PROBABILISTIC ATTENTION BASED ON GAUSSIAN PROCESSES FOR DEEP MULTIPLE INSTANCE LEARNING

PREPRINT

Arne Schmidt, Pablo Morales-Álvarez, Rafael Molina *

ABSTRACT

Multiple Instance Learning (MIL) is a weakly supervised learning paradigm that is becoming increasingly popular because it requires less labeling effort than fully supervised methods. This is especially interesting for areas where the creation of large annotated datasets remains challenging, as in medicine. Although recent deep learning MIL approaches have obtained state-of-the-art results, they are fully deterministic and do not provide uncertainty estimations for the predictions. In this work, we introduce the Attention Gaussian Process (AGP) model, a novel probabilistic attention mechanism based on Gaussian Processes for deep MIL. AGP provides accurate bag-level predictions as well as instance-level explainability, and can be trained end-to-end. Moreover, its probabilistic nature guarantees robustness to overfitting on small datasets and uncertainty estimations for the predictions. The latter is especially important in medical applications, where decisions have a direct impact on the patient's health. The proposed model is validated experimentally as follows. First, its behavior is illustrated in two synthetic MIL experiments based on the well-known MNIST and CIFAR-10 datasets, respectively. Then, it is evaluated in three different real-world cancer detection experiments. AGP outperforms state-of-the-art MIL approaches, including deterministic deep learning ones. It shows a strong performance even on a small dataset with less than 100 labels and generalizes better than competing methods on an external test set. Moreover, we experimentally show that predictive uncertainty correlates with the risk of wrong predictions, and therefore it is a good indicator of reliability in practice. Our code is publicly available.

Keywords A

ttention Mechanism, Multiple Instance Learning, Gaussian Processes, Digital Pathology, Whole Slide Images.

1 Introduction

Machine learning classification algorithms have achieved excellent results in many different applications [1, 2, 3, 4, 5]. However, these algorithms need large datasets that must be labelled by an expert. Such labelling process often becomes the bottleneck in real-world applications. In the last years, Multiple Instance Learning (MIL) has become a very popular weakly supervised learning paradigm to alleviate this burden. In MIL, instances are grouped in bags, and only bag labels are needed to train the model [6].

*This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627 (CLARIFY Project), from the Spanish Ministry of Science and Innovation under project PID2019-105142RB-C22, and by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades under the project P20_00286. P. Morales-Álvarez acknowledges funding from the University of Granada postdoctoral program "Contrato Puente". A. Schmidt and R. Molina are with the Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain (email: {arne, rms}@decsai.ugr.es). P. Morales-Álvarez is with the Department of Statistics and Operations Research, University of Granada, Granada, Spain.

MIL is especially interesting for the medical field [7, 6, 8, 9]. As an example, consider the problem of cancer detection in histopathological images, where the goal is to predict whether a given image contains cancerous tissue or not (binary classification problem). Since these images are extremely large (in the order of gigapixels), they cannot be completely fed to a classifier (typically a deep neural network). Therefore, the classical approach is to split the image in many smaller patches and train a classifier at patch level. Unfortunately, this requires that an expert pathologist labels *every patch* as cancerous or not, which is a daunting, time-consuming, expensive and error-prone task [10]. In the MIL setting, each image is considered as a bag that contains many different instances (its patches). Importantly, since MIL only requires bag labels, the workload for the pathologist is reduced to labelling each image (and not every single patch).

Different underlying classification methods have been proposed for the MIL problem. Early approaches relied on traditional methods such as support vector machines [11], expectation maximization [12] or undirected graphs [13]. In recent years, many approaches are based on deep learning models due to their flexibility and their capacity to learn complex functions [14, 15]. In particular, the current state-of-the-art is given by an attention-based deep learning model originally introduced in [16]. The idea of attention-based MIL is to predict an attention weight for each instance, which determines its influence on the final bag prediction.

The attention mechanism has several advantages: it can be trained end-to-end with deep learning architectures because it is differentiable, it provides accurate bag-level predictions, and it provides explainability at instance-level (by looking at the instances with higher attention). Due to its success, it has been adapted and extended several times [17, 18, 19, 20, 21]. However, all these attention-based MIL approaches are based on *deterministic* transformations (usually one or two fully connected layers to calculate the attention weights). This has several drawbacks, such as the lack of uncertainty estimation and the overfitting to small datasets. These limitations can be addressed by introducing a probabilistic model as a backbone for the MIL attention module. Moreover, such a sound probabilistic treatment leads to better predictive performance, see e.g. [22, 23, 24, 25].

In this work, we introduce a novel probabilistic attention mechanism based on Gaussian processes for MIL, which will be referred to as AGP. It leverages a Gaussian process (GP) to obtain the attention weight for each instance. GPs are powerful Bayesian models that can describe flexible functions and provide accurate uncertainty estimation due to their probabilistic nature (their most relevant properties will be reviewed in Section 2). Moreover, AGP uses variational inference to ensure a probabilistic treatment of the estimated parameters. We experimentally evaluate AGP on different datasets, including an illustrative MNIST-based MIL problem, CIFAR-10, and three real-world prostate cancer classification tasks. Specifically, we show that: 1) AGP outperforms state-of-the-art and related MIL methods, 2) the estimated uncertainty can be used to identify which model predictions should be disregarded or double-checked, 3) higher attention is assigned to the most relevant instances of each bag (e.g. cancerous patches in images), and 4) AGP generalizes better than competitors to different datasets (and the estimated uncertainty reflects this extrapolation).

In machine learning literature, some related approaches have used GPs in the context of MIL, such as GPMIL [26] or VGPMIL [27]. These methods rely on a sparse GP for the instance classification, followed by an (approximated) maximum function for the final bag prediction. In contrast to our approach, both GPMIL and VGPMIL focus on instance-level predictions which are later combined for bag predictions, they are limited due to the simplicity of the instance aggregation (approximated max-aggregation), and cannot perform end-to-end training with deep learning models. In [9] we proposed the combination of a deep learning attention model and VGPMIL, but as a two-stage approach: first, the attention mechanism serves to train the feature extractor; in a second step, the VGPMIL model is applied to the extracted features. Although this approach showed promising results, it did not overcome all the aforementioned limitations of VGPMIL. In particular, the attention mechanism is discarded after the first training phase and not used at all for the final predictions, so the model cannot fully leverage the advantages of combining an attention mechanism and GPs. So three major challenges remain unsolved: (i) end-to-end training, (ii) uncertainty estimation, and (iii) multiclass classification. Our proposed AGP model provides all three features, as described below.

For a different scenario, time series prediction, the combination of GP and attention has recently shown promising results [28]. For this problem, the GP replaced the final regression layer while the attention weights were calculated deterministically, which is a major difference to our work. Another existing approach combines an attention mechanism with GPs for channel attention [29]. Here, the GPs model correlations in activation maps of convolutional neural networks (CNNs) to estimate the channel attention weights. The channel attention helps improve the CNNs performance for visual tasks. However, to the best of our knowledge, our approach is the first one that estimates the attention weights with GPs in the context of deep MIL. Moreover, the novel method fully leverages the strengths of GPs for the attention estimation, including the strong function regression capabilities and uncertainty estimation. Indeed, the proposed AGP model does not only provide accurate predictions, outperforming existing state-of-the-art methods, but also provides an estimation of the uncertainty introduced by the attention. At the same time, our model inherits all

positive properties of deterministic attention modules, such as explainability on instance level (as later discussed in section 4.2) and end-to-end training with deep learning feature extractors.

The paper is organized as follows. Section 2 describes the probabilistic model and inference used by AGP, preceded by the notation and background on the attention mechanism and GPs. Section 3.1 includes an illustrative and visual MIL experiment using MNIST. In section 4, we carry out three experiments on prostate cancer classification. We not only report a strong performance of the AGP model, but also analyze the probabilistic predictions for these real-world datasets. Finally, section 5 summarizes the main conclusions.

2 Methodology

In this section we present the theoretical framework for AGP. First, we describe some required background, the MIL notation (section 2.1), the attention mechanism (section 2.2), and the basics on (sparse) Gaussian Processes (section 2.3). Then, section 2.4 focuses on the description of AGP, including the probabilistic modelling, the variational inference, and how to make predictions.

2.1 Multiple Instance Learning (MIL) for Cancer Classification

In the classical MIL setting we assume that instances $\chi \in \mathbb{R}^D$ are grouped into bags $\mathcal{X}_b = \{\chi_{b1}, \chi_{b2}, .., \chi_{bN_b}\}$, where the number of instances N_b in each bag b can vary. Notice that this notation is not the most standard in MIL, where X_b is usually used for bags and x_{bi} for instances. However, we will use these letters for the input of the GP in Section 2.3. Therefore, to avoid confusion, we have chosen to use $\mathcal{X}_b$ and χ_{bi} for bags and instances, respectively. Each instance has a (binary) label $y_{bi} \in \{0, 1\}$ that remains unknown. The bag label T_b is known, and it is given by:

$$T_b = 0 \;\Leftrightarrow\; \forall i = 1, ..., N_b,\; y_{bi} = 0, \tag{1}$$

$$T_b = 1 \;\Leftrightarrow\; \exists i \in \{1, ..., N_b\} : y_{bi} = 1. \tag{2}$$

In cancer classification with histopathological images, the MIL setting considers a Whole Slide Image (WSI) as a bag $\mathcal{X}_b$ and its diagnosis as the bag label T_b. Since the complete WSI is too big to be processed by a common convolutional neural network, it is sliced into patches that form the instances. As MIL only requires bag labels for training, the local annotation of the patches by expert pathologists is not necessary. This provides huge benefits in terms of time and cost of the labeling process. Apart from the binary classification (cancerous vs. non-cancerous), we are also interested in the cancer class of the WSI. This class can be determined based on the features of the cancerous (positive) patches. In this setting, equation (1) still applies for non-cancerous (negative) WSIs, while for cancerous WSIs we want to specify the class $T_b = c$, with $c \in \{c_1, c_2, .., c_K\}$ representing each one of the possible K cancer classes based on the cancerous areas.

