

Miguel López-Pérez[a], Alba Morquecho[b], Arne Schmidt[b], Fernando Pérez-Bueno[d], Aurelio Martín-Castro[c], Javier Mateos[b], Rafael Molina[b]

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

[b]*Department of Computer Science and Artificial Intelligence, Universidad de Granada, Granada, Spain*

[c]*Department of Pathology, Virgen de las Nieves University Hospital, 18014 Granada, Spain*

[d]*Basque Center on Cognition, Brain and Language, Donostia - San Sebastián, Spain.*

Abstract

Background: Currently, prostate cancer (PCa) diagnosis relies on the human analysis of prostate biopsy Whole Slide Images (WSIs) using the Gleason score. Since this process is error-prone and time-consuming, recent advances in machine learning have promoted the use of automated systems to assist pathologists. Unfortunately, labeled datasets for training and validation are scarce due to the need for expert pathologists to provide ground-truth labels.

This work was supported in part by FEDER/Junta de Andalucía under project P20_00286, grant PID2022-140189OB-C22 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU". The work by Miguel López-Pérez and Fernando Pérez-Bueno was supported by the grants JDC2022-048318-I and JDC2022-048784-I, respectively, funded by MICIU/AEI/10.13039/501100011033 the European Union "NextGenerationEU"/PRTR.

Email addresses: mlopper3@upv.es (Miguel López-Pérez), e.amorq@go.ugr.es (Alba Morquecho), arne@decsai.ugr.es (Arne Schmidt), fperezbueno@bcbl.eu (Fernando Pérez-Bueno), aureliomartincastro@gmail.com (Aurelio Martín-Castro), jmd@decsai.ugr.es (Javier Mateos), rms@decsai.ugr.es (Rafael Molina)

Method: This work introduces a new prostate histopathological dataset named CrowdGleason, which consists of 19,077 patches from 1,045 WSIs with various Gleason grades. The dataset was annotated using a crowdsourcing protocol involving seven pathologists-in-training to distribute the labeling effort. To provide a baseline analysis, two crowdsourcing methods based on Gaussian Processes (GPs) were evaluated for Gleason grade prediction: SVGPCR, which learns a model from the CrowdGleason dataset, and SVGPMIX, which combines data from the public dataset SICAPv2 and the CrowdGleason dataset. The performance of these methods was compared with other crowdsourcing and expert label-based methods through comprehensive experiments.

Results: The results demonstrate that our GP-based crowdsourcing approach outperforms other methods for aggregating crowdsourced labels ($\kappa = 0.7048\pm0.0207$) for SVGPCR vs.($\kappa = 0.6576\pm0.0086$) for SVGP with majority voting). SVGPCR trained with crowdsourced labels performs better than GP trained with expert labels from SICAPv2 ($\kappa = 0.6583 \pm 0.0220$) and outperforms most individual pathologists-in-training (mean $\kappa = 0.5432$). Additionally, SVGPMIX trained with a combination of SICAPv2 and CrowdGleason achieves the highest performance on both datasets ($\kappa = 0.7814 \pm 0.0083$ and $\kappa = 0.7276 \pm 0.0260$).

Conclusions: The experiments show that the CrowdGleason dataset can be successfully used for training and validating supervised and crowdsourcing methods. Furthermore, the crowdsourcing methods trained on this dataset obtain competitive results against those using expert labels. Interestingly, the combination of expert and non-expert labels opens the door to a future of massive labeling by incorporating both expert and non-expert pathologist annotators.

Keywords: Computational pathology, Crowdsourcing, Prostate cancer, Gleason grade, Gaussian processes, Medical image analysis

1. Introduction

Prostate cancer is a prevalent cancer and the fifth leading cause of cancer-related deaths worldwide [1]. Timely and precise diagnosis is crucial for effective treatment and reducing mortality rates [2]. Currently, the gold standard for diagnosis and prognosis is to analyze a biopsy of prostate tissue by the Gleason grading (GG) system which assesses the cancer stage and aggressiveness based on gland morphology. However, the assessment of GG is inherently subjective with high intra- and inter-observer variability [3, 4].

Computer-Aided Diagnosis (CAD) systems assist pathologists and aim to minimize human variability in decision-making. These systems utilize WSIs and computer vision and machine learning (ML) algorithms to detect and grade cancerous regions. The main bottleneck in training and validating ML methods for GG prediction is the scarcity of large-scale public datasets [5]. Creating these datasets is costly and time-consuming and, together with the scarcity of expert pathologists, explain why there are few annotated datasets and even fewer public datasets.

Crowdsourcing has emerged as a cost-effective and efficient method for labeling histopathological datasets by leveraging a large pool of annotators with varying levels of expertise [6, 7]. While crowdsourcing has shown success in tasks like nuclei detection [8] and cancer cell identification [9], the labels generated are frequently noisy, limiting their direct application to complex tasks such as GG. To address this challenge, probabilistic models like GPs have become popular [10, 11]. GPs for crowdsourcing have demonstrated excellent performance in various tasks [12, 13, 14] and have been successfully applied in histopathological image classification studies, including breast cancer [15, 16] and skin cancer detection [17]. These methods offer competitive performance compared to methods trained with expert labels, indicating that crowdsourced labeling of histopathological images could be a feasible option for cancer classification with minimal reliance on expert pathologists. Regarding GG classification, there are no previous studies involving non-expert annotators. However, it has been shown that learning from the opinion of multiple expert pathologists, despite high inter- and intra-observer variability, results in strong performance when effectively modeling this variability [18, 19].

The objective of this work is twofold. First, we present and make publicly available the first prostate dataset labeled by non-experts for GG prediction. Second, we explore the learning from crowds framework with this novel dataset, assessing and analyzing two state-of-the-art methods based on GP for crowdsourcing. This paper also demonstrates the viability of integrating this new dataset with existing datasets containing expert labels, to create a larger and more diverse dataset. Our experiments indicate that the noisy non-expert labels from the presented dataset can improve previous models in the literature. Below, we outline our contributions in detail:

- Introduction of new crowdsourcing protocol for the annotation of patches from WSIs, outlined in Fig. 1, which cheapens and speeds up the label-

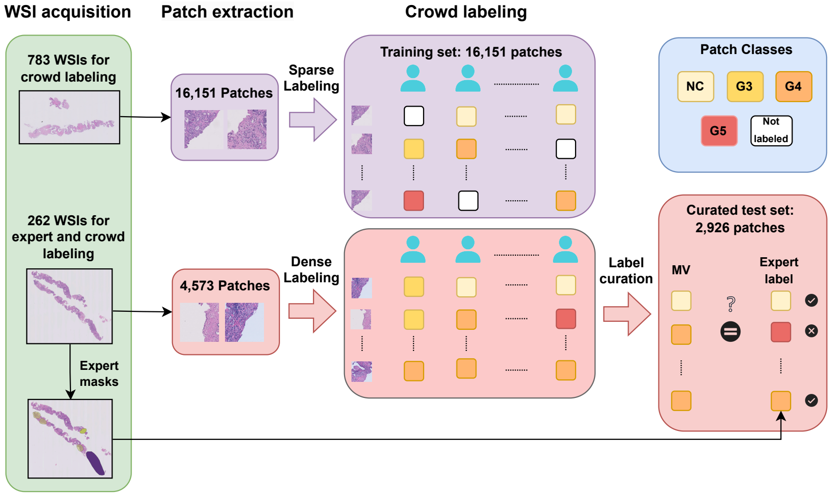


Figure 1: Dataset creation and annotation protocol. We collect 1,045 WSIs, of which 783 are used exclusively for crowd labeling and 262 for crowd and expert labeling. We divide all WSIs into patches and distribute them among the non-expert annotators to obtain the training set. We create a curated test set with patches where the experts and majority of the non-experts agree.

ing process by dramatically reducing the intervention of expert pathologists.

- Creation of a new dataset, called CrowdGleason, comprising 19,077 patches from 1,045 WSIs of PCa with different GG. This dataset was annotated by seven pathologist-in-training without expert supervision. Note that not all annotators labeled all patches. To the best of our knowledge, this is the first PCa dataset annotated by non-expert pathologists.

- Development of a curated test set annotated by each pathologist-in-training and two expert PCa pathologists to evaluate automated ML methods and assess bias, expertise, and discrepancies between participants.

- Comprehensive experiments to evaluate two GP-based methods for GG prediction: SVGPCR [7] and SVGPMIX [14]. SVGPCR learns from the CrowdGleason dataset, while SVGPMIX combines expert labels from the public SICAPv2 [20] dataset with the CrowdGleason dataset. Results demonstrate that these GP-based crowdsourcing methods outperform popular techniques for label aggregation, with SVGPMIX achieving the best performance in both datasets.

The remainder of the work is organized as follows. Section 2 describes related work. Section 3 presents the CrowdGleason dataset and its annotation protocol. Section 4 describes the experimental setup and the methods evaluated. The experimental results are shown in Sect. 5, and Sect. 6 discusses them. Finally, Sect. 7 presents the conclusions and future work.

