A Structured Latent Model for Ovarian Carcinoma Subtyping from Histopathology Slides

Aïcha BenTaieb[a], Hector Li-Chang[b], David Huntsman[b], Ghassan Hamarneh[a]

[a]*Department of Computing Science, Medical Image Analysis Lab, Simon Fraser University, Burnaby, Canada*
[b]*Departments of Pathology and Laboratory Medicine and Obstetrics and Gynaecology, University of British Columbia, Vancouver, Canada*

Abstract

Accurate subtyping of ovarian carcinomas is an increasingly critical and often challenging diagnostic process. This work focuses on the development of an automatic classification model for ovarian carcinoma subtyping. Specifically, we present a novel clinically inspired contextual model for histopathology image subtyping of ovarian carcinomas. A whole slide image is modelled using a collection of tissue patches extracted at multiple magnifications. An efficient and effective feature learning strategy is used for feature representation of a tissue patch. The locations of salient, discriminative tissue regions are treated as latent variables allowing the model to explicitly ignore portions of the large tissue section that are unimportant for classification. These latent variables are considered in a structured formulation to model the contextual information represented from the multi-magnification analysis of tissues. A novel, structured latent support vector machine formulation is defined and used to combine information from multiple magnifications while simultaneously operating within the latent variable framework. The structural and contextual nature of our method addresses the challenges of intra-class variation and pathologists' workload, which are prevalent in histopathology image classification. Extensive experiments on a dataset of 133 patients demonstrate the efficacy and accuracy of the proposed method against state-of-the-art approaches for histopathology image classification. We achieve an average multi-class classification accuracy of 90%, outperforming existing works while obtaining substantial agreement with six clinicians tested on the same dataset.

Keywords: Ovarian carcinoma, subtyping, digital pathology, machine learning, support vector machines, latent representation.

1. Introduction

According to the World Health Organization, ovarian cancer is the fifth most common cancer type worldwide and its outcomes are the poorest among women (Prat, 2012). Clinical differences between histologic subtypes of ovarian cancer have long been recognized, but it is only recently that pathologists have been able to define carcinomas in a way that correlates well with clinical and molecular differences (Prat, 2012; Racoceanu and Capron, 2016). Currently, five main histologic types of ovarian carcinomas (cancers derived from epithelial cells) have been identified (figure 1): high-grade serous (HGSC), endometrioid (EN), clear cell (CC), mucinous (MC) and low-grade serous (LGSC). It is now recognized that these ovarian carcinoma subtypes can not

