IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 January 2023, accepted 16 January 2023, date of publication 18 January 2023, date of current version 24 January 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3237990*

RESEARCH ARTICLE

# Deep Gaussian Processes for Classification With Multiple Noisy Annotators. Application to Breast Cancer Tissue Classification

**MIGUEL LÓPEZ-PÉREZ[1], PABLO MORALES-ÁLVAREZ[2], LEE A. D. COOPER[3,4], RAFAEL MOLINA[1], (Life Senior Member, IEEE), AND AGGELOS K. KATSAGGELOS[4,5], (Life Fellow, IEEE)**

[1]Department of Computer Science and Artificial Intelligence, University of Granada, 18010 Granada, Spain
[2]Department of Statistics and Operations Research, University of Granada, 18010 Granada, Spain
[3]Department of Pathology, Northwestern University, Chicago, IL 60611, USA
[4]Center for Computational Imaging and Signal Analytics, Northwestern University, Chicago, IL 60611, USA
[5]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA

Corresponding author: Miguel López-Pérez (mlopez@decsai.ugr.es)

This work was supported in part by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 under Project PID2019-105142RB-C22, and in part by the Fondo Europeo de Desarrollo Regional (FEDER)/Junta de Andalucia-Consejeria de Transformacion Economica, Industria, Conocimiento y Universidades under Project P20_00286.

**ABSTRACT** Machine learning (ML) methods often require large volumes of labeled data to achieve meaningful performance. The expertise necessary for labeling data in medical applications like pathology presents a significant challenge in developing clinical-grade tools. Crowdsourcing approaches address this challenge by collecting labels from multiple annotators with varying degrees of expertise. In recent years, multiple methods have been adapted to learn from noisy crowdsourced labels. Among them, Gaussian Processes (GPs) have achieved excellent performance due to their ability to model uncertainty. Deep Gaussian Processes (DGPs) address the limitations of GPs using multiple layers to enable the learning of more complex representations. In this work, we develop Deep Gaussian Processes for Crowdsourcing (DGPCR) to model the crowdsourcing problem with DGPs for the first time. DGPCR models the (unknown) underlying true labels, and the behavior of each annotator is modeled with a confusion matrix among classes. We use end-to-end variational inference to estimate both DGPCR parameters and annotator biases. Using annotations from 25 pathologists and medical trainees, we show that DGPCR is competitive or superior to Scalable Gaussian Processes for Crowdsourcing (SVGPCR) and other state-of-the-art deep-learning crowdsourcing methods for breast cancer classification. Also, we observe that DGPCR with noisy labels obtains better results (F1 = 81.91%) than GPs (F1 = 81.57%) and deep learning methods (F1 = 80.88%) with true labels curated by experts. Finally, we show an improved estimation of annotators' behavior.

**INDEX TERMS** Crowdsourcing, deep Gaussian processes, digital pathology, breast cancer.

## I. INTRODUCTION

Machine learning (ML) classification algorithms have achieved very promising results in the field of digital pathology [1], [2], [3], [4]. These methods extract knowledge from data that has been previously labeled by an expert pathologist. However, modern ML models require a *large amount* of labeled data to perform well. Given the enormous workload of pathologists, the labeling process has become one of the most important bottlenecks in real practice [5], [6]. To address this issue, *crowdsourcing* has emerged as an alternative labeling approach in the last few years. The idea in crowdsourcing is to share the labeling effort among many different annotators who may not be experts and may have different degrees of expertise [7], [8], [9]. Currently, the use of crowdsourcing approaches in the medical field is a topic of significant

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

interest. A multitude of studies have used crowdsourced data to examine a wide range of problems [10], [11], [12].

Crowdsourced labels are inherently noisy, thus ML methods must be adapted to cope with this new scenario. A first approach would be to aggregate the crowdsourced labels to yield a set of noise-free labels (e.g. majority voting) and then use a standard ML classification method. However, as explained in [8], this approach usually performs worse than modeling the confusion of each annotator as part of the training process. In this work, we focus on the latter. Currently, the most successful crowdsourcing approaches are based on deep learning (DL) [13], [14] and Gaussian Processes (GPs) [7], [15], [16]. DL methods provide excellent predictive performance due to their hierarchical architecture that allows for learning complex features [17]. GPs are sound probabilistic methods that excel at uncertainty estimation, which is very valuable in the noisy crowdsourcing scenario [18], [19]. In the ML community, Deep Gaussian Processes (DGPs) represent a state-of-the-art method that leverages the strengths of both DL and GPs. The idea behind DGPs is to build a deep model by stacking various layers of GPs. Therefore, DGPs model flexible and complex functions like deep models while preserving the uncertainty estimation capability of GPs [20], [21].

In this work, we adapt DGPs to learn from crowdsourced labels and apply the new method to breast cancer detection. We call our method DGPCR (Deep Gaussian Processes for Crowdsourcing). To the best of our knowledge, DGPCR is the first extension of DGPs to the crowdsourcing setting in any area of application. DGPCR assumes that there exists an unknown ground truth label for each instance, which is modeled with a DGP. The crowdsourced labels are modeled from such ground truth through a per-annotator confusion matrix. These matrices codify the degree of expertise of each annotator for each class. DGP parameters and confusion matrices are estimated end-to-end by doubly stochastic variational inference [20]. Therefore, in addition to making predictions on previously unseen instances, DGPCR can estimate the reliability of each annotator as well as the ground truth labels.

To better understand the behavior of the novel DGPCR, we first conduct two controlled experiments. First, we use a fully synthetic 1D dataset, simulating data and annotators, to show the effectiveness of the method in a simple crowdsourcing problem. Then, we address a semi-synthetic problem using the well-known MNIST dataset, where we simulated only the crowdsourcing annotations. We consider five synthetic annotators with different known reliabilities, and we check that DGPCR is able to accurately estimate such reliabilities. We also show that DGPCR performance on the test set is superior to its shallow GP-based counterpart.

Then, we apply DGPCR to solve a real-world medical imaging problem. The data used here comes from an international study where pathology experts and non-experts annotated, following a crowdsourcing process, breast cancer tissue regions from the TCGA Breast Cancer cohort [7], [22]. TCGA is the well-known "Cancer Genome Atlas Program" [23]. In total, there are 161 rectangular regions of interest (ROI), from which 79607 patches were extracted. These patches are considered as training/testing instances, and features are obtained from them through a deep neural network. We will deal with a multiclass problem in which each patch belongs to one of three classes: tumor, stroma, and immune infiltrate.

The experimental results on this dataset show that DGPCR is competitive or superior to other state-of-the-art crowdsourcing methods based on both DL and GPs. We also show that, as theoretically expected, DGPCR performance is upper bounded by that of DGP-Gold (that is, a DGP trained with true expert labels), and it is lower bounded by that of DGP-MV (that is, a DGP trained with the naive majority voting of the crowdsourced labels). Moreover, DGPCR obtains slightly better results than DL and GPs trained with true expert labels. We also report enhanced estimation of annotator reliabilities (behavior), as well as good performance in the minority class across different training sizes. The reported statistics are illustrated through an insightful visualization of the predictions.

In summary, the main contributions of this paper are:

- We formulate Deep Gaussian Processes for Crowdsourcing, a novel method to integrate the benefits of DL and GPs in crowdsourcing. To the best of our knowledge, this is the first time that DGPs are extended to the crowdsourcing setting in any application domain.
- We illustrate the behavior of the method with controlled experiments. We use a fully synthetic experiment with a 1D dataset (where we simulate the data and the annotators) and a semi-synthetic one using MNIST (where we only simulate the annotators).
- We apply the new method to a real-world problem of histology breast cancer images annotated by medical students. We show promising results compared to state-of-the-art crowdsourcing methods. We also discuss the power of crowdsourcing labeling in medical imaging and future opportunities.

The rest of the paper is organized as follows. Section II reviews the most relevant related work. Section III explains the newly developed method, providing a general overview (Section III-A) as well as details on the probabilistic model (Section III-B) and variational inference (Section III-C). Section V contains the synthetic experiment with MNIST. Section VI analyzes the real-world problem involving breast cancer images. Finally, Section VII presents the main conclusions and some future outlooks.

## II. RELATED WORK

To set a richer context for DGPCR, this section reviews the most related approaches used for crowdsourcing. There are two contemporary approaches in the literature: i) combining the noisy labels and using a classification algorithm, and ii) utilizing the multiple labels during the learning process. The first one often involves using a weighted majority

vote, with some works using Decision Trees to estimate the annotators' weights [24] and others utilizing label propagation/augmentation techniques to incorporate information from similar instances [25], [26]. The second one treats ground-truth labels as latent variables and maps them to noisy annotations using a confusion matrix per annotator. These confusion matrices encode the annotators' expertise and biases. In this work, we focus on this line of action. In this context, we distinguish two kinds of approaches: non-probabilistic ones (mainly based on deep learning) and probabilistic ones (mainly based on Gaussian Processes).