2. Related work

Public datasets are essential to develop precise ML methods for GG prediction. Hence, Sect. 2.1 delves into the current publicly available PCa datasets and Sect. 2.2 provides an overview of the core work of crowdsourcing and its applications in the context of histopathology.

2.1. Public PCa histopathological datasets

The current publicly available PCa histopathological datasets have been typically created by staining tissue biopsies with hematoxylin and eosin (H&E) and scanning them as WSI for histopathological examination. In clinical practice, a WSI usually contains one or a few tissue samples. The use of Tissue Micro Arrays (TMAs) allows many tissue samples to be arranged on a grid and processed simultaneously to obtain a single slide. These datasets are labeled at pixel, patch, or WSI levels. The labeling process at pixel level consists of manually delineating tumor areas and assigning GG classes. This meticulous procedure provides comprehensive tumor information but it is time-consuming. In WSI level labeling, pathologists assign a label to the entire image without specific tumor location information. Patch level labeling divides WSIs into small regions, named patches, and a label is assigned to a selected set of patches, thus reducing the need to examine the entire WSI.

We briefly examine popular public datasets for GG prediction, including Arvaniti, SICAP, GLEASON2019, and PANDA. Table 1 provides an overview of these datasets and our proposed CrowdGleason. Arvaniti *et al.* [21] dataset comprises TMAs annotated at pixel level by an expert pathologist, while SICAPv1 [11] and SICAPv2 [20] datasets offer pixel-level anno-

Table 1: Publicly available datasets for GG prediction. MA refers to multiple annotators.

	Biopsy	# Samples	Annotations	Experts	MA
Arvaniti [21]	TMA	895	Pixel-level	Yes	No
SICAPv1 [11]	WSI	79	Pixel-level	Yes	No
SICAPv2 [20]	WSI	182	Patch-/pixel-level	Yes	No
GLEASON19 [18]	TMA	331	Pixel-level	Yes	Yes
PANDA [22]	WSI	12,625	WSI-/pixel-level	Yes	No
CrowdGleason (proposed)	WSI	1,045	Patch-level	No	Yes

tations on WSIs. The WSIs were downsampled at 10× magnification and divided into patches of 512^2 pixels with 50% overlap obtaining patch-level annotations. To our knowledge, SICAPv2 is the largest fully annotated dataset at patch level in the literature.

Challenges, such as Gleason2019 and PANDA, have been a popular way of promoting research in GG prediction by providing benchmark datasets for evaluating ML algorithms. The Gleason2019 challenge [18] dataset provides TMA images annotated by a panel of 5 expert pathologists, and the PANDA Challenge [22] dataset includes WSIs annotated at the WSI level by consensus among a large panel of highly experienced expert pathologists, with some samples annotated at pixel level.

2.2. Crowdsourcing

To the best of our knowledge, all works for GG prediction from patches addressed the problem with ground-truth labels provided by either a single expert or a panel of expert pathologists. Crowdsourcing presents an opportunity to scale datasets by engaging non-expert annotators in computational pathology-related tasks [8]. Various studies explored the use of labels from non-expert annotators for tasks like mitosis detection [23] or histopathological image classification [24]. Previous works [25, 26] have demonstrated promising results in the field of histopathology using crowdsourcing, but they required strong supervision from senior pathologists to review the annotations provided by the crowd. To reduce the need for expert supervision, label aggregation techniques [27] have been developed to automatically curate crowdsourcing labels, enabling the creation of datasets suitable for ML without expert supervision. Various label aggregation methods have been proposed, including majority voting (MV) and more elaborated methods that consider the biases of the different annotators, yielding a better-calibrated set of training labels [7]. They include Dawid-Skene (DS) [28], GLAD [29] and MACE [30] models.

Recent studies show that jointly learning ground-truth labels, annotator expertise, and the latent classifier leads to superior performance [31]. Models like SVGPCR [7] have successfully combined sparse GPs with a crowdsourcing probabilistic framework, demonstrating competitive performance to GPs trained with expert labels in breast cancer detection from histopathological images [32]. Moreover, SVGPMIX method [14] is the first probabilistic approach based on GPs for fusing expert and non-expert labels, leveraging the confidence offered by expert labels and the larger volume of data provided by non-expert annotators. To the best of our knowledge, this model has not yet been applied in the biomedical domain.

3. CrowdGleason dataset

The dataset presented in this paper, named CrowdGleason, has been partially annotated by different pathologists in-training with varying degrees of expertise. A subset of the dataset was also annotated by expert pathologists, which helped to obtain a test set.

3.1. Data acquisition and annotation by expert pathologists

To create CrowdGleason, 1,045 WSIs of H&E-stained prostate tissue samples from different patients, were collected by medical experts from the archive of the Hospital Universitario San Cecilio (HUSC) in Granada. All WSIs were digitally scanned at 40× magnification factor. Two expert pathologists exhaustively annotated 262 of those 1,045 WSIs at the pixel level. Each image was annotated by only one of the pathologists independently, using the online application described in [20]. Experts thoroughly marked all pathological areas with their GG and delineated artifacts.

3.2. Patch extraction

All WSIs were divided into patches of size 2048×2048 pixels at a magnification of 40x, without overlapping. This size and magnification were selected in agreement with expert pathologists to provide sufficient context and detail to facilitate the identification of cancerous lesions. Patches containing less than 20% of tissue were discarded, as they do not contain enough tissue to make an accurate diagnosis. Tissue presence was detected by thresholding the magenta channel by the Otsu method [33]. From images with pathological areas marked by the experts, we selected patches containing at least

15% of pathological tissue, labeled with GG of the area marked by the expert: Gleason grade 3 (G3), Gleason grade 4 (G4), or Gleason grade 5 (G5). Patches containing more than one pathological area were discarded since it was not possible to assign a single label to the patch. From images labeled as non-cancerous (NC) by the expert pathologist, on the other side, we could use all tissue to extract patches. To reduce the number of candidate patches, we discarded patches having less than 30% of tissue. A total of 4573 patches, which form the so called *expert-labeled set*, were obtained from the 262 images annotated by experts.

For the remaining 783 images not annotated by expert pathologists, a large number of patches without ground-truth labels were extracted to be annotated by non-expert pathologists at a later stage. To expedite the labeling process, we reduced the number of patches. As gland structure is crucial in PCa diagnosis, we chose patches with a substantial presence of nuclei as representative of tissue with glands. Since nuclei stain with hematoxylin, which is prominent in the cyan component, to extract patches rich in nuclei, we selected those patches where at least 40% of the tissue's pixels had a high cyan value. Still the number of patches was overwhelming. Due to the huge class imbalance in histopathological data, with large areas of non-pathological tissue, and cancerous tissue that is only sparsely represented, to select a set of patches representing the different PCa grades, we proceeded as follows. We trained the classification algorithm in [34] that combines semi-supervised and multiple instance learning on the public PANDA dataset for PCa classification, following the successful setting in [34]. Note that the PANDA dataset is labeled at the WSI level; hence, standard supervised learning techniques

Table 2: Metrics for the patch selection algorithm in a small set of patches extracted from the expert-labeled set.

Accuracy	F1 score	Kappa
0.732	0.648	0.606

cannot be used. Although segmentation masks are provided for some images, they can be only used to develop strategies for selecting the most significant subsamples of the images [35]. The algorithm in [34] uses an EfficientNet-B5 neural network architecture [36] that was trained with a learning rate of 0.01 for 10 epochs on the classes NC, G3, G4, and G5. To validate this approach, we classified a small set of patches extracted from the expert-labeled set. Note that these patches were only used to obtain the metrics shown in Table 2 and were not used in model training. These figures of merit show that the method is good enough to distinguish patches from the different classes. Using the learned model, the patches from non-annotated WSIs were classified in the class with the highest probability, selecting a total of 16,151 patches. This collection constitutes a roughly balanced set that will serve as *training set* of our study. Since this set was designed for annotation by non-expert pathologists, the labels provided by the classification algorithm were discarded after patch selection and not further used in this study.

3.3. Annotation by non-expert pathologists

Seven pathologists in-training with different expertise levels participated in the annotation of the dataset. Two of them were in their fourth year of medical residency, two in their third, two in their second, and one was a first-year medical resident. Following the Spanish Official Specialist Train-

Table 3: Distribution of labels in the expert-labeled set labeled by each resident pathologist and the expert. We refer to each annotator as A#, where the number is an anonymized ID.