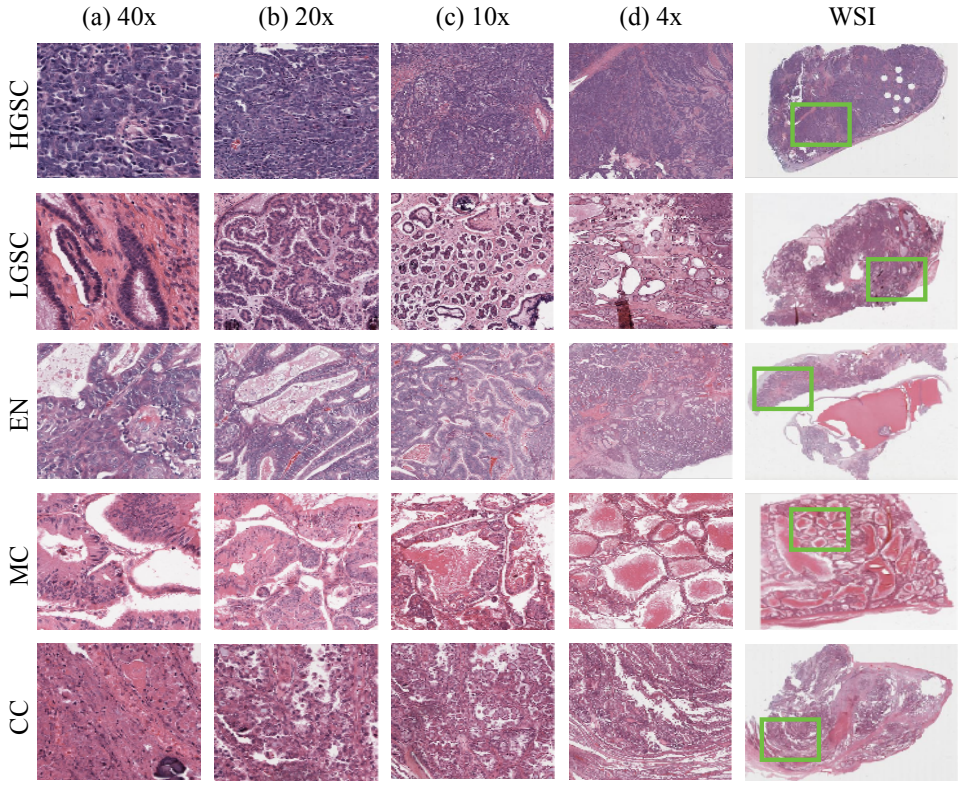

Figure 1: Whole slide images of ovarian carcinoma subtypes. HGSC: High Grade Serous Carcinoma, LGSC: Low Grade Serous Carcinoma, EN: Endometrioid carcinoma, MC: Mucinous Carcinoma, CC: Clear cell Carcinoma

. Columns from left to right correspond to the appearance of tissues at a selected region on the WSI (green box) for decreasing magnification levels.

(and should not) be treated equivalently and necessitate accurate classification for a personalized treatment.

Despite recent advances in the understanding of these histotypes, patients suffering from ovarian carcinomas still have poor prognostic rates. The success of cell-type-specific chemotherapy regimens and personalized treatments is contingent on a reliable and accurate subtyping or characterization of these cell-types from tissue sections.

Presently, clinical diagnosis of ovarian cancer involves the subtyping of ovarian carcinomas and is derived from the microscopic analysis of tissue sections, either from biopsies or resection specimens, that are mounted on glass slides, stained with hematoxylin and eosin (H&E) (Lalwani et al., 2011), and examined using light microscopy. Digitized tumor biopsies or whole slide images (WSI) are used by a small number of research centres and clinical laboratories, but their use (so-called "digital pathology") is expected to increase over time. Staining is used to highlight nuclei and the cellular content known as cytoplasm (the cells main biological components, which are naturally transparent) with various shades of blue and red.

During diagnosis, pathologists scan the tissue biopsy under a microscope seeking relevant abnormalities to diagnose each carcinoma type. These abnormalities appear at multiple magnifications. At lower magnifications, tissues organization resulting from cell proliferation leads to specific architectural patterns that are recognized as malignant and suggest cancer subtypes. At high magnification, cellular appearances confirm the histologic subtype, and nuclear shape and size are often seen as indicators of the risk of cancer progression. Thus, a clinical diagnosis or subtyping of ovarian carcinomas is the outcome of a visual-cognitive combination of these magnification-specific cues extracted from tissues.

The nature of the diagnostic procedure implies an inherent element of interpretation and hence subjectivity, and major errors can occur in pathology that have the potential of being undetected without appropriate safeguards (such as automatic review of new malignant diagnoses). The rate of major errors has been estimated to be in the range of 1.5 to 5% (Frable, 2006). In a recent study (Gavrielides et al., 2015) involving 114 patients and three expert pathologists, it was found that pathologists disagree on ovarian cancer cell-type classification on average 13% of the time, with a maximum disagreement on MC (21.4%) and EN (10%) cases. In practice, pathologists often end up scanning large amount of tissues for a diagnosis. When multiple tissue sections are not available, molecular features are required (e.g. p53 staining). Finally, challenging cases often require additional resource-intensive tests (e.g. immunohistochemistry) or asking for expert pathologists' opinion before agreeing on a diagnosis. Consequently, there is currently a need for faster, more robust, and reproducible systems that would complement and assist pathologists and clinicians during the diagnosis of ovarian carcinomas (Hipp et al., 2011).