2.2 Attention Mechanism

As mentioned in the introduction, AGP leverages a probabilistic GP-based attention mechanism to aggregate the information of the different instances in a bag. This is inspired by the deterministic attention introduced in [16]. Specifically, the model proposed in [16] consists of three main components: the feature extractor f_{fe}, the attention mechanism, and the final classification layer f_{cl}. The feature extractor is given by a convolutional neural network followed by some fully connected layers. It is used to process each instance, resulting in one feature vector $h_{bi} = f_{fe}(\chi_{bi})$ per instance, with $h_{bi} \in \mathbb{R}^P$. Then, the attention mechanism estimates a *deterministic* attention weight per instance a_{bi} based on these features. Specifically, the used mapping is:

$$a_{bi} = \frac{\exp\{w^\top \tanh(V h_{bi})\}}{\sum_j \exp\{w^\top \tanh(V h_{bj})\}}, \tag{3}$$

where the vector $w \in \mathbb{R}^{L\times 1}$ and the matrix $V \in \mathbb{R}^{L\times P}$ are optimized during training. Finally, the classification is done using the average of the extracted features, weighted by the attention values:

$$\hat{T}_b = f_{cl}\left(\textstyle\sum_i h_{bi} a_{bi}\right). \tag{4}$$

Our goal is now to replace the deterministic attention mechanism given in equation (3) by a GP model. The GP is a probabilistic method that is able to give a better estimation of the attention weights. Moreover, it allows for capturing the uncertainty introduced by the attention mechanism.

2.3 (Sparse) Gaussian Processes

Gaussian Processes are stochastic processes where the output distribution is assumed to be multivariate Gaussian [30]. They can be used to estimate an objective function f in a probabilistic way: the GP defines a prior distribution over functions (whose properties depend on the type of kernel used), and the posterior is computed given such prior and the observed data [31]. The major drawback of GPs is that the computation of the posterior is not scalable, because it involves inverting a matrix of size $N \times N$, with N the number of datapoints [32]. This is clearly relevant for our MIL scenario, where there exist typically plenty of instances.

To overcome this limitation, different types of Sparse Gaussian Processes (SGPs) have been introduced in the last years [33, 34, 35, 36]. Here we will follow the approach in [34], since it allows for training in batches. The idea behind SGP is to define the GP posterior distribution on a set of M inducing point locations $Z = \{z_i\}_{i=1}^{M}$, instead of doing it on the N real instances $X = \{x_i\}_{i=1}^{N}$. The amount of inducing points is taken $M \ll N$, and their location must be representative for the training distribution (in fact, they can be optimized during training, as we will do in AGP).

The formulation of SGP is as follows. Let U be the output of the GP at the inducing point locations Z, and F the output at the datapoints X. The SGP model is given by

$$p(U|Z) = \mathcal{N}(U|0, K_{ZZ}), \tag{5}$$

$$p(F|U, Z, X) = \mathcal{N}(F|K_{XZ}K_{ZZ}^{-1}U, \hat{K}), \tag{6}$$

where $K_{AB} := k(A, B)$ is the matrix obtained by applying the GP kernel function on A and B. Moreover, we have

$$\hat{K} = K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}. \tag{7}$$

To perform inference, a posterior Gaussian distribution $q(U) = \mathcal{N}(U|\mu_u, \Sigma_u)$ is used on the inducing points. Therefore, the parameters to be estimated during training are μ_u, Σ_u, the kernel parameters, and the inducing points locations.

Given a test point, the prediction of the SGP model is a random variable (and not a single deterministic value). The mean of the random variable provides the value to regress, while the standard deviation provides the uncertainty. Finally, in this section we have written X for the input to the GP, as is standard in the GP literature. However, we want to stress that in our case the input to the GP will be given by the features extracted in some previous step, and not the raw input itself.

2.4 Probabilistic attention based on Gaussian Process (AGP)

Probabilistic modelling. The AGP model is depicted in Figure 1. It is a combination of a deterministic convolutional network that serves as a feature extractor f_{fe}, an SGP to estimate the attention, and a deterministic fully connected layer for the final classification f_{cl}. Next, we describe the different components using Figure 1 as reference.

Remember that $\mathcal{X}_b = \{\chi_{b1}, .., \chi_{bN_b}\}$ describe the instances in one bag b. First, the feature extractor f_{fe} is applied. It consists of a CNN and two fully connected layers with ReLu activation and 128 and 64 units, respectively. The choice of the CNN backbone depends on the task. For cancer classification, we will use EfficientNetB5 as the CNN backbone [37]. The output of the feature extractor are high-level feature vectors $H_b = \{h_{b1}, .., h_{bN_b}\}$ where each $h_{bi}, i = 1, .., N_b$ has 64 dimensions:

$$H_b = f_{fe}(\mathcal{X}_b). \tag{8}$$

We focus now on the attention module. Here, we first apply to each instance the same fully connected layer with sigmoid activation and 32 units. This further reduces the dimensionality, resulting in the SGP input feature vectors $X_b = \{x_{b1}, .., x_{bN_b}\}$ of 32 dimensions each. This alleviates the optimization of the inducing point locations, which are defined in the input space. Moreover, the sigmoid function guarantees values between 0 and 1, which facilitates the initialization of the inducing point locations. In summary, each vector h_{bi} with 64 components is transformed into a vector x_{bi} with 32 components. With these feature vectors, the SGP model described in Section 2.3 regresses the values $F_b = \{f_{b1}, .., f_{bN_b}\}$, which are normalized through a softmax (SM) layer to calculate the attention weights $A_b = \{a_{b1}, .., a_{bN_b}\}$:

$$a_{bi} = \frac{\exp\{f_{bi}\}}{\sum_j \exp\{f_{bj}\}}. \tag{9}$$

Importantly, note that the attention weights are random variables, since they are computed from the SGP output. During inference, we will use Monte Carlo sampling to approximate their distribution.

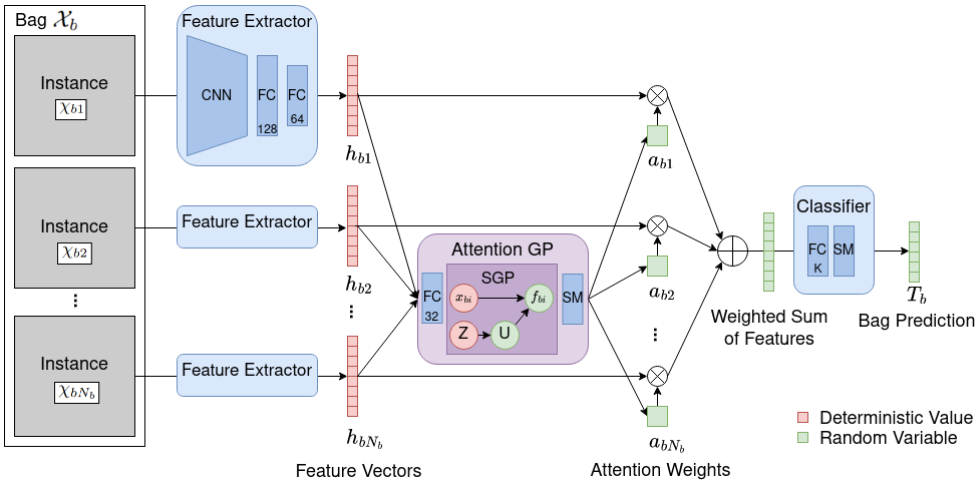

Figure 1: The AGP model architecture. The feature extractor consists of a Convolutional Neural Network (CNN) and two Fully Connected Layers (FC). The attention module incorporates another FC layer, the Sparse Gaussian Process (SGP), and a softmax (SM) function. The final classification is performed by another fully connected layer (where the dimensionality depends on the number of classes) and a softmax activation. While the feature vectors are deterministic values, the use of a (sparse) GP makes the attention weights and the final prediction random variables.

Finally, the classifier f_{cl} is used to obtain the bag label T_b. The bag label T_b follows a categorical distribution with K classes, $T_b \sim \text{Cat}(p_1, \ldots, p_K)$, $\sum_k p_k = 1$. These probabilities are computed by applying the classifier f_{cl} over the average of the feature vectors H_b weighted by the attention weights A_b:

$$(p_1, \ldots, p_K) = f_{cl}\left(\textstyle\sum_i h_{bi} a_{bi}\right). \tag{10}$$

As shown in Figure 1, the classifier f_{cl} consists of one fully connected layer with one unit per class and a softmax activation function. Again, since the attention weights A_b are random variables, the probabilities $p_1, \ldots, p_K$ are random variables whose distribution will be estimated through Monte Carlo sampling. This probabilistic nature will allow for computing uncertainty estimation in the predictions.