Class	A1	A2	A3	A4	A5	A6	A7	Expert	Total
NC	1891	2290	2968	2983	2941	2868	389	2438	14439
G3	1024	1267	507	601	763	693	1402	1498	5233
G4	1155	674	808	663	413	666	1513	449	4737
G5	503	341	281	308	431	317	1074	188	2752
Total	4573	4572	4564	4555	4548	4544	4378	4573	27161

ing Program in Anatomic Pathology, third and fourth-year pathologists in-training have completed specific training in the subspecialty of Uropathology, which includes the study of the prostate, with an approximate duration of 2-3 months. First and second-year students do not have specific training in this subspecialty. In collaboration with expert pathologists, we designed an annotation protocol based on the well-known PCa grading of the Gleason Score, originally introduced by Donald F. Gleason [37] for grading prostatic carcinoma based solely on the architectural pattern of the tumor. Non-expert pathologists were instructed to label as NC, G3, G4, or G5 the patches in the expert-labeled and training sets, described in the previous section, rather than thoroughly examine the WSI and exhaustively delineate the tumor areas. Patches that could not be labeled due to the presence of artifacts, blurriness, tissue from other organs, folded tissue, etc. have been labeled as "Other". Patches with more than one GG, which have not a clear label, have also been labeled as "Other".

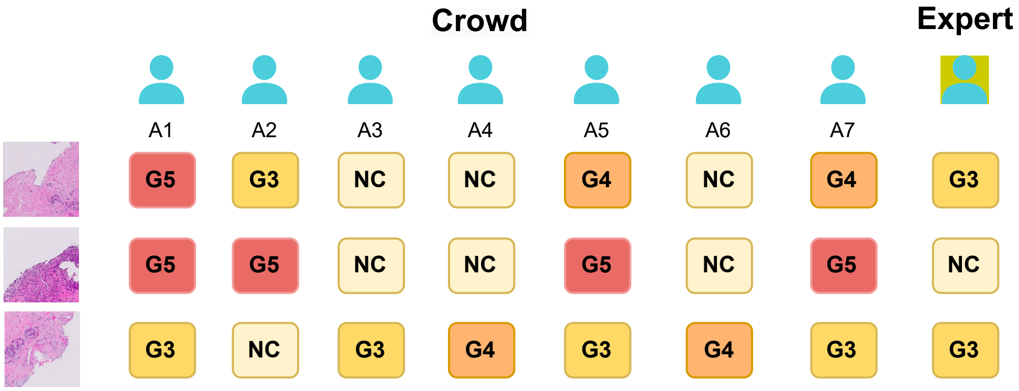


Figure 2: Examples of patches and annotations by each participant.

All pathologists in-training labeled the 4,573 patches in the expert-labeled set. Table 3 presents a summary of the distribution of the labels in this set. Note that some patches were labeled as “Other” by some residents and, hence, the total of each column may not add up to the total number of patches. An example of patches and the annotations provided by the crowd and the experts is shown in Fig. 2. To minimize the influence of inherent pathologist variability in labeling, we created a *curated test set* where the ground-truth label for each sample was established by consensus between the majority of pathologists in-training and the expert pathologist. Using this curated set, whose distribution of samples for each class is shown in Table 4, we will estimate the degree of reliability of each resident as well as evaluate ML methods. Recall that expert intervention has only been necessary for the creation of the curated test set, not for the training set.

The 16,151 patches in the training set were labeled, on average, by more than two resident pathologists, with each pathologist in-training labeling approximately 5,000 patches. Table 5 summarizes the training set. Patches

Table 4: Distribution of patches in each class of the curated test set.

NC	G3	G4	G5	Total
2157	548	164	57	2926

Table 5: Distribution of patches in the training set labeled by each resident pathologist. We refer to each annotator as A#, where the number is an anonymized ID.

Class	A1	A2	A3	A4	A5	A6	A7	Total
NC	2165	2747	3659	4186	3476	3452	866	20551
G3	1462	1618	604	149	1024	667	682	6206
G4	995	510	544	580	405	641	2580	6255
G5	400	205	140	111	250	337	677	2120
Total	5022	5080	4947	5026	5155	5097	4805	35132

were provided to the residents in 4 batches of approximately equal size over a 6-month period.

Finally, as a post-processing step, the patches of both the training set and the curated test set were downsampled using bicubic interpolation to a size of 512 × 512 pixels. This is equivalent to obtaining the patches at a magnification factor of 10×, and it was necessary to accommodate the patches into the GPU memory.

In summary, the **CrowdGleason** consists of a crowdsourcing annotated training set with 16,151 patches of size 512 × 512 extracted from 783 WSIs, annotated by one or more of the seven pathologists in-training, and a curated test set with 2,926 patches of size 512 × 512 extracted from other 262 WSIs, annotated by expert pathologists and all the pathologists in-training.

Ground-truth labels for the curated test set were obtained by consensus between the expert pathologists and the majority of the pathologists in-training.

3.4. Ethical Consent and Data Availability

The Research Ethics Committee of the Universidad de Granada approved the study with code 4096/CEIH/2024 as part of the project P20_00286, funded by FEDER/Junta de Andalucía, following the principles established in international and national biomedical international and national legislation in the field of biomedicine, bioethics and bioethics, as well as the rights derived from the protection of personal data.

The complete CrowdGleason dataset will be available in Figshare upon acceptance of the paper.

4. Materials and Methods

Datasets. We present and utilize the dataset, described in Sec. 3.1, and combine it with the public dataset SICAPv2 [20]. We normalize both datasets via the BKSVD method [38], and use them for training and evaluation to demonstrate the utility of the proposed CrowdGleason dataset with respect to another popular dataset in the literature. Our approach also allows for the generalization ability of the classifiers on an external cohort. The task is to learn the GG of each patch, i.e. to classify the patches as 'NC', 'G3', 'G4' and 'G5'.

Feature extraction. A reduced set of features is extracted and used as input to the GP-based methods. For this, we utilize the 18-layer variant of ResNet [39], i.e., ResNet18, pre-trained on Imagenet and fine-tuned on SICAPv2. We use the output of the last convolutional layer as a feature extractor. Since SICAPv2 is the largest publicly available PCa dataset with patch labels, we assume that the learned feature extractor generalizes well to other datasets. The experimental results will validate this assumption. We utilize these 512-dimensional feature vectors as input to the GP classifiers and also report the results of the end-to-end training of ResNet18 for comparison. We perform five independent runs of all the presented experiments, including the mean performance and 95% confidence interval (CI). Note that for each run, we also run the feature extractor to obtain a different set the features and ensure the robustness of the whole pipeline proposed in this work. To train the network, we use the SGD optimizer with a learning rate of 10^{-3}, a momentum of 0.9, and a batch size of 32 patches. Common data augmentation transformations, such as horizontal and vertical flips, blur, and brightness, contrast, hue, and saturation variations, are applied to the training dataset. The CNN is implemented using Pytorch 2.0.1 and is run on an NVIDIA GeForce RTX 3090 GPU.

Supervised Learning: Gaussian Processes. We use the features extracted by ResNet18 and the ground-truth label as inputs of a stochastic variational Gaussian process (SVGP) model to learn the GG from the SICAPv2 dataset. SVGPs are scalable GP models that use variational inference to approximate the posterior distribution. See a detailed description in [40] and an intuitive review in [32]. We utilize a squared exponential kernel to compute the correlation matrix. We initialize the kernel hyperparameters, lengthscale and variance, to 2. We train the SVGP for 50 epochs and save the parameters that obtain the best Cohen's Quadratic Kappa (κ) on the validation set. In all cases, we use the Adam Optimizer with a learning rate of 10^{-2} and a batch size of 128. Based on the experimental results (see Sect. 5.1), we fixed a value of 512 inducing points, which provides a good trade-off between the generalization and complexity of the model, for all GP based methods. SVGP is implemented using GPflow 1.2.0 and is run on an NVIDIA GeForce RTX 3090 GPU. The code will be released in GitHub upon acceptance of the paper.

Label aggregation. We utilize and compare Majority Voting (MV), DS [28], MACE [30], and GLAD [29] aggregation methods to curate the multiple noisy labels available in the CrowdGleason dataset. The aggregated labels can be used as the single ground-truth label and, therefore, used by supervised learning methods. All methods, implemented in the popular Python library for crowdsourcing tasks *crowd-kit* [41], are run with the default hyperparameters. The aggregated labels and the features extracted by ResNet18 are fed to SVGP to learn a GG classifier.

Crowdsourcing models. We utilize, as an enhancement of the label aggregation methods, the learning from crowds model based on GPs, SVGPCR [7]. This model extends the GPs to the crowdsourcing scenario and jointly learns the expertise of the annotators and the GP classifier. The main assumption is that multiple annotators provide noisy labels that are corrupted observations of the ground-truth label. This corruption is modeled with a confusion matrix for each annotator, which reflects the probability of providing a given label for each ground-truth class (as in the DS model [28]). Once trained, the model can predict ground-truth labels in unseen instances using the GP classifier.

Furthermore, we analyze how crowdsourcing labeled datasets can be used in conjunction with expert labeled datasets to learn a classifier. For this purpose, we use the SVGPMIX model [14]. This model, used here for the first time in medical imaging, generalizes SVGPCR to cases where labels have been provided either by a noisy annotator or by an expert. We utilize this model to study the combination of CrowdGleason with SICAPv2. SVGPCR and SVGPMIX use the same training procedure and software framework as supervised GPs.

