A New Optical Density Granulometry-Based Descriptor for the Classification of Prostate Histological Images Using Shallow and Deep Gaussian Processes

Ángel E. Esteban[a], Miguel López-Pérez[b,*], Adrián Colomer[a], María A. Sales[c], Rafael Molina[b], Valery Naranjo[a]

[a] *Institute of Research and Innovation in Bioengineering, I3B, Polytechnic University of Valencia, Valencia, Spain*
[b] *Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain*
[c] *Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain*

Abstract

Background and objective:

Prostate cancer is one of the most common male tumors. The increasing use of whole slide digital scanners has led to an enormous interest in the application of machine learning techniques to histopathological image classification. Here we introduce a novel family of morphological descriptors which, extracted in the appropriate image space and combined with shallow and deep Gaussian process based classifiers, improves early prostate cancer diagnosis.

Method:

We decompose the acquired RGB image in its RGB and optical density hematoxylin and eosin components. Then, we define two novel granulometry-based descriptors which work in both, RGB and optical density, spaces but perform better when used on the latter. In this space they clearly encapsulate knowledge used by pathologists to identify cancer lesions. The obtained features become

*Corresponding author. The first two authors contributed equally.
This work was supported by the the Spanish Ministry of Economy and Competitiveness through project DPI2016-77869. The Titan V used for this research was donated by the NVIDIA Corporation.
Email addresses: ngeesav@i3b.upv.es (Ángel E. Esteban), mlopez@decsai.ugr.es (Miguel López-Pérez), adcogra@i3b.upv.es (Adrián Colomer), salesman@gva.es (María A. Sales), rms@decsai.ugr.es (Rafael Molina), vnaranjo@dcom.upv.es (Valery Naranjo)

the inputs to shallow and deep Gaussian process classifiers which achieve an accurate prediction of cancer.

Results:

We have used a real and unique dataset. The dataset is composed of 60 Whole Slide Images. For a five fold cross validation, shallow and deep Gaussian Processes obtain area under ROC curve values higher than 0.98. They outperform current state of the art patch based shallow classifiers and are very competitive to the best performing deep learning method. Models were also compared on 17 Whole Slide test Images using the FROC curve. With the cost of one false positive, the best performing method, the one layer Gaussian process, identifies 83.87% (sensitivity) of all annotated cancer in the Whole Slide Image. This result corroborates the quality of the extracted features, no more than a layer is needed to achieve excellent generalization results.

Conclusion:

Two new descriptors to extract morphological features from histological images have been proposed. They collect very relevant information for cancer detection. From these descriptors, shallow and deep Gaussian Processes are capable of extracting the complex structure of prostate histological images. The new space/descriptor/classifier paradigm outperforms state-of-art shallow classifiers. Furthermore, despite being much simpler, it is competitive to state-of-art CNN architectures both on the proposed SICAPv1 database and on an external database.

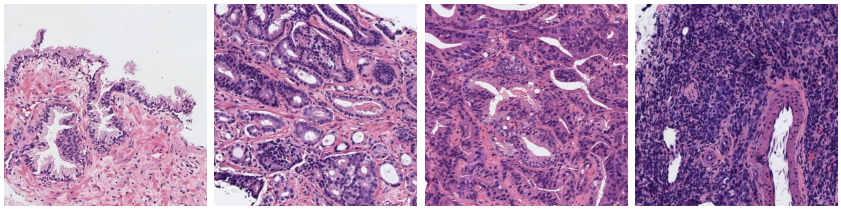
Keywords: Prostate cancer, Histopathological Images, Gaussian Processes, Variational Inference, Granulometries, Deep Gaussian Processes.

---

1. Introduction

According to the World Health Organization, prostate cancer is the most common non-cutaneous cancer in men [1]. A histological diagnosis of prostate cancer is almost always required prior to instituting therapy for any stage of the disease. Pathologists determine the grade of cancer based on the formation,

disposition, and structure of the glands (nuclei, lumen, cytoplasm and stroma) in the tissue, scoring the samples between 1 to 5, following the Gleason grading system [2], see Figure 1.



(a) (b) (c) (d)

Figure 1: Examples of Gleason grades of histological images: (a) benign; (b) grade 3; (c) grade 4; (d) grade 5.

Tissue histopathological slides can nowadays be acquired and digitally stored thanks to the advent of whole slide digital scanners. The widespread use of such scanners has led to an increasing interest on applying machine learning techniques to classify these images, for a review of this topic, see [3]. Due to the large resolution of the images obtained under the microscope, evaluating each single diagnostic test manually is a very time-consuming task. This fact encourages the research on CAD algorithms that decrease pathologists workload by recognizing obviously benign cases so that experts can focus on the delicate ones [4].

In digital brightfield microscopy, tissues are usually stained before digitization and evaluation by pathologists. Hematoxylin and Eosin (H&E) are probably the most widely used combination of stains. Since Color Deconvolution (CD), that is, H&E separation, is a very important preprocessing step, several methods have been developed (see [5] for a recent review). One of the first CD methods, which is widely used, was proposed by Ruifrok et al. [6]. This is a supervised method where the stain color vectors are obtained by measuring the relative absorption of each stain in single-stained images. These color vectors are used on all the WSI images to obtain their RGB and Optical Density (OD) space H&E images. CAD algorithms based on hand-driven approaches use RGB

space H&E images, while deep learning approaches work directly with the orignal RGB images. In this paper we will show that the selection of the space where H&E are represented significantly affects the performance of classifiers.

Two approaches are currently being used in the literature to detect tumorous prostatic tissues. One is based on segmenting the images and identifying the regions of interest (ROIs), while the other utilizes patches for classification purposes. In this work, we follow the second approach: the entire whole slide image (WSI) is split into patches and each one is analyzed independently. While pathologists use several scales (magnification factors), most machine learning algorithms use a single one. Gupta et al. [7] compare different scales for training and test in breast histology. They conclude that with suitable features together with an ensemble classifier framework, such as bagging or boosting, the classification can be made largely magnification invariant. For a selected magnification factor and patch size, a feature extraction process to encode the relevant information of the images must be carried out.

Nowadays, the remarkable progress in the deep learning field allows to automatically compute high-abstraction feature maps by means of neural networks based on stacks of convolutional blocks (a.k.a. convolutional neural networks or CNNs). CNNs are being successfully applied in many computer vision tasks. In the particular case of histological images, CNNs have also benefit of the automatic feature extraction for the classification of different tumoral patterns in diverse organs [8]. Le Hou et al. [9] use a CNN for path-based classification which achieves good results discriminating different cancer subtypes in WSIs. The BACH challenge[1] resulted in several works [10, 11, 12] in which the different types of breast cancer including in-situ carcinoma, invasive tumor, and benign tumor were automatically identified by means of well-known CNN architectures: Inception v3, Xception and ResNet. A fine-tuning process of the same architectures was carried out by Ferlaino et al. [13] to robustly localize and classify placental cells using histological images. Shallu et al. [14] demonstrated

[1] https://iciar2018-challenge.grand-challenge.org/

that transfer learning is better than training from scratch in breast cancer histological image classification, obtaining very good accuracy with the VGG16 and VGG19 architectures. In prostate cancer histology, CNNs have recently been utilized for semantic segmentation grading [15, 16]. These methods provide for each pixel its probability of belonging to each class.

According to Komura et al. [3], the relevant information to classify histological images is related to texture and morphology. Although CNNs are able to learn these feature representations, textural and morphological tissue properties can also be manually captured by a suitable hand-crafted descriptor avoiding specific hardware requirements and reducing computational cost. Therefore, the information (descriptors) extracted from each patch becomes the key to a successful tissue classification. Generic descriptors, such as HOG [17], LBP [18], SIFT [19] or Gabor filters [20] are frequently used for prostate cancer detection. Kumar et al. [21] show that LBP are as good as deep features and dictionaries with the benefits of easy computation and low dimensionality. Recent works in the field [22, 23, 24, 25] also indicate that descriptors based on structural and morphological properties of the prostatic tissue could outperform those based on standard features. It is also possible to combine a convolutional neural network with handcrafted features as Zhou et al. [26] but it is not widely used in the literature.

In a hand-driven learning paradigm, once a descriptor has been selected, a suitable classifier must be chosen. Although ensemble classifiers as Random Forests [27], Adaboost [19] or Xgboost [28] have been used, it could be said that Support Vector Machine (SVM) is the preferred classifier [23, 29, 30]. Unfortunately, nonparametric probabilistic models which take into account the uncertainty of the predictions, particularly Gaussian Processes (GPs) [31], which are in the state-of-art in classification, have been less used. It has long been known that neural networks with an *infinite* amount of hidden layers are equivalent to Gaussian Processes with a certain covariance kernel. GPs have the advantage of been nonparametric, unlike neural networks that have to learn a large number of parameters in order to have a sufficiently complex model. GPs allow us to

use a sound framework with a well defined inference procedure. Prior models in the form of different kernels can be used to encapsulate knowledge on the problem at hand. Model parameters can be automatically estimated without hand-tuning and predictions go beyond point estimates to provide very important information on uncertainty. They are starting to be used in histological image classification. Kandemir et al. [32] proposed a multi-instance relational learning based on GPs for histhopathology images. For the multi-instance purpose, they process each image as a bag and each patch as an instance. In order to capture the differences in cell formations caused by the disease status, they also introduce relational learning between instances and add relational side information from the spatial positions of segmented cells. More recently, with the purpose of facing more complex models, Deep Gaussian Processes (DGPs) [33] have been proposed. Unlike deep learning that requires a large dataset to learn a good model, DGPs can be applied with success even when data is scarce. In the last years, the ML community has experienced a remarkable interest in DGPs which are a hierarchical extension of GPs. Roughly speaking, they are deep architectures (like CNNs) whose layers are modelled by probabilistic GPs. This brings all the advantages of using GPs and provides much more power to approximate complex patterns in data. Results are really promising, surpassing CNNs in several problems. Unfortunately, in spite of its representation power, there are hardly any works in histopathology that make use of DGPs, see, however, Kandemir et al. [34] who apply a two-layer DGP model in histopathology cancer classification using an asymmetric transfer learning approach. The dataset used was built from two different tissues: breast and esophagus.

Once a patch classifier has been learned (using either hand-crafted or learned features), an image level evaluation is needed for prostate cancer diagnosis. Some works utilize a multiple instance learning approach and provide an overall WSI diagnosis, see Campanella et al. [35]. Another approach, which is frequently followed, is presented in Litjens et al. [8]. For each pixel, the probability of being cancerous is estimated from the patch probabilities, constructing a heat map for the WSI. This probability map is then thresholded to classify

every WSI pixel as cancerous or benign.

In this work we approach the classification of prostate histological images by first calculating the OD of each WSI to then estimate its H&E concentration components (we will show that OD is a better space than RGB for feature extraction and classification tasks). Hand-crafted features, which are expected to capture the expertise of pathologists, are then extracted from patches of these two concentration components. Finally, patches are classified using single-layer and multilayer Gaussian processes into benign and cancerous classes. We also carry out a validation at WSI level. We predict the per pixel probability of being cancerous and validate the obtained probability map. GPs and DGPs perform similarly and they are competitive to the tested shallow and deep classifiers. In other words, the quality of our OD extracted features does not require more than a single-layer GP to outperform the best performing classifiers.