### A. NON-PROBABILISTIC RELATED METHODS

Several crowdsourcing works have focused on how to adapt existing ML methods when multiple annotators label data. Raykar et al. [27] proposed a crowdsourcing classification method based on logistic regression. This method jointly learns the annotators' expertise and a latent classifier. Following an Expectation-Maximization (EM) scheme, they iteratively estimated the annotators' reliability and the classifier's coefficients. This method was applied to prostate cancer classification where there was a great amount of disagreement between expert pathologists [28]. However, this linear classifier can not achieve satisfying performance compared to other ML methods. To overcome this limitation, DL methods have been adapted to this crowdsourcing scenario. AggNet [13] considered a deep neural network (DNN) as the latent classifier, and a probabilistic noise model estimated the annotators' reliability. This method also used EM for the learning process. Lately, CrowdLayer [14] also included a DNN as the latent classifier. This time, they estimated the confusion matrix within the forward pass of the network to model the noisy observation. This characteristic enabled end-to-end training with stochastic gradient descent leading to better and faster convergence than EM.

### B. PROBABILISTIC GAUSSIAN PROCESSES RELATED METHODS

Recently, probabilistic Gaussian Processes reported great performance in several classification problems, including digital pathology, being competitive with DL-based methods [29]. These methods are usually more reliable than deterministic DL methods due to their probabilistic formulation. They are not likely to overfit and generalize well to unseen data. Also, they are well-calibrated. All these properties encouraged their adaptation to the crowdsourcing scenario, where they have achieved very competitive results. Namely, Variational Gaussian Processes for Crowdsourcing (VGPCR) addressed different tasks using variational inference with the mean-field approximation [15]. They showed clear superiority against deterministic crowdsourcing methods. However, the training was not end-to-end. They iteratively updated the coefficients of the GP and the confusion matrices. Also, this method was not scalable. Then, Morales-Álvarez et al. [19] proposed Scalable Gaussian Processes for Crowdsourcing (SVGPCR) overcoming the two main limitations of VGPCR. They performed stochastic variational inference enabling end-to-end learning using stochastic gradient descent and at the same time, they achieved scalability. They applied this method to glitch detection in gravitational waves with great results against various state-of-the-art methods in crowdsourcing. Recently, this method was extended to accommodate the situation in which a small number of expert labels is available concurrently with the labels from less experienced annotators generated by the crowdsourcing process [16]. In the medical imaging field, SVGPCR was applied to breast cancer classification with promising results compared to other related methods [7]. However, no probabilistic deep methods have been proposed for crowdsourcing problems, and this is the gap that our method intends to fill.

**TABLE 1. Summary of the most related methods for crowdsourcing classification.**

| | End-to-End | Deep | Probabilistic |
|---|---|---|---|
| Raykar [27] | ✗ | ✗ | ✗ |
| AggNet [13] | ✗ | ✓ | ✗ |
| Crowdlayer [14] | ✓ | ✓ | ✗ |
| VGPCR [15] | ✗ | ✗ | ✓ |
| SVGPCR [19] | ✓ | ✗ | ✓ |
| DGPCR (ours) | ✓ | ✓ | ✓ |

Table 1 summarizes the main properties of the algorithms reviewed here. In the experimental evaluation, we will compare against the most advanced methods in each family, i.e. the deep learning-based AggNet and CrowdLayer and the GP-based SVGPCR.

## III. METHODOLOGY

In this section, we introduce the proposed methodology (DGPCR). Section III-A provides a general overview, and Sections III-B and III-C introduce the details of the probabilistic model and the variational inference, respectively.

### A. OVERVIEW OF THE METHOD

Figure 1 shows the pipeline for DGPCR. The inputs are i) features extracted from a pretrained VGG16, and ii) crowdsourced labels provided by annotators with varying degrees of expertise. DGPCR learns a latent DGP classifier for the ground truth of the instances and a confusion matrix for each annotator. In addition to ground truth predictions, DGPCR can make predictions on the annotator's behavior by combining the latent classifier with the estimated confusion matrices. The confusion matrices are valuable on their own, as they estimate how good each annotator is and which classes they are prone to confuse. This can be further used to enhance the training provided to each annotator. DGPCR is trained end-to-end by maximizing the objective described in eq. (7). Our implementation leverages GPU acceleration through GPflow [30], more specifically GPflow 1.2.0, a library on top of Tensorflow dedicated to GPs. The code is publicly available on GitHub: https://github.com/wizmik12/DGPCR.

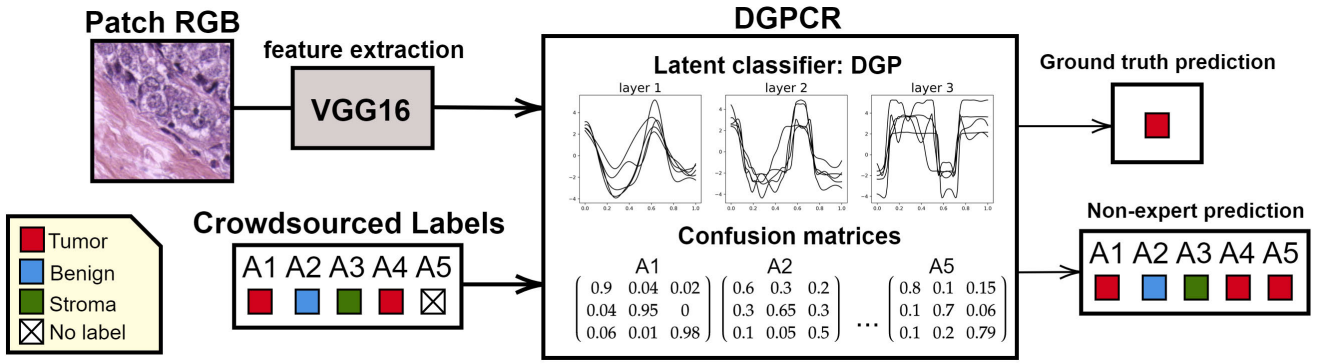



**FIGURE 1.** Pipeline for the proposed method. A1 through A5 represent five crowdsourcing (non-expert) annotators. The input data is given by features extracted from an RGB patch plus non-expert crowdsourced labels for such a patch. With this information, DGPCR estimates a latent DGP classifier and confusion matrices describing each annotator's behavior. In the test stage, DGPCR can predict the expert label for previously unseen patches, as well as the predictions that each annotator would give for the such patch.

### B. PROBABILISTIC MODEL

Let us assume a $K$-class crowdsourcing classification problem. There are $N$ training data points, and $A$ annotators label the instances. We denote the training set as $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})\} = \{(\mathbf{x}_n, \mathbf{y}_n^a) : \ n = 1, \ldots, N; \ a \in A_n\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ is a vector of features and $\mathbf{y}_n^a$ is the label provided by the $a$-th annotator for the $n$-th instance. $A_n \subseteq \{1, \ldots, A\}$ is the set of annotators who labelled the $n$-th instance. Notice that, in general, not all annotators label every instance. We represent the crowdsourced labels $\mathbf{y}_n^a$ with a one-hot encoded vector. That is, if the $a$-th annotator assigns the $k$-th class to the $n$-th instance, then $\mathbf{y}_n^a = \mathbf{e}_k$, a $K$-dimensional vector with all zeros except for the $k$-th position, where there is a one. Figure 2 depicts the probabilistic graphical model for DGPCR, which we describe next.

#### 1) INTRODUCING THE CONFUSION MATRICES

Inspired by [19], [27], and [31], we assume an (unknown) true label $\mathbf{z}_n \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ for each instance. Then, the crowdsourced labels depend on this true label and on the degree of expertise of each annotator. We model the expertise of each annotator $a$ with a confusion matrix $\mathbf{R}^a = (r_{ij}^a)_{1 \leq i,j \leq K}$. Each element $r_{ij}^a \in [0, 1]$ represents the probability that the $a$-th annotator labels as class $i$ an instance whose real class is $j$. We also assume that every annotator labels the different instances independently. Mathematically, this is given by

$$\mathrm{p}(\mathbf{Y}|\mathbf{Z}, \mathbf{R}) = \prod_n \prod_{a \in A_n} (\mathbf{y}_n^a)^\intercal \mathbf{R}^a \mathbf{z}_n, \tag{1}$$

where we write $\mathbf{Z}$ for all the $\mathbf{z}_n$'s and $\mathbf{R}$ for all the confusion matrices $\mathbf{R}^a$, $a = 1, \ldots, A$. We use prior (independent) Dirichlet distributions for the behavior of annotators, i.e.

$$\mathrm{p}(\mathbf{R}) = \prod_{a=1}^{A} \prod_{j=1}^{K} \mathrm{Dir}(\mathbf{r}_j^a | \alpha_{1j}^a, \ldots, \alpha_{Kj}^a). \tag{2}$$

This distribution is conjugate to the categorical one in eq. (1), which eases subsequent computations. Also, such Dirichlet prior can be used to incorporate prior knowledge that may be available for the annotator's behavior. In the default case where there is no prior knowledge, which we will assume in our experiments, we can set $\alpha_{ij}^a = 1$ and we obtain a uniform prior distribution.