Evaluation Metrics. To assess the quality of the learned models, we report the numerical results of three different metrics: Accuracy, Cohen's Quadratic Kappa (κ), and the F1-score. The accuracy is the rate of success of the classifiers. The F1 can be defined per class as the harmonic mean between precision and recall. We only report multiclass F1, which can be defined as the average across class-wise F1. Recall that this score is of special importance in imbalanced scenarios, which are common in medical imaging. Finally, the κ score is increasingly popular for GG assessment [20, 42, 21, 43]. It measures the level of agreement between the output of the classifier and the ground-truth label [44]. We can also use it to measure the agreement between annotators. The κ metric ranges from -1 to 1, being directly proportional to the level of agreement between observers (-1 complete disagreement, 0 no agreement beyond what would be expected by chance, 1 total agreement). It is commonly argued that a moderate agreement is achieved if κ is higher than 0.6, whereas a strong agreement is attained when κ is higher than 0.8. Furthermore, this metric also penalizes disagreements depending on class differences (in a quadratic manner). For example, a disagreement between classes NC and G5 implies a stronger penalization than between classes NC

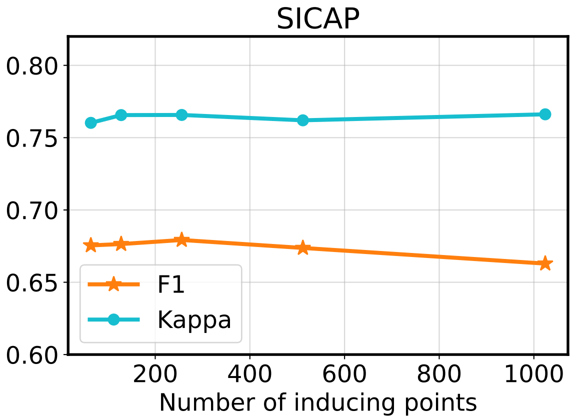


Figure 3: Variation of the F1 and Kappa scores with the number of inducing points for the SVGP model trained on the SICAPv2 dataset. These results are reported on the validation set.

and G3.

5. Experimental Results

In this section, we report the results of a set of experiments. They compare different GP-based approaches that learn from expert labels, crowdsourcing labels, and a combination of both.

5.1. Experiment 1: Expert labels

In this experiment, we present the results of models trained on expert SICAPv2 labels. The model is validated using the validation set of SICAPv2. For comparison purposes, we also train the CNN-based method ResNet18 with the same data.

To select the number of inducing points for the SVGP method, we run the method with several values: 64, 128, 256, 512, and 1024. Figure 3 shows

Table 6: Results of the methods trained on SICAPv2 (expert labels) when tested on SICAPv2 and CrowdGleason test sets.

	SICAPv2			CrowdGleason		
Method	Accuracy	F1 score	Kappa	Accuracy	F1 score	Kappa
ResNet18	**0.7648 ± 0.0102**	**0.7145 ± 0.0143**	0.6611 ± 0.0149	**0.8839 ± 0.0122**	**0.6698 ± 0.0317**	**0.7095 ± 0.0426**
SVGP-SICAP	0.7515 ± 0.0048	0.6912 ± 0.0119	**0.7736 ± 0.0139**	0.8736 ± 0.0075	0.6628 ± 0.0061	0.6583 ± 0.0220

Table 7: Results of the methods trained on CrowdGleason (crowdsourcing labels) when tested on SICAPv2 and CrowdGleason test sets

	SICAPv2			CrowdGleason		
Method	Accuracy	F1 score	Kappa	Accuracy	F1 score	Kappa
ResNet18-MV	0.5910 ± 0.0568	0.5251 ± 0.0694	0.4139 ± 0.0763	0.8815 ± 0.0149	0.6791 ± 0.0316	0.6958 ± 0.0415
SVGP-DS	0.4960 ± 0.0233	0.4345 ± 0.0282	0.4965 ± 0.0283	0.8402 ± 0.0121	0.5499 ± 0.0360	0.6152 ± 0.0236
SVGP-MACE	0.4980 ± 0.0212	0.4345 ± 0.0263	0.4759 ± 0.0347	0.8486 ± 0.0070	0.5363 ± 0.0300	0.5574 ± 0.0325
SVGP-GLAD	0.4909 ± 0.0256	0.4342 ± 0.0344	0.5052 ± 0.0533	0.8539 ± 0.0091	0.5410 ± 0.0260	0.5776 ± 0.0338
SVGP-MV	0.6861 ± 0.0138	0.6331 ± 0.0169	0.6242 ± 0.0277	0.8649 ± 0.0016	0.6287 ± 0.0123	0.6576 ± 0.0086
SVGPCR	**0.7123 ± 0.0072**	**0.6850 ± 0.0075**	**0.6953 ± 0.0176**	**0.9023 ± 0.0037**	**0.7068 ± 0.0142**	**0.7048 ± 0.0207**

that the F1 and Kappa scores are stable in the SICAPv2 validation set across different numbers of inducing points. This result means that the information can be summarized in a few points of the feature space, and adding more flexibility does not improve the performance. Furthermore, we can see that a large number of inducing points does not lead to overfitting. Hence, we fix a value of 512 inducing points for all experiments as a trade-off between complexity and generalization capability.

Table 6 shows the results of the SVGP method trained on SICAPv2 (denoted as SVGP-SICAP) tested on the SICAPv2 and CrowdGleason curated test sets. Additionally, Table 6 includes the test results of the end-to-end trained ResNet18 network for comparative analysis. The SVGP-SICAP classifier achieves better figures-of-merit for the Kappa score in SICAPv2 than using ResNet18 and is competitive in the rest of the metrics.

Table 8: Results of the methods trained on SICAPv2 and CrowdGleason (expert and crwodsourcing labels, respectively) combined when tested on SICAPv2 and CrowdGleason test sets.

	SICAPv2			CrowdGleason		
method	Accuracy	F1 score	Kappa	Accuracy	F1 score	Kappa
ResNet18-MV	**0.7743 ± 0.0056**	**0.7216 ± 0.0152**	0.6748 ± 0.0085	0.8804 ± 0.0192	0.6843 ± 0.0355	0.7042 ± 0.0381
SVGP-MV	0.6861 ± 0.0138	0.6331 ± 0.0169	0.6242 ± 0.0277	0.8649 ± 0.0016	0.6287 ± 0.0123	0.6576 ± 0.0086
SVGPMIX	0.7660 ± 0.0056	0.7137 ± 0.0119	**0.7814 ± 0.0083**	**0.9027 ± 0.0096**	**0.7176 ± 0.0270**	**0.7276 ± 0.0260**

5.2. Experiment 2: Crowdsourcing labels

In this experiment, we train the methods with the CrowdGleason dataset. The dataset is split into 13824 training samples and 2327 validation samples. For validation, we use the MV strategy for aggregating the labels.

We first use different label aggregation strategies (DS, MACE, GLAD, and MV) to train the SVGP classifier. Note that the input features, as we have already indicated, are extracted using ResNet18 trained on SICAPv2. Results in Table 7 show that MV produces the best result among the label aggregation strategies followed by DS. For comparison purposes, we also report the results on ResNet18 trained end-to-end with the MV labels. Finally, we report the results of the SVGPCR method trained with the crowdsourcing labels of CrowdGleason. From Table 7 it is clear that SVGPCR outperforms the rest of the methods in the literature in both datasets. The MV aggregation strategy can reduce the bias of the annotations but the noisy labels hinder the classifier's performance.

5.3. Experiment 3: Combining expert and crowdsourcing labels

Until now, information from experts and crowds was not used simultaneously. In this experiment, we explore the possibility of enhancing expert-labeled datasets with crowdsourcing-labeled ones. We add the CrowdGleason training set to the SICAPv2 training set for this purpose. For supervised methods (SVGP and ResNet18), we consider MV aggregation, since it achieved the best results in experiment 2. All methods use SICAPv2 as the validation set since it already provides ground-truth labels.

Results are shown in Table 8. SVGPMIX outperforms the competing methods on the SICAPv2 and CrowdGleason test datasets, showing that the combination of expert and crowdsourcing labels is feasible and beneficial. Although ResNet18-MV achieves a slightly higher F1 score value than SVGPMIX on SICAPv2 (F1=0.7137 ± 0.0119 vs. F1=0.7216 ± 0.0152), its Kappa value is much lower ($\kappa = 0.6748 \pm 0.0085$) compared to SVGPMIX ($\kappa = 0.7814 \pm 0.0083$). We observe a similar behavior in the SVGP-MV performance. We believe that this is due to the presence of noisy labels in the combined dataset. In contrast, SVGPMIX achieves a satisfying Kappa value on the SICAPv2 ($\kappa = 0.7814 \pm 0.0083$) and CrowdGleason ($\kappa = 0.7276 \pm 0.0260$) datasets, demonstrating its robustness.