The rest of the paper is organized as follows, in section 2 we introduce and describe a new WSI database of histological prostate images which has been manually annotated by experts[2]. In section 3, we explain how the CD task is performed on each WSI and describe how to obtain its RGB and OD H&E representations. In section 4, we motivate and define our new two morphological descriptors, we explain how the proposed framework, to discriminate between cancer and benign tissue in prostate, tries to mimic the way of analysis of a pathologist. In section 5, we provide an introduction to GPs and its hierarchical extension, DGP, in supervised learning. In section 6 we carry out a comparative study of several classifiers using the proposed features in a real clinical database provided by pathologists from the Hospital Clínico of Valencia. The performed experiments show that the classifier based on GP and deep GP together with the proposed features extracted in the OD space outperforms the current state of the art shallow classifiers and it is competitive to state-of-art deep convolutional neural netwok classifiers. In the experimental discussion we provide an insightful analysis. We use the area under the curve (AUC) for the evaluation of patch

[2]The dataset will be made public upon acceptance of the paper.

classification and FROC for diagnosis (detection) of prostate cancer in Whole Slide Image. We also analyze its complexity and computational cost compared to CNNs. Besides, to assess the robustness, we use the database proposed in [15, 29] for external validation. Finally, in Section 7 we summarize the conclusions extracted from our experimental results.

2. Material: SICAP database

The lack of large and public databases of prostate histopathological images has prevented researchers from a rigorous and meaningful comparison of supervised learning methods on these images. To the best of our knowledge, only three public databases containing histological prostate images are available. The first one, which is the result of a joint work by the National Cancer Institute and the National Human Genome Research Institute, both from United States, has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. However, the fact of not providing pixel-wise annotations along with a large amount of missing labels makes this database [3] inappropriate to validate new methodologies. The second one, the public database released by the authors of [36], is composed by 886 images and their corresponding pixel-wise annotations according to the Gleason scale. Unfortunately, only isolated tissue spots, representing characteristic patterns, are provided which prevents a patch size comparison and a full WSI classification. The third one, a database used in [15, 29] is composed by 625 different grade patches with a pixel-wise mask provided by pathologists. No WSIs are provided.

In this work, we present the SICAPv1 database, publicly available at `https://cvblab.synology.me/PublicDatabases/SICAPv1.zip`. It was obtained by a team of pathologists working at the Hospital Clínico of Valencia. Biopsies of 48 different patients were processed, hematoxylin and eosin stained and then digitized using the *Ventana iScan Coreo* scanner at 40x magnification. The

[3] `https://portal.gdc.cancer.gov/`

Table 1: SICAPv1 database description. Number of training WSIs and number of $512^2/1024^2$ associated patches.

	Benign	**Grade 3**	**Grade 4**	**Grade 5**	**Pathological**
#WSIs	17	18	15	10	43
#512^2 patches	6725	380	589	173	1142
#1024^2 patches	1909	113	181	50	344

database consists of 79 WSI: 19 correspond to benign prostate tissue biopsies (negative class) and 60 to pathological prostate tissue biopsies (positive class). Note that the entire dataset was divided into two subsets, 60 WSI (17 benign and 43 pathological) were used to learn the models and the remaining 19 images (two benign, seven diagnosed as grade 3, eight corresponding to grade 4 and two grade 5 WSIs) to test them. The malignant regions of the pathological images were carefully pixel-wise annotated by an expert team of pathologists. For this purpose, experts manually annotated the relevant tumoral areas using an online in-house application based on the OpenSeadragon functional core [37].

In order to automatically analyse these gigapixel images, the images were downsampled from 40× to 10× and divided in patches with a 50% overlap. To test the influence of the patch size, different sizes were selected: 512^2 and 1024^2, resulting on the two different datasets detailed in Table 1. Note that malignant patches were extracted from the annotated tumoral areas in the positive class images. Patches less than 25% inside a malignant area were not considered. And benign patches were extracted from benign WSIs.

3. Color deconvolution

For each WSI, the three-channel image information is the RGB intensity detected by a brightfield microscope observing a stained prostate histological slide. H&E are the stains usually used in pathology: Hematoxylin highlights the nuclei in purple and Eosin the stroma and cytoplasm in pink. Each $M \times N$

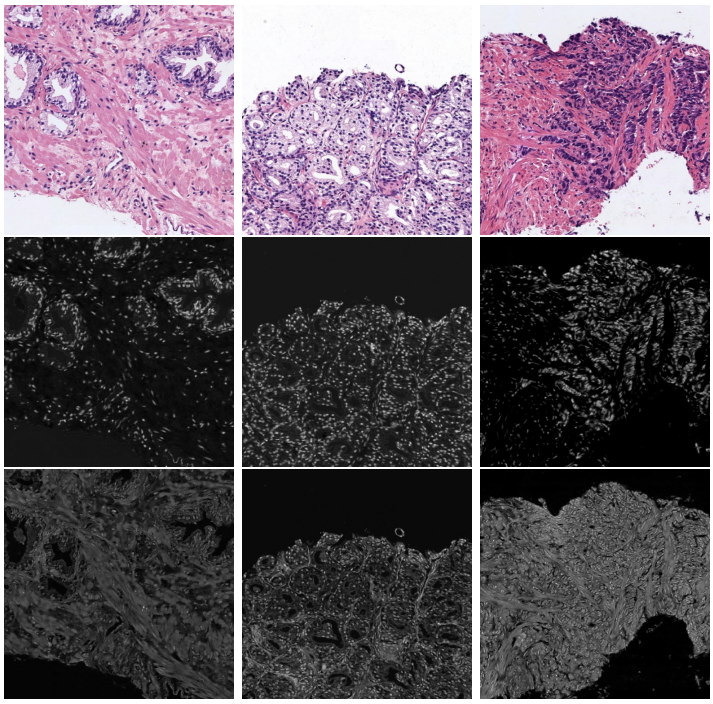
image is denoted by $\mathbf{I}$ with columns $\mathbf{i}_c = (i_{1c}, \ldots, i_{MN_c})^T,\ c \in \{R, G, B\}$.

We follow the color deconvolution approach described in [6]. According to the Lambert-Beer's law we can express the OD for channel c of the slide as $\mathbf{y}_c = -\log(\mathbf{i}_c/\mathbf{i}_c^0) \in \mathbb{R}^{MN\times 1}$, where $i_c^0 = 255$ is the incident light and division inside the logarithm is performed element-wise. Slides are stained using $\mathbf{n}_s = 3$ stains, $s \in \{H, E, Res\}$ (to obtain a unique stain decomposition we consider a third stain which represents the residual part) then the observed OD multichannel $\mathbf{Y} = [\mathbf{y}_R, \mathbf{y}_G, \mathbf{y}_B] \in \mathbb{R}^{MN\times 3}$ can be decomposed as a matrix multiplication $\mathbf{Y}^T = \mathbf{M}\mathbf{C}^T$, where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3] \in \mathbb{R}^{MN\times 3}$ is the stain concentration matrix, with $\mathbf{c}_s$, the s-th column of $\mathbf{C}$, containing at each pixel position the concentration of stain color s and $\mathbf{M} \in \mathbb{R}^{3\times 3}$ denoting the normalized stain matrix of the fixed form exposed in [6]. Notice that the s-th column of $\mathbf{M}$, $\mathbf{m}_s$, denotes the specific color of stain s.

The stain concentration matrix can then be recovered using $\mathbf{C}^T = \mathbf{M}^{-1}\mathbf{Y}^T$. Concentrations are transformed back to color (RGB) images using $\mathbf{y}_s^{sep} = \exp(-\mathbf{m}_s\mathbf{c}_s^T), s \in \{H, E\}$. Features are usually extracted from the single channel images $\exp(-\mathbf{c}_s), s \in \{H, E\}$ in the so called RGB space. In this work, we propose to perform this step in the OD space where stains are linearly separable, that is, directly on $\mathbf{c}_s, s \in \{H, E\}$. Figure 2 shows three different images from three different biopsies (and patients), one benign and two pathological, and their corresponding OD concentrations, Hematoxylin in the first row and Eosin in the second one. OD Hematoxylin captures nuclei infomation while OD Eosin contains information on stroma and cytoplasms.

4. Granulometry-based descriptors

Granulometry is a technique based on mathematical morphology. Size distributions of different elements in an image are obtained applying a series of morphological opening (or closing) operations with increasing-size structuring elements. The obtained size distribution provides shape and size information. In this paper, we propose the use of the classic formulation of granulometry



(a) (b) (c)

Figure 2: Hematoxylin (second row) and Eosin (third row) optical densities for three samples: a) Benign; b) and c) pathological.

as a new descriptor used in histhological images and define a new variant for prostate cancer classification which makes use of morphological reconstruction. The two proposed descriptors are explained below.

4.1. Granulometry-based descriptor

Based on a pyramid of morphological operators, granulometry calculates the size distribution of bright and dark objects present in an image. Let $\mathbf{z}$ be either a whole gray level image or an image patch. We can define a morphological descriptor, using the opening operator $\gamma_i(\mathbf{z})$ applied to the image $\mathbf{z}$ with a SE (window) of size i. This opening operator can be expressed as the combination of an erosion ($\epsilon_i(\mathbf{z})$) followed by a dilation ($\delta_i(\mathbf{z})$), both with the SE of size i. When this opening is computed with a SE of increasing size (λ), we

obtain a morphological opening pyramid (or granulometry profile) which can be formalized as:

$$\Pi_\gamma(\mathbf{z}) = \{\Pi_{\gamma\lambda} : \Pi_{\gamma\lambda} = \gamma_\lambda(\mathbf{z}), \forall\lambda \in [0, s, 2s, ..., n_{max}]\}. \quad (1)$$

where n_{max} represents the maximum size of the structuring element, and the sizes increase in steps s.

Making use of the opening pyramid (Π_γ), the granulometry curve or pattern spectrum of $\mathbf{z}$, $PS_\Gamma(\mathbf{z}, n)$, can be defined as:

$$PS_\Gamma(\mathbf{z}, n) = \frac{m(\Pi_{\gamma n}(\mathbf{z})) - m(\Pi_{\gamma n+1}(\mathbf{z}))}{m(\mathbf{z})},\ n \geq 0 \quad (2)$$

where $m(\mathbf{z})$ is the Lebesgue measure of $\mathbf{z}$ and it is computed as the area of $\mathbf{z}$ in the binary case and the volume in the gray-scale case (sum of pixel values).

$PS_\Gamma(\mathbf{z}, n)$ (also called size density of $\mathbf{z}$) maps each size n to a measure of the bright image structures with this size: loss of bright image structures between two successive openings. It is a probability density function (a histogram) in which a large impulse in the pattern spectrum at a given scale indicates the presence of many image structures at that scale.