#### 2) MODELING THE UNDERLYING TRUTH WITH A DGP

The true underlying labels $\mathbf{Z}$ are modeled from the input $\mathbf{X}$ with a DGP of $L$ layers [20]. For this, we introduce latent variables $\{\mathbf{F}^l\}_{l=1}^L$, where each $\mathbf{F}^l$ follows a GP prior independently across dimensions, with input locations given by the outputs of the previous layer $l-1$. We write $f_{n,d}^l$ for the latent variable of the $n$-th instance in the $d$-th dimension of the $l$-th layer (each layer has $D^l$ units, $d = 1, \ldots, D^l$). Since the last layer defines the output and we are considering $K$ classes, we have $D^L = K$.

The true label $\mathbf{z}_n$ is defined from the last layer of the DGP $\mathbf{f}_{n,:}^L$ with a multinomial distribution $\mathrm{p}(\mathbf{z}_n|\mathbf{f}_{n,:}^L)$ that depends on the chosen likelihood. In this paper, we use the popular softmax likelihood. Because of the computational cost of vanilla DGPs, which is $\mathcal{O}(N^3)$, we resort to the well-known sparse model [20], [32]. In brief, this approximation introduces $M^{l-1} \ll N$ inducing locations $\tilde{\mathbf{X}}^{l-1}$ at each layer $l$ with inducing values $\mathbf{U}^l$. These values are realizations from the same GP as $\mathbf{F}^l$ and summarize the information contained in the $N$ training points at the $l$-th layer. Mathematically, this is given by the probability distribution

$$\mathrm{p}(\mathbf{Z}, \{\mathbf{F}, \mathbf{U}^l\}_{l=1}^L) = \underbrace{\prod_n \mathrm{p}(\mathbf{z}_n|\mathbf{f}_{n,:}^L)}_{\text{likelihood}} \times \underbrace{\prod_l \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l; \mathbf{F}^{l-1}, \tilde{\mathbf{X}}^{l-1}) \mathrm{p}(\mathbf{U}^l; \tilde{\mathbf{X}}^{l-1})}_{\text{DGP prior}}, \tag{3}$$

where the semicolon notation indicates the inputs of each function. Notice also that we are writing $\mathbf{F}^0$ for the input $\mathbf{X}$.

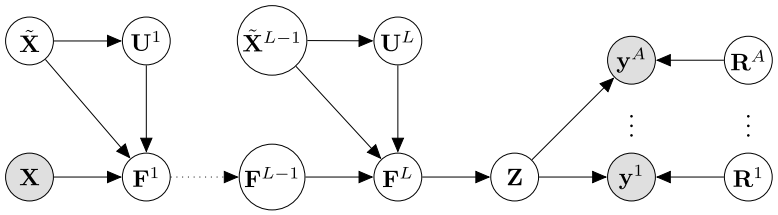



**FIGURE 2.** Probabilistic graphical model for an $L$-layer DGPCR. Dark (resp. light) circles are used for observed (resp. latent) variables.

#### 3) SUMMARIZING THE MODEL

In total, the joint probabilistic model for DGPCR is given by

$$\begin{aligned} &\mathrm{p}(\mathbf{Y}, \mathbf{Z}, \mathbf{R}, \{\mathbf{F}^l\}_{l=1}^L, \{\mathbf{U}^l\}_{l=1}^L) \\ &\quad = \mathrm{p}(\mathbf{Y}|\mathbf{Z}, \mathbf{R}) \cdot \mathrm{p}(\mathbf{R}) \cdot \mathrm{p}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L), \end{aligned} \tag{4}$$

where $\mathrm{p}(\mathbf{Y}|\mathbf{Z}, \mathbf{R})$, $\mathrm{p}(\mathbf{R})$ and $\mathrm{p}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_l)$ are given by eqs. (1), (2) and (3), respectively. As mentioned before, Figure 2 shows the probabilistic graphical model for DGPCR.

### C. VARIATIONAL INFERENCE

#### 1) MOTIVATION FOR VARIATIONAL INFERENCE

To obtain the posterior distribution over the latent parameters, we need to integrate out $\mathbf{Z}$, $\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L$ and $\mathbf{R}$ in eq. (4). Since this is analytically intractable, we resort to doubly stochastic variational inference to approximate the computations [20]. The idea is to propose a parametric posterior distribution $\mathrm{q}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L, \mathbf{R})$ to approximate the true posterior $\mathrm{p}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L, \mathbf{R}|\mathbf{Y})$. To optimize the parameters of the parametric posterior, the Kullback-Leibler (KL) divergence with respect to the true posterior is minimized. The KL divergence is a metric that quantifies how different two distributions are, it is always non-negative, and vanishes if and only if both distributions coincide, see e.g. [33].

#### 2) THE PROPOSED POSTERIOR DISTRIBUTION

Here we propose the following factorization for the posterior distribution:

$$\begin{aligned} \mathrm{q}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L, \mathbf{R}) = \mathrm{q}(\mathbf{Z})\mathrm{q}(\{\mathbf{F}^l\}_{l=1}^L|\{\mathbf{U}^l\}_{l=1}^L) \\ \times \mathrm{q}(\{\mathbf{U}^l\}_{l=1}^L)\mathrm{q}(\mathbf{R}). \end{aligned} \tag{5}$$

The details for each factor are as follows. The distribution on the true labels $\mathbf{Z}$ is given by categorical distributions, $\mathrm{q}(\mathbf{Z}) = \prod_{n=1}^N \mathbf{z}_n^\top \mathbf{q}_n$. The probability for each instance, $\mathbf{q}_n$, is a variational parameter to be estimated. Namely, $\mathbf{q}_n$ is a $K$-dimensional vector containing the probabilities that the $n$-th instance belongs to each one of the $K$ classes (in particular, all the values in $\mathbf{q}_n$ add up to one). The distribution $\mathrm{q}(\{\mathbf{F}^l\}_{l=1}^L|\{\mathbf{U}^l\}_{l=1}^L)$ is considered to be equal to the prior $\mathrm{p}(\{\mathbf{F}^l\}_{l=1}^L|\{\mathbf{U}^l\}_{l=1}^L)$. As discussed in previous work [20], [21], this ultimately allows for efficient mini-batch training. For the inducing point distribution, $\mathrm{q}(\{\mathbf{U}^l\}_{l=1}^L)$ is a multivariate Gaussian distribution where we have to estimate the mean vectors $\mathbf{m}_d^l$ and the covariance matrices $\mathbf{S}_d^l$ for each unit $d$ in each layer $l$. Finally, for the confusion matrix distribution, we assume posterior Dirichlet distributions

$$\mathrm{q}(\mathbf{R}) = \prod_{a=1}^A \prod_{j=1}^K \mathrm{Dir}(\mathbf{r}_j^a|\tilde{\alpha}_{1j}^a, \ldots, \tilde{\alpha}_{Kj}^a). \tag{6}$$

All the variational parameters which $\mathrm{q}(\mathbf{Z}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L, \mathbf{R})$ depends on are collectively denoted by $\mathbf{V}$, i.e., $\mathbf{V} = \{\mathbf{q}_n, \mathbf{m}_d^l, \mathbf{S}_d^l, \tilde{\boldsymbol{\alpha}}_j^a\}$.

#### 3) THE RESULTING ELBO AND TRAINING PROCEDURE

Following the variational inference procedure, minimizing the KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO) [33]. In our case, the ELBO is given by:

$$\begin{aligned} \mathrm{ELBO} = &\sum_{n,k} \sum_{a \in A_n} q_{nk} \mathbb{E}_{\mathrm{q}(\mathbf{r}_k^a)} \left[\log \mathrm{p}(\mathbf{y}_n^a|\mathbf{e}_k, \mathbf{r}_k^a)\right] \\ &+ \sum_{n,k} q_{nk} \mathbb{E}_{\mathrm{q}(\mathbf{f}_{n,:}^L)} \left[\log \mathrm{p}(\mathbf{e}_k|\mathbf{f}_{n,:}^L)\right] - \sum_{n,k} q_{nk} \log q_{nk} \\ &- \sum_l \mathrm{KL}(\mathrm{q}(\mathbf{U}^l)||\mathrm{p}(\mathbf{U}^l)) - \sum_{a,k} \mathrm{KL}(\mathrm{q}(\mathbf{r}_k^a)||\mathrm{p}(\mathbf{r}_k^a)), \end{aligned} \tag{7}$$

The ELBO is composed of five interpretable terms. The first term encodes fidelity to the noisy observed data. The second term ensures that the DGP predicts well the distribution of the latent ground-truth labels. The third term imposes informativeness on the distribution of the ground-truth labels. And the last two terms encode fidelity to the prior distributions on the DGP and the confusion matrices, respectively. Due to the chosen posterior distribution, all these terms (except the second one) can be computed in closed form. For the expectation in the second one, we leverage Monte Carlo samples. The ELBO is maximized w.r.t. the variational parameters $\mathbf{V}$, the inducing point locations $\tilde{\mathbf{X}}$, and the DGP kernel hyperparameters, which will be denoted $\boldsymbol{\Theta}$. For clarity, the training process is summarized in Algorithm 1.

#### 4) MAKING PREDICTIONS

Finally, for a previously unseen $\mathbf{x}^*$, we can predict both its true label and the label that each annotator would assign to it (recall Figure 1). For the former, we must propagate $\mathbf{x}^*$ through the DGP with the estimated parameters. Specifically, we have that the prediction on the last layer is a mixture of

**Algorithm 1** DGPCR Training Procedure

**Input:** Instances $\mathbf{X} = \{\mathbf{x}_n : n = 1, \ldots, N\}$ (e.g. extracted features from image patches); crowdsourcing labels $\mathbf{Y} = \{\mathbf{y}_n^a : n = 1, \ldots, N; a \in A_n\}$, number of iterations *Iter*.