Once we have described how AGP processes one bag $\mathcal{X}_b$, the joint full probabilistic model is

$$p(T, F, U) = p(T|F)p(F|U)p(U). \tag{11}$$

Here, we have written $T = \{T_1, \ldots, T_B\}$ for the collection of all the bag labels, and analogously for F. The inducing points U and their locations Z are global for all the bags because the feature space is the same for all instances of all bags and the SGP should be able to generalize to unseen bags. Notice that, to lighten the notation, we are not writing explicitly the dependency on all the variables. For instance, $p(U) = p(U|Z)$ depends on the inducing point locations Z; $p(F|U) = p(F|U, Z, X)$ also depends on Z and the SGP input X; and $p(T|F) = p(T|F, X)$ depends on X and depends on F only through A (recall eqs. (9)-(10)). Also, we are not writing explicitly the dependency on other parameters such as all the neural network weights (which are collectively denoted as W) and the SGP kernel parameters (which are denoted as θ).

Variational inference. To perform inference in AGP, we need to obtain the posterior distribution $p(F, U|T)$ and the learnable parameters W, θ, Z. Since eq. (11) is not analytically tractable, we resort to variational inference [38]. Namely, variational inference considers a parametric posterior distribution and finds the parameters that minimize the distance to the true posterior in the Kullback-Leibler divergence sense (by maximizing the log evidence lower bound, ELBO). In our case, we consider the distribution $q(F, U) = p(F|U)q(U)$, where $p(F|U)$ equals the (prior) conditional distribution in eq. (6) and $q(U) = \mathcal{N}(U|\mu_u, \Sigma_u)$ is a (multivariate) Gaussian with mean vector μ_u and covariance matrix Σ_u, both to be estimated during training (variational parameters).

With this choice, the ELBO to be maximized is

$$\log p(T) \geq \text{ELBO} = \mathbb{E}_{q(F)} \log p(T|F) - \text{KL}(q(U)||p(U)), \tag{12}$$

where $q(F) = \int p(F|U)q(U)dU$ is a Gaussian distribution since both $p(F|U)$ and $q(U)$ are Gaussian, recall eqs. (5)-(6). Specifically, we have $q(F) = \mathcal{N}(F|K_{XZ}K_{ZZ}^{-1}\mu_u, K_{XX} - K_{XZ}K_{ZZ}^{-1}(K_{ZZ} - \Sigma_u)K_{ZZ}^{-1}K_{ZX})$. The KL divergence $\text{KL}(q(U)||p(U))$ can be calculated in closed-form, since both distributions are Gaussian too. In practice, this term

Algorithm 1 AGP training procedure

Input: Instances $\{\chi_{bi}\}_{i=1,..,N_b}$ (e.g. image patches) for each bag $b = 1, \ldots, B$; bag labels $\{T_b\}$; number of epochs E.
Output: Optimal model parameters $Z, \mu_u, \Sigma_u, \theta, W$.

```
for e = 1 to E (all epochs) do
  for b = 1 to B (all bags) do
    Predict features H_b ← f_fe(X_b).
    Apply fully connected layer in the attention module, i.e. X_b ← f_FC(H_b).
    Calculate SGP output p(F_b|U, Z, X_b) (eq. 6).
    Draw S Monte-Carlo samples F̃_b^s ~ p(F_b|U, Z, X_b).
    Calculate log likelihood (LL) term following eq. (14).
    Calculate KL term in eq. (12) in closed-form.
    Calculate loss as L = −LL + KL.
    Update Z, μ_u, Σ_u, θ, W with ∇L using Adam.
  end for
end for
return Optimal model parameters Z, μ_u, Σ_u, θ, W.
```

acts as a regularizer for the SGP model, since it encourages the posterior on the inducing points to stay close to the prior. To calculate the other term (log-likelihood), we have

$$\mathbb{E}_{q(F)} \log p(T|F) = \textstyle\sum_b \mathbb{E}_{q(F_b)} \log p(T_b|F_b), \tag{13}$$

since we naturally assume that bag labels are independent. Although the terms $\mathbb{E}_{q(F_b)} \log p(T_b|F_b)$ cannot be obtained in closed-form, they can be approximated by Monte Carlo integration:

$$\mathbb{E}_{q(F_b)} \log p(T_b|F_b) \approx \frac{1}{S} \textstyle\sum_s \log p_{T_b}^{(s)}. \tag{14}$$

Here, the subindex T_b indicates that we take the class probability that corresponds to the (observed) bag label T_b (recall from eq. (10) that there exists one p_k for each class). The S samples $\{p_{T_b}^{(s)}\}_s$ are obtained by sampling $F_b^{(s)}$ from the Gaussian $q(F_b)$ with the reparametrization trick [39] and propagating through the rest of the network (that is, applying eqs. (9)-(10)). Notice that maximizing the log-likelihood term is equivalent to minimizing the standard cross-entropy between the estimated class probabilities and the ground truth vector (in a one-hot encoding), as shown in [22].

In summary, AGP training consists in maximizing the ELBO in eq. (12) with respect to the variational parameters (μ_u and Σ_u), the neural network parameters W, the SGP kernel parameters θ and the inducing point locations Z. To do so, we use stochastic optimization with the Adam algorithm and mini-batches [40]. In the experiments, each mini-batch is given by all the instances of one bag. Notice that the proposed model and inference allow for end-to-end training through the ELBO maximization. Algorithm 1 summarizes the training process.

Predictions. After training is completed, we are interested in predicting the class label T_b^* for a previously unseen bag $\mathcal{X}_b^*$. The prediction of the AGP model is given by a K-class categorical distribution with class scores $p_1, \ldots, p_K$ which are random variables. The mean of such random variables, $\overline{p}_k$, represents the predicted probabilities per class (and the predicted class is the one with highest probability, i.e. $T_b^* = \arg\max_k \overline{p}_k$). Additionally, the standard deviation of each class probability provides its degree of uncertainty. The total uncertainty for the bag prediction is defined as the mean of the standard deviations for each class. Notice that this approximation follows popular existing literature [34, 22].

To calculate the predictive random variables p_k, we have

$$p(T_b^*) = \textstyle\int p(T_b^*|F_b) q(F_b) dF_b. \tag{15}$$

Similarly to eq. (13), this cannot be obtained in closed-form, since it requires integrating out the neural network f_{cl}. Following the same idea as there, we approximate the predictive distribution by Monte Carlo sampling. Namely, we take S samples from $q(F_b)$ and propagate them through the rest of the network (eqs. (9)-(10)) to obtain S samples $p_k^{(s)}$ for each class $k = 1, \ldots, K$. The mean and standard deviation for each p_k are computed empirically based on these samples. Analogously, we have S samples $a_{bi}^{(s)}$ for the predictive distribution over the attention weights for each instance inside the bag. We will see that these values provide explainability on which instances are the most relevant to obtain the bag prediction. Using just $S = 20$ samples works well in practice.

Implementation.

Algorithm 1 provides an overview of the implementation. It summarizes the main steps to train the model. In practice, the model is implemented with the deep learning libraries tensorflow (version 2.3.0) and its extension tensorflow-probability (version 0.11.1). The class *tensorflow_probability.layers.VariationalGaussianProcess()* is specially useful for the implementation of the SGP. It allows for an efficient, parallel execution of the algorithm on the GPU, as well as end-to-end training. Another benefit is the easy integration in other existing deep learning projects. The complete code for AGP is publicly available. [2]

Another key component for the implementation is the reparametrization trick to sample from $q(F_b)$, recall eq. (14). The (Gaussian) output distribution of the SGP is 'reparametrized' into one deterministic part and one probabilistic part to perform backpropagation through the probabilistic layer. Namely, the random vector $F_b \sim \mathcal{N}(\mu, \Sigma)$ can be split into $(\mu + L\epsilon) \sim \mathcal{N}(\mu, \Sigma)$, where L is the Cholesky factor of Σ and $\epsilon \sim \mathcal{N}(0, I)$. For MC sampling, a random sample $\hat{\epsilon} \sim \mathcal{N}(0, I)$ is drawn to obtain a sample from F_b, i.e. $\hat{F}_b = \mu + L\hat{\epsilon}$. This allows the backpropagation of the gradient through μ and L, while the random variable ϵ is independent from the model parameters. The reparametrization trick is already implemented in the above mentioned tensorflow-probability library. For further details we refer the interested reader to the original work [39].

3 Synthetic Experiments

In this section we evaluate our method on two synthetic MIL problems. First, Section 3.1 shows a visual experiment based on MNIST that helps better understand the proposed method. Then, Section 3.2 provides a more sophisticated multi-class problem where we compare our method against a wide range of state-of-the-art baselines.

3.1 An illustrative example: MNIST bags

The goal of this section is to illustrate the behavior of AGP in a simple and intuitive example. Specifically, we analyze two aspects of AGP: 1) its predictive performance. We will see that bag-level predictions are correct and high attention weights are given to positive instances. 2) The information provided by the estimated uncertainty (i.e. the standard deviation of the predictions). We will see that high uncertainty is assigned to bags that are difficult to classify.