To assess the statistical significance of our results, we apply the Almost Stochastic Order (ASO) test [45, 46] (implemented in the deep significance library[1]) on the five random runs of both SVGP-SICAP and the proposed SVGPMIX models. The test was performed on the F1 score metric since it takes into account the imbalanced scenario presented in this paper. The ASO test outputs a value between 0 and 1 indicating the degree of violation in stochastic order, where a value below 0.5 indicates that the SVGPMIX

[1] https://deep-significance.readthedocs.io/en/latest/

model performs statistically better than SVGP-SICAP. Using ASO with a confidence level $\alpha = 0.05$, we found the score distribution of SVGPMIX to be stochastically dominant over SVGP-SICAP ($\epsilon_{\min} = 0.0615$ in SICAPv2 and $\epsilon_{\min} = 0.0$ in CrowdGleason). In conclusion, the proposed CrowdGleason dataset outperforms the model trained on SVGP-SICAP.

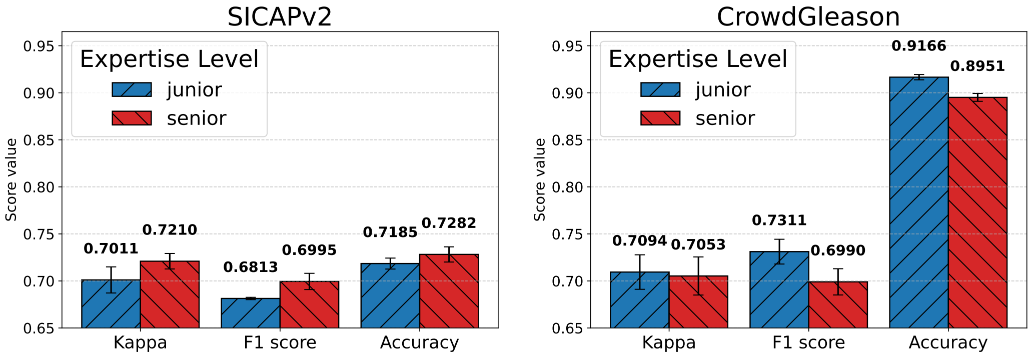
5.4. Ablation studies on the quality of the labels and the number of annotators

We have seen how well the models perform on the test set but have not analyzed in-depth the role of the non-expert annotators in the crowdsourcing model. In this subsection, we provide an ablation study on the crowdsourcing models, highlighting the impact of crowdsourced annotations into the final performance. First, we assess the effect of experience on pathologists in training by dividing them into two groups: junior (residents in their first or second year) and senior (third or fourth year residents). We train the SVGPCR model using the same configuration as in previous experiments, but in two different settings: (i) using only samples labeled by junior participants and (ii) using only samples labeled by senior participants. This experiment is conducted over five independent runs. The results are shown in Figure 4 which includes the mean performance and 0.95 CI. The performance is comparable across both datasets; however, the model trained with junior-labeled samples performs better on the CrowdGleason dataset, while the model trained with senior-labeled samples excels in the SICAPv2 dataset. Since the junior residents were specifically trained using the CrowdGleason dataset, the model trained with their annotations tends to overfit. In contrast, the senior participants, with greater experience, are able to recognize a broader range of patterns, allowing the model to generalize more effectively to the SICAPv2 dataset.

Secondly, in the ablation study on the number of annotators, we investigate how many annotators are sufficient to achieve a satisfactory crowdsourcing model. For this, we trained the SVGPCR model with varying numbers of annotators. For each number of annotators, we performed eight independent runs, randomly sampling different subsets of annotators. For each subset, we run the models five times to ensure stability and consistency. Figure 5 illustrates the results for the SICAPv2 and CrowdGleason datasets, showing the mean performance, the 95% CI, and the SVGPCR performance with all annotators. As we increase the number of annotators, the CI narrows, indicating that the models become more stable and less dependent on the specific annotators selected. On both test datasets, the performance of the models trained with subsets of annotators overlaps with that of the model trained with all annotators. This suggests that fewer annotators can achieve comparable performance. Overall, for this experiment, about five annotators appear sufficient to achieve satisfactory results, although increasing the number of annotators leads to a more stable performance since the results are highly influenced by the expertise of the selected annotators.

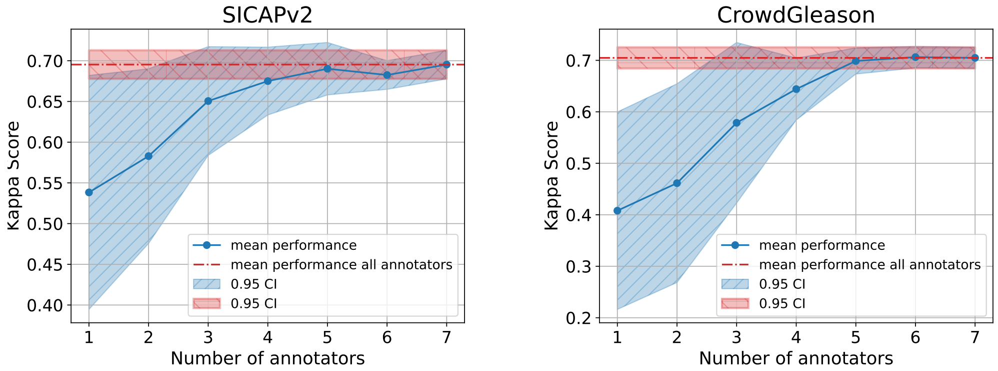
5.5. Ablation study on the impact of expert label datasets in the crowdsourcing scheme

This section analyzes the impact of the expert labels from SICAPv2 on the SVGPMIX model. We investigate how many expert-labeled samples are necessary to achieve a robust SVGPMIX model. For this, we trained the SVGPMIX model using CrowdGleason and different proportions of expert-labeled samples from SICAPv2. For each proportion, we performed eight in-



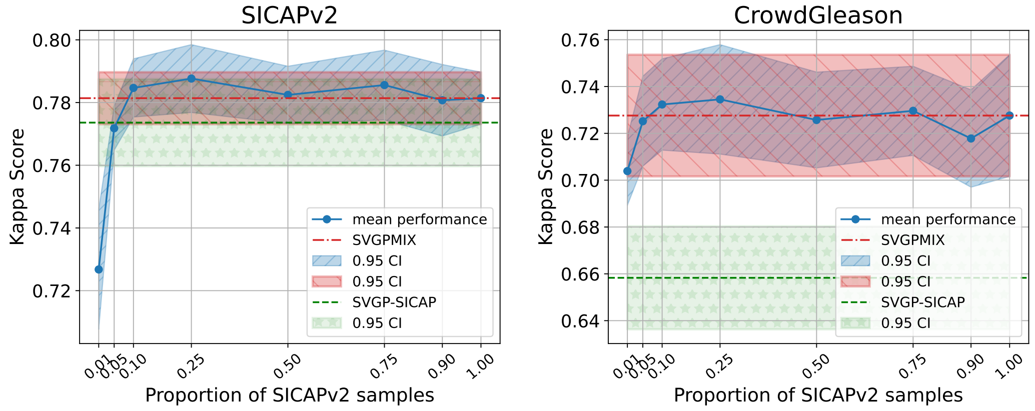
(a) (b)

Figure 4: Results of the SVGPCR model trained with labels provided by junior participants (first and second year) or senior participants (third and fourth year).



(a) (b)

Figure 5: Results of the SVGPCR model varying the number of annotators.



(a)

(b)

Figure 6: Results of the SVGPMIX model varying the proportion expert-labeled samples from SICAPv2.

dependent runs, randomly sampling different subsets of expert-labeled data. For each sampled dataset, we run the model five times to ensure stability and consistency. Figure 6 illustrates the results for each dataset, showing the mean performance, the 95% CI, and both SVGPMIX and SVGP-SICAP performance with all expert-labeled samples.

Unlike crowdsourced-labeled samples, increasing the number of expert-labeled samples does not significantly narrow the CI, as expert labels tend to have less variability and are inherently more robust. Notably, when 10% of the expert samples are used, the model stabilizes (and also surpasses the results from SVGP-SICAP in the SICAPv2 dataset), indicating that the samples are highly informative for training the crowdsourcing model. Beyond this point, adding more samples does not provide additional benefits, demonstrating that the model can perform effectively with a relatively small amount of expert data.

5.6. Analysis of annotator behavior

We measure the performance of each non-expert annotator by means of the Kappa score. The figures-of-merit, shown in Table 9, indicate the degree of agreement between each annotator and the curated test set. The best-performing annotator is A4 ($\kappa = 0.7765$), while A7 presents the lowest agreement ($\kappa = 0.0899$). The disparity of performance among annotators highlights the crowd heterogeneity and complexity of the task.

We further depict the per-class behavior of the annotators in Figure 7. The confusion matrices are normalized row-wise for better visualization and comparison purposes. These matrices can be understood as an estimation (on the test set) of the annotators' expertise. The crowdsourcing methods aim to estimate these confusion matrices from the noisy labeled training set. Recall that the ground-truth labels are not observed for these models. Figures 8 and 9 show the estimated confusion matrices estimated by SVGPCR and SVGPMIX, respectively. These matrices closely approximate the annotators' behavior, emphasizing the excellent performance of the crowdsourcing methods.