**Output:** Variational parameters $\mathbf{V} = \{\mathbf{q}_n, \mathbf{m}_d^l, \mathbf{S}_d^l, \tilde{\boldsymbol{\alpha}}_j^a\}$, inducing point locations $\tilde{\mathbf{X}}$, DGP kernel hyperparameters $\boldsymbol{\Theta}$.

Initialize $\mathbf{q}_n$ and $\tilde{\boldsymbol{\alpha}}_j^a$ according to the frequencies in training data.

Initialize $\mathbf{m}_d^l$ and $\mathbf{S}_d^l$ with the corresponding values of the prior $p(\mathbf{u}_d^l)$.

Initialize $\tilde{\mathbf{X}}$ with K-means clustering.

Initialize all the variances and lengthscales in $\boldsymbol{\Theta}$ to two.

**for** $i = 1$ to *Iter* **do**

Calculate ELBO in eq. (7).

Update $\mathbf{V}$, $\tilde{\mathbf{X}}$ and $\boldsymbol{\Theta}$ using Adam optimizer.

**end for**

**return** Optimal model parameters $\mathbf{V}$, $\tilde{\mathbf{X}}$ and $\boldsymbol{\Theta}$.

Gaussians:

$$q(\mathbf{f}_*^L) \approx \frac{1}{S}\sum_{s=1}^{S} q(\mathbf{f}_*^L|\mathbf{m}^L, \mathbf{S}^L; \mathbf{f}_*^{(s)^{L-1}}, \tilde{\mathbf{X}}^{L-1}), \tag{8}$$

where we use $S$ samples from the posterior. For the latter, we combine the predicted true label with the estimated confusion matrices.

## IV. A FULLY SYNTHETIC EXAMPLE ON 1D DATASET

This section presents a fully synthetic example. The goal is to show that DGPCR is able to predict the annotators' expertise, leading to high predictive performance. In next sections, we will use more complex datasets and a wide range of baselines. We first describe the experimental framework in Section IV-A, and then the obtained results in Section IV-B.

### A. EXPERIMENTAL FRAMEWORK

#### 1) DATA DESCRIPTION

This experiment uses a 1D synthetic dataset for binary classification. A cosine function produces the labels: the label is 1 where the cosine function is positive and it is 0 where the cosine is negative. We sample 200 points for training and 100 for test uniformly distributed in the interval (−4,4). Figure 3 illustrates the data.

#### 2) DESCRIPTION OF THE ANNOTATORS

We simulate five synthetic annotators with behaviors that one can find in real-world problems. Figure 4 shows the labels provided each annotator. We define them by their specificity and sensitivity, which are assumed to be the same. Therefore, notice that the three first annotators can be considered as "experts" with varying degrees of precision (0.95, 0.9, and 0.6). The fourth annotator is a "spammer", as they label randomly regardless of the true label (probability of 0.5 for each class). Notice that the third "expert" is just slightly better than the spammer. The behavior of the last annotator is known as "adversarial" since they have learned a wrong concept. Namely, in this case, they swap both classes with a probability. That is, its specificity and sensitivity are equal to 0.1. Notice that, whereas no knowledge can be extracted from a "spammer" annotator, whose labels are pure noise, very valuable knowledge can be obtained from an "adversarial" one, *as long as* its confusion matrix is correctly identified. This is because annotator 5 produces annotations with systematic errors, in contrast to the random labels of annotator 4.

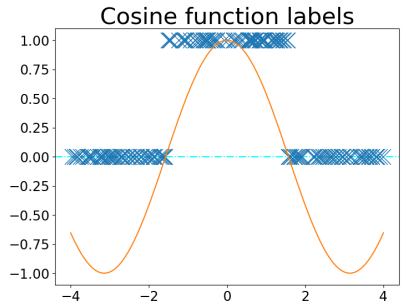



**FIGURE 3.** 1D dataset with labels produced by a cosine function for the fully synthetic experiment.

**TABLE 2.** Real and DGPCR average prediction and standard deviation of the specificity and sensitivity for the five simulated annotators in the fully synthetic experiment through 10 runs.

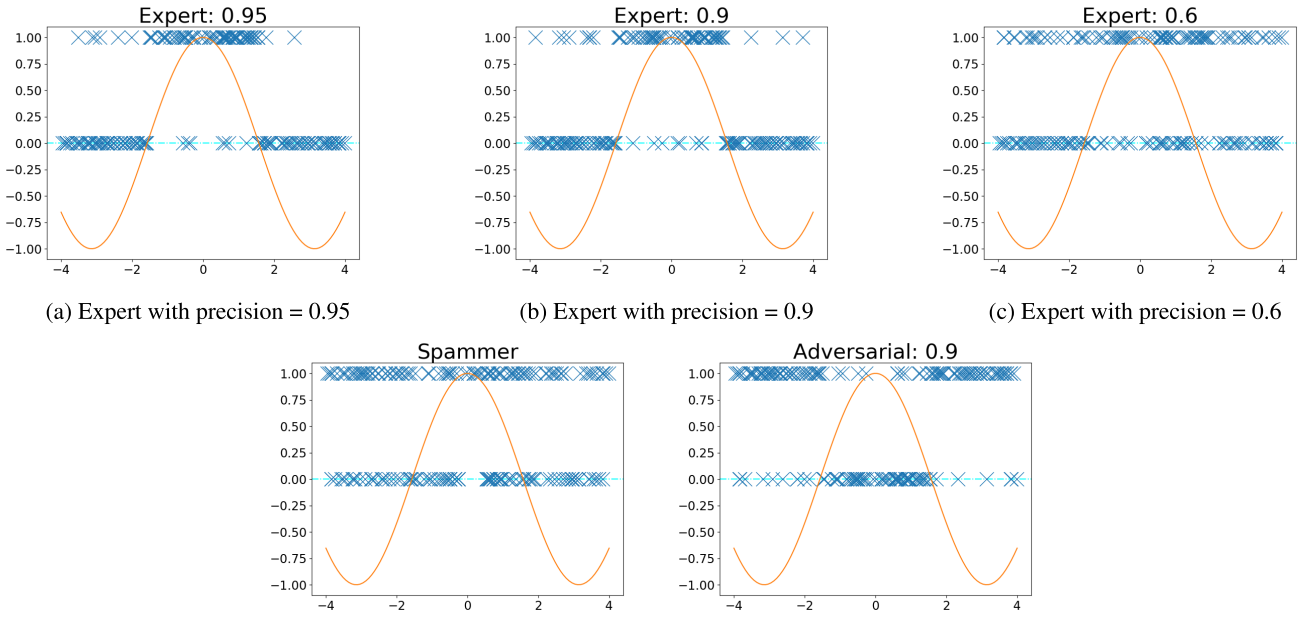
| | Real | | DGPCR prediction | |
|---|---|---|---|---|
| | Specificity | Sensitivity | Specificity | Sensitivity |
| Annotator 1 | 0.95 | 0.95 | 0.89±0.01 | 0.88±0.01 |
| Annotator 2 | 0.9 | 0.9 | 0.86±0.017 | 0.84±0.02 |
| Annotator 3 | 0.6 | 0.6 | 0.58±0.02 | 0.57±0.02 |
| Annotator 4 | 0.5 | 0.5 | 0.50±0.03 | 0.49±0.03 |
| Annotator 5 | 0.1 | 0.1 | 0.15±0.02 | 0.19±0.02 |

#### 3) EXPERIMENTAL DETAILS

We design a simple DGPCR of 2 layers. We use M = 64 inducing points and a batch size of 128. The ELBO is optimized using Adam and a learning rate of $10^{-2}$. We optimize the GP methods through 2,000 iterations. We use a Squared Exponential (SE) kernel. When predicting, we propagate S = 100 samples. We trained the method in the CPU. We repeat the experiment 10 times, sampling a different synthetic dataset each time.

### B. RESULTS AND DISCUSSION

The DGPCR method achieves an accuracy of 99% ± 0.66 and a log-loss of 0.0186 ± 0.0204. It is capable of classifying well the test set through the 10 runs. Furthermore, the value of the log-loss reveals that the predicted scores are well-separated for both classes. For a better understanding of the crowdsourcing scenario, we reported the predicted reliability of the simulated annotators in Table 2. These results suggest that the method is capable of estimating the three different kinds of annotators and provides an accurate prediction. Notice that the estimation is not exact. This fact is due to the prior distribution of the annotators' reliability, which acts as a regularizer. This characteristic that arises from the




(a) Expert with precision = 0.95 (b) Expert with precision = 0.9 (c) Expert with precision = 0.6

(d) Spammer (e) Adversarial

**FIGURE 4.** Noisy labels provided by five different simulated annotators for the fully synthetic classification problem.

bayesian framework is of vital importance in real problems with limited data. This experiment confirms the effectiveness of our proposed method on a fully synthetic problem. We will further validate our method in more complex scenarios in the following sections.