Nonetheless, designing such automatic systems is a challenging task, as ovarian carcinomas are diverse and exhibit large intra-class variation. Besides, WSIs represent a computational hurdle as they contain large amount of information but may be composed of only a small number of important regions, while the remaining parts are irrelevant for classification. In practice, pathologists can easily spot these irrelevant regions (e.g fibrous tissue or apoptotic cells common to many types of cancers) and discard them during their analysis. However, building a computational model that can correctly identify and categorize these regions of interest without the need for extra manual annotation is challenging. Arguably, such a model must reason about which *combination* of spatial regions, and at what magnification levels, diagnostically relevant evidences occur.

Ovarian carcinomas are the result of an abnormal growth of epithelial cells. Epithelial tissues are formed by an ensemble of similar cells whose core is a nuclei and a cytoplasm enclosed in a membrane. Figure 1 presents examples of clinico-pathologic features observed on each ovarian carcinoma subtype at different levels of magnification. First, at 40x and 20x, an abnormally high proliferation of nuclei is observed (figure 1-a,b). This cellular growth further causes characteristic glandular organization and a solid appearance to the tissue that can be visualized under 10x and 4x magnification (figure 1-c,d). Analysis of tissues at an isolated magnification can rarely lead to an effective characterization of the tumour type. In the case of HGSC and LGSC (figure 1), the highest magnification is ambiguous as it shows similar nuclei grade and proliferation for both carcinoma subtypes. At the other end of the spectrum, the lowest magnification shows how these nuclei formed cells that organize into glands with specific patterns distinctive of each tumour subtype (e.g. micropapillary in the case of HGSC vs. [macro]papillary for LGSC). At 40x, EN, CC and MC show subtle variations with a few to many mitotic figures and a lack of nuclear atypia. However, at 20x, papillary patterns with little cell stratification are often observed in CC tissues while in MC tissues cells appear disorganized and form irregular glands with prominent foldings (Soslow, 2008). Finally, EN cases can also contain CC cells which complicates the

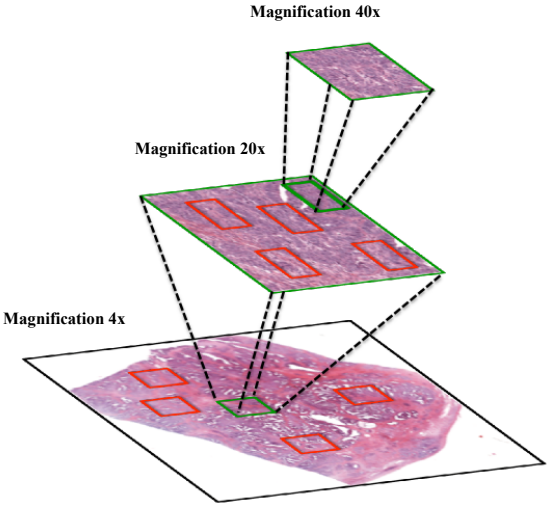

Figure 2: Contextual representation of a tissue slide. We use patches extracted at different magnifications and representing different fields of view to detect salient (green box) regions of interest in the tissue. These discriminative regions are used to infer a class label for the WSI while non-discriminative regions (red box) are discarded.

diagnosis of EN carcinomas. The distinction of EN from CC is generally based on the nuclear features of CC cells observed in EN tissues and the architectural features of CC carcinomas. While clinicians implicitly combine the contextual information gathered from multiple magnification levels, building a model that can correctly encode such structure is not straightforward.

We present a novel *contextual* model for ovarian carcinomas cell-types classification. We use the term context to describe the aforementioned multi-magnification appearance of the biologic construction of ovarian tumour tissues. Our model uses a structured latent variable framework to localize discriminative regions within tissue sections. The proposed contextual encoding is implemented via a pyramidal organization of regions (i.e tiled regions at highest magnifications come from different spatial locations of a region at lower magnification) and matched to training patches of the same cell-type (figure 2). Our method belongs to the category of weakly-supervised machine-learning approaches and does not require extra annotations of salient regions on WSI. The image of an unseen tissue sample is represented as a composition of related training images and a carcinoma subtype is determined by the composition that best predicts the given test image. While many works in histopathology image analysis investigate the tasks of cancer characterization (e.g. identifying malignant vs benign tumour), staging and grading; this work focuses on the task of cancer subtyping and its challenges.