The well-known MNIST dataset [41] contains 60000 training and 10000 test images. To define a MIL problem, we randomly group these images into bags of 9 instances each. We define the digit "0" as the positive class, and the rest of digits as negative class. Thus, a bag is positive if at least one of the nine digits in that bag is a "0". Otherwise, the bag is negative. We choose "0" because it can be mistaken with "6" or "9". Similar procedures to define a MIL problem on MNIST have been used in previous work [16]. The resulting MIL dataset has 6667 bags for training (2558 negative, 4109 positive), which are made up of 60.000 instances (54077 negative, 5923 positive). For testing it has 1112 bags (404 negative, 708 positive), which are made up of 10.000 instances (9020 negative, 980 positive). For all splits, around 40% of the bags have one positive instance, 17% two, 4% three and 1% four and the rest are negative bags.

As the given problem is less complex than the cancer classification task, we simplify the feature extractor. Namely, it contains one convolutional layer (4 filters, 3x3 convolutions) and one fully connected layer (64 units). The rest of the model remains as described in Figure 1. We train it end-to-end with cross-entropy and the Adam optimizer with a learning rate of 0.0001, for 5 epochs.

Regarding the predictive performance, AGP achieves 98.02% test accuracy at bag level. This is slightly better than when using deterministic attention (i.e. A-Det, which obtains 97.82%). Figures 2a and 2b show the predictions obtained by AGP for a negative and a positive bag, respectively. Notice that AGP learns to discriminate between positive and negative instances, assigning a high attention weight to the digit "0" in the positive bag. Also, notice that the standard deviation for both the attention weights and the bag prediction is low (i.e. the algorithm is confident on the decision).

Next, we analyze the role of the uncertainty by visualizing the prediction for ambiguous bags. In Figures 2c and 2d we see two examples of predictions with high standard deviations. These high standard deviations originate in ambiguous instances that lead to an uncertain final bag prediction. In Figure 2c, there is a '9' which is visually similar to a '0' because one line is (almost) missing. The model assigns a high attention but also a high standard deviation to this instance. The final bag prediction is false positive, but the high standard deviation of the bag prediction indicates a high uncertainty. In Figure 2d, two digits are ambiguous (those corresponding to the items $(1, 2)$ and $(3, 3)$ of the 3×3 matrix of digits). They are assigned a higher attention but again a slightly higher standard deviation than the other instances. The final negative bag prediction is correct, but the high standard deviation reflects a high uncertainty.

[2] `https://github.com/arneschmidt/attention_gp`

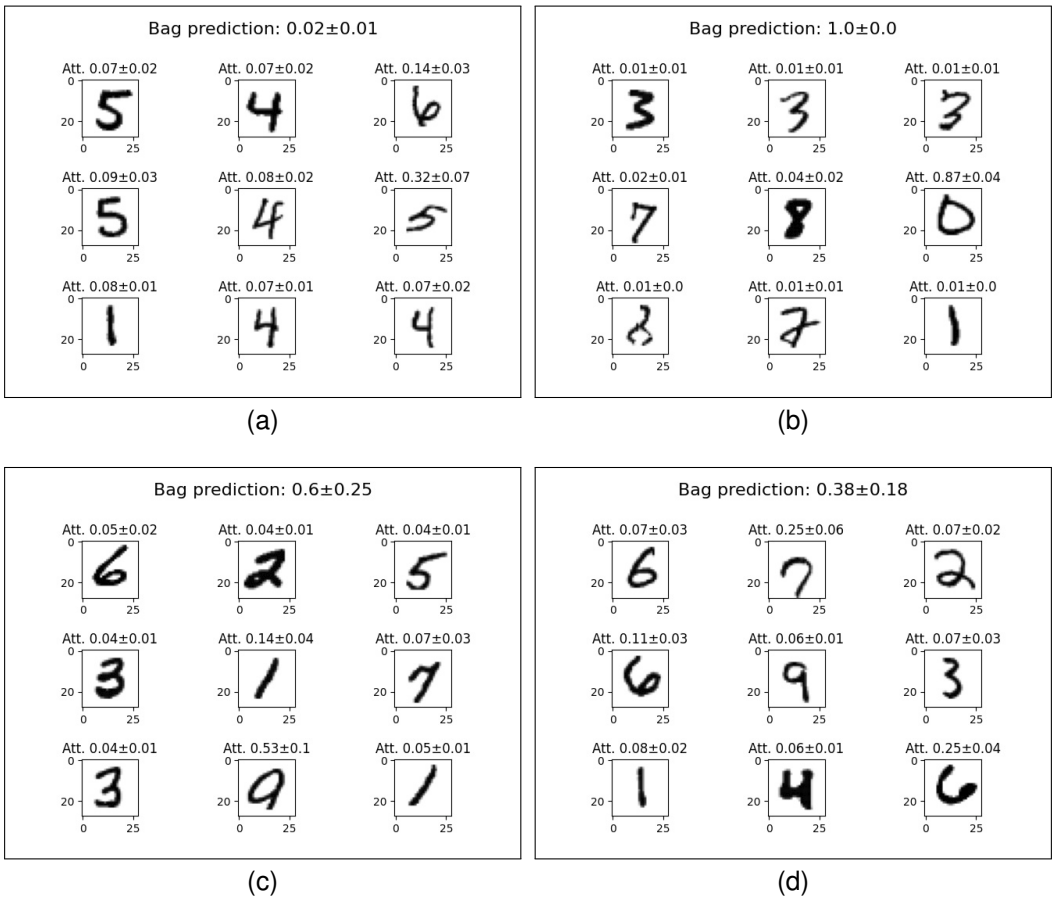

Figure 2: AGP classification of MNIST bags. In this dataset, the number "0" represents a positive instance, and all other digits are considered negative instances. We show the AGP bag prediction (probability to be positive) and corresponding attention weights for each instance. As these estimations are random variables in the AGP model, we report the mean and standard deviation. The top two figures show confident predictions for a negative (a) and a positive (b) bag. The bottom figures, (c) and (d), show two unconfident predictions with high standard deviations. Both bags are negative, but the model misclassifies (c) due to an ambiguous digit.

Finally, notice that this qualitative observation on the uncertainty can also be confirmed statistically: while correctly classified bags have an average standard deviation of 0.006, the average standard deviation of incorrectly classified bags is more than ten times higher (0.065). Therefore, a high standard deviation indicates a high risk of a wrong prediction.

3.2 Evaluation on CIFAR-10

In this experiment with the CIFAR-10 dataset [42] we want to compare different deterministic and probabilistic approaches for a more difficult multi-class MIL problem.

The CIFAR-10 dataset consists of 32x32 images containing 10 different classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks). The dataset contains 10.000 test images and we split the remaining images into 50.000 for training and 10.000 for validation. Originally, all labels of the images are known, but we create a multi-class MIL problem for our use-case. As in the MNIST experiment, we use bags of nine instances each. In this case, we select two positive classes while all other classes are negative, as explained next in more detail. Each bag has either the label 'airplane', 'car' or 'negative'. Negative bags contain only negative instances (i.e. birds, cats, deer, dogs, frogs, horses, ships, and/or trucks). Bags labelled as 'airplane' contain at least one image of airplane, while the remaining instances are negative. Similarly, each bag with the label 'car' contains at least one image of cars (and the rest of instances are negative). We choose to have an equal distribution of each class on the bag-level for training (1481 bags per class,

Table 1: Results for Cifar-10 experiments with bags of 9 images and three classes. The experiment was repeated in 10 independent runs, we report the mean and standard error.

Method	Type	Acc. mean	Acc. S.E.	F1 mean	F1 S.E.
Mean-Agg	Det.	0.732	0.008	0.730	0.009
A-Det	Det.	0.735	0.003	0.735	0.003
A-Det-Gated	Det.	0.730	0.005	0.730	0.005
AGP	Prob.	**0.749**	0.008	**0.750**	0.008

4443 bags in total), validation (370 bags per class, 1110 bags in total) and test (370 bags per class, 1110 bags in total). As the instances are drawn randomly, the exact amount per class vary in this setup. On average, 5.67% of the instances are from the 'airplane' class, 5.67% from the 'car' class, and the rest are negative instances.

We use the model architecture described in section 2 and depicted in Figure 1. For the feature extraction, we choose a CNN backbone that consists of 3x3 convolutions with relu activation and max pooling with a stride of 2x2. The exact layers are: two convolutional layers with 32 filters, max pooling, two convolutional layers with 64 filters, max pooling, two convolutional layers with 128 filters and max pooling. The fully connected layers have 128 and 64 units, respectively. The whole architecture is trained end-to-end. We use the Adam optimizer with a learning rate of 0.0001, cross-entropy, and 15 training epochs.

We compare our method against three state-of-the-art deterministic baselines that only differ in the MIL aggregation mechanism. In all the cases, we use the same feature extractor architecture, hyperparameters and iterations. The compared methods are:

- **Mean Aggregation (Mean-Agg)**. Instead of using an attention module, we aggregate the extracted features from each instance by taking their mean. This mean vector is then used for the final classification.
- **Attention Deterministic (A-Det)**. The attention module as proposed in [16] is used. Their attention weights and the final prediction are deterministic values.
- **Attention Deterministic Gated (A-Det-Gated)**. The advanced attention module proposed in [16] as an extension to A-Det is used. The gating mechanism was introduced to allow the algorithm to efficiently learn more complex relationships between instances.
- **Attention Gaussian Process (AGP)**. The probabilistic model proposed in this work, as described in section 2.