## V. AN ILLUSTRATIVE SEMI-SYNTHETIC EXAMPLE ON MNIST

This section focuses on a controlled experiment where we can simulate the behavior of the crowdsourcing annotators, and then check that DGPCR is able to accurately estimate it. We first describe the experimental framework in Section V-A, and then the obtained results in Section V-B.

### A. EXPERIMENTAL FRAMEWORK

#### 1) DATA DESCRIPTION

This experiment uses the well-known MNIST database, where the goal is to classify hand-written digits into ten different classes (from 0 to 9).

#### 2) DESCRIPTION OF THE ANNOTATORS

Following the previous experiment, we simulate five synthetic annotators with different paradigmatic behaviors that one can find in real-world problems. The first row in Figure 5 shows the confusion matrices for each annotator. Recall that the element $(i, j)$ of the matrix represents the probability that the annotator labels as class $i$ an instance whose real class is $j$. Therefore, notice that the three first annotators can be considered as "experts" with varying degrees of precision (0.95, 0.9, and 0.5). The fourth annotator is a "spammer" one, as they label randomly regardless of the true label (probability of 0.1 for each class, recall that MNIST has 10 classes).

**TABLE 3.** Performance of different GP and deep GP crowdsourcing methods on the test set for MNIST.

| | Accuracy | Log loss | CM error |
|---|---|---|---|
| SVGPCR | 96.99 | 0.1075 | 0.08762 |
| DGPCR2 | 97.82 | 0.0784 | 0.08747 |
| DGPCR3 | **98.02** | **0.0712** | **0.08745** |

The last annotator is "adversarial" since they have learned a wrong concept. Namely, in this case, they are confidently classifying the digit 0 as a 5, the 1 as a 6, etc.

#### 3) EXPERIMENTAL DETAILS

In the experimental validation, we try DGPCR with 2 and 3 layers (they will be called DGPCR2 and DGPCR3, respectively). We use $M = 100$ inducing points and a batch size of 1000. The ELBO is optimized using Adam and a learning rate of $10^{-2}$. We optimize the GP methods through 20,000 iterations. The dimensionality of the latent space is 30, and we leverage a SE kernel. When predicting, we propagate $S = 100$ samples. We trained the methods in an NVIDIA TITAN X (Pascal) GPU device with 12 Gb memory.

### B. RESULTS AND DISCUSSION

#### 1) DGPCR IS ABLE TO IDENTIFY THE BEHAVIOR OF THE ANNOTATORS

Figure 5 shows the estimations provided by DGPCR2 and DGPCR3 for the annotator's behavior (confusion matrices). More specifically, the depicted values for DGPCR are the expectations of the posterior Dirichlet distributions $q(\mathbf{R}^a)$. We observe a very accurate prediction for all types of annotators. In particular, identifying the spammer annotator allows DGPCR to discard the information provided by them, which

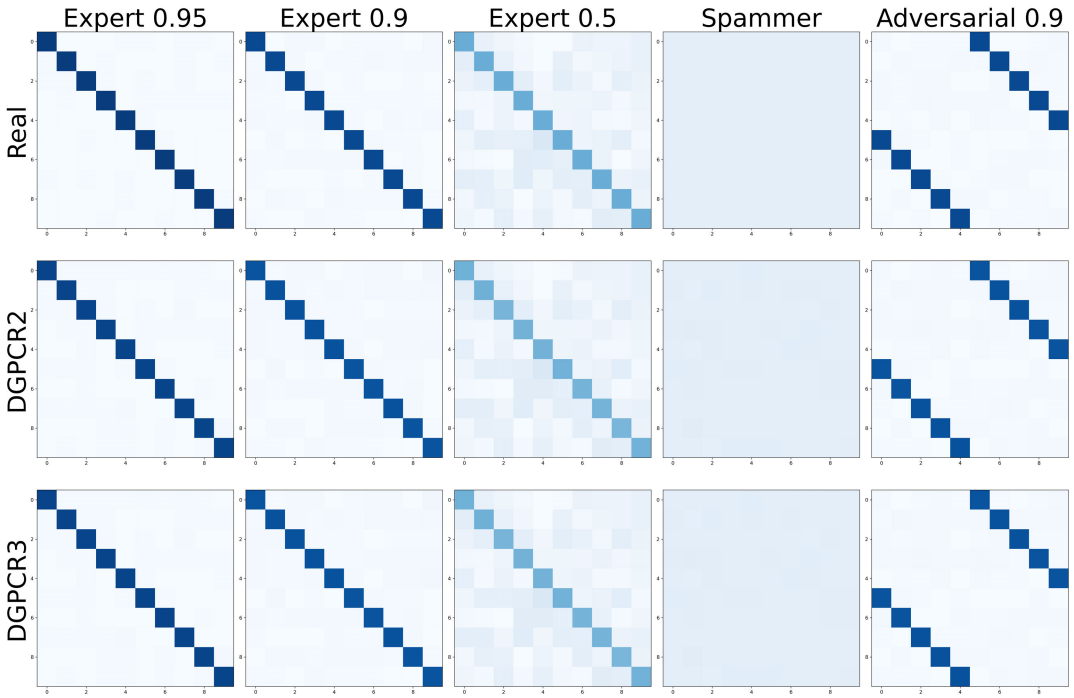



**FIGURE 5.** Real and predicted confusion matrices for the toy example with MNIST. This problem groups 5 types of annotators. The three firsts are experts with varying reliability: 0.95, 0.9 and 0.5. The fourth is a spammer which randomly annotates any class and the last one is an adversarial one.

is pure noise. Likewise, identifying the adversarial nature of the fifth annotator allows DGPCR to extract knowledge from it. In total, thanks to the accurate prediction of annotators' biases, DGPCR has access to valuable information from the noisy labels, as demonstrated next.

#### 2) DGPCR REACHES A HIGH PREDICTIVE PERFORMANCE ON THE TEST SET

In spite of being trained with noisy labels, Table 3 shows a high test performance for DGPCR. Namely, its predictions are correct 97.82% of times for $L = 2$ layers, and 98.02% of times for $L = 3$. The log-loss is also interestingly low compared to the value obtained by SVGPCR (the shallow counterpart of DGPCR, based on plain GPs instead of DGPs, which is included here as a baseline). Notice that the log-loss is the average negative log-likelihood for the test data (the lower the better). It takes into account the uncertainty of the predictions (unlike the accuracy, which only accounts for the mean of the prediction). For completeness and to numerically support the findings in Figure 5, the last column of Table 3 shows low values for the confusion matrix (CM) error. This error is the mean absolute error between the estimated CM and the true CM for all the annotators.

Finally, it is important to stress that DGPCR has no information about the ground-truth label or the annotators' expertise. It automatically estimates the confusion matrices and learns the latent DGP to make new predictions. In the following subsection, we will see how this method can be applied to a real-world problem of histology breast cancer classification.

## VI. CLASSIFICATION OF HISTOLOGICAL BREAST CANCER IMAGES

This section is devoted to the real-world application of our method to histological breast cancer images. Specifically, Section VI-A introduces the experimental framework. Then, Section VI-B presents and discusses the main results.

### A. EXPERIMENTAL FRAMEWORK

#### 1) DATA DESCRIPTION

We evaluate our methodology on a histopathology image dataset collected from the TCGA Breast Cancer cohort [22]. It contains 161 ROIs from 151 different WSIs (Whole Slide Images) collected in 18 institutes. It was originated from an international study where 2 senior pathologists provided expert labels and 20 medical students, which were non-pathologists, provided crowdsourced annotations. The interested reader is referred to [7] for the full details on the annotation protocol. The images have color variations which may downgrade the performance of systems [34]. Thus, we apply color normalization [35] to minimize differences among institutes and crop the WSIs in patches of $224 \times 224$ size. We divide the dataset into train and test sets. The test set contains the images annotated by the senior pathologists, whose label is considered the ground truth. The train set contains the crowdsourcing labels provided by the students (a total amount of 108495 crowdsourcing labels are available). In total, we obtain 75243 patches for the train and 4364 for the test. These are patches from three different classes: tumor (train: 37260, test: 2692), stroma (train: 27668, test: 1196) and immune infiltrate (train: 10315, test: 476).

**TABLE 4.** Performance of different state-of-the-art crowdsourcing methods on the test set for breast histology cancer images. The per-class results are given in terms of the F1-Score. The confusion matrix (CM) error is reported only for those methods where it can be computed, see its definition in the text.

| | Global results | | | | Per-class results | | |
|---|---|---|---|---|---|---|---|
| | F1-Score | Log-Loss | AUC | CM error | Tumor | Stroma | Infiltrate |
| AggNet [13] | 79.98 | 0.6814 | 92.87 | - | 89.52 | 76.40 | 74.03 |
| CL-VW [14] | 80.72 | 0.4911 | 92.64 | - | 89.32 | 75.59 | 77.26 |
| CL-VWB [14] | 81.79 | 0.5536 | 93.01 | - | 90.20 | 78.12 | 77.04 |
| CL-MW [14] | 81.58 | 0.4963 | 93.17 | 3.0922 | **90.58** | 77.26 | 76.89 |
| SVGPCR [19] | 81.47 | 0.3983 | 93.60 | 2.0637 | 90.32 | **78.72** | 75.37 |
| DGPCR-2 | 81.60 | **0.3961** | **93.71** | **2.0329** | 90.13 | 78.06 | 76.61 |
| DGPCR-3 | 81.85 | 0.3974 | 93.69 | 2.0377 | 90.19 | 78.15 | 77.22 |
| DGPCR-4 | **81.91** | 0.3978 | 93.69 | 2.0374 | 90.29 | 78.17 | **77.27** |

This constitutes a moderately imbalanced scenario where immune infiltrate is the minority class. All the methods are tested and assessed on the test set.