The main contributions of this paper are in the theoretical development and formulation of a novel learning algorithm that mimics the reasoning of expert clinicians and pathologists for the analysis of multi-magnification histopathology slides, and the design of an effective feature learning strategy for which we present a complete validation with comparisons to multiple baseline works as well as trained clinicians. Finally, discriminative regions are highlighted to the user on the whole slide image, which is useful for the user's confirmation and comprehension of how the automatic method arrived at its decision.

2. Related works

Generally, automatic histopathology image classification approaches follow the archetype of feature extraction followed by classification using a trained classifier (Gurcan et al., 2009; Veta et al., 2014; Irshad et al., 2014). Typically, existing works either attempt to design new features (Petushi et al., 2006; Basavanhally et al., 2013; Doyle et al., 2012; Kothari et al., 2013) that are related to the specific histology task and adopt well-established classifiers (e.g. SVM (Kothari et al., 2013; BenTaieb et al., 2016), Boosting (Doyle et al., 2012)) or, focus on the classifier-design (DiFranco et al., 2011; Petushi et al., 2006; Basavanhally et al., 2013; Zhang et al., 2013) while using standard features (e.g. color (DiFranco et al., 2011), texture (Wang and Yu, 2013), segmentation (Zhang et al., 2013)). Color-based features combined with Haralick texture features have shown to be successful in predicting breast (Zhang et al., 2013), prostate (DiFranco et al., 2011), ovarian (BenTaieb et al., 2016) and lung (Wang and Yu, 2013) cancer from non-cancerous tissues. More elaborate features have been designed for specific applications (Roux et al., 2013). For example, Petushi et al. (Petushi et al., 2006) found the amount of nuclei with dispersed chromatin to be a relevant marker for differentiating grades of breast cancer.

While appropriate for individual tasks, feature design typically requires a large amount of labeled data and these human-engineered features are often unable to capture the complex visual variations found in histopathology images (Gutiérrez et al., 2013; Qureshi et al., 2009). More recently, feature learning methods have shown to be successful in classifying cancerous from non-cancerous tissues (Chang et al., 2013), mitotic cells (Sirinukunwattana et al., 2015) as well as ovarian carcinoma subtypes (BenTaieb et al., 2015). These methods overcome the limitations of human-engineered features by automatically identifying patterns (or features) that collectively form a compact and meaningful representation of the data, with no need for expert input or labeled examples. Also, recent works (BenTaieb et al., 2015) have shown that the learnt features can capture complex visual patterns with cell-like shapes and nuclei structures that are biologically relevant for tissue analysis.

Inspired by pathologists procedure that employ a multi-magnification approach to analyze tissue slides (Krupinski et al., 2006; Roa-Peña et al., 2010); a few works in automatic histopathology classification have focused on designing (Basavanhally et al., 2013; Doyle et al., 2012) or learning (Romo et al., 2014) magnification-specific features. The proposed methods mimic the pathologist‘s diagnosis by analyzing tissues from the lowest magnification levels in terms of texture and color appearance, and use the higher magnification levels to collect more detailed information such as nuclei and cell abundance.

Fewer works focused on the classifier design. Doyle et al. (Doyle et al., 2012) make use of a boosted Bayesian classifier operating on patches from multiple magnifications, to automatically detect prostate cancer regions and their Gleason grades. A boosting classifier was also adopted by Basavanhally et al. (Basavanhally et al., 2013) in order to classify low and high grades of breast cancer from quantitative features extracted at different scales (i.e different image sizes). Other works have used a patch-based representation where patches were discriminative regions of interest gathered from annotated data (Kothari et al., 2012; Xu et al., 2012, 2014). Given labeled samples of cancerous and non-cancerous regions in tissue sections, Xu et al. (Xu et al., 2012, 2014) show the importance of localizing discriminative regions to learn a weakly-supervised classifier. They train a multiple instance learning model based on a patch representation of the WSI to classify cancerous regions then further cluster them into different subtypes of colon cancer.