By duality, a closing, $\varphi_i(\mathbf{z})$ is defined as the dilation of $\mathbf{z}$ followed by an erosion, both with a SE of size i. In the same way, a morphological closing pyramid is an anti-granulometry profile and can be computed on the image performing repeated closings with a SE of increasing size (λ) defined as:

$$\Pi_\varphi(\mathbf{z}) = \{\Pi_{\varphi\lambda} : \Pi_{\varphi_\lambda} = \varphi_\lambda(\mathbf{z}), \forall\lambda \in [0, ..., n_{max}]\} \quad (3)$$

The concept of pattern spectrum extends to the anti-granulometry curve $PS_\Phi(\mathbf{z})$ with respect to the family of closings Φ:

$$PS_\Phi(\mathbf{z}, -n) = \frac{m(\Pi_{\varphi n}(\mathbf{z})) - m(\Pi_{\varphi n-1}(\mathbf{z}))}{m(\mathbf{z})},\ n \geq 0. \quad (4)$$

Notice that this spectrum characterises the size of image structures with low level intensities.

Both granulometry and anti-granulometry descriptors are concatenated to construct the final descriptor (*Gran*).

4.2. Geodesic Granulometry-based descriptor

In this work, we introduce a variant of the granulometry, named geodesic granulometry, which is based on geodesic transformations.

A geodesic transformation involves two images: a marker image (or patch) $\mathbf{y}$ and a reference image $\mathbf{z}$. The *geodesic dilation* is the iterative unitary dilation of $\mathbf{z}$ with respect to $\mathbf{y}$, that is:

$$\delta_{\mathbf{y}}^{(n)}(\mathbf{z}) = \delta_{\mathbf{y}}^{(1)}\delta_{\mathbf{y}}^{(n-1)}(\mathbf{z}), \text{ being } \delta_{\mathbf{y}}^{(1)}(\mathbf{z}) = \delta_B(\mathbf{z}) \wedge \mathbf{y}. \tag{5}$$

The *reconstruction by dilation* is the successive geodesic dilation of $\mathbf{z}$ regarding $\mathbf{y}$ up to idempotence, that is:

$$R_{\mathbf{y}}^{\delta}(\mathbf{z}) = \delta_{\mathbf{y}}^{(i)}(\mathbf{z}), \text{ so that } \delta_{\mathbf{y}}^{(i)}(\mathbf{z}) = \delta_{\mathbf{y}}^{(i+1)}(\mathbf{z}). \tag{6}$$

The *reconstruction by erosion* can be obtained as its dual operator:

$$R_{\mathbf{y}}^{\varepsilon}(\mathbf{z}) = [R_{\mathbf{y}^c}^{\delta}(\mathbf{z}^c)]^c, \tag{7}$$

being $\mathbf{z}^c$ the complement image (or patch).

The reconstruction by dilation removes from the reference $\mathbf{z}$ the bright objects unconnected with the marker $\mathbf{y}$. The underlying idea on which the new descriptor is based is to only consider in the granulometry spectrum the objects totally removed in each opening (closing) step. Using $\gamma(\mathbf{z})$ as indicated in Equation (1) can lead to the inclusion in the pattern spectrum of fragments of objects partially removed in the process. To solve this shortcoming, we modify the granulometry profile (Equation (1)) by using the geodesic opening given by $\gamma^r(\mathbf{z}) = R_{\gamma(\mathbf{z})}^{\delta}(\mathbf{z})$. By duality, the proposed geodesic closing, to be used in the computation of the anti-granulometry profile, (Equation (3)) is $\varphi^r(\mathbf{z}) = R_{\varphi(\mathbf{z})}^{\varepsilon}(\mathbf{z})$. The new geodesic granulometry descriptors will be denoted $PS_{\Gamma}^r(\mathbf{z}, n)$ and $PS_{\Phi}^r(\mathbf{z}, -n)$, respectively.

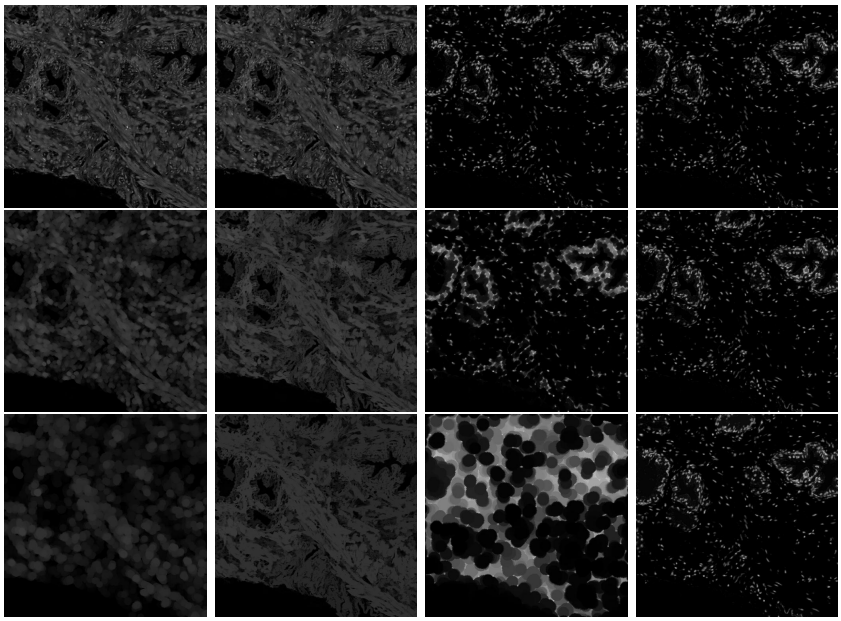
Both geodesic descriptors are concatenated to construct the final descriptor ($GeoGran$).

4.3. Granulometry profiles for prostate cancer detection

The proposed framework, to discriminate between cancer and benign tissue in prostate, tries to mimic the way of analysis of a pathologist. Basically, the cancer destroys the tissue structure. A benign tissue is formed by glands, each of them with a lumen surrounded by cytoplasm and nuclei, distributed in a background of stroma (which also contains sparsely distributed nuclei) (Figure 2(a)). As cancer progresses, glands begin to proliferate and merge, destroying the structure of benign tissues. Cytoplasm and lumens disappear and stroma is invaded by nucleis. Figure 2, (first row), shows three different cancer stages ((a) benign, (b) grade 3, (c) grade 5). To capture in a descriptor the tissue structure, we propose to use PS_Φ with H as input image. This encodes the structure of the glands by recovering the structure of the nuclei which formed the gland frontiers (those that enclosed their lumen and cytoplasm). The granulometric profiles, Π_φ, for the three image examples are shown in Figures 3(c), 4(c) and 5(c). To capture stroma information, PS_Γ is applied on the E component. Figures 3(a), 4(a) and 5(a) show the Π_γ profiles for the three examples. Figures 3, 4 and 5 also depict in columns (b) and (d) the geodesic profiles Π^r_γ and Π^r_φ, respectively. Note that Π^r_φ (columns (d)), for the three cases, shows that the results for different steps (different sizes of SEs) of the granulometric profile do not change. This suggests that stroma information more accurately extracted in PS^r_Γ, is the most relevant information to discriminate between pathological and benign tissues (as results presented in the experimental section corroborate).

5. Probabilistic model and inference

In this section we provide a brief introduction to the use of GPs and DGPs in supervised learning. An in depth study of these models can be found in [31] and [38]. Let us assume that we have n labeled training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector, $y_i \in \{0, 1\}$ for a binary classification problem, and $y_i \in \mathbb{R}$ for a regression one. We use either $y_i = f_i + \epsilon_i$ or $\mathrm{p}(y_i|f_i) = \sigma^{y_i}(f_i)\sigma^{1-y_i}(f_i)$ depending on whether we are dealing with a re-



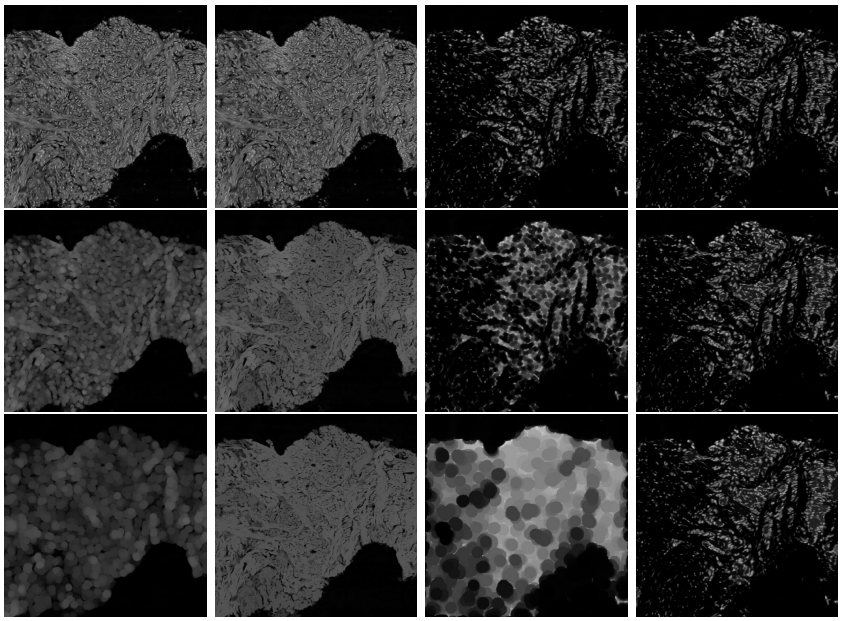
(a) (b) (c) (d)

Figure 3: Granulometry profiles (steps $s = 1, 4, 16$) for image (a) in Figure 2: (a) Π_φ; (b) Π^r_φ; (c) Π_γ; (d) Π^r_γ.

gression or classification problem, respectively. We assume that the noise in the regression problem is uncorrelated Gaussian of variance ρ^2 and $\sigma(\cdot)$ denotes the sigmoid function. We have used f_i instead of $f(\mathbf{x}_i)$ for simplicity. Notice that to tackle both problems we need to model the behavior of the function $f(\cdot)$ on seen and unseen samples $\mathbf{x}$.