As shown in Table 1, AGP outperforms all the baselines, including the state-of-the-art attention mechanism A-Det-Gated. This suggests that our probabilistic attention is able to accurately assign attention weights to different instances. Indeed, this can be explained by the good regression capabilities of GPs, as known from previous studies [33, 35, 36]. To illustrate the learning process, we plot a training/validation curve of the AGP model in Figure 3. As seen in the plot, the model is robust to overfitting as the validation accuracy remains stable once it converges.

Finally, as a first approach towards future work, we also investigated other options to implement a probabilistic attention mechanism, based on Bayesian Neural Networks (BNNs) instead of the SGP. Leaving the rest of the architecture as shown in Figure 1, we first exchange the AGP attention mechanism by two fully connected Bayesian layers with weights following a Gaussian distribution [43] (32 and 1 units for the layers, respectively). In the same experiment setup, this model achieved 0.642 accuracy and 0.644 F1-score, quite far from AGP. Similarly, we tested a model based on MC dropout [22] with an attention mechanism composed by one fully connected layer with 32 units, a Bayesian dropout layer, and a fully connected layer with 1 unit. The dropout probability was set to 0.5. This model obtained better results, achieving 0.74 accuracy and 0.739 F1-score. However, this is still lower than AGP (0.749 accuracy, 0.750 F1-score), which will be the focus in terms of probabilistic methods in the rest of this paper.

4 Experiments on prostate cancer classification

In this section, we evaluate AGP on the real-world problem of cancer classification. This is a very timely problem, since the development of computer aided diagnosis tools is attracting plenty of attention due to the large workload that pathologists are experiencing in the last years [8, 44]. For easier reproducibility, we use publicly available datasets. We will focus on prostate cancer, although our model is agnostic to the cancer type and can be applied to other cancer classification tasks.

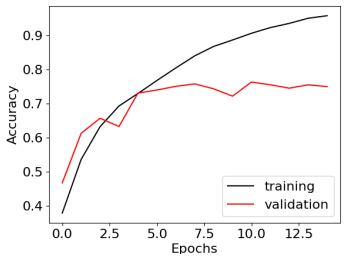

Figure 3: Training and validation accuracy for the AGP model in the CIFAR-10 experiment. Although the amount of labeled bags is low, the model is robust to overfitting as the validation accuracy remains stable.

In the rest of this section, we present the datasets used (SICAPv2 and PANDA), the implementation details, and the baselines used for comparison. Then, Section 4.1 focuses on SICAPv2 data, Section 4.2 focuses on PANDA data, and Section 4.3 evaluates the ability to extrapolate from one dataset to the other. The goal of these experiments is not only to show the strong performance of AGP on real-world data, but also to highlight the usefulness of the probabilistic output to estimate the predictive reliability.

Datasets. We use two publicly available datasets: SICAPv2 and PANDA. The extracted biopsies (WSIs) are classified by pathologists based on the appearance and quantity of cancerous tissue. There are two different scales: the Gleason Score and the ISUP grade. For further background on both scales, we refer the interested reader to [45]. In our experiments, we use the Gleason Score for SICAPv2 and the ISUP grade for PANDA. The reason for this is twofold. First, to compare our results with previous literature (for which we need to use the same grading scale as them). Second, to show that the proposed method is robust to the grading scale used (obtaining good results in both scenarios).

SICAPv2[3] consists of 155 biopsies (WSIs). The class distribution of the assigned Gleason Score (GS) is the following: Non-Cancerous: 36, GS6: 14, GS7: 45, GS8: 18, GS9: 35, GS10: 7. The dataset is already split into four cross-validation folds, which contain between 86 and 97 WSIs for training and a separate set for testing. The publishers of the data distributed the biopsies so that the class proportions are reflected in each of the train and test splits. For more details, see [46]. The PANDA dataset[4] consists of 10616 WSIs and was presented at the MICCAI 2020 conference as a challenge. The total number of WSIs for each ISUP grade is the following: Non-Cancerous: 2892, G1: 2666, G2: 1343, G3: 1242, G4: 1249, G5: 1224. As the test set of PANDA is not publicly available, we use the train/validation/test split proposed in [47], which has 8469, 353 and 1794 WSIs, respectively. Again, each split follows the overall class proportions.

For both datasets, a 10x magnification is used and the WSIs are split into 512x512 patches with a 50% overlap. The patch-level annotations of the datasets are discarded for our experiments, since our model only requires bag labels for training.

Implementation Details. The AGP architecture used for these experiments is depicted in Figure 1, and explained in Section 2.4. Here we provide the rest of the details for full reproducibility. We use 64 inducing points with 32 dimensions each, whose locations are initialized with random values between 0.3 and 0.7 (because this is typically the range of values initially obtained by the previous sigmoid layer). For Monte-Carlo integration, we draw 20 samples for training and for testing. We use a class balanced loss with cross-entropy and the Adam optimization algorithm. We set the learning rate to 0.001 for the first 10 epochs, and use learning rate decay afterwards with the factor $e^{-0.1}$ per epoch. The total number training epochs is set to 100. Finally, as it is computationally unfeasible to train the feature extractor on all patches of a (huge) WSI at once, we first extract the high-level features of each patch. We train the CNN and the first fully connected layer with the method proposed in [48], using only WSI labels. The obtained 128 dimensional feature vectors per patch are then used to train the last fully connected layer of the feature extractor and the rest of the model.

Baselines. In all the experiments, we compare mean aggregation as a baseline and three state-of-the-art MIL approaches trained with the same feature vectors, hyperparameters and iterations. The only variation is the attention mechanism. Additionally, in each experiment we compare with other related approaches that have used the same data. For details about the approaches (Mean-Agg, A-Det, A-Det-Gated) please recall the bullet points in Section 3.2.

[3] Available at: https://data.mendeley.com/datasets/9xxm58dvs3/1

[4] Available at: https://www.kaggle.com/c/prostate-cancer-grade-assessment

Table 2: Ablation studies with the SICAPv2 dataset. We study the effect of three important components: the feature vector dimension, the number of inducing points for the SGP, and the activation function used inside the attention module. Bold letters indicate the configuration used in the final setup.

Method	κ mean	κ S.E.
AGP-feat-dim-32	0.832	0.004
AGP-feat-dim-64	**0.847**	**0.001**
AGP-feat-dim-128	0.835	0.001
AGP-feat-dim-256	0.844	0.001
AGP-ind-points-16	0.832	0.004
AGP-ind-points-32	0.840	0.002
AGP-ind-points-64	**0.847**	**0.001**
AGP-ind-points-128	0.689	0.007
AGP-relu	0.675	0.008
AGP-sigmoid	**0.847**	**0.001**
AGP-tanh	0.830	0.007

Table 3: Results for SICAPv2 dataset. We report the mean and standard error of Cohen's quadratic kappa (κ) for 4 independent runs, with a four-fold cross-validation in each run. The last two methods do not report the standard error.

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.800	0.041
A-Det	MIL	0.770	0.008
A-Det-Gated	MIL	0.814	0.007
AGP	MIL	**0.847**	0.001
Arvaniti et al. [50] [49]	Supervised	0.769	N.A.
Silva-Rodríguez et al. [49]	Supervised	0.818	N.A.

Evaluation metric. As common in prostate cancer classification tasks [49, 47, 50], we report the performance in terms of quadratic Cohen's kappa, which measures the agreement between the labels provided by pathologists and the model's predictions. A kappa value of 0 means no agreement (random predictions) and a kappa value of 1 means complete agreement. In all cases, we show the mean and the standard error of the results over several independent runs.

4.1 SICAPv2 Results

The SICAPv2 experiment is used to test our model on a very small dataset, where there is a high risk of overfitting. Recall that the training set for each cross-validation fold has less than 100 WSIs, and correspondingly there are less than 100 labels for the MIL models.

As a first part of the SICAPv2 experiment, we perform an ablation study to show the effect of different hyperparameters that are important in the proposed attention module. We identify three hyperparameters that are especially interesting for this analysis: the dimension of the feature vectors h, the number of inducing points of the SGP, and the activation function inside the attention module (the one before the SGP). We separately vary these hyperparameters while all the other values are set to default (as described in Section 4, *Implementation Details* paragraph).