To the best of our knowledge, existing classifiers proposed for the detection of salient regions

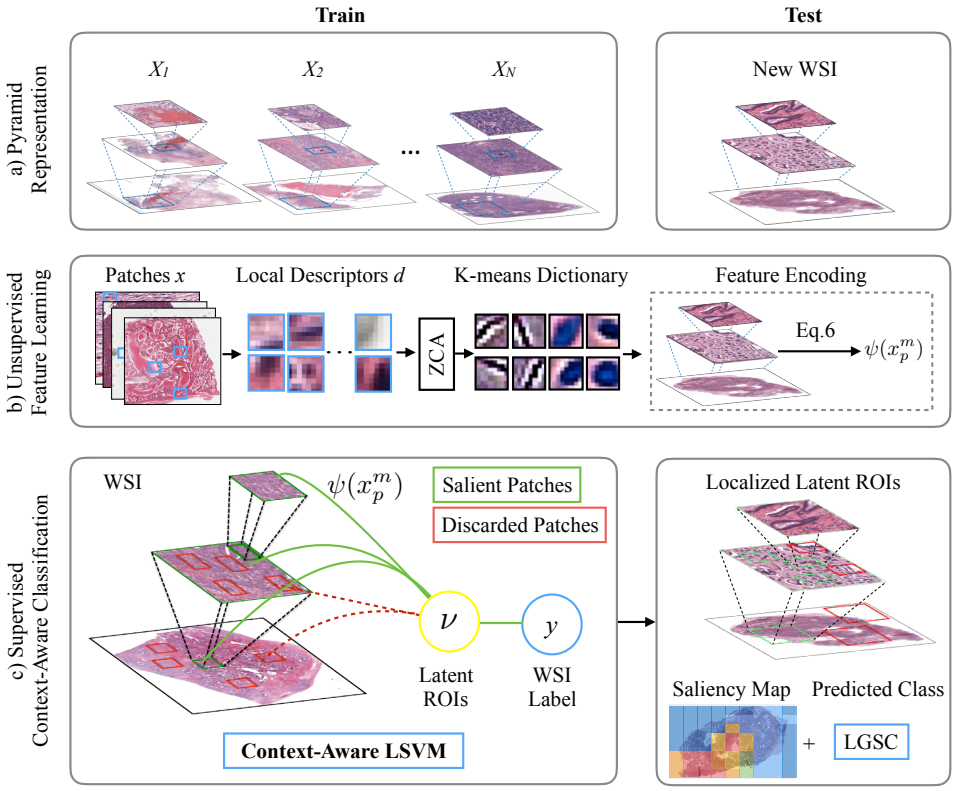

Figure 3: Proposed pipeline for ovarian carcinomas subtypes classification. (a) First, for each WSI we construct a multi-magnification pyramid. (b) Second, we use K-means dictionary learning to learn feature representations for each multi-magnification patch as described in section 3.3. (c) Finally, we train the context-aware LSVM framework (section 3) to identify salient regions within the WSI and infer a class label for the whole tissue slide. The trained model is applied to unseen tissue slides and outputs a carcinoma subtype as well as a saliency map that serves as estimated evidence for the predicted class label. Colors in the saliency map represent salient regions detected by the proposed classifier (red being the most discriminative region).

of interest in WSI, either require extra-annotated data (e.g. segmentations (Doyle et al., 2012; Barker et al., 2016), labeled regions of interest (Artan et al., 2010)), specific multi-magnification features (Basavanhally et al., 2013), or do not handle the structural relationship between different magnification levels (Xu et al., 2012, 2014; Barker et al., 2016). In contrast, we propose a unified framework that handles the structural and latent information embedded in large-scale histopathology images. Our method does not require extra supervision, considers multiple magnifications and scales, and generalizes to different feature types. We apply our method to ovarian carcinomas diagnosis and achieve superior classification accuracy compared to competing methods. Figure 3 depicts an overview of our WSI analysis pipeline. The following sections describe the details of our pipeline, our implementation and its validation.