5.1. Single-layer Gaussian Process

In a GP based formulation of a supervised problem we assume that the distribution of $\mathbf{f} = (f_1, \ldots, f_n)^{\mathrm{T}}$ given $\mathbf{X}$ is a multivariate normal, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where the zero mean is assumed for simplicity and $\sigma_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. where $k(\cdot, \cdot)$ is a kernel function. The use of kernel functions will guarantee that $\mathbf{\Sigma}$ is always a semidefinite positive matrix (independently of the number of samples and the features in $\mathbf{X}$). In this paper we use the squared exponential kernel (SE), also



(a) (b) (c) (d)

Figure 4: Granulometry profiles (steps $s = 1, 4, 16$) for image (b) in Figure 2: (a) Π_φ; (b) Π^r_φ; (c) Π_γ; (d) Π^r_γ.

known as Radial Basis Function (RBF), defined as:

$$k(\mathbf{x}, \mathbf{x}') = C \exp(-\gamma ||\mathbf{x} - \mathbf{x}'||^2). \tag{8}$$

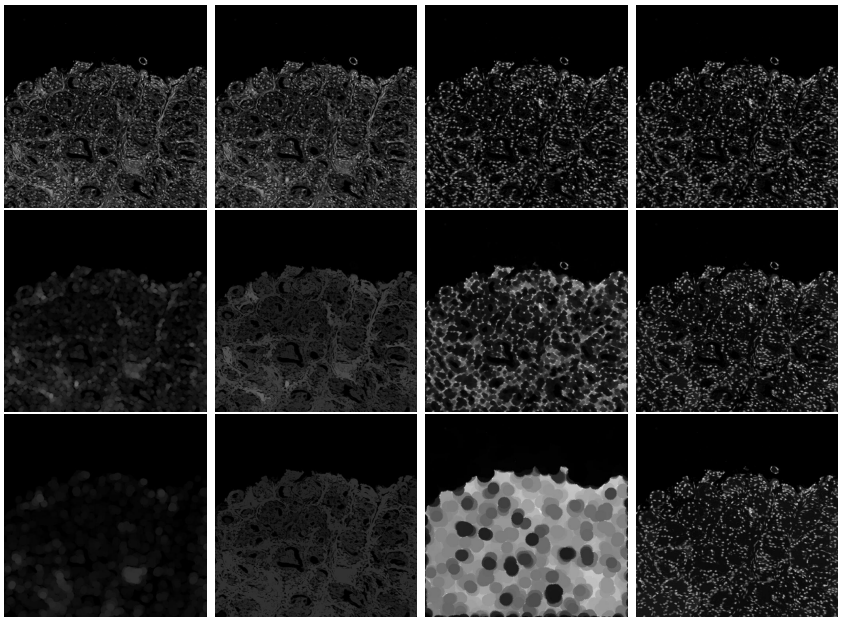
where the parameters C and γ will be estimated from the observations (the learning task).

Now we have all the ingredients we need to model our supervised learning problem using GPs. Given $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ we write

$$\mathrm{p}(\mathbf{y}, \mathbf{f}) = \prod_{i=1}^{n} \mathrm{p}(y_i|f_i)\mathrm{p}(\mathbf{f}|\mathbf{X}) \tag{9}$$

and proceed with the learning and inference tasks. We first learn the model parameters (C, γ and for a regression problem ρ^2 as well) by maximizing on them the marginal log-likelihood, that is,

$$\log \mathrm{p}(\mathbf{y}) = \log \int \mathrm{p}(\mathbf{y}|\mathbf{f})\mathrm{p}(\mathbf{f}|\mathbf{X})\mathrm{d}\mathbf{f} \tag{10}$$



(a) (b) (c) (d)

Figure 5: Granulometry profiles (steps $s = 1, 4, 16$) for image (c) in Figure 2: (a) Π_φ; (b) Π^r_φ; (c) Π_γ; (d) Π^r_γ.

which will allow us to calculate $\mathrm{p}(\mathbf{f}|\mathbf{y})$ and finally perform inference: given a new feature vector $\mathbf{x}^*$, we calculate

$$\mathrm{p}(f_*|\mathbf{y}, \mathbf{x}_*, \mathbf{X}) = \int \mathrm{p}(f_*|X, \mathbf{x}_*, \mathbf{f})\mathrm{p}(\mathbf{f}|X, \mathbf{y})d\mathbf{f} \tag{11}$$

which will allow us to predict $\mathbf{y}_{\mathbf{x}^*}$. There are two problems that must be faced when using GP in supervised learning. The first one, which is easier to handle, comes from the fact that in classification problems the prior distribution is not conjugate for the observation model. That is usually handled by maximizing a lower bound of the marginal likelihood in eq. 10. This will also have the effect of obtaining an approximation to $\mathrm{p}(\mathbf{f}|\mathbf{y})$ but not the real one, however, this problem is less relevant than the second one. Maximizing eq. 10 requires inverting a matrix the size of the number of samples (an $\mathcal{O}(n^3)$ operation) which is prohibitive for large datasets.

The most popular approach to dealing with the computational burden of

GPs is to introduce $m \ll n$ *inducing points* $\mathbf{u} = (u_1, \ldots, u_m)$ which the inference is based on. These are GP realizations at the *inducing locations* $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\} \subset \mathbb{R}^d$, just like $\mathbf{f}$ is at the inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ [39], in other words, $\mathbf{u} = f(\mathbf{Z})$. We can rewrite the joint distribution as

$$\mathrm{p}(\mathbf{y}, \mathbf{f}, \mathbf{u}) = \underbrace{\prod_{i=1}^{N} \mathrm{p}(y_i|f_i)}_{\text{likelihood}} \underbrace{\mathrm{p}(\mathbf{f}|\mathbf{u}; \mathbf{X}, \mathbf{Z})\mathrm{p}(\mathbf{u}; \mathbf{Z})}_{\text{GP prior}} \tag{12}$$

where a semicolon is used to specify the inputs of the GP, this will clarify multilayer-models notation.

Notice that we have overloaded the notation a bit to make clear the introduction of the inducing points but no changes in the modelling have been introduced since $\mathrm{p}(\mathbf{f}) = \int \mathrm{p}(\mathbf{f}|\mathbf{u}; \mathbf{X}, \mathbf{Z})\mathrm{p}(\mathbf{u}; \mathbf{Z})\mathrm{d}\mathbf{f}$.

Equipped with this decomposition, we go back to the marginal likelihood function in eq. 10 and use Jensen's inequality to, following the approach in [40], write

$$\log \mathrm{p}(\mathbf{y}) \geq \int \mathrm{q}(\mathbf{u})\mathrm{p}(\mathbf{f}|\mathbf{u}; \mathbf{Z}) \log \frac{\mathrm{p}(\mathbf{y}|\mathbf{f})\mathrm{p}(\mathbf{f}|\mathbf{u}; \mathbf{X}, \mathbf{Z})\mathrm{p}(\mathbf{u}; \mathbf{Z})}{\mathrm{p}(\mathbf{f}|\mathbf{u}; \mathbf{X}, \mathbf{Z})\mathrm{q}(\mathbf{u})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f}. \tag{13}$$

Now the optimization process becomes more involved. We have to estimate, together with the model parameters (C, γ and for a regression problem ρ^2 as well), the parameters of the distribution $\mathrm{q}(\mathbf{u})$ which is usually assumed to be a multivariate Gaussian, and the inducing point locations $\mathbf{Z}$. The benefit is that this learning process has become $\mathcal{O}(nm^2)$. Finally, $\mathrm{q}(\mathbf{u})$ is used, instead of $\mathrm{p}(\mathbf{f}|\mathbf{y})$, in eq. 11 for the inference (testing) process.

5.2. Deep Gaussian Processes

In standard (single-layer) GPs, the output of the GP is directly used to model the observed response $\mathbf{y}$. However, this output could be used to define the input locations of another GP. If this is repeated L times, we obtain a hierarchy of GPs that is known as a Deep Gaussian Process (DGP) with $L+1$ layers. DGPs were first introduced in [33], they can be used for regression and classification

problems by placing appropriate likelihoods (like the ones introduced at the beginning of this section) after the last layer.

Unfortunately, exact inference in DGP is intractable (beyond the the computationally expensiveness of GPs and the non-conjugacy of the prior), as it involves integrating out latent variables that are used as inputs in the next layer (i.e. they appear inside a complex kernel matrix). To overcome this, again m inducing points $\mathbf{u}^l$ at inducing locations $\mathbf{z}^{l-1}$ are introduced at each layer l. We write the joint distribution of the observation and DGP as

$$\mathrm{p}(\mathbf{y}, \{\mathbf{f}, \mathbf{u}^l\}_{l=1}^L) = \underbrace{\prod_{i=1}^N \mathrm{p}(y_i|f_i^L)}_{\text{likelihood}} \underbrace{\prod_{l=1}^L \mathrm{p}(\mathbf{f}^l|\mathbf{u}^l;\mathbf{f}^{l-1},\mathbf{z}^{l-1})\mathrm{p}(\mathbf{u}^l;\mathbf{z}^{l-1})}_{\text{DGP prior}}. \tag{14}$$

Here, $\mathbf{f}^0 = \mathbf{X}$, and each factor in the product is the joint distribution over $(\mathbf{f}^l, \mathbf{u}^l)$ of a GP in the inputs $(\mathbf{f}^{l-1}, \mathbf{z}^{l-1})$, but rewritten with the conditional probability given $\mathbf{u}^l$. For notation simplicity, in this description the dimension of the hidden layers has been fixed to one. This can be generalized straightforwardly, in this case $\mathbf{f}^l, \mathbf{u}^l$ and $\mathbf{z}^{l-1}, l = 1, \ldots, L$ will be matrices of the appropriate sizes, see see [33, 38].

To train the model, we follow the approach in [38] where the authors use the Jensen's inequality, with the posterior distribution approximation

$$\mathrm{q}(\{\mathbf{f}^l, \mathbf{u}^l\}_{l=1}^L) = \prod_{l=1}^L \mathrm{p}(\mathbf{f}^l|\mathbf{u}^l;\mathbf{f}^{l-1},\mathbf{z}^{l-1})\mathrm{q}(\mathbf{u}^l). \tag{15}$$

where $\mathrm{q}(\mathbf{u}^l) = \mathcal{N}(\mathbf{u}^l|\mathbf{m}^l, \mathbf{S}^l)$, to write

$$\begin{aligned}
\log \mathrm{p}(\mathbf{y}) \geq & \int \prod_{l=1}^L \mathrm{p}(\mathbf{f}^l|\mathbf{u}^l;\mathbf{f}^{l-1},\mathbf{z}^{l-1})\mathrm{q}(\mathbf{u}^l) \\
& \times \log \frac{\prod_{i=1}^N \mathrm{p}(y_i|f_i^L) \prod_{l=1}^L \mathrm{p}(\mathbf{f}^l|\mathbf{u}^l;\mathbf{f}^{l-1},\mathbf{z}^{l-1})\mathrm{p}(\mathbf{u}^l;\mathbf{z}^{l-1})}{\prod_{l=1}^L \mathrm{p}(\mathbf{f}^l|\mathbf{u}^l;\mathbf{f}^{l-1},\mathbf{z}^{l-1})\mathrm{q}(\mathbf{u}^l)} \prod_l \mathrm{d}\mathbf{u}^l \mathrm{d}\mathbf{f}^l \\
= & \sum_{i=1}^n \mathbb{E}_{\mathrm{q}(f_i^L)}[\log \mathrm{p}(y_i|f_i^L)] - \sum_{l=1}^L \mathrm{KL}(\mathrm{q}(\mathbf{u}^l)||\mathrm{p}(\mathbf{u}^l;\mathbf{z}^{l-1})).
\end{aligned} \tag{16}$$

Now the optimization process of the above Evidence Lower Bound (ELBO) becomes even more involved. We have to estimate, together with the model

parameters for each layer, the parameters of the distributions $q(\mathbf{u}^l)$ and the inducing point locations $\mathbf{z}^l$.

The second term is tractable, as the KL divergence between Gaussians is known. However, the expectation involves the marginals of the posterior at the last layer, $q(f_i^L)$. As we will now see, although this distribution is analytically intractable, it can be sampled efficiently using univariate Gaussians.