#### 2) EXPERIMENTAL DETAILS

We use a pretrained VGG16 to extract features for the proposed DGPCR, recall Figure 1. After the last convolutional layer, we apply average pooling with a $7 \times 7$ window to reduce the number of features to a vector of 512 components. To maximize the ELBO, recall eq. (7), we use Adam optimizer with a learning rate of $10^{-2}$. We performed 31, 000 iterations. We utilize 100 inducing points for the sparse GPs, and the minibatch size is set to 1000. The latent dimension of the hidden layers is 10. When predicting, we propagate $S = 100$ samples. We implemented the proposed DGPCR in GPflow 1.2.0. We trained the methods in an NVIDIA TITAN X (Pascal) GPU device with 12 Gb memory.

### B. RESULTS AND DISCUSSION

In this section, we evaluate the performance of DGPCR on the aforementioned breast cancer problem. We analyze five different research questions, which are discussed in the following five sections, respectively.

#### 1) COMPARISON TO STATE-OF-THE-ART CROWDSOURCING METHODS

Here we show that DGPCR performance on the test set is slightly but consistently better than that of other state-of-the-art crowdsourcing methods. Table 4 compares DGPCR (with two, three and four layers) to five state-of-the-art crowdsourcing methods. These are based on deep learning (AggNet [13]; CL-VW, CL-VWB, CL-MW [14]) and GPs (SVGPCR [19]), which are precisely the core components of DGPs, recall also Section II.

In global results, DGPCR obtains slightly superior performance in different types of metrics. Namely, the F1-Score does not consider the uncertainty in the predictions and is a trade-off between Recall and Precision, which is very relevant in this imbalanced scenario. The log-loss considers the uncertainty in the predictions (it is just the negative log-likelihood of the test data, the lower the better). The AUC (area under the ROC curve) is a threshold-free metric commonly used in machine learning.

**TABLE 5.** F1-Score for DL, GP, and DGP trained with three different types of labels: crowdsourced (CR), majority voting (MV), and gold. Details in the text.

| | MV | CR | GOLD |
|---|---|---|---|
| DL | 79.75 | 79.98 | 80.88 |
| GP | 79.19 | 81.47 | 81.57 |
| DGP-2 | 81.51 | 81.60 | 82.21 |
| DGP-3 | 80.92 | 81.85 | 81.97 |
| DGP-4 | 80.67 | 81.91 | 82.01 |

Moreover, in this imbalanced scenario, it is particularly important to analyze the performance in the minority class (immune infiltrate). We observe that DGPCR-4 obtains the best result in the minority class. The closest method (CL-VW) gets significantly worse performance in the other classes (especially in stroma).

#### 2) EVALUATING THE ESTIMATION OF ANNOTATORS BEHAVIOR

As explained in Section III and illustrated in Section V, DGPCR estimates a per-annotator confusion matrix (CM) that describes their behavior on the different classes. To evaluate the quality of this estimation, the fourth column of Table 4 shows the CM error for all those methods that estimate an analogous CM. As before, this error is the mean absolute error between the estimated CM and the true CM (which can be approximated based on the true labels provided by expert pathologists). We observe that DGP obtains the best result, with a significant difference against DL-based CL-MW. This can be also visualized in Figure 6, which shows the actual CMs estimated for five different annotators.

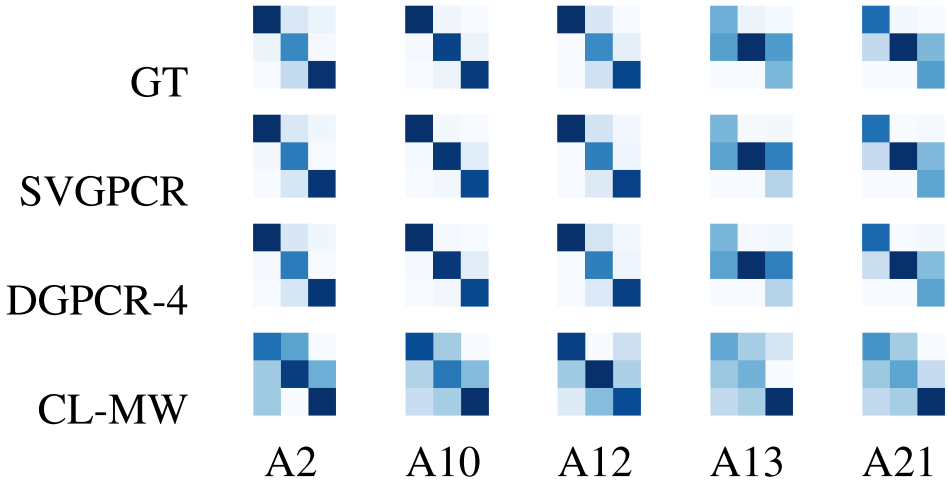



**FIGURE 6.** First row shows the true confusion matrices (CM). Rows 2-4: CMs estimated by different methods. Each column is an annotator. Displayed values are probabilities in the range [0, 1] (the darker the color, the closer to 1). In each CM, the first, second, and third rows/columns correspond to the classes tumor, stroma, and immune infiltrate, respectively.

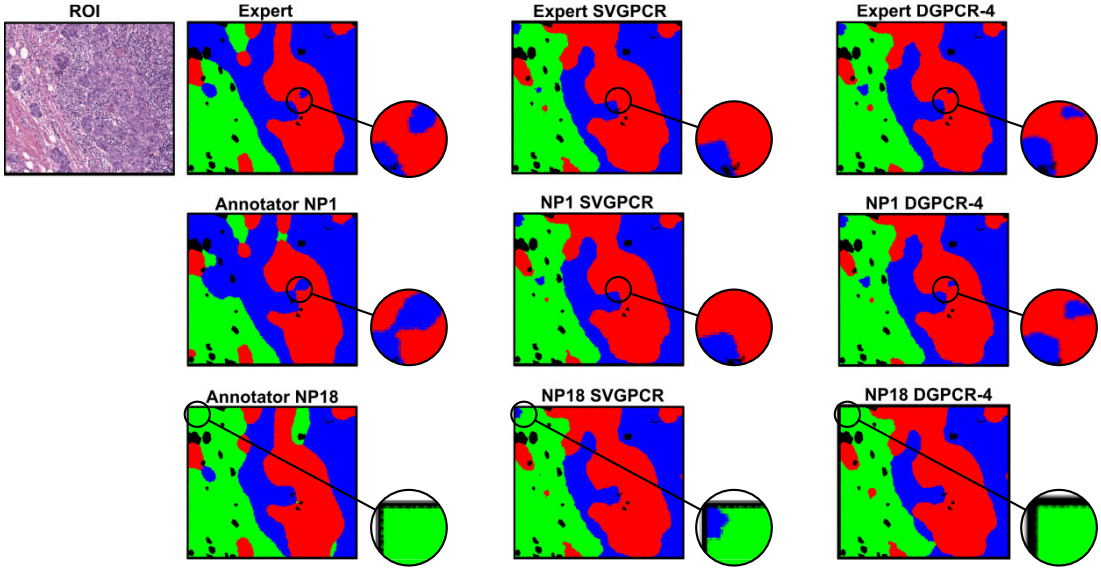



**FIGURE 7.** Segmentations for the ROI shown in first column. Second column shows the masks provided by an expert pathologist and two non-expert annotators. Third and fourth columns show the predictions obtained by SVGPCR and DGPCR-4 in each case. Colors represent different classes (red: tumor, green: stroma, blue: immune infiltrate).

### 3) COMPARISON TO THEORETICAL LOWER AND UPPER BOUNDS

Since DGPCR uses noisy crowdsourced labels, in theory its performance should be upper bounded by a standard DGP trained with expert (GOLD) labels. Analogously, its performance should be (lower) bounded by a standard DGP trained with the naive majority voting (MV) strategy, that is, considering as true label the one that was assigned by most annotators. Indeed, one of our main hypotheses is that, by adapting machine learning methods to the crowdsourcing paradigm, we can overcome naive methods like MV and obtain results that are very close to the ideal (but non-affordable) setting where all the expert labels are available (GOLD).

Table 5 shows the F1-Score results of DGP trained under these three different paradigms (when using 2, 3, and 4 layers). This confirms the hypothesized bounds, which reinforces the consistency of the proposed methodology. Moreover, Table 5 includes analogous results for DL and GP. Importantly, notice that DGP with crowdsourced labels obtains better results than GP and DL with gold ones. Here, we use a VGG-16 net for DL.