Notation	Definition
$\mathcal{X}$	Set of WSI composed of one tissue slide per patient.
N	Total number of WSI (equals the number of patients in our case) in the set $\mathcal{X}$.
ν	Set of binary latent variables (indicators of which patches are discriminative).
y	Ground truth class label for a given WSI.
n	Indexing variable over the set of WSI, refers to the n^{th} WSI or patient.
x	Patch from a given WSI.
P	Total number of patches extracted from a WSI.
p	Indexing variable, refers to the p^{th} patch.
M	Total number of magnifications.
m	Indexing variable, refers to the m^{th} magnification level.
V	Constant describing the total number of patches selected at a given magnification.
w	Classifier's parameters, learned using eq.(4).
$\phi(\mathcal{X}, y, \nu)$	Joint feature vector describing the relation among $\mathcal{X}$, y and ν.
$\psi(x)$	Feature vector representation of the input patch x.
$f_w(X, y)$	Scoring function that sets the latent variables ν given the model's parameters w.
y^*	Predicted class label.
D	Learned dictionary of K-means centroids.
d	Local descriptors extracted from WSI patches to learn the K-means dictionary D.
w_d, h_d	Width and height of d^{th} local descriptor.
W_p, H_p	Width and height of p^{th} patch x.
$\|K\|$	Number of elements in the dictionary D.

Table 1: Summary of most used notations as they appear in the method section.

3. Proposed method

Our goal in this paper is to develop a novel weakly supervised learning framework for ovarian carcinomas subtype classification, i.e only WSI labels, indicating the presence (not the location) of a particular ovarian carcinoma within the imaged tissue, are provided. The model should produce accurate classification of tissue images as well as regions of interest within each WSI that captures the discriminative essence of the tumour subtype.

Motivated by pathologists' diagnostic procedure, our method is based on the assumption that only a few salient regions of interests (ROIs) exists within the large WSI and that these regions contain discriminative features in, at least, one magnification level. We seek for these regions of interest at different spatial locations, fields of view (scale) and magnifications of the tissue slide. By introducing latent variables in our proposed method, salient ROIs are detected within the WSI. Furthermore, each ROI is analyzed at multiple magnifications through the use of a structured formulation on the latent variables. We now provide the details of the proposed learning framework referred to as the context-aware classification model.

3.1. Notation

To facilitate the reader's comprehension, we summarized the most frequently used notations in Table 1. We are given a set $\mathcal{X}$ of N WSI corresponding to N different patients (or training instances, as each WSI corresponds to a single patient) and their corresponding class labels

$y \in \mathcal{Y} = \{0, 1, 2, 3, 4\}$ referring, respectively, to each carcinoma: HGSC, LGSC, EN, MC and CC. We assume a data instance is an observed 2D color image or WSI from a single patient and is composed of a set of patches extracted at different spatial locations. Each instance is also associated with latent variables ν that capture some unobserved information about the data. Here, this information corresponds to a subset of salient and discriminative patches or ROI.

For each instance X, patches are extracted at multiple magnifications of an optical microscope built up as pyramids of image series collected from different locations (emulating panning, see figure 2). Concretely, the histopathology image data of a patient n is represented by $X^{(n)}$ which is composed of $[x_1^1, x_2^1, \ldots, x_1^2, x_2^2, \ldots, x_i^j, \ldots, x_P^M]^{(n)}$, sorted from lowest to highest magnification patches where M is the highest magnification (figure 2). Here, x_i^j refers to the i^{th} patch extracted at magnification j. P is the total number of patches extracted from a WSI and M is the highest magnification level. Patches are represented using a feature vector $\psi(x_i^j)$. Latent variables $\nu = [\nu_1^1, \nu_2^1, \ldots, \nu_1^2, \nu_2^2, \ldots, \nu_i^j, \ldots, \nu_P^M]^{(n)}$ are associated with each patch from a WSI such that $\nu_i^j \in \{0, 1\}$ is a binary variable that indicates whether patch x_i^j is selected. In the next sections, we describe the different components of the proposed context-aware classification model. Note that the method is not bound to a specific feature representation ψ and although we present in subsection 3.3 an example of feature learning strategy computationally effective for large scale image analysis, we tested our proposed context-aware classification model on different state-of-the-art feature representations (section 4).