Marginalizing out the inducing points in eq. (15), the posterior for the GP layers $\{\mathbf{f}^l\}_{l=1}^L$ is

$$q(\{\mathbf{f}^l\}_{l=1}^L) = \prod_{l=1}^{L} q(\mathbf{f}^l|\mathbf{m}^l, \mathbf{S}^l; \mathbf{f}^{l-1}, \mathbf{z}^{l-1}) = \prod_{l=1}^{L} \mathcal{N}(\mathbf{f}^l|\tilde{\boldsymbol{\mu}}^l, \tilde{\boldsymbol{\Sigma}}^l), \tag{17}$$

where the vector $\tilde{\boldsymbol{\mu}}^l$ is given by $[\tilde{\boldsymbol{\mu}}^l]_i = \mu_{\mathbf{m}^l,\mathbf{z}^{l-1}}(f_i^{l-1})$ and the $n \times n$ matrix $\tilde{\boldsymbol{\Sigma}}^l$ by $[\tilde{\boldsymbol{\Sigma}}^l]_{ij} = \Sigma_{\mathbf{S}^l,\mathbf{z}^{l-1}}(f_i^{l-1}, f_j^{l-1})$. The specific form of the functions $\mu_{\mathbf{m}^l,\mathbf{z}^{l-1}}$ and $\Sigma_{\mathbf{S}^l,\mathbf{z}^{l-1}}$ can be found in [38, Eqs. (7-8)]. Although the distribution in eq. (17) is fully coupled between layers (and thus the posterior in the last layer is analytically intractable), the i-th marginal at each layer $\mathcal{N}(f_i^l|[\tilde{\boldsymbol{\mu}}^l]_i, [\tilde{\boldsymbol{\Sigma}}^l]_{ii})$ only depends on the corresponding i-th input of the previous layer. This allows one to recursively sample $\hat{f}_i^1 \to \hat{f}_i^2 \to \cdots \to \hat{f}_i^L$ from all the layers up to the last one by means of univariate Gaussians. Specifically, $\varepsilon_i^l \sim \mathcal{N}(0,1)$ is first sampled and then for $l = 1, \ldots, L$:

$$\hat{f}_i^l = \mu_{\mathbf{m}^l,\mathbf{z}^{l-1}}(\hat{f}_i^{l-1}) + \varepsilon_i^l \cdot \sqrt{\Sigma_{\mathbf{S}^l,\mathbf{z}^{l-1}}(\hat{f}_i^{l-1}, \hat{f}_i^{l-1})}. \tag{18}$$

In summary, the expectation $\mathbb{E}_{q(f_i^L)}[\log p(y_i|f_i^L)]$ in the ELBO (see eq. (16)) can be approximated with a Monte Carlo sample generated with eq. (18). Since the ELBO factorizes across data points and the samples can be drawn independently for each point i, scalability is achieved through sub-sampling the data in mini-batches. The complexity to evaluate the ELBO and its gradients is $\mathcal{O}(nm^2L)$. The code is integrated within GPflow (a GP framework built on top of Tensorflow) and is publicly available[4].

[4] https://github.com/ICL-SML/Doubly-Stochastic-DGP

To predict in a new x_*, eq. (18) is used to sample S times[5] from the posterior up to the $(L-1)$-th layer using the test location as initial input. This yields a set $\{f_*^{L-1}(s)\}_{s=1}^S$ with S samples. Then, the density over f_*^L is given by the Gaussian mixture (recall that all the terms in eq. (17) are Gaussians):

$$\mathrm{q}(f_*^L) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{q}(f_*^L|\mathbf{m}^L, \mathbf{S}^L; f_*^{L-1}(s), \mathbf{z}^{L-1}).$$

6. Experiments

In this section we carry out an exhaustive evaluation of the proposed classification approach which, as we have already indicated, is based on the use of GPs and DGPs and granulometry profiles on OD H&E images. First, we compare the classification performance of GPs with the most popular shallow classifiers using classical texture descriptors, granulometry profiles and a combination of them extracted from OD H&E images. To show the importance of the space where images are represented, we replicate the experiments using RGB H&E images. Once we show that features should be extracted from OD H&E images and that our approach is the best performing one when only shallow classifiers are used, we proceed to compare it to state-of-art deep learning strategies based on a variety of pre-trained CNNs. To demonstrate the generalization capability of the patch-wise trained model, we carry out a validation at WSI level (for the test set). We predict the per pixel probability of being cancerous and validate the obtained probability map.Despite being much simpler, GPs and DGPs perform similarly and they are also competitive to the tested deep classifiers. In other words, the quality of our OD extracted features does not require more than a single layer GP toobtain excellent results. Finally, an external validation has been carried out to assess the competitiveness of the proposed descriptor together with the GP classifier against other models.

[5] Results become stable after a few samples. Here, S was set to 100.

Table 2: Performance of descriptors and classifiers in RGB space with a 512^2 patch size.

AUC	RF	GP	XgBoost
LBP	0.6663 ± 0.1400	0.7003 ± 0.1190	0.6728 ± 0.1279
LBPV	0.7695 ± 0.0565	0.8243 ± 0.0891	0.7912 ± 0.0674
Gran	0.8549 ± 0.0856	0.8984 ± 0.0641	0.8778 ± 0.0735
GeoGran	0.9089 ± 0.0494	0.8910 ± 0.0599	0.9095 ± 0.0454
GranLBP	0.8331 ± 0.0949	0.9111 ± 0.0492	0.8551 ± 0.0842
GranLBPV	0.8758 ± 0.0611	0.9280 ± 0.0349	0.8908 ± 0.0509
GeoGranLBP	0.8958 ± 0.0566	0.9014 ± 0.0507	0.9048 ± 0.0469
GeoGranLBPV	0.9174 ± 0.0351	$\mathbf{0.9307 \pm 0.0307}$	0.9273 ± 0.0329

6.1. Feature extraction

As feature descriptors we computed the morphological descriptors PS_Φ and PS_Γ on H and E, respectively, and their geodesic versions PS^r_Φ and PS^r_Γ. PS_Φ and PS^r_Φ with SE of increasing size in steps of $s = 2$ from 0 to $n_{max} = 24$, and in steps of $s = 4$ for PS_Γ and PS^r_Γ from 0 to $n_{max} = 48$. Note that we use *Gran* and *GeoGran* labels to denote PS and PS^r descriptors, respectively. Besides that, to capture the texture information we use the uniform and rotationally invariant Local Binary Patterns (LBP) [41] as baseline descriptor (with neighbourhood of $R = 1$ and $P = 8$) and the combination of it with a contrast measure, according to the work of Guo et al. [42], obtaining an additional Local Binary Pattern Variance ($LBPV$) descriptor. The different combinations of descriptors have been labelled as *GranLBP*, *GranLBPV*, *GeoGranLBP* and *GeoGranLBPV*.

6.2. Comparison of shallow classifiers

To demonstrate the superiority of nonparametric probabilistic models based on GPs and morphological features we compare GPs with different state-of-art shallow classifiers on different extracted features. We compare the performance of the models on OD and RGB spaces, testing two patch sizes, 512^2 and 1024^2.

Table 3: Performance of descriptors and classifiers in OD space with a 512^2 patch size.

AUC	RF	GP	XgBoost
LBP	0.9300 ± 0.0603	0.9253 ± 0.0635	0.9262 ± 0.0615
LBPV	0.9351 ± 0.0373	0.9443 ± 0.0314	0.9421 ± 0.0243
Gran	0.9323 ± 0.0453	0.9516 ± 0.0346	0.9461 ± 0.0322
GeoGran	0.9690 ± 0.0303	0.9636 ± 0.0242	0.9688 ± 0.0249
GranLBP	0.9436 ± 0.0640	0.9581 ± 0.0422	0.9541 ± 0.0524
GranLBPV	0.9370 ± 0.0340	0.9696 ± 0.0175	0.9573 ± 0.0206
GeoGranLBP	0.9666 ± 0.0408	0.9669 ± 0.0283	0.9700 ± 0.0304
GeoGranLBPV	0.9692 ± 0.0241	$\mathbf{0.9807 \pm 0.0097}$	0.9747 ± 0.0170

Table 4: Performance of descriptors and classifiers in RGB space with a 1024^2 patch size.

AUC	RF	GP	XgBoost
LBP	0.6279 ± 0.1751	0.6900 ± 0.1841	0.6460 ± 0.1660
LBPV	0.7517 ± 0.0847	0.8222 ± 0.1169	0.7638 ± 0.0934
Gran	0.8018 ± 0.1166	0.8785 ± 0.0525	0.8177 ± 0.1071
GeoGran	0.9269 ± 0.049	0.9242 ± 0.0398	0.9242 ± 0.0425
GranLBP	0.7910 ± 0.1379	0.8780 ± 0.0512	0.7955 ± 0.1437
GranLBPV	0.8471 ± 0.0820	$\mathbf{0.9447 \pm 0.0252}$	0.8536 ± 0.0708
GeoGranLBP	0.9079 ± 0.0675	0.9062 ± 0.0462	0.9146 ± 0.0478
GeoGranLBPV	0.9338 ± 0.0339	0.9293 ± 0.0510	0.9289 ± 0.0347

Table 5: Performance of descriptors and classifiers in OD space with a 1024^2 patch size.

AUC	RF	GP	XgBoost
LBP	0.9433 ± 0.0615	0.9353 ± 0.0661	0.9350 ± 0.0640
LBPV	0.9244 ± 0.0671	0.9684 ± 0.0217	0.9419 ± 0.0575
Gran	0.9408 ± 0.0493	0.9635 ± 0.0320	0.9590 ± 0.0448
GeoGran	0.9826 ± 0.0237	0.9824 ± 0.0165	0.9814 ± 0.0256
GranLBP	0.9525 ± 0.0654	0.9647 ± 0.0488	0.9578 ± 0.0603
GranLBPV	0.9318 ± 0.0480	0.9736 ± 0.0211	0.9553 ± 0.0386
GeoGranLBP	0.9760 ± 0.0366	0.9800 ± 0.0230	0.9800 ± 0.0277
GeoGranLBPV	0.9789 ± 0.0187	$\mathbf{0.9855 \pm 0.0089}$	0.9764 ± 0.0218

We use variational inference on a single-layer GP classifier with a RBF kernel. We utilize a sparse model with 800 inducing points when the patch size is 512^2. For 1024^2 patch size we do not utilize inducing points. For comparison, we use Random Forest (RF) and Extreme Gradient Boosting (XgBoost). These tree-based ensemble models can capture complex patterns in data. They are state-of-art shallow classifiers.

For each classifier we applied a five-fold cross-validation to validate and compare the performance of the proposed granulometry descriptors (using the described classifiers). Patches coming from the same image and the same patient were assigned to the same fold. Consequently, we avoided correlation between training and test sets which would distort the results. Due to the nature of prostatic images, the amount of benign instances is significantly greater than the cancerous ones. To deal with this imbalanced scenario, we built several classifiers with the positive instances and a subset of the negative ones so that each classifier faces a balanced problem being the final prediction the average of the predictions of each classifier. The evaluation metric we selected to compare the performance of different methods is the area under the ROC curve (AUC).