### 4) VISUALIZING THE PREDICTIONS

The numerical results obtained so far are well illustrated in Figure 7. This figure focuses on an ROI annotated by all the participants. Notice that the segmentations predicted by SVGPCR and DGPCR-4 are obtained by aggregating the predictions obtained at the patch level.

The first row shows the analyzed ROI, the mask provided by the expert pathologist, and the predictions obtained by SVGPCR and DGPCR-4. In spite of working at the patch level, both methods capture well the structure of the different

**TABLE 6.** Average and 0.95 confidence interval of macro F1 score with reduced subsets of the training set. Each column refers to a percentage of the original training set size. Every experiment has been repeated three times using different subsets.

| | 1% | 5% | 10% | 25% | 50% | 75% |
|---|---|---|---|---|---|---|
| CL-MW | 0.7442 ± 0.0232 | 0.7479 ± 0.009 | 0.7533 ± 0.0057 | 0.7861 ± 0.0195 | 0.8062 ± 0.0042 | 0.8074 ± 0.0049 |
| CL-VW | 0.7437 ± 0.0221 | 0.746 ± 0.0099 | 0.7537 ± 0.0111 | 0.776 ± 0.0091 | 0.7931 ± 0.0070 | 0.8022 ± 0.0091 |
| CL-VWB | 0.7448 ± 0.0187 | 0.7508 ± 0.0102 | 0.7623 ± 0.0115 | 0.7867 ± 0.0112 | 0.8018 ± 0.0038 | 0.8085 ± 0.0049 |
| SVGPCR | **0.7545 ± 0.0169** | **0.7810 ± 0.0068** | **0.7837 ± 0.0095** | 0.7980 ± 0.0083 | 0.8075 ± 0.0027 | 0.8130 ± 0.0031 |
| DGPCR-2 | 0.7491 ± 0.0215 | 0.7522 ± 0.0101 | 0.7801 ± 0.0117 | 0.8004 ± 0.0049 | 0.8136 ± 0.002 | 0.8158 ± 0.0049 |
| DGPCR-3 | 0.7500 ± 0.0223 | 0.7521 ± 0.0103 | 0.7815 ± 0.0116 | **0.8020 ± 0.0084** | 0.8139 ± 0.0016 | 0.8196 ± 0.0023 |
| DGPCR-4 | 0.7495 ± 0.0233 | 0.7522 ± 0.0101 | 0.7808 ± 0.0098 | 0.8008 ± 0.0084 | **0.8159 ± 0.0035** | **0.8204 ± 0.004** |

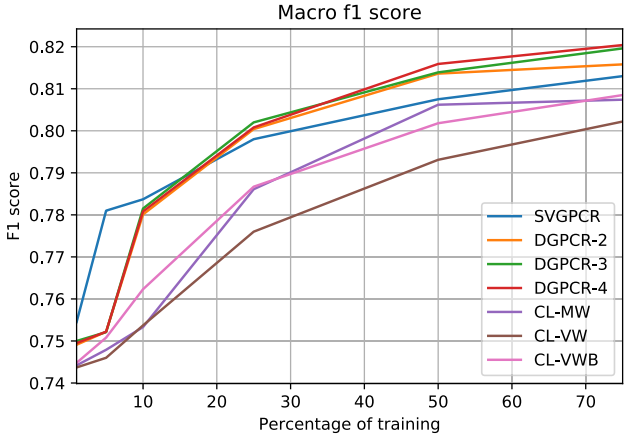



**FIGURE 8.** Average macro F1 score (axis-y) using subsets of the training set (given a percentage; axis-x). We see that GP-based methods are more robust to small amounts of crowdsourced labeled data. Furthermore, DGPCR methods perform quite well through different sizes unlocking their full potential with more data available. However, DL methods fail considerably when data is reduced.

classes. Notice that, as shown by the previous numerical results, DGPCR-4 is sharper in the minority class (see the blue isles in the green and red areas, which are better captured by DGPCR-4). The second and third rows show the segmentation provided by two crowdsourcing annotators, and the predictions obtained by SVGPCR and DGPCR-4 for those annotators. Again, we observe accurate predictions in general, with DGPCR-4 being finer in the minority class (specifically, see again the blue isle in the red area in the second row; and the blue isle that SVGPCR predicts wrongly in the green area in the third row, top left corner).

#### 5) ROBUSTNESS TO THE SIZE OF THE TRAINING SET

Finally, we assess the generalization capability and robustness of DGPCR against the lack of labeled data, which is a typical scenario in medical imaging. We consider three different subsets for each size to measure the variability of the performance, reporting the average and 0.95 confidence interval of the three runs. Following Section VI-B1 we compare with SVGPCR and DL methods (for the latter we focus on CL methods, which have obtained better results so far).

Figure 8 shows the results graphically. We observe a gap between GP-based and DL-based methods, confirming that probabilistic methods can generalize better even when data is scarce. Shallow SVGPCR is the best with little data, but DGPCR performs reasonably well across different settings. Furthermore, as data increases, DGPCR takes advantage of its complex architecture exploiting the data available. In conclusion, DGPCR performs well even when training data is reduced, showing how DGPCR combines the advantages of both SVGPCR and DL methods.

Table 6 shows these results with a 0.95 confidence interval. In addition to the conclusions already drawn, we can observe the stability of the GP-based methods in different training subsets. In general, these methods outperform DL methods for every training size. Specifically, DGPCR outperforms the rest, with non-overlapping confidence intervals (including the shallow SVGPCR), when the data available is enough (i.e., higher than 25%).

## VII. CONCLUSION

Crowdsourcing can be an effective approach for generating labeled data at scale for medical applications. ML models trained on crowdsourced data, however, should ideally address the noise introduced by less experienced annotators and the biases of individual annotators. While probabilistic methods can effectively model crowdsourcing, many of these methods cannot learn complex representations required in problems like image classification or segmentation. DGPCR addresses this challenge by combining the advantages of deep learning and probabilistic methods. Specifically, it combines the capabilities of complex function modeling with uncertainty quantification to provide a robust solution to crowdsourcing tasks. This is the first step towards the end-to-end training of deterministic feature extractors and probabilistic classifiers in crowdsourcing scenarios.

Our DGPCR method can infer an estimated ground truth on unseen instances and can generate predictions that reflect the biases of individual annotators. We evaluated DGPCR in MNIST and a real-world breast cancer classification problem, showing competitive or superior performance to state-of-the-art crowdsourcing methods. The performance of DGPCR trained on noisy labels is similar to training with expert labels. DGPCR was compared to alternatives for overall performance, performance on minority classes, robustness to adversarial annotators, and training set size efficiency. While the additional parameters introduced in DGPCR require larger training sets, they produce higher performance in most tasks.

There are still some open questions in crowdsourcing. Future work should address how much labeled data or which

overlap between experts and non-experts would lead to satisfying results. Despite these limitations, this paper opens the door to more robust and competitive classifiers in crowdsourcing scenarios, which are of great interest to the medical imaging community. We consider that data labeled by multiple pathologists are needed to tackle inter-observer variability and individual biases. This approach can lead to a consensus and leverage noisy labels provided by generalists and pathology trainees. This tool can also train novel pathologists and medical students, boosting their performance. Ultimately, this approach will help to achieve more robust clinical systems in digital pathology.

## REFERENCES

[1] Y. Yari, T. V. Nguyen, and H. T. Nguyen, "Deep learning applied for histological diagnosis of breast cancer," *IEEE Access*, vol. 8, pp. 162432–162448, 2020.

[2] I. Hirra, M. Ahmad, A. Hussain, M. U. Ashraf, I. A. Saeed, S. F. Qadri, A. M. Alghamdi, and A. S. Alfakeeh, "Breast cancer classification from histopathological images using patch-based deep learning modeling," *IEEE Access*, vol. 9, pp. 24273–24287, 2021.

[3] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.

[4] A. Pedersen, M. Valla, A. M. Bofin, J. P. De Frutos, I. Reinertsen, and E. Smistad, "FastPathology: An open-source platform for deep learning-based research and decision support in digital pathology," *IEEE Access*, vol. 9, pp. 58216–58229, 2021.

[5] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

[6] N. G. Laleh, H. S. Muti, C. M. L. Loeffler, A. Echle, O. L. Saldanha, F. Mahmood, M. Y. Lu, C. Trautwein, R. Langer, B. Dislich, and R. D. Buelow, "Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102474.

[7] M. López-Pérez, M. Amgad, P. Morales-Álvarez, P. Ruiz, L. A. D. Cooper, R. Molina, and A. K. Katsaggelos, "Learning from crowds in digital pathology using scalable variational Gaussian processes," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, Jun. 2021.

[8] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101759.

[9] Z. Su, T. E. Tavolara, G. Carreno-Galeano, S. J. Lee, M. N. Gurcan, and M. K. K. Niazi, "Attention2Majority: Weak multiple instance learning for regenerative kidney grading on whole slide images," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102462.

[10] C. J. Sudha and Y. Sneha, "Classification of medical images using deep learning to aid in adaptive big data crowdsourcing platforms," in *ICT With Intelligent Applications*. Berlin, Germany: Springer, 2022, pp. 69–77.