As shown in Table 2, the AGP model is robust against variations of the feature vector dimensionality, but the model with 64 feature dimensions slightly outperforms the others. For the number of inducing points we see a robust performance for less inducing points (16-32-64), but a high number (128) led to instabilities in model training. Indeed, we had to reduce the learning rate to 0.0001 (from 0.001) to obtain convergence, but we still observe a remarkable performance drop. We believe that, as fully connected layers have reduced the complexity and dimensionality of the features, a relatively small amount of inducing points allows precise predictions. Finally, the same convergence problems appear if a ReLu activation function is used before the SGP (we again used a lower learning rate of 0.0001 in this case to get convergence). The tanh function, which is more similar to the sigmoid function, shows a robust performance. Interestingly, these results suggest that limiting the input range to the SGP is clearly beneficial, as the

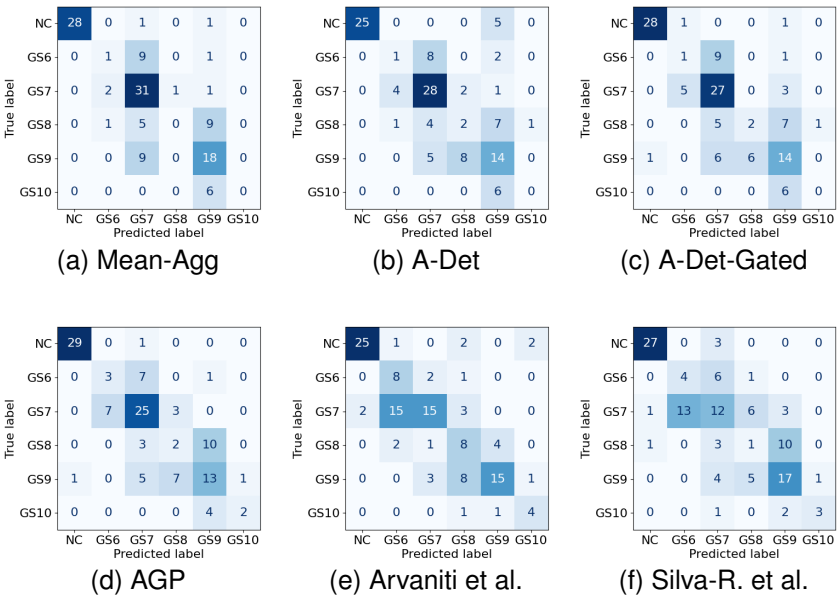

Figure 4: Confusion matrices for the 4-fold cross-validation of SICAPv2 with the classes "non-cancerous" (NC) and Gleason Score 6 to 10 (GS6-GS10).

sigmoid and tanh functions output values in the range $(0, 1)$ and $(-1, 1)$, respectively. The relu function has outputs in the range $[0, \infty)$, which makes it harder to find adequate inducing point locations for the SGP.

After the ablation study, the best performing model is compared to other state-of-the-art methods. Notice that the best performing configuration is precisely the one that was described in Section 4, *Implementation Details* paragraph. The results in Table 3 show that the AGP model outperforms all other MIL approaches, including existing attention based methods A-Det and A-Det-Gated, which can be considered state of the art for MIL. Also, the Cohen's quadratic kappa value of 0.847 is a remarkable one for such a small dataset. Furthermore, we see that AGP outperforms the existing supervised methods Silva-Rodríguez et al. [49] and Arvaniti et al. [50], which use all patch-level annotations to train the feature extractor (reported by [49] for this dataset). This can be partly explained by the stronger focus on patch-level predictions instead of bag-level predictions of these approaches (for bag-level predictions, they implement a simple aggregation method). Interestingly, notice that the AGP model provides an accurate WSI diagnosis without local annotations.

We also report the confusion matrices for all the compared models, see Figure 4. We see that AGP is strong in distinguishing cancerous from non-cancerous WSIs: there is only one false positive and one false negative in AGP predictions. The other two approaches that show a comparable performance in the binary (cancerous vs non-cancerous) classification task, Mean-Agg and A-Det-Gated, show a major systematic error: they misclassified all the WSIs with Gleason Score 10.

The training process of the proposed AGP model (with previously extracted features) takes less than 2 minutes for the SICAPv2 dataset, and is negligible in comparison to the training of the feature extractor ($\sim$ 7 hours in this case [48]). The test time is on average 2.1 seconds for a complete WSI. This computing time mainly corresponds to the feature extraction ($\sim$ 7 seconds [48]), while the attention mechanism and final classification together can be executed in less than 0.1 seconds per WSI. The runtime bottleneck is therefore the feature extractor, and not the proposed probabilistic attention module. This efficiency is an important benefit for the clinical practice, as well as other areas where speed in prediction is paramount.

4.2 PANDA Results

In this experiment, we show that AGP also outperforms other approaches in a large real-world problem. Moreover, by visually inspecting the predictions, we check that the AGP attention mechanism allows for identifying cancerous regions. Finally, we analyze the relevance of the probabilistic predictions provided by AGP, which can be used to detect wrong predictions. Also, as explained before, the different grading scale used here (ISUP scale) shows the robustness of AGP.

Table 4: Results for PANDA dataset. We report the mean and standard error of Cohen's quadratic kappa (κ) for 4 independent runs. The last method does not report the standard error.

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.803	0.003
A-Det	MIL	0.811	0.004
A-Det-Gated	MIL	**0.816**	0.004
AGP	MIL	**0.817**	0.003
Silva-Rodríguez et al. [47]	MIL	0.793	N.A.

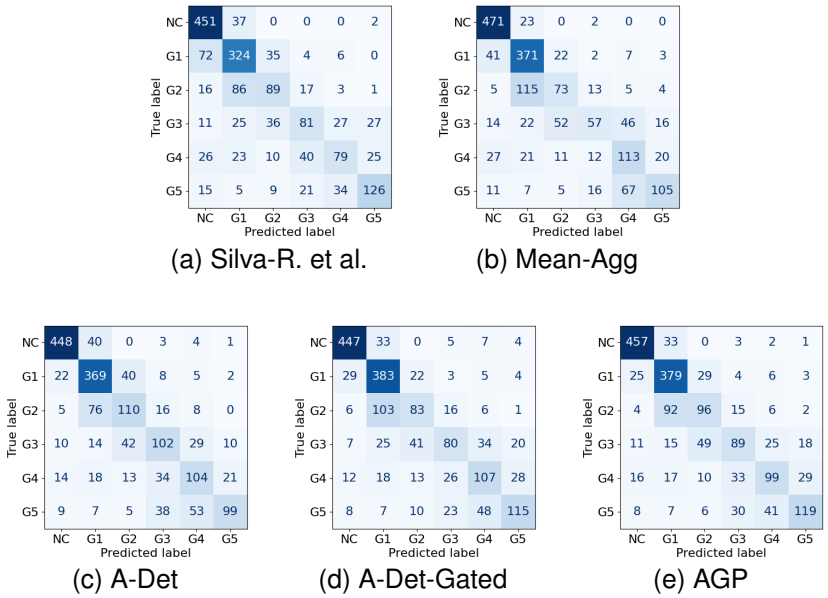

Figure 5: Confusion matrices for the PANDA test set. The six classes are "non-cancerous" (NC) and ISUP grades 1 to 5 (G1-G5).

In Table 4 we see that the AGP performance is superior to the other MIL approaches. We have included Silva-Rodríguez et al. [49], a recent MIL method that was evaluated on the same dataset. Although the difference in performance is small in some cases (e.g. against A-Det-Gated), we will see that the probabilistic nature of AGP provides additional benefits (such as the degree of uncertainty on the predictions).

The confusion matrices, see Figure 5, show again that the AGP model is very strong at differentiating cancerous from non-cancerous WSIs. Although the models of Silva-Rodriguez et al. [49] and Mean-Agg have less false positives (see the first row of the matrices), these models suffer from more false negatives (see first column of the matrices). This can also be confirmed by the binary F1 score (cancerous vs. non-cancerous): AGP outperforms all other approaches with a binary F1 score of 0.960 (Silva-R: 0.927, Mean-Agg: 0.950, A-Det: 0.958, A-Det-Gated: 0.956).

Next, we illustrate how the AGP attention mechanism provides an explainable prediction at instance level. The top row in Figure 6 shows one test WSI example (it has two pieces of tissue, a big one on the left and a small one on the right). In the second row, the cancerous areas are colored in green (this example has been manually segmented by an expert pathologist for this evaluation; recall that AGP only uses bag-level labels). The third row shows the areas with high attention weights predicted by AGP highlighted in green. For this figure, the attention weights were predicted by the model for each patch of the image (remember that the patches are of 512x512 resolution with 50% overlap). To obtain the heatmap for the complete WSI, linear interpolation was performed between the grid of patches. We find that the parts with high attention correspond to areas of the WSI that are most affected by cancer. Other parts that are non-cancerous, such as the whole piece on the right, are not assigned high attention weights. This means that AGP works as expected and the final prediction is based on discriminative areas. The attention helps the pathologist verify the prediction and further inspect the affected tissue. Moreover, it might even point out cancerous regions that the pathologist may have missed.