Tables 2, 3 (512^2) and 4, 5 (1024^2) summarize the obtained results. Analysing

all the tables, we observe that, in both spaces, key tumoral information is better encoded by morphological than by texture features. More in depth, *LBPV* and *GeoGran* perform better than *LBP* and *Gran* in both spaces. Regarding the classifiers, GPs discriminate better than the others for all patch sizes and spaces.

For every descriptor and classifier, the results obtained in the OD space are superior to those achieved in the RGB space. This is the space used by the majority of current state-of-the-art methods. Moreover, texture and morphological information for classification purposes are better captured in the OD space.

In summary, for both patch sizes, the best results are obtained in the OD space when *GeoGranLBPV* are the input to a GP classifier. The obtained AUCs are 0.9807 (512^2) and 0.9855 (1024^2). This fact suggests that texture and morphology features provide complementary information to characterize prostatic tumoral tissues. In the coming section we compare GPs and DGPs, using the best performing features, to CNNs.

6.3. Comparison of deep classifiers

The previous experiment indicates that the proposed geodesic granulometries (*GeoGran*) in combination with texture information (*LBPV*) allows us to create a descriptor *GeoGranLBPV* able to accurately classify histopathological tissues using GPs. We now compare GPs and DGPs used on *GeoGranLBPV* extracted from OD images to CNNs used on raw images. Three of the most well-known deep convolutional neural networks for image classification: VGG19 [43], Xception [44] and Inception v3 [45] are utilized. The main reason to select these CNNs was their wide use in the detection of tumoral tissues in histological images [46, 10, 11, 12, 13, 14].

For this comparison, the cross validation setup used for shallow classifiers was utilized. Together with the two GPs described in the previous section, a three-layer DGP classifier [38] with RBF kernel was used on the extracted features. Our model employs 100 inducing points per layer. Although with shallow GPs we achieved a very good performance, the DGP is used here as a nonparametric

Table 6: Empirically-tuned hyperparameters for Inception v3, Xception and VGG19.

Architecture	**Layer name**	**Optimizer**	**Learning rate**
VGG19	'block3_conv1'	Stochastic Gradient Descent	$1 \cdot 10^{-4}$
Inception v3	'mixed7'	Nesterov Adam	$1 \cdot 10^{-5}$
Xception	'add10'	Stochastic Gradient Descent	$1 \cdot 10^{-4}$

multi-layer classification model to carry out a comparison between the deep structure of VGG19, Xception, and Inception v3 and a GP based counterpart.

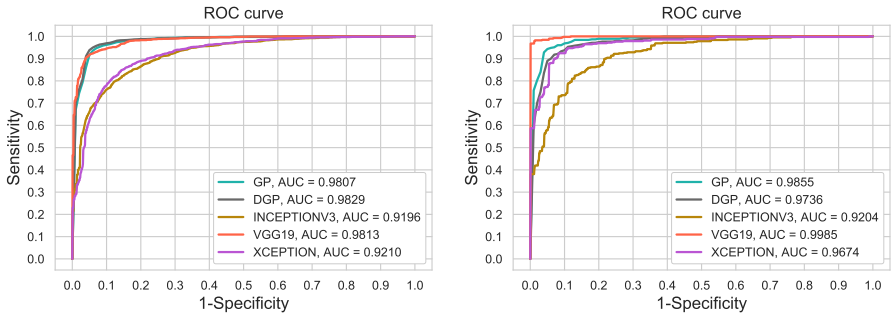
The parameters of the CNN were optimized following the procedures described in Table 6. In this experiment, due to the reduced number of samples of our data set, we fine-tuned the architectures, initializing them with the best weights obtained in the ImageNet challenge [47] and re-trained them using our raw RGB histological images as input. The re-training process was performed using the binary cross entropy loss function, from the layers indicated in Table 6 to the end of the networks. Early stopping, with fifteen epochs of patience value, was used to prevent overfitting. Synthetic data was automatically created using data augmentation methods (i.e. rotating, flipping, rescaling, translating, etc.) and a batch size of 16 samples, constrained by the available memory of the NVIDIA Titan V GPU utilized in this work, was used.

Table 7: Performance of Deep Classifiers for 512^2 patch size.

	Inception v3	VGG19	Xception	DGP
AUC	0.9196 ± 0.0302	0.9813 ± 0.0068	0.921 ± 0.026	$\mathbf{0.9829 \pm 0.0092}$

The average metric values for the five-fold comparison of deep models are reported in Tables 7 (512^2 patch size) and 8 (1024^2 patch size). As it can be observed from these tables, the morphological and textural information encoded by our proposed hand-crafted descriptor compares well to the automatic features directly learned by the CNNs from the data.

For 512^2 patch size (see Table 7), the hand-driven learning by DGP outper-



(a) (b)

Figure 6: ROC curve plot for all deep classifiers together with the best performing shallow one: for (a) 512^2 and (b) 1024^2 patch size.

Table 8: Performance of Deep Classifiers for 1024^2 patch size.

	Inception v3	VGG19	Xception	DGP
AUC	0.9204 ± 0.0525	$\mathbf{0.9985 \pm 0.0009}$	0.9674 ± 0.0194	0.9736 ± 0.0239

forms Inception v3 and Xception models in terms of AUC values by 6.33% and 6.19%, respectively. Additionally, the proposed methodology performs similarly to VGG19. The obtained AUC is 0.9829 which is slightly better than the one obtained by the shallow GP (0.9807), this suggests that our hand-crafted features are good enough to perform an excellent classification and they do not require more than the use of a well grounded nonparametric single layer classifier with no parameter tuning. Figure 6a shows the ROC curves corresponding to these deep classifiers together with the single layer GP used in the previous section for the 512^2 case.

When the patch size is 1024^2, see Table 8, VGG19 outperforms the rest of the deep classifiers. Its corresponding AUC is 0.9985 which is slightly better than the ones obtained by our DGP (0.9736) and GP (0.9855). Figure 6b shows the ROC curves corresponding to these deep classifiers together with the one-layer GP used in the previous section for the 1024^2 case. Note again our approach does not seem to need more than a layer to obtain excellent results.

Table 9: Analysis of the patch-wise (512^2) computational cost for the deep models and shallow GPs. Time measured for Deep GPs and GPs includes the feature extraction and classification steps. Note that CNNs were trained and tested in a Titan V GPU while these tasks were performed in the CPU for GPs and DGPs.

Time (sec.)	VGG19	Inception v3	Xception	Deep GPs	GPs
Training	28742.71	24321.12	23441.33	14587.1362 + 4431.1845 = 19018.3207	14587.1362 + 550.2587 = 15137.3949
Inference	0.8522	0.7873	0.7177	1.5753 + 0.0357 = 1.611	1.5753 + 0.0003 = 1.581

Regarding computational cost, the proposed methodology needs less time than the deep learning-based approaches in the training stage (see Table 9). It is important to remark that CNN models require specific hardware to be trained in an affordable time interval while GPs and DGPs just need a CPU to be trained. Due to this fact the inference phase in a CNN model requires less time than the proposed hand-driven approach. The computational time analysis was performed on an Intel i7@3.10 GHz of 16 GB of RAM with an NVIDIA GeForce Titan V to train VGG19, Inception v3, and Xception CNNs. Python 3.5 was the language used and the libraries GPflow and Keras were used for GPs and DGPs and deep learning methods, respectively.

6.4. Whole Slide Image evaluation

Our ultimate goal is to provide pathologists with useful tools for WSI analysis. With this aim, we extend the patch-wise classification model to WSI classification, trying to identify cancerous areas in unseen WSIs. Following the approach in [8], we split each biopsy of the WSIs into overlapping patches. For each pixel, we estimate the probability of being cancerous by bilinearly interpolating the predicted probabilities of the four closest patches (in terms of euclidean distance to the center of the patches). With this pixel-wise classification, we obtain a probability map per each biopsy of a WSI (see Figure 8(b)). To assess the generalization capability of our model we used the 19 WSIs in the test set: 17 maligns and 2 benigns. The magnification factor was, like during training, 10×. The overlap between patches was 75%, for both, 512^2 and 1024^2, patch sizes. We compare GP and DGP + *GeoGranLBPV* extracted in the OD

space to the models obtained by fine-tuning the three CNNs. All patch-wise models were trained using the 60 images in the training set. For WSI based evaluation, the free-response receiver operating characteristic (FROC) curve, defined as sensitivity versus the average number of false-positives per image, was used. After CAMELYON16 challenge [6], FROC is widely used for image level cancer detection evaluation.

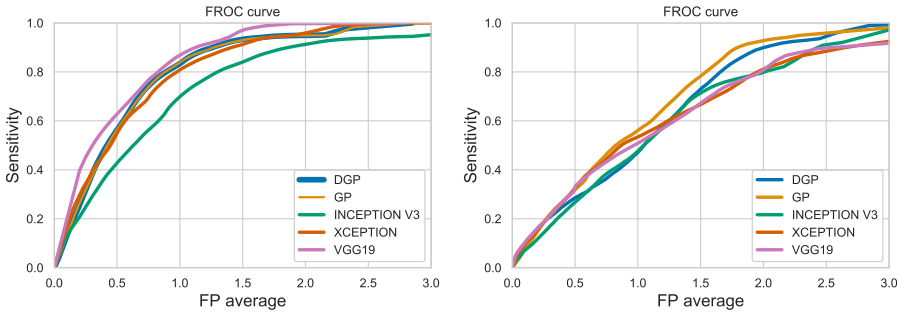
Table 10 shows, for both patch sizes, the sensitivity of each model for 1, 2 and 3 false cancerous regions. The results have been averaged over the 17 malign testing WSIs: these WSIs contain both benign and malign glands in addition to different cancer grades. These images present a high inflammation so it is a challenging task to detect well the benign glands. All models (CNN-based together with GP and DGP) generalize worse for 1024^2 patch size. This is probably due to the reduced sample size which may lead to overfitting during training and poor generalization during testing. Notice, however, that for this reason, the probabilistic and nonparametric nature of our GP and DGP models leads to a better generalization capability for this size. For a 512^2 patch size, we see that VGG19 performs slightly better than GP and DGP while Xception is a bit worse. Inception v3 generalizes poorly compared to the rest. Indeed, VGG19, GP and DGP are the only methods that detect all cancer pixels with a cost of 3 false positives areas for each pixel correctly classified. Figure 7 depicts the FROC for all compared models (512^2 and 1024^2 patch size) and clearly shows that our approach is competitive to state-of-art CNN architectures.

In Figure 8, for 512^2 patch size, we can compare the probability maps obtained by the best performing model (GP) (Figure 8(a)), and the cancerous regions annotated by the pathologist (Figure 8(b)). The probability maps are represented as heat maps, where red and blue colors indicate the highest and the lowest probabilities of being cancerous, respectively. The zoomed in regions show that the highest probabilities (redish colors) obtained by our model are in agreement with the cancerous areas marked by the experts while at the bound-

[6] https://camelyon16.grand-challenge.org/Home/

Table 10: Sensitivity for 1, 2 and 3 false positives for 512^2 and 1024^2 patch sizes.