[11] R. Böhm, C. Betsch, Y. Litovsky, P. Sprengholz, N. T. Brewer, G. Chapman, J. Leask, G. Loewenstein, M. Scherzer, C. R. Sunstein, and M. Kirchler, "Crowdsourcing interventions to promote uptake of COVID-19 booster vaccines," *eClinicalMedicine*, vol. 53, Nov. 2022, Art. no. 101632.

[12] M. Amgad, L. A. Atteya, H. Hussein, K. H. Mohammed, E. Hafiz, M. A. Elsebaie, A. M. Alhusseiny, M. A. Al Moslemany, A. M. Elmatboly, P. A. Pappalardo, and R. A. Sakr, "NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer," *GigaScience*, vol. 11, pp. 1–12, May 2022.

[13] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1313–1321, May 2016.

[14] F. Rodrigues and F. Pereira, "Deep learning from crowds," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[15] P. Ruiz, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos, "Learning from crowds with variational Gaussian processes," *Pattern Recognit.*, vol. 88, pp. 298–311, Apr. 2019.

[16] P. Ruiz, P. Morales-Álvarez, S. Coughlin, R. Molina, and A. K. Katsaggelos, "Probabilistic fusion of crowds and experts for the search of gravitational waves," *Knowl.-Based Syst.*, vol. 261, Feb. 2023, Art. no. 110183.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[18] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2. Cambridge, MA, USA: MIT Press, 2006.

[19] P. Morales-Alvarez, P. Ruiz, S. Coughlin, R. Molina, and A. K. Katsaggelos, "Scalable variational Gaussian processes for crowdsourcing: Glitch detection in LIGO," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1534–1551, Mar. 2022.

[20] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4588–4599.

[21] D. H. Svendsen, P. Morales-Álvarez, A. B. Ruescas, R. Molina, and G. Camps-Valls, "Deep Gaussian processes for biogeophysical parameter retrieval and model inversion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 68–81, Aug. 2020.

[22] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. A. Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, and J. Ahmed, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019.

[23] C. Hutter and J. C. Zenklusen, "The cancer genome atlas: Creating lasting value beyond its data," *Cell*, vol. 173, no. 2, pp. 283–285, 2018.

[24] W. Yang, C. Li, and L. Jiang, "Learning from crowds with decision trees," *Knowl. Inf. Syst.*, vol. 64, no. 8, pp. 2123–2140, Aug. 2022.

[25] L. Jiang, H. Zhang, F. Tao, and C. Li, "Learning from crowds with multiple noisy label distribution propagation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6558–6568, Nov. 2022.

[26] Z. Chen, L. Jiang, and C. Li, "Label augmented and weighted majority voting for crowdsourcing," *Inf. Sci.*, vol. 606, pp. 397–409, Aug. 2022.

[27] V. Raykar, S. Yu, L. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.

[28] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," *Med. Image Anal.*, vol. 50, pp. 167–180, Jan. 2018.

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

[30] D. G. Matthews, G. Alexander, M. V. D. Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. Leon-Villagra, Z. Ghahramani, and J. Hensman, "GPflow: A Gaussian process library using tensorflow," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1299–1304, 2017.

[31] F. Rodrigues, M. Lourenço, B. Ribeiro, and F. C. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2409–2422, Dec. 2017.

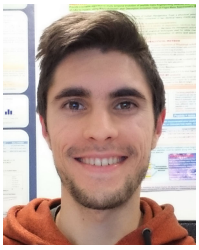
[32] P. Morales-Álvarez, P. Ruiz, R. Santos-Rodríguez, R. Molina, and A. K. Katsaggelos, "Scalable and efficient learning from crowds with Gaussian processes," *Inf. Fusion*, vol. 52, pp. 110–127, Dec. 2019.

[33] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, "Advances in variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2008–2026, Aug. 2019.

[34] N. Kanwal, F. Perez-Bueno, A. Schmidt, K. Engan, and R. Molina, "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review," *IEEE Access*, vol. 10, pp. 58821–58844, 2022.

[35] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Sep./Oct. 2001.

**MIGUEL LÓPEZ-PÉREZ** received the B.Sc. degree in mathematics and the M.S. degree in data science and computer engineering from the Universidad de Granada, Granada, Spain, in 2017 and 2018, respectively, and the joint Ph.D. degree from the Universidad de Granada under the supervision of Prof. Molina and Northwestern University under the supervision of Prof. Katsaggelos, in 2022. He has visited Northwestern University under the supervision of Prof. Katsaggelos for predoctoral research stays. Currently, he is working as a Postdoctoral Researcher with the Visual Information Processing Group, Department of Computer Science and Artificial Intelligence, Universidad de Granada. His research interests include the use of Bayesian modeling, especially the Gaussian processes and their application to weakly supervised learning, working usually with medical imaging problems. He also contributes to a podcast called The Fluxions on AI and mathematics.



**PABLO MORALES-ÁLVAREZ** received the B.Sc. degree in mathematics from the University of Granada (UGR), Spain, in 2014, and the Ph.D. degree from UGR under the supervision of Prof. Rafael Molina and Northwestern University, USA, under the supervision of Prof. Aggelos K. Katsaggelos, in 2020. Funded by the highly competitive Ph.D. fellowship from La Caixa Banking Foundation. During his Ph.D. degree, he visited several research groups, including the Machine Learning Group, University of Cambridge, U.K., where he worked for five months with Prof. José Miguel Hernández-Lobato. From October 2020 to May 2021, he did postdoctoral research at Microsoft Research Cambridge under the supervision of Cheng Zhang. He serves as an Assistant Professor at the Department of Statistics and Operations Research, UGR. He has published in top machine learning conferences and journals, such as NeurIPS 2022, ICLR 2021, and IEEE TRANSACTIONS PAMI 2022. His research interests include probabilistic machine learning methods, specially (deep) Gaussian processes and Bayesian neural networks. He has also served as a Reviewer for top-tier conferences (ICML 2020, NeurIPS 2020, ICLR 2021, and NeurIPS 2021), and *Science* journal. He has obtained the SCIE-FBBVA Award 2022 to the Most Outstanding Young Spanish Researchers in Computer Science. He obtained the First End of Studies Award by the Spanish Ministry of Science to the Most Outstanding Undergraduate in Mathematics in Spain.

**LEE A. D. COOPER** received the Ph.D. degree in electrical and computer engineering from The Ohio State University, in 2009. He joined the Biomedical Informatics Faculty, Emory University, in 2012, where he was jointly appointed with the Department of Biomedical Engineering, Georgia Institute of Technology. He joined the Department of Pathology, Northwestern University, in 2019, as an Associate Professor and the Director of Computational Pathology.

**RAFAEL MOLINA** (Life Senior Member, IEEE) received the M.Sc. degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from the University of Granada, Granada, Spain, in 1979 and 1983, respectively. He was the Dean of the School of Computer Engineering, University of Granada, from 1992 to 2002, where he became a Professor in computer science and artificial intelligence, in 2000. He was the Head of the Department of Computer Science and Artificial Intelligence, University of Granada, from 2005 to 2007. He has coauthored an article that received the Runner-Up Prize from the Reception for Early Stage Researchers at the House of Commons, in 2007, the Best Student Paper from the IEEE International Conference on Image Processing, in 2007, the ISPA Best Paper, in 2009, and the EUSIPCO 2013 Best Student Paper. His research interests include Bayesian modeling and inference in image restoration (applications to astronomy and medicine), super-resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, supervised learning, and crowdsourcing. He has served as an Associate Editor for *Applied Signal Processing*, from 2005 to 2007, and the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2010 to 2014. Since 2011, he has been serving as an Area Editor for *Digital Signal Processing*.

**AGGELOS K. KATSAGGELOS** (Life Fellow, IEEE) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1981 and 1985, respectively. In 1985, he joined the Department of Electrical Engineering and Computer Science, Northwestern University, where he is currently a Professor Holder of the Joseph Cummings Chair. Previously, he was the Holder of the Ameritech Chair of Information Technology and the AT&T Chair. He is also a member of the Academic Staff, NorthShore University Health System, an affiliated Faculty Member of the Department of Linguistics and he has an appointment with the Argonne National Laboratory. He has authored or coauthored extensively in the areas of multimedia signal processing and communications, computational imaging, and machine learning, including more than 250 journal articles, 600 conference papers, and 40 book chapters, and he is the holder of 30 international patents. He is the coauthor of *Rate-Distortion Based Video Compression* (Kluwer, 1997), *Super-Resolution for Images and Video* (Claypool, 2007), *Joint Source-Channel Video Transmission* (Claypool, 2007), and *Machine Learning Refined* (Cambridge University Press, 2016). He has supervised 57 Ph.D. theses. Among his many professional activities, he was the Editor-in-Chief of the *IEEE Signal Processing Magazine* (1997–2002), a BOG Member of the IEEE Signal Processing Society (1999–2001), a member of the Publication Board of the PROCEEDINGS OF THE IEEE (2003–2007), and a member of the Award Board of the IEEE Signal Processing Society. He is a fellow of the SPIE (2009), EURASIP (2017), and OSA (2018). He was a recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), the IEEE Signal Processing Society Best Paper Award (2001), the IEEE ICME Paper Award (2006), the IEEE ICIP Paper Award (2007), the ISPA Paper Award (2009), and the EUSIPCO Paper Award (2013). He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008).