6. Discussion

Our experiments have shown (see Tables 6 and 7) that SVGP improves the performance of ResNet18 tested on SICAPv2 and is competitive or outperforms ResNet18 when tested on the new CrowdGleason. These results confirm the potential of GPs to perform GG classification. The SVGPCR classifier, used in the learning from crowds framework, achieved a value of $\kappa = 0.7048 \pm 0.0207$ and $\kappa = 0.6953 \pm 0.0176$ on CrowdGleason and SICAPv2

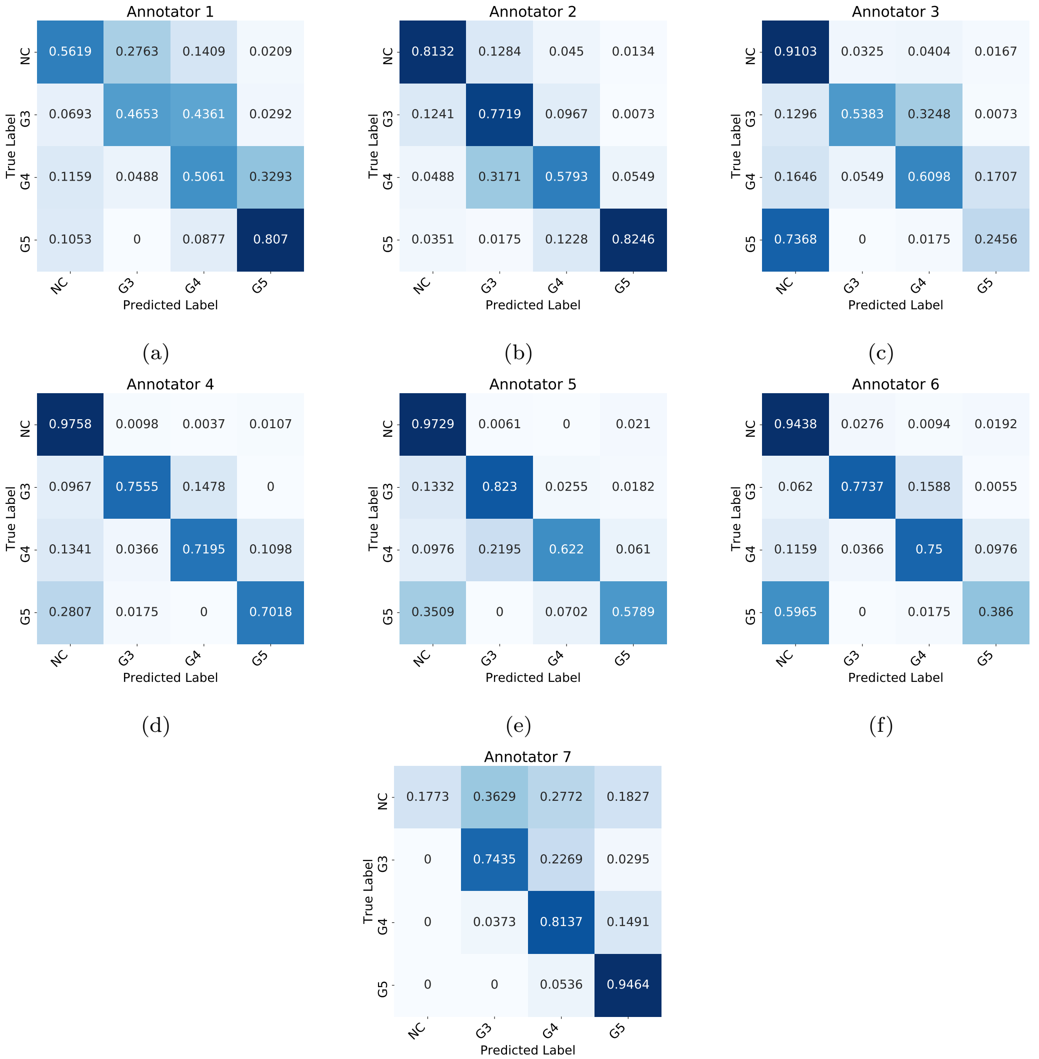
Table 9: Cohen's Quadratic Kappa (κ) coefficient of non-expert annotators on the CrowdGleason curated test set.

	A1	A2	A3	A4	A5	A6	A7
κ	0.4120	0.6283	0.5394	0.7765	0.7040	0.6520	0.0899

test sets, respectively (see Table 7), outperforming label aggregation strategies, such as MV, DS, MACE, and GLAD. The best label aggregation model (i.e., MV) obtains $\kappa = 0.6576 \pm 0.0086$ and $\kappa = 0.6242 \pm 0.0277$ (see Table 7) for CrowdGleason and SICAPv2 test sets, respectively. This significant difference highlights the enormous impact of noisy labels provided by non-expert annotators on the models performance and the need to use a suitable model to learn from crowds. Furthermore, the SVGPCR results are competitive with SVGP trained on SICAPv2 with expert labels that obtain $\kappa = 0.6583 \pm 0.0220$ and $\kappa = 0.7736 \pm 0.0139$ (see Table 6). Regarding the F1 metric, SVGPCR can even improve the performance of SVGP trained with expert labels on both test datasets. These results align with previous works in crowdsourcing ([12, 15, 16, 18]), and validate the use of the proposed CrowdGleason dataset for further studies on crowdsourcing and GG.

We have also explored the combination of the SICAPv2 dataset with our dataset. Recall that learning a model with samples from two different centers is difficult due to the heterogeneity between samples and labels. Additionally, noisy labels from non-expert annotators introduce noise into the dataset, which worsens the classifier performance. See, for instance, the decrease from $\kappa = 0.7736 \pm 0.0139$ of SVGP-SICAP in Table 6 to $\kappa = 0.6242 \pm 0.0277$ of SVGP-MV in Table 8. In this work, we propose using SVGPMIX to address this issue. SVGPMIX extends SVGPCR to the scenario where some labels are given by one expert and the rest are given by multiple non-experts. In this case, SVGPMIX improves the results of both SVGP-SICAP and SVGPCR on both datasets (see Table 8). Specifically, it achieves $\kappa = 0.7276 \pm 0.0260$ and $\kappa = 0.7814 \pm 0.0083$ on CrowdGleason and SICAPv2, respectively. Remarkably, SVGPMIX achieves stable results with only 10% percent of samples labeled by expert pathologists. The results obtained by both SVGPCR and SVGPMIX are within the range of results reported in the literature for GG classification. For example, Marrón-Esquivel *et al.* [42] reported $\kappa = 0.826$, Xiang *et al.* [47] reported $\kappa = 0.81$ and Arvaniti *et al.* [21] reported $\kappa = 0.49$ and $\kappa = 0.53$ for two different pathologists.

During the study, we have observed great variability between annotators that is even more accentuated when they have little experience in the area. Table 9 shows that the results obtained by non-experts in the CrowdGleason curated test set are very dissimilar ranging from $\kappa = 0.09$ to $\kappa = 0.78$. In general, we observe a lower mean agreement with the test set ($\kappa = 0.5432$) than that observed in other works involving only expert pathologists. For example, in [42] the authors reported $\kappa = 0.6946$ among expert pathologists, and in [21] two expert pathologists scored $\kappa = 0.71$. The annotators classified non-cancerous patches relatively well, but had more confusion between classes G3 and G4 (see Fig. 7). SVGPCR and SVGPMIX automatically estimate these confusion matrices from the noisy training data. The results in Figures 8 and 9 show that the estimated matrices capture the behavior of the noisy annotators. For instance, both models capture the higher sensitivity in the G4 and G5 grades of annotator 7. Furthermore, these models



(a) (b) (c)

(d) (e) (f)

(g)

Figure 7: Normalized confusion matrices of the seven annotators in the CrowdGleason curated test set.