Sensitivity	512^2			1024^2		
	1 FP	2 FP	3 FP	1 FP	2 FP	3 FP
GP	0.8387	0.9489	**1**	**0.5606**	**0.9277**	0.9804
DGP	0.8340	0.9492	**1**	0.4710	0.8993	**0.9920**
Inception v3	0.6985	0.9125	0.9519	0.4763	0.7981	0.9715
Xception	0.8081	0.9589	0.9984	0.5342	0.8115	0.9248
VGG19	**0.8610**	**0.9972**	**1**	0.5084	0.8089	0.9171

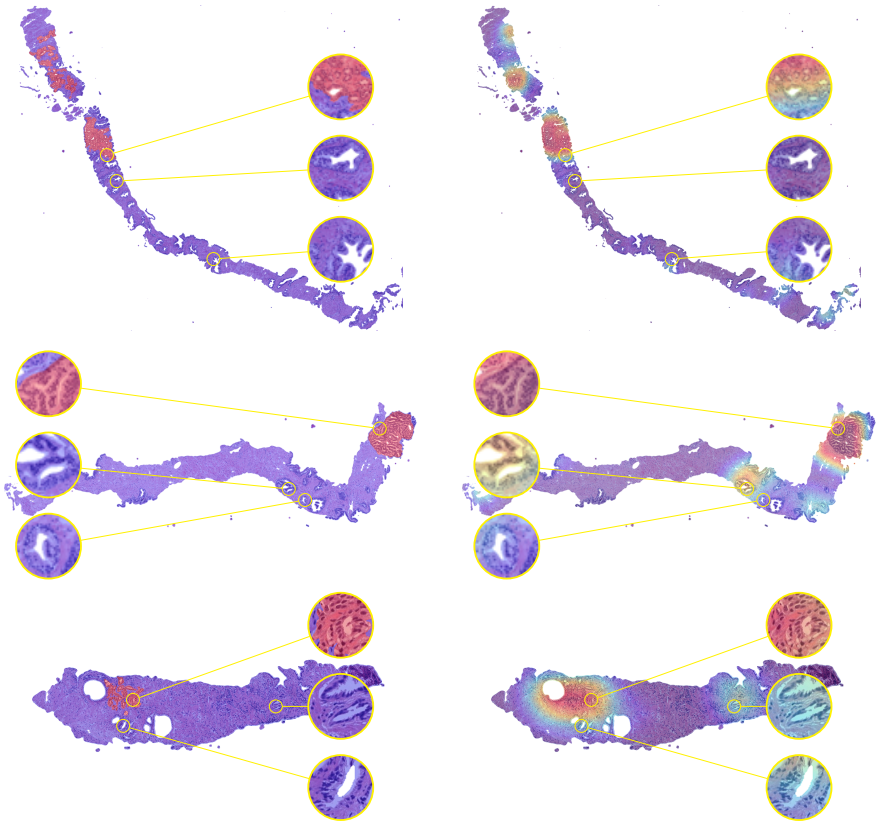


(a) (b)

Figure 7: FROC for CNN-based and GP and DGP models: (a) 512^2 and (b) 1024^2 patch size.

ary the probability decreases. Besides, the proposed model can discriminate successfully whether a gland is benign or malign in the same WSI giving zero or low probability to benign glands. For a more complete study, in Figure 9, we show the prediction of the proposed GP model in 3 regions of the two benign samples in the test subset. Since the heat maps give to each image a very low probability of being cancerous, this model does not suffer from false positives in benign WSIs.

Regarding computational cost and model complexity, taking into account the patch-wise average time (see Table 9) and the average number of patches resulting from all the biopsies contained in the testing WSIs (see Section 2), we can calculate the average time to predict a new WSI. Xception is the fastest



(a) (b)

Figure 8: WSI validation: (a) Cancerous areas annotated by the pathologists (ground truth); (b) Probability maps (heat maps) obtained by the proposed GP model.

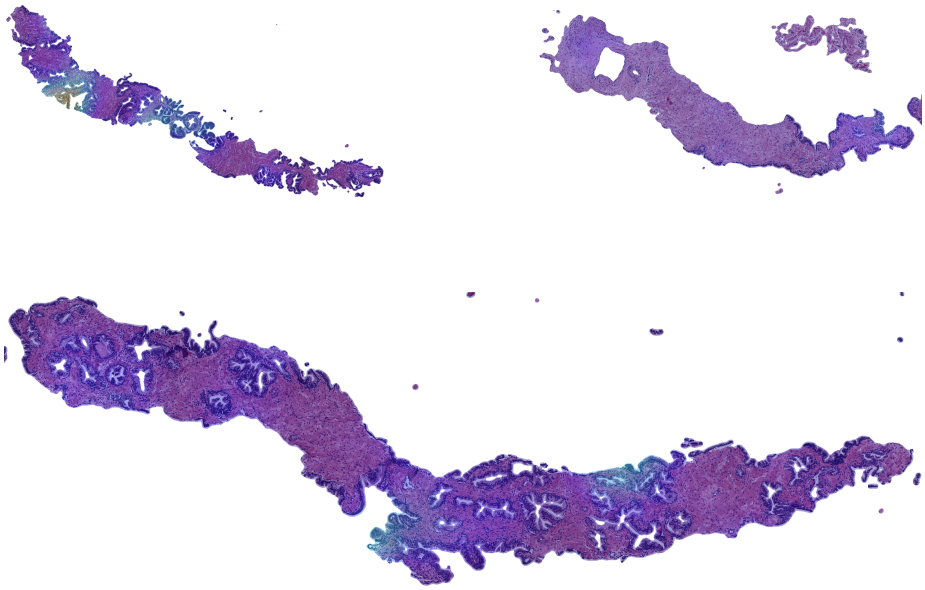


Figure 9: WSI validation: prediction in benign WSIs of the proposed GP model

model in obtaining the probability map for a WSI, in particular, the expected time ranges from 4.3 to 5.7 minutes depending on whether the WSI is composed of three or four biopsies. The Xception fine-tuning process is performed on 8,406,458 trainable parameters and the storage space of the model is 147.6 MB. Inception v3 model has 12,816,002 trainable parameters and the storage space of the model is 186.2 MB. The inference time ranges from 4.7 to 6.24 minutes. VGG19 takes around 5.1 to 6.8 minutes for WSIs with three and four biopsies, respectively. The fine-tuning process is performed on 130,923,522 trainable parameters and the storage space of the model is 1.02 GB. The models with the highest ability of generalization, i.e. models based on gaussian processes, spend around 9.3 and 12.7 minutes to compute the resulting probability map for a WSI composed of three and four biopsies, respectively. The number of GP and DGP parameters is 2,672,008 and 339,644 (due to the use of a less number of inducing points for DGP), respectively. The storage space is 20.88 MB for GP model and 10.10 MB for DGP model. As we have already indicated, notice that DL-based methods are computed in a Titan V GPU while

our hand-driven learning approaches are run in a i7 core.

Analysing the obtained computational cost, the model complexity and the performance of the models on new samples (see Table 10), we conclude that the proposed approaches based on GPs reach an interesting trade-off between these three capabilities. It is important to highlight that the task of diagnosing biopsies is an offline process and spending six additional minutes (additional DGP computational time in comparison to Inception v3 for a WSI with four biopsies) pays off due to the increased sensitivity. See in Table 10 the 24% improvement for 1FP for 512^2 patch size. In addition, the GP based models are, with regard to number of parameters and space, four (GPs) and five (DGP) times (DGPs) less expensive than the best CNN-based approach (VGG19).

6.5. Validation on an external data

To analyze and corroborate the robustness and generalization power of the proposed methodology, we also evaluate all the models on an external database. We have used the prostate cancer database proposed by Gertych et al. [15, 29]. This database includes 625 patches with different grades and combinations of them. No spatial information of these patches in the WSI is provided. The size of the patches at $20\times$ magnification is 1201^2. Each patch has a mask with annotation provided by pathologists (see Figure 10). This mask indicates the class of each pixel: stroma, benign or malign (distinguishing between grade 3, 4, and 5).

The GP model was trained using the SICAPv1 database and tested on the Gertych et al. [15, 29] database. Since we use for training 512^2 patches at $10\times$ magnification, we downsampled the test patches to a $10\times$ magnification and cropped the central region of 512^2 size. We labelled each patch of the test set as benign if there are no malign pixels in the image. Patches with more than 20% malign pixels (this information is provided by the mask) are classified as malign (for the binary classification approach proposed). This results in 593 patches of which 244 are benign and 349 are pathological.

The obtained results are reported in Tables 11 and 12 for the OD and RGB

Table 11: Performance of descriptors and classifiers in the OD space on the external database

AUC	RF	GP	XgBoost	DGP
LBP	0.8490	0.7529	0.7464	0.7833
LBPV	0.8415	0.6869	0.8593	0.6867
Gran	0.8572	0.8775	0.8851	0.8156
GeoGran	0.8828	**0.9249**	0.8636	0.8471
GranLBP	0.8629	0.8624	0.8643	0.6913
GranLBPV	0.8494	0.7998	0.8811	0.8850
GeoGranLBP	0.8757	0.8766	0.8754	0.8221
GeoGranLBPV	0.8872	0.7645	0.8365	0.8010

Table 12: Performance of descriptors and classifiers in RGB space on the external database

AUC	RF	GP	XgBoost	DGP
LBP	0.3444	0.3336	0.7051	0.2840
LBPV	0.6122	0.3116	0.7285	0.6597
Gran	0.7251	0.6473	0.7367	0.5928
GeoGran	0.8674	0.7130	0.8507	0.8026
GranLBP	0.5536	0.1214	0.7292	0.2728
GranLBPV	0.6346	0.3048	0.6622	0.8310
GeoGranLBP	0.8597	0.2756	0.8101	0.8158
GeoGranLBPV	0.8746	0.8097	0.8392	**0.8902**

Table 13: Performance of Deep Classifiers on the external database.

	Inception v3	VGG19	Xception
AUC	0.8846	**0.9714**	0.8670

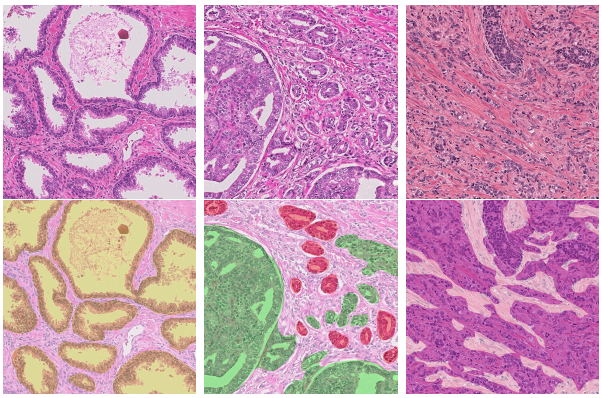


Figure 10: Patches from the external database [15, 29]. The colored masks indicate the annotated classes by the pathologist in this database: white (stroma), yellow (benign), red (grade 3), green (grade 4) and purple (grade 5).

spaces, respectively. The morphological features (*Gran* and *GeoGran*) outperform those based on texture (*LBP* and *LBPV*) in both RGB and OD spaces independently of the chosen classifier. Furthermore, in almost all cases, the OD space outperforms the RGB space. In this experiment combining texture and morphological descriptors does not achieve better results except in a few cases, for example, *GeoGranLBPV* + DGP in RGB space which obtains the best result in this space. However, the proposed descriptor based on geodesic granulometry *GeoGran* using GP as the classifier in the OD space outperforms the rest with an AUC of 0.9249.