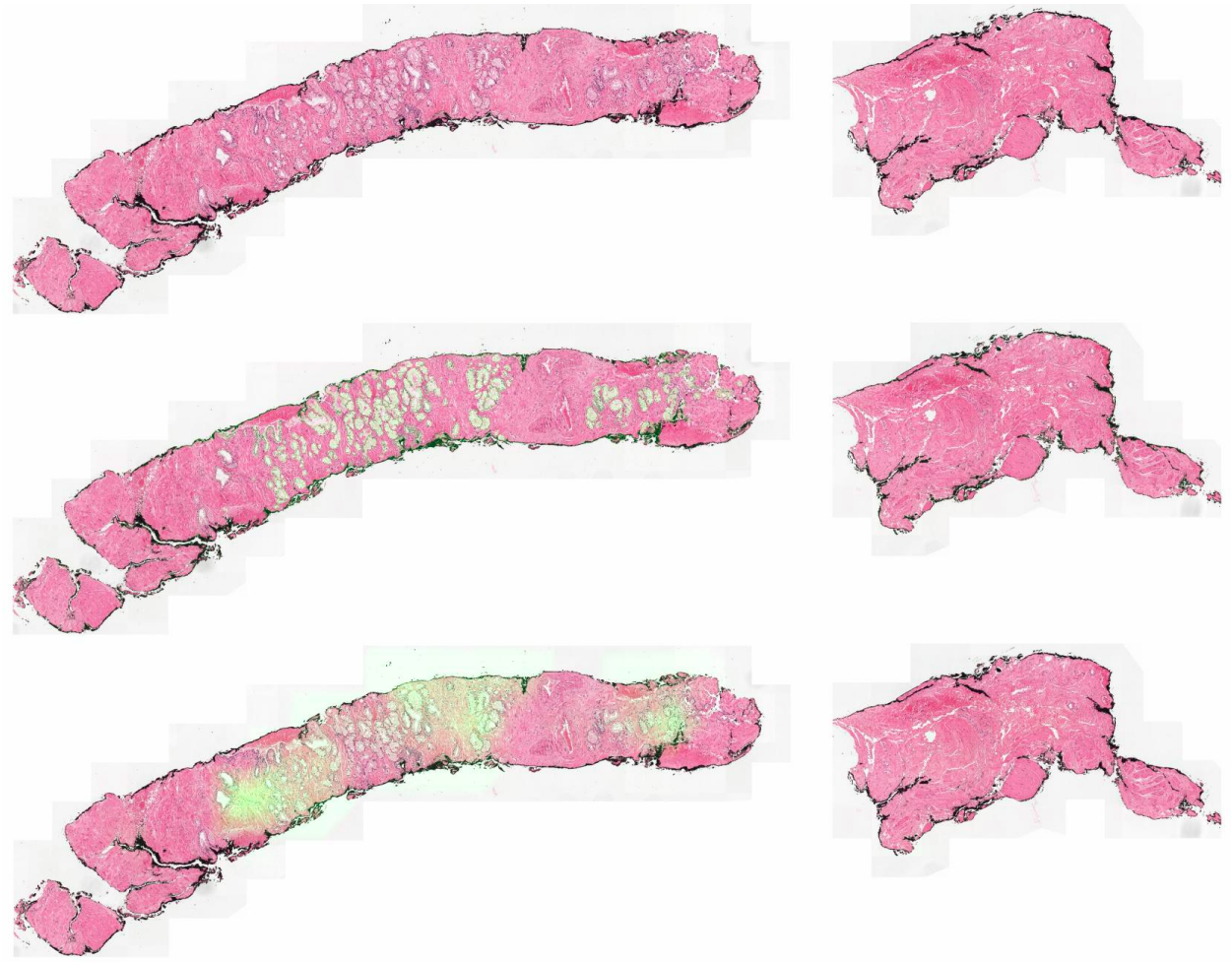

Figure 6: Attention weights for a test WSI of the PANDA dataset. The top image shows the original WSI, the middle image shows the cancerous areas marked by an expert pathologist (in green) and the bottom image shows the areas of high attention as predicted by the model (in green). The predicted attention weights were normalized and interpolated from the patch coordinates by linear interpolation. As can be seen in the image, the model successfully assigns high attention weights to the discriminative areas of the WSI. This improves explainability of the models prediction and helps the pathologist to find suspicious areas.

Although the attention mechanism works well for most WSIs (as supported by the superior predictive performance), we observed that for some WSIs the attention does not capture all the important areas. In Figure 7 we show the same plots as in Figure 6 for a failure case of the attention mechanism (top: WSI, middle: annotation, bottom: attention estimation). In this case, not all the cancerous areas are assigned a high attention weight. However, the attention is still useful here as a source of explainability. It highlights all the areas which the classification is based on. Notice also that the highlighted areas are indeed tumorous, and the correct class (ISUP grade 4) is predicted.

Finally, we focus on the probabilistic bag predictions that provide not only a class score, given by the mean, but also the predictive uncertainty, given by the standard deviation. Figure 8 shows a histogram over the predictive uncertainty (i.e. the standard deviation) for all the AGP bag predictions. The green (resp. red) bars indicate the number of correctly (resp. incorrectly) predicted bags whose standard deviation falls in a certain range. It is clearly visible that correct predictions tend to have a lower standard deviation than the incorrectly classified bags. In other words: a high standard deviation correlates with a high risk of a wrong classification. This suggests that the standard deviation provides a useful measure of the predictive reliability. In fact, if we only take the reliable predictions with a std. below 0.02, the Cohen's kappa value for the PANDA test set rises to 0.864 for the AGP model (from 0.817 in Table 4). Therefore, in practice, the uncertainty estimation can help the pathologists decide when the models' prediction should be disregarded or double-checked.

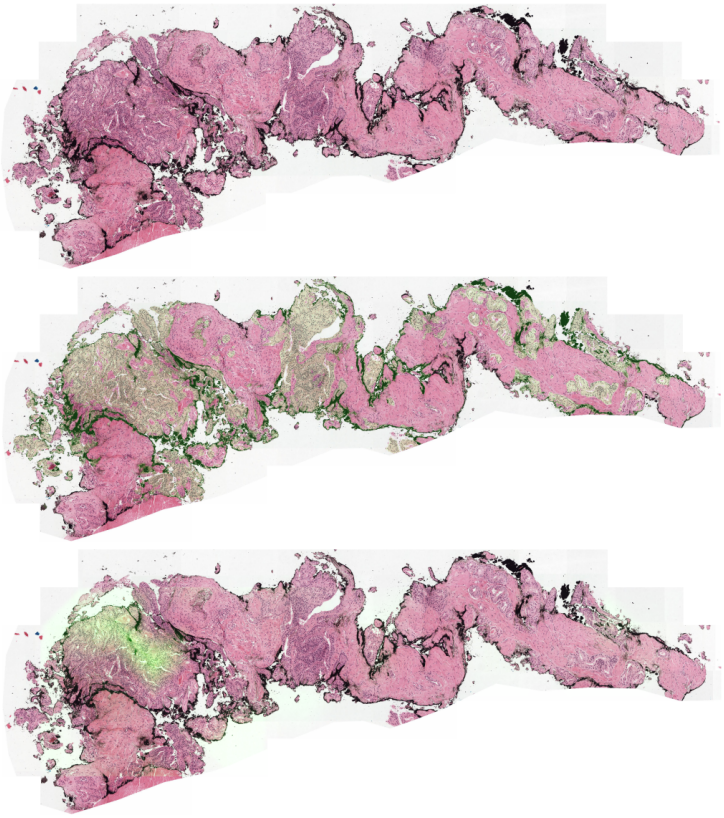

Figure 7: Example of an inaccurate assignment of attention weights. The top image shows the original WSI, the middle image shows the cancerous areas marked by an expert pathologist (in green), and the bottom image shows the areas of high attention as predicted by the model (in green). Although in most cases the areas of high attention correspond to the tumorous areas, recall Figure 6, we observed some inaccurate cases as the presented in this image. Here, the correct class (ISUP grade 4) is predicted, but the attention mechanism does not capture all the cancerous tissue parts.

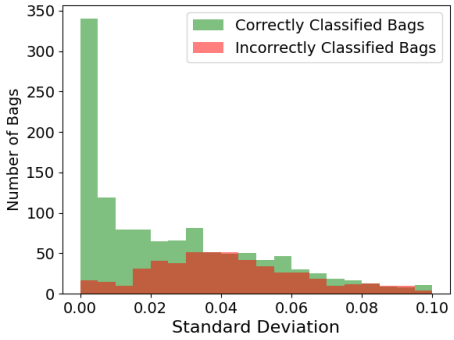

Figure 8: Distribution of the predicted standard deviations. Bags (images) with a low standard deviation are likely to be classified correctly, while a high standard deviation indicates a high risk of a wrong prediction. The standard deviations are divided into bins of width 0.005, and the y-axis shows the number of bags with the corresponding std.

Table 5: Models trained on the PANDA dataset, tested on SICAPv2 with ISUP grading scale. The mean and standard error of Cohen's quadratic kappa (κ) are reported for four independent test runs. The last method does not report the standard error.

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.911	0.007
A-Det	MIL	0.903	0.004
A-Det-Gated	MIL	0.910	0.007
AGP	MIL	**0.920**	0.001
Silva-Rodríguez et al. [47]	MIL	0.885	N.A.

Table 6: The average standard deviation of correct and wrong bag predictions for AGP trained on the PANDA dataset. The first (resp. second) row shows the result when using PANDA (resp. SICAPv2) as test set. The values, which reflect the uncertainty in the prediction, get higher for wrong predictions and when testing on a different test set.

	Average uncertainty	
Test set	Correct predictions	Wrong predictions
PANDA	0.029	0.046
SICAPv2	0.045	0.051

4.3 External Validation with PANDA and SICAPv2

The third experiment tests the generalization capability for the models evaluated in Section 4.2. Similar to [47], we take the models trained on the PANDA dataset, and use all images of SICAPv2 as an external test set (since the training is done under ISUP grading, SICAPv2 uses ISUP grading here too; therefore, results are not comparable to those obtained in Section 4.1). Table 5 shows the results. Again, AGP outperforms the rest of approaches, and achieves a remarkable Cohen's quadratic kappa value of 0.92.

Similar to Section 4.2, it is worth analyzing the standard deviation of the predictions. In addition to distinguishing between correctly and incorrectly classified images, there is now an additional dimension to consider: whether the test images follow the same distribution as the training ones or not (i.e. whether we are using PANDA or SICAPv2 as test set, respectively).

Table 6 shows the average predicted standard deviation of correct and wrong predictions for each of the test sets. The main findings are twofold: (1) the output standard deviation is higher for wrong predictions, and (2) the output standard deviation is higher for data originating from a different data distribution. This shows that the output uncertainty is not only helpful to identify model failures for in-distribution data, but also accounts for uncertainty added by a data shift. [51] This can help to determine images that might be out of scope for the model and should be handled with caution.