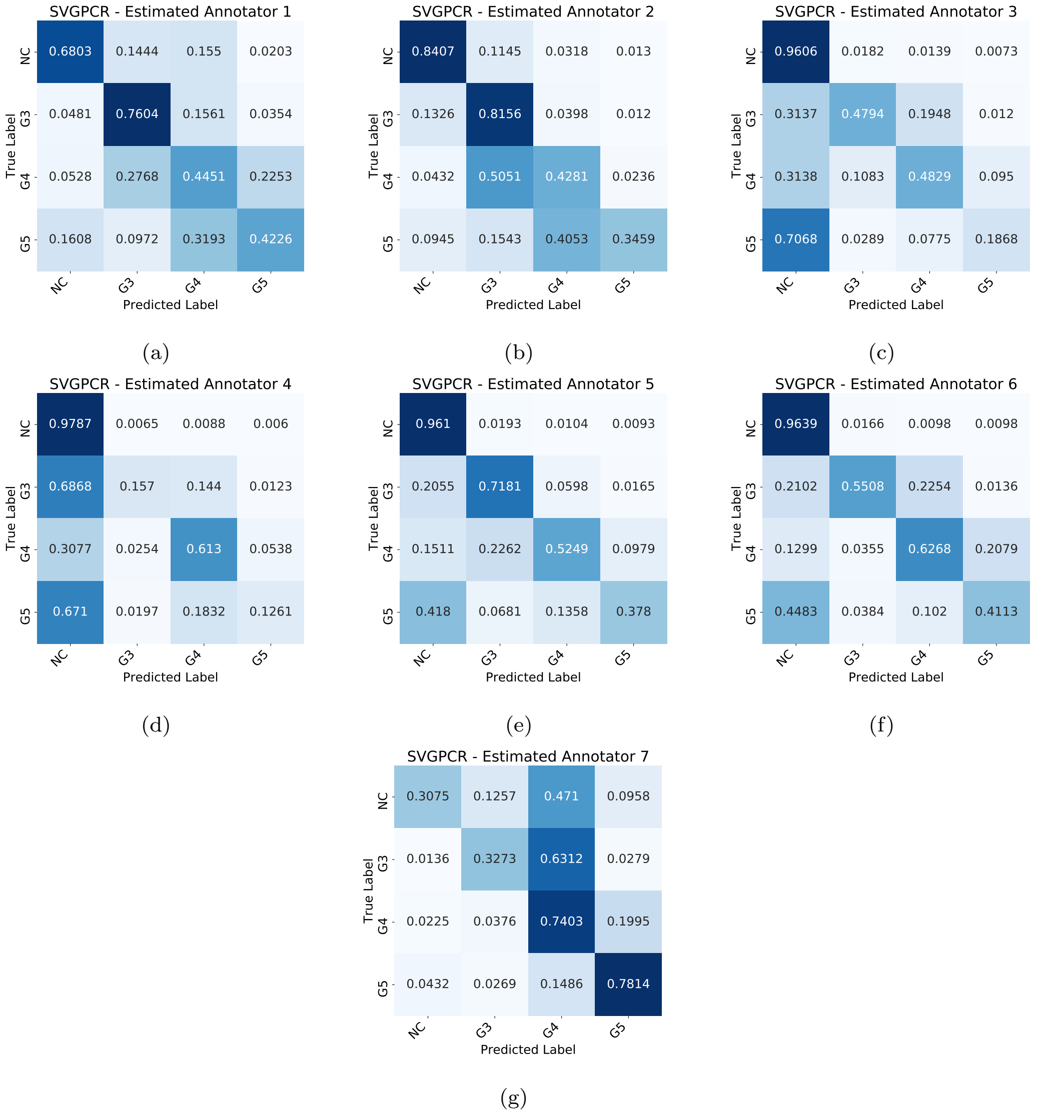
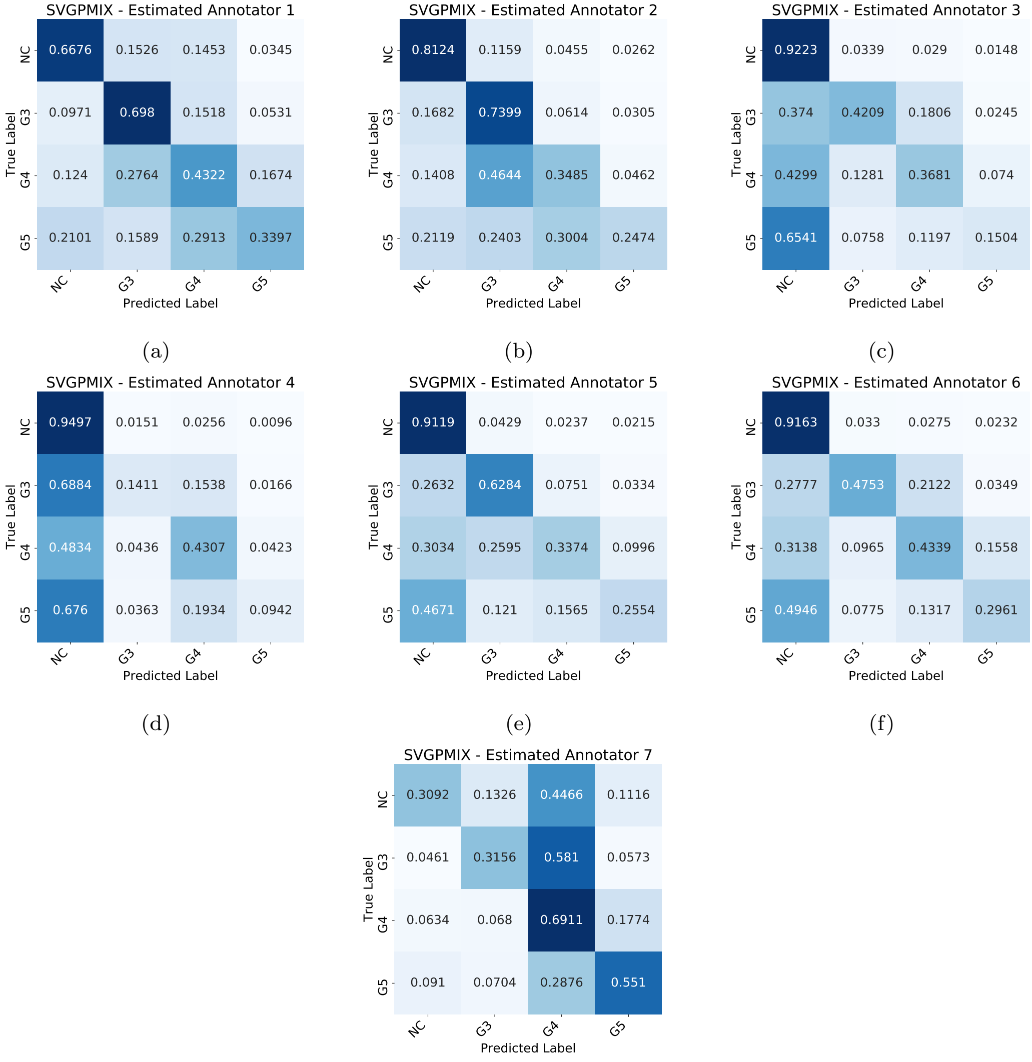


Figure 8: Estimated confusion matrices for the seven annotators by the SVGPCR model.



SVGPMIX - Estimated Annotator 1

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.6676	0.1526	0.1453	0.0345
G3	0.0971	0.698	0.1518	0.0531
G4	0.124	0.2764	0.4322	0.1674
G5	0.2101	0.1589	0.2913	0.3397

(a)

SVGPMIX - Estimated Annotator 2

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.8124	0.1159	0.0455	0.0262
G3	0.1682	0.7399	0.0614	0.0305
G4	0.1408	0.4644	0.3485	0.0462
G5	0.2119	0.2403	0.3004	0.2474

(b)

SVGPMIX - Estimated Annotator 3

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.9223	0.0339	0.029	0.0148
G3	0.374	0.4209	0.1806	0.0245
G4	0.4299	0.1281	0.3681	0.074
G5	0.6541	0.0758	0.1197	0.1504

(c)

SVGPMIX - Estimated Annotator 4

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.9497	0.0151	0.0256	0.0096
G3	0.6884	0.1411	0.1538	0.0166
G4	0.4834	0.0436	0.4307	0.0423
G5	0.676	0.0363	0.1934	0.0942

(d)

SVGPMIX - Estimated Annotator 5

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.9119	0.0429	0.0237	0.0215
G3	0.2632	0.6284	0.0751	0.0334
G4	0.3034	0.2595	0.3374	0.0996
G5	0.4671	0.121	0.1565	0.2554

(e)

SVGPMIX - Estimated Annotator 6

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.9163	0.033	0.0275	0.0232
G3	0.2777	0.4753	0.2122	0.0349
G4	0.3138	0.0965	0.4339	0.1558
G5	0.4946	0.0775	0.1317	0.2961

(f)

SVGPMIX - Estimated Annotator 7

True Label \ Predicted Label	NC	G3	G4	G5
NC	0.3092	0.1326	0.4466	0.1116
G3	0.0461	0.3156	0.581	0.0573
G4	0.0634	0.068	0.6911	0.1774
G5	0.091	0.0704	0.2876	0.551

(g)

Figure 9: Estimated confusion matrices for the seven annotators by the SVGPMIX model.

also capture the behavior of the annotators when labeling samples as 'NC'. The models correctly estimate that annotators 1 and 7 have the lowest sensitivity in this class (as seen in Fig. 7). Note also that SVGPCR achieves a better concordance (Kappa value) on the test set than most pathologists in-training, as seen in Tables 7 and 9. This means that SVGPCR outperforms each pathologist in-training individually. As an additional result, SVGPCR trained with less experienced annotators tends to overfit more to the CrowdGleason dataset, see Figure 4. This may be due to the lack of specific training in prostate cancer for less experienced participants.