These results indicate the robustness and generalization capabilities of the proposed morphological descriptor on different datasets. They also indicate that texture based features perform worse. This may have been exacerbated by the fact that white balancing was not performed on the second dataset since only patches were provided. We also verified that the OD space is more informative than the RGB one for most of the descriptors/classifiers used in the four studies carried out in this work. Furthermore, the GP is the classifier which shows the best performance.

Finally, for a complete comparison, the performance of deep neural networks

in this database is reported in Table 13. We can see that VGG19 obtains the best results. Notice, however, that the size of this model exceeds the Gigabyte in contrast to GP models which can be stored in much smaller disks (21 MB). Notice also that VGG19 is a well established architecture while the best DGPs is still work in progress. Regarding the other architectures (i.e. Inception v3 and Xception), our proposed descriptor *GeoGran* performs better using the probabilistic classifier based on a single-layer GP on the OD space, improving by a 4% and 6%, respectively. This demonstrates the competitive ability to capture cancer patterns with respect to state-of-art CNNs, even in databases that have never been seen by the classifier.

7. Conclusions and future work

In this work, we have proposed a novel descriptor to characterize and differentiate benign and pathological regions in histological prostate images. This descriptor registers the granularity of the tissue elements without previous segmentation.

We have shown that features should be extracted from OD H&E images, where our OD geodesic granulometry descriptor reveals the importance of the stroma identifying cancer. We have also shown that GP is the best performing classifier when only shallow classifiers are used. The best performing features ($GeoGranLBPV$) and the best performing shallow classifier (GP) together with its multilayer version (DGP) have then been compared to state-of-art deep learning strategies based on a variety of pre-trained CNNs. To analyze the generalization capability of the patch-wise trained model, we have carried out a validation at WSI level. We have predicted the per pixel probability of being cancerous and validate the obtained probability map. GPs and DGPs perform similarly and, furthermore, they are also competitive to the tested deep classifiers identifying successfully cancer in WSIs. To assess the robustness and generalization capabilities of the proposed descriptor, an external database has been utilized. The obtained results corroborate the quality of the proposed de-

scriptor when combined with a GP based classifier. In summary, we have shown that our OD extracted features do not require more than a single layer GP to outperform the best performing shallow classifiers and to be competitive to deep classifiers.

Additionally, we have created a public database (SICAPv1) that includes original WSIs and labels annotated by expert pathologists.

As future work, the use of geodesic granulometries and multi-class DGP for the automatic detection of Gleason grade in histopathological images will be addressed. Moreover, new annotated images will be added to SICAPv1.

References

[1] W. H. Organization, Global cancer observatory (2018).
URL http://gco.iarc.fr/

[2] D. F. Gleason, Histologic grading of prostate cancer: A perspective, Human Pathology 23 (3) (1992) 273 – 279, the Pathobiology of Prostate Cancer-Part 1.

[3] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, Computational and Structural Biotechnology Journal 16 (2018) 34 – 42.

[4] S. Wang, K. Burtt, B. Turkbey, P. L. Choyke, R. M. Summers, Computer aided-diagnosis of prostate cancer on multiparametric mri: A technical review of current research, in: BioMed research international, 2014.

[5] S. Roy, A. K. Jain, et al., A study about color normalization methods for histopathology images, Micron 114 (2018) 42–61.

[6] A. C. Ruifrok, D. A. Johnston, Quantification of histochemical staining by color deconvolution, Analytical and quantitative cytology and histology 23 (4) (2001) 291—299.

[7] V. Gupta, A. Bhavsar, Breast cancer histopathological image classification: Is magnification important?, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 769–776.

[8] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen van de Kaa, P. Bult, B. van Ginneken, J. van der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis 6 (2016) 26286.

[9] L. Hou, D. Samaras, T. M. Kurç, Y. Gao, J. E. Davis, J. H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2016, pp. 2424–2433.

[10] S. Kwok, Multiclass classification of breast cancer in whole-slide images, in: Image Analysis and Recognition, Springer International Publishing, Cham, 2018, pp. 931–940.

[11] I. Koné, L. Boulmane, Hierarchical resnext models for breast cancer histology image classification, in: Image Analysis and Recognition, Springer International Publishing, Cham, 2018, pp. 796–803.

[12] B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, Ensemble network for region identification in breast histopathology slides, in: Image Analysis and Recognition, Springer International Publishing, Cham, 2018, pp. 861–868.

[13] M. Ferlaino, C. A. Glastonbury, C. Motta-Mejia, M. Vatish, I. Granne, S. Kennedy, C. M. Lindgren, C. Nellåker, Towards deep cellular phenotyping in placental histology, CoRR abs/1804.03270.

[14] Shallu, R. Mehra, Breast cancer histology images classification: Training from scratch or transfer learning?, ICT Express 4 (4) (2018) 247 – 254.

[15] N. Ing, Z. Ma, J. Li, H. Salemi, C. W. Arnold, B. S. Knudsen, A. Gertych, Semantic segmentation for prostate cancer grading by convolutional neural networks, in: SPIE Medical Imaging, Vol. 10581, 2018.

[16] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, B. S. Knudsen, A. Gertych, C. W. Arnold, Path r-cnn for prostate cancer diagnosis and gleason grading of histological images, IEEE Transactions on Medical Imaging 38 (2018) 945–954.

[17] R. Srivastava, R. Kumar, S. Srivastava, Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features, Journal of Medical Engineering.

[18] O. Oğuz, A. E. Çetin, R. c. Atalay, Classification of hematoxylin and eosin images using local binary patterns and 1-d sift algorithm, in: IWCIM, Vol. 2, 2018.

[19] L. Gorelick, O. Veksler, M. Gaed, J. A. Gómez, M. Moussa, G. Bauman, A. Fenster, A. D. Ward, Prostate histopathology: Learning tissue component histograms for cancer detection and classification, IEEE Transactions on Medical Imaging (TMI) 32 (10) (2013) 1804–1818.

[20] M. T. Farooq, A. Shaukat, U. Akram, O. Waqas, M. Ahmad, Automatic gleason grading of prostate cancer using gabor filter and local binary patterns, in: 2017 40th International Conference on Telecommunications and Signal Processing (TSP), 2017, pp. 642–645.

[21] M. Dinesh Kumar, M. Babaie, S. Zhu, S. Kalra, H. R. Tizhoosh, A comparative study of cnn, bovw and lbp for classification of histopathological images, in: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 2017, pp. 1–7.

[22] K. Nguyen, A. Sarkar, A. K. Jain, Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei, IEEE Transactions on Medical Imaging (TMI) 33.

[23] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, J. Tomaszeweski, Automated grading of prostate cancer using architectural and textural image features, in: 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2007, pp. 1284–1287.

[24] J. T. Kwak, S. M. Hewitt, Multiview boosting digital pathology analysis of prostate cancer, Computer Methods and Programs in Biomedicine 142 (2017) 91–99.

[25] J. Ren, E. Sadimin, D. J. Foran, X. Qi, Computer aided analysis of prostate histopathology images to support a refined gleason grading system, in: Proceedings Volume 10133, Medical Imaging 2017: Image Processing, 2017.

[26] N. Zhou, A. Fedorov, F. Fennessy, R. Kikinis, Y. Gao, Large scale digital prostate pathology image analysis combining feature extraction and deep neural network, arXiv e-prints (2017) arXiv:1705.02678.

[27] M. Valkonen, K. Kartasalo, K. Liimatainen, M. Nykter, L. Latonen, P. Ruusuvuori, Metastasis detection from whole slide images using local features and random forests, Cytometry. Part A : the journal of the International Society for Analytical Cytology 91.

[28] A. Pimkin, G. Makarchuk, V. Kondratenko, M. Pisov, E. Krivov, M. Belyaev, Ensembling neural networks for digital pathology images classification and segmentation, in: Image Analysis and Recognition - 15th International Conference, ICIAR 2018, 2018, pp. 877–886.

[29] A. Gertych, N. Ing, Z. Ma, T. J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M. B. Amin, B. S. Knudsen, Machine learning approaches to analyze histological images of tissues from radical prostatectomies, Computerized Medical Imaging and Graphics 46 (2015) 197 – 208, information Technologies in Biomedicine.

[30] K. Rajpoot, N. Rajpoot, Svm optimization for hyperspectral colon tissue cell classification, in: Medical Image Computing and Computer-Assisted

Intervention (MICCAI) 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 829–837.

[31] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2006.

[32] M. Kandemir, C. Zhang, F. A. Hamprecht, Empowering multiple instance histopathology cancer diagnosis by cell graphs, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, Springer International Publishing, Cham, 2014, pp. 228–235.

[33] A. Damianou, N. Lawrence, Deep Gaussian processes, in: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, Vol. 31 of Proceedings of Machine Learning Research, PMLR, Scottsdale, Arizona, USA, 2013, pp. 207–215.

[34] M. Kandemir, Asymmetric transfer learning with deep gaussian processes, in: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37 of Proceedings of Machine Learning Research, PMLR, Lille, France, 2015, pp. 730–738.

[35] G. Campanella, V. Werneck Krauss Silva, T. J. Fuchs, Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology, arXiv e-prints (2018) arXiv:1805.06983.

[36] E. Arvaniti, K. S. Fricker, M. Moret, N. J. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, M. Claassen, Automated gleason grading of prostate cancer tissue microarrays via deep learning, Scientific Reports.

[37] Openseadragon, http://openseadragon.github.io/, accessed: 10-07-2018.

[38] H. Salimbeni, M. Deisenroth, Doubly stochastic variational inference for deep gaussian processes, in: Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4588–4599.

[39] M. Bauer, M. van der Wilk, C. Rasmussen, Understanding probabilistic sparse Gaussian process approximations, in: Advances in Neural Information Processing Systems, 2016, pp. 1533–1541.

[40] M. Titsias, Variational learning of inducing variables in sparse gaussian processes, in: Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Vol. 5 of Proceedings of Machine Learning Research, PMLR, 2009, pp. 567–574.

[41] M. Pietikäinen, A. Hadid, G. Zhao, T. Ahonen, Computer Vision Using Local Binary Patterns, Springer, 2011.

[42] Z. Guo, L. Zhang, D. Zhang, Rotation invariant texture classification using lbp variance (lbpv) with global matching, Pattern Recognition 43 (3) (2010) 706 – 719.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR), 2015.

[44] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[46] N. Coudray, P. Santiago Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nature Medicine 24.

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet

Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252.