5 Conclusions

We have proposed AGP, a novel probabilistic attention mechanism based on GPs for deep multiple instance learning (MIL). We have evaluated AGP in a wide range of experiments, including real-world cancer detection tasks. The novel attention module is capable of accurately assigning attention weights to the instances, and outperforms state-of-the-art deterministic attention modules. Furthermore, it provides important advantages due to its probabilistic nature. For instance, it addresses the problem of reliability in safety critical environments such as medicine: the probabilistic output of our model can be used to estimate the uncertainty on each prediction. For future research, we plan to explore the use of deep GPs (instead of GPs) to further improve the performance. Also, the promising results of AGP encourage the application of GP-based attention to other recent methods such as transformer networks, self-attention or channel attention. Moreover, alternative probabilistic attention mechanisms based on other Bayesian approaches (instead of GPs) can be explored.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[2] Z. Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[3] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2063–2079, 2018.

[4] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2021.

[5] P. Morales-Álvarez, P. Ruiz, S. Coughlin, R. Molina, and A. K. Katsaggelos, "Scalable variational gaussian processes for crowdsourcing: Glitch detection in ligo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1534–1551, 2022.

[6] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 213–234, 2017.

[7] J. Melendez, B. van Ginneken, P. Maduskar, R. H. H. M. Philipsen, K. Reither, M. Breuninger, I. M. O. Adetifa, R. Maane, H. Ayles, and C. I. Sánchez, "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 179–192, 2015.

[8] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019. [Online]. Available: http://www.nature.com/articles/s41591-019-0508-1

[9] Y. Wu, A. Schmidt, E. Hernández-Sánchez, R. Molina, and A. K. Katsaggelos, "Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2021, pp. 582–591.

[10] O. Ciga, T. Xu, S. Nofech-Mozes, S. Noy, F.-I. Lu, and A. L. Martel, "Overcoming the limitations of patch-based learning to detect cancer in whole slide images," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 15, 2003. [Online]. Available: https://proceedings.neurips.cc/paper/2002/file/3e6260b81898beacda3d16db379ed329-Paper.pdf

[12] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Conference on Neural Information Processing Systems - NIPS*, vol. 14, 2002. [Online]. Available: https://proceedings.neurips.cc/paper/2001/file/e4dd5528f7596dcdf871aa55cfccc53c-Paper.pdf

[13] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *International Conference on Machine Learning - ICML*, 2009, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1553374.1553534

[14] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Asian Conference on Machine Learning - ACML*, ser. Proceedings of Machine Learning Research, vol. 95, 2018, pp. 662–677. [Online]. Available: https://proceedings.mlr.press/v95/yan18a.html

[15] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[16] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning - ICML*, 2018, pp. 2127–2136.

[17] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2021, pp. 14 313–14 323. [Online]. Available: https://ieeexplore.ieee.org/document/9578683/

[18] 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 Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021. [Online]. Available: https://doi.org/10.1038/s41551-020-00682-w

[19] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Weakly labelled AudioSet tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019. [Online]. Available: http://arxiv.org/abs/1903.00765

[20] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[21] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3033–3044, 2020.

[22] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning - ICML*, vol. 48, 2016, pp. 1050–1059.

[23] V. B. Alex Kendall and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference - BMVC*, 2017, pp. 57.1–57.12. [Online]. Available: https://dx.doi.org/10.5244/C.31.57

[24] P. Ruiz, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos, "Learning from crowds with variational gaussian processes," *Pattern Recognition*, vol. 88, pp. 298–311, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318304060

[25] P. Morales-Álvarez, D. Hernández-Lobato, R. Molina, and J. M. Hernández-Lobato, "Activation-level uncertainty in deep neural networks," in *International Conference on Learning Representations - ICLR*, 2020.

[26] M. Kim and F. De la Torre, "Multiple instance learning via gaussian processes," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 1078–1106, 2014. [Online]. Available: https://doi.org/10.1007/s10618-013-0333-y

[27] M. Haussmann, F. A. Hamprecht, and M. Kandemir, "Variational bayesian multiple instance learning with gaussian processes," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2017, pp. 810–819. [Online]. Available: https://ieeexplore.ieee.org/document/8099576/

[28] K. Chen and C.-G. Lee, "Attentive gaussian processes for probabilistic time-series generation," *ArXiv*, vol. abs/2102.05208, 2021.

[29] J. Xie, Z. Ma, D. Chang, G. Zhang, and J. Guo, "GPCA: A probabilistic framework for gaussian process embedded channel attention," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2021.

[30] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press, 2006.

[31] P. Morales-Álvarez, A. Pérez-Suay, R. Molina, and G. Camps-Valls, "Remote sensing image classification with large-scale gaussian processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1103–1114, 2017.

[32] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 68–81, 2020.

[33] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, vol. 18, 2006. [Online]. Available: https://proceedings.neurips.cc/paper/2005/file/4491777b1aa8b5b32c2e8666dbe1a495-Paper.pdf

[34] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," in *International Conference on Artificial Intelligence and Statistics - AIStat*, vol. 38, 2015, pp. 351–360. [Online]. Available: https://proceedings.mlr.press/v38/hensman15.html

[35] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When gaussian process meets big data: A review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

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

[37] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning - ICML*, 2019.

[38] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[39] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *International Conference on Neural Information Processing Systems - NIPS*, 2015, p. 2575–2583.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations - ICLR*, 2015.

[41] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[42] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[43] M. Vladimirova, J. J. Verbeek, P. Mesejo, and J. Arbel, "Understanding priors in bayesian neural networks at the unit level," in *ICML*, 2019.

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

[45] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey, "The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system," *American Journal of Surgical Pathology*, vol. 40, no. 2, pp. 244–252, 2016. [Online]. Available: https://journals.lww.com/00000478-201602000-00010

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

[47] J. Silva-Rodríguez, A. Colomer, J. Dolz, and V. Naranjo, "Self-learning for weakly supervised gleason grading of local patterns," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3094–3104, 2021. [Online]. Available: http://arxiv.org/abs/2105.10420

[48] A. Schmidt, J. Silva-Rodríguez, R. Molina, and V. Naranjo, "Efficient cancer classification by coupling semi supervised and multiple instance learning," *IEEE Access*, vol. 10, pp. 9763–9773, January 2022.

[49] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105637, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016926072031470X

[50] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports*, vol. 8, no. 1, p. 12054, 2018. [Online]. Available: http://www.nature.com/articles/s41598-018-30535-1

[51] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems - NeurIPS*, vol. 32, 2019.

Arne Schmidt studied Mathematics and received his Bachelor degree at the Freie Universität Berlin in 2015 and his Master degree at the Technische Universität Berlin in 2018. He worked as a software engineer specialized in deep learning at Astrofein GmbH, Fraunhofer Heinrich Hertz Institut and TomTom. In 2020, he started his Ph.D. under the supervision of Prof. Rafael Molina at the University of Granada. It is part of the EU-funded H2020-MSCA-ITN CLARIFY, an interdisciplinary, international project on digital pathology for cancer classification. His research interests are probabilistic deep learning, multiple instance learning, active learning and crowdsourcing with applications to medical images.

Pablo Morales-Álvarez received the B.Sc. degree in Mathematics from the University of Granada (UGR), Spain, in 2014. He obtained the First End of Studies Award by the Spanish Ministry of Science to the most outstanding undergraduate in mathematics in Spain. Then, he received the M.Sc. degrees in Mathematical Physics (2015) and Data Science (2016), both from UGR. Funded by the highly competitive Ph.D. fellowship from La Caixa Banking Foundation, he got his Ph.D. degree in 2020 at UGR under the supervision of Prof. Rafael Molina (UGR) and Prof. Aggelos K. Katsaggelos (Northwestern University, USA). During his PhD he visited several research groups, including the Machine Learning Group at the University of Cambridge (UK), 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 (UK) with Cheng Zhang. He has published in top machine learning conferences and journals, such as ICLR 2021, NeurIPS 2022 and IEEE Trans. PAMI. He has also served as a reviewer for top-tier conferences (ICML 2020, NeurIPS 2020, ICLR 2021, NeurIPS 2021), and journals (Science). He has obtained the SCIE-FBBVA Award 2022 to the Most Outstanding Young Spanish Researchers in Computer Science. His main research interests are probabilistic machine learning methods, specially (Deep) Gaussian Processes and Bayesian Neural Networks.

Rafael Molina received the M.Sc. degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from the University of Granada, Granada, Spain, in 1979 and 1983, respectively. He was the Dean of the Computer Engineering School, University of Granada, from 1992 to 2002, where he became a Professor of computer science and artificial intelligence in 2000. He was the Head of the Computer Science and Artificial Intelligence Department, University of Granada, from 2005 to 2007. He has coauthored an article that received the runner-up prize at

reception for early stage researchers at the House of Commons in 2007. He has coauthored an awarded Best Student Paper at the IEEE International Conference on Image Processing in 2007, the ISPA Best Paper in 2009, and the EUSIPCO 2013 Best Student Paper. His research interest focuses mainly on using Bayesian modeling and inference in image restoration (applications to astronomy and medicine), super-resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, supervised learning, and crowdsourcing. Dr. Molina has served as an Associate Editor for Applied Signal Processing from 2005 to 2007 and the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He has been serving as an Area Editor for Digital Signal Processing since 2011.