Our crowdsourced dataset, CrowdGleason, serves as a valuable benchmark for training models that should be robust to label noise and high inter-observer variability. It can also be used to study types of errors and class confusions among non-expert pathologists. As demonstrated in our paper, this dataset can enhance previous models trained on expert-labeled data, leading to improved generalization. However, the main disadvantage of using our dataset is the high level of label noise; employing standard supervised models with these labels will likely result in poor performance. Therefore, it is crucial to use appropriate machine learning models designed to learn from crowds, as outlined in our study. These models can be further enhanced by incorporating feature-dependent information on the annotators, as they may be more prone to making errors when specific features are present. These include architectural characteristics such as gland size, arrangement in groupings and/or fusions, appearance of lumens, and loss of basal cells. Nuclear characteristics include nuclear size, staining intensity, size and number of nucleoli, or presence of mitosis. Cytoplasmic characteristics involve their shape, quantity, and staining. Luminal characteristics may include the presence of different materials.

7. Conclusions

In this work, we propose a novel crowdsourcing protocol to scale up the labeling of prostate histopathological images. As a result, we present the new CrowdGleason dataset labeled by seven pathologists in-training at the patch level. To the best of our knowledge, this is the most extensive dataset with patch-level annotations and the first with non-expert annotations for PCa. We conducted comprehensive experiments utilizing this new dataset and the previous SICAPv2, labeled by a PCa expert pathologist.

Despite the high disagreement between non-expert annotators, experiments show that crowdsourcing methods trained with the proposed CrowdGleason obtain competitive results against using expert labels on different test sets. Remarkably, the learning from crowds method performs better than most of the pathologists in-training on the test set. We have demonstrated that while results from five non-expert annotators are satisfactory, the performance becomes more stable as the number of annotators providing labels increases.

We have also proposed a method to augment SICAPv2 with the proposed CrowdGleason dataset and achieved better results than those obtained using only one dataset. Furthermore, we have shown that the combined model requires only 10% of expert-labeled samples to achieve a satisfactory performance. The combination of a small number of expert and non-expert labels paves the way for future large-scale labeling efforts by integrating both expert and non-expert pathologist annotators. CrowdGleason can be leveraged in future works for training or validating methods and augmenting existing datasets.

Although this work presents important findings and motivates the use of crowdsourcing to scale up the labeling of histopathological datasets, there are still very interesting open research questions. For example, how large the dataset has to be, how many samples have to be annotated by each annotator or how large and diverse the pool of participants has to be. Also, the presented methods for learning from crowds estimate a confusion matrix per annotator. However, it is not feature-dependent (i.e., architectural, nuclear, cytoplasmic, or luminal characteristics). A valuable future direction is to study how these features influence non-expert behavior in PCa diagnosis and use this information within the crowdsourcing model.

References

[1] P. Rawla, Epidemiology of prostate cancer, World Journal of Oncology 10 (2) (2019).

[2] J. L. Mohler, E. S. Antonarakis, A. J. Armstrong, A. V. D'Amico, B. J. Davis, T. Dorff, J. A. Eastham, C. A. Enke, T. A. Farrington, C. S. Higano, et al., Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology, Journal of the National Comprehensive Cancer Network 17 (5) (2019) 479–505.

[3] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, J. I. Epstein, Interobserver reproducibility of Gleason grading

of prostatic carcinoma: General pathologist, Human Pathology 32 (1) (2001) 81–88.

[4] P. A. Rodriguez-Urrego, A. M. Cronin, H. A. Al-Ahmadie, A. Gopalan, S. K. Tickoo, V. E. Reuter, S. W. Fine, Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies, Human Pathology 42 (1) (2011) 68–74.

[5] P. Ball, Is AI leading to a reproducibility crisis in science?, Nature (2023) 22–25.

[6] H. Su, J. Deng, L. Fei-Fei, Crowdsourcing annotations for visual object detection, in: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 40–46.

[7] P. Morales-Álvarez, P. Ruiz, S. Coughlin, R. Molina, A. K. Katsaggelos, Scalable variational Gaussian processes for crowdsourcing: Glitch detection in LIGO, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (3) (2022) 1534–1551.

[8] 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.

[9] 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.

[10] C. K. Williams, C. E. Rasmussen, Gaussian processes for machine learning, Vol. 2, MIT Press Cambridge, MA, 2006.

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

[12] F. Rodrigues, F. Pereira, B. Ribeiro, Gaussian process classification and active learning with multiple annotators, in: International conference on Machine Learning, PMLR, 2014, pp. 433–441.

[13] P. Morales-Álvarez, P. Ruiz, R. Santos-Rodríguez, R. Molina, A. K. Katsaggelos, Scalable and efficient learning from crowds with Gaussian processes, Information Fusion 52 (2019) 110–127.

[14] P. Ruiz, P. Morales-Álvarez, S. Coughlin, R. Molina, A. K. Katsaggelos, Probabilistic fusion of crowds and experts for the search of gravitational waves, Knowledge-Based Systems (2022) 110183.

[15] 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) 11612.

[16] 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.

[17] R. del Amor, J. Pérez-Cano, M. López-Pérez, L. Terradez, J. Aneiros-Fernandez, S. Morales, J. Mateos, R. Molina, V. Naranjo, Annotation protocol and crowdsourcing multiple instance learning classification of skin histological images: The CR-AI4SkIN dataset, Artificial Intelligence in Medicine 145 (2023) 102686.

[18] 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.

[19] A. Schmidt, P. Morales-Álvarez, R. Molina, Probabilistic attention based on Gaussian processes for deep multiple instance learning, IEEE Transactions on Neural Networks and Learning Systems (2023).

[20] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, V. Naranjo, Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection, Computer Methods and Programs in Biomedicine 195 (2020) 105637.

[21] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, M. Claassen, Au-

tomated Gleason grading of prostate cancer tissue microarrays via deep learning, Scientific Reports 8 (1) (2018) 12054.

[22] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, et al., Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge, Nature Medicine 28 (1) (2022) 154–163.

[23] 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.

[24] 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.

[25] 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.

[26] 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).

[27] D. Difallah, A. Checco, Aggregation techniques in crowdsourcing: Multiple choice questions and beyond, in: Proc. of the International Conference on Information & Knowledge Management, 2021, p. 4842–4844.

[28] 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.

[29] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, P. Ruvolo, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems, Vol. 22, 2009, p. 2035–2043.

[30] 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, 2013, pp. 1120–1130.

[31] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, Journal of Machine Learning Research 11 (4) (2010).

[32] M. Lopez-Perez, L. Garcia, C. Benitez, R. Molina, A contribution to deep learning approaches for automatic classification of volcano-seismic events: Deep Gaussian processes, IEEE Transactions on Geoscience and Remote Sensing 59 (5) (2021) 3875–3890.

[33] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[34] A. Schmidt, J. Silva-Rodriguez, R. Molina, V. Naranjo, Efficient cancer classification by coupling semi supervised and multiple instance learning, IEEE Access 10 (2022) 9763–9773.

[35] G. Litjens, H. Pinckaers, K. Kartasalo, M. Eklund, P. Ruusuvuori, P. Ström, S. Dane, W. Bulten, Prostate cANcer graDe Assessment (PANDA) challenge (2020).
URL https://kaggle.com/competitions/prostate-cancer-grade-assessment

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

[37] D. F. Gleason, G. T. Mellinger, null null, Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging, Journal of Urology 111 (1) (1974) 58–64.

[38] F. Pérez-Bueno, J. Serra, M. Vega, J. Mateos, R. Molina, A. K. Katsaggelos, Bayesian K-SVD for H&E blind color deconvolution. Applications to stain normalization, data augmentation, and cancer classification, Computerized Medical Imaging and Graphics 97 (2022) 102048.

[39] 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.

[40] J. Hensman, A. Matthews, Z. Ghahramani, Scalable variational Gaussian process classification, in: G. Lebanon, S. V. N. Vishwanathan (Eds.), Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, Vol. 38, 2015, pp. 351–360.

[41] D. Ustalov, N. Pavlichenko, B. Tseitlin, Learning from crowds with Crowd-Kit (2023). arXiv:2109.08584.
URL `https://arxiv.org/abs/2109.08584`

[42] J. M. Marrón-Esquivel, L. Duran-Lopez, A. Linares-Barranco, J. P. Dominguez-Morales, A comparative study of the inter-observer variability on Gleason grading against deep learning-based approaches for prostate cancer, Computers in Biology and Medicine 159 (2023) 106856.

[43] 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, et al., Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, The Lancet Oncology 21 (2) (2020) 222–232.

[44] J. Cohen, Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit, Psychological Bulletin 70 (4) (1968) 213.

[45] E. Del Barrio, J. A. Cuesta-Albertos, C. Matrán, An optimal transportation approach for assessing almost stochastic order, in: The Mathematics of the Uncertain, Springer, 2018, pp. 33–44.

[46] R. Dror, S. Shlomov, R. Reichart, Deep dominance - how to properly compare deep neural models, in: A. Korhonen, D. R. Traum, L. Màrquez

(Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 2773–2785.

[47] J. Xiang, X. Wang, X. Wang, J. Zhang, S. Yang, W. Yang, X. Han, Y. Liu, Automatic diagnosis and grading of prostate cancer with weakly supervised learning on whole slide images, Computers in Biology and Medicine 152 (2023) 106340.