3.2. Context-Aware LSVM Classifier

3.2.1. Scoring function: finding salient regions

Each WSI $X^{(n)}$ is to be classified with a carcinoma subtype y. We formulate the learning model, with parameters w, for scoring a tissue slide $X^{(n)}$ with a label y using a linear scoring function denoted by $w^T \phi(X^{(n)}, y, \nu)$.

During training, we learn the set of weights w which parameterize the scoring function and the latent variables ν identifying the ROI. The weights w per class label y are defined such that $w = [w_{(y)}^1, \ldots, w_{(y)}^M]$ for patches of all M magnifications. Specifically, we learn a set of weights for each class and each magnification level which allows us to identify discriminative features from multiple magnification levels. In the scoring function, $\phi(X^{(n)}, y, \nu)$ is a potential function that allows for different components of w to be active for different class labels (e.g. we learn a linear model for each class y) given the limited set of discriminative ROIs. Thus, the scoring function measures the compatibility of a class label y with the WSI $X^{(n)}$ given latent variables ν.

We consider a low-magnification patch x_p^m as discriminative (i.e $\nu_p^m = 1$), if at least one of its respective higher magnification patches contains discriminative information, and is thus selected. This condition induces a hierarchy structure between latent variables at different magnification levels. Also, we assume that only a subset of spatial locations is considered as discriminative over the whole tissue. Therefore, the following constraints are imposed on the selection of binary variables: $\forall m_1 < m_2 \in M, \nu^{m_1} \leq \nu^{m_2}$ and $\forall m \in M, \sum_{p=1}^{P} \nu_p^m \triangleq V$. V is a user-defined variable that corresponds to the number of patches to select at a given magnification.

Formally, given $\psi(x_p^m)$ the visual feature representation of the p^{th} patch extracted at magnification m for patient $X^{(n)}$, the scoring function for patient $X^{(n)}$ is defined as follows:

$$w^T \phi(X^{(n)}, y, \nu) = \sum_{m=1}^{M} \sum_{p=1}^{P} \left[{w_{(y)}^m}^T \psi(x_p^m) \right] \nu_p^m \tag{1}$$

Note how we have multi-magnification scores (e.g. ${w^m_{(y)}}^T \psi(x^m_p)$) and the sum of lower magnification scores appears to reflect the information gathered from their respective higher magnification patches through the constraints defined on binary operators ν^m_p.

3.2.2. Learning formulation

We find the set of patches ν^M_p and their corresponding lower magnification representation (M being the highest magnification) that maximizes the prediction score for a given patient $X^{(n)}$. Intuitively, by maximizing the prediction score, we seek the latent variables (thus the spatial regions in the WSI) that allow the model to predict a class label that agrees with the expert-provided carcinoma subtype. This maximization step can be related to pathologist's diagnostic approach during which they scan the WSI seeking relevant cues (salient regions) that are compatible or agrees the most with their predicted diagnostic. The function $f_w(X^{(n)}, y)$, sets the latent variables ν by optimizing the following equation:

$$f_w(X^{(n)}, y) = \max_{\nu} w^T \phi(X^{(n)}, y, \nu) \quad \forall y \in \mathcal{Y}. \tag{2}$$

Formally, we score the set of multi-magnification patches $X^{(n)}$ of each patient according to the model's parameters w using eq.(2) where the scoring function $w^T \phi(X^{(n)}, y, \nu)$ is defined in eq.(1). To infer the latent variable $\nu^* = \arg\max_{\nu} w^T \phi(X^{(n)}, y, \nu)$, as there is a dependency between latent variables at different magnifications, we have to infer latent variables for each magnification sequentially. Considering that ν^j_i is binary for any patch i at magnification j, we first infer latent variables at the highest magnification M such that $\nu^M_i = 1$ for the V patches x^M_i with maximal score $w^M \psi(x^M_i)$. Then, for all lower magnification patches containing the highest magnification patches selected previously by ν^M_i, we infer the latent variables $\nu^m_i = 1$ where $m < M$ for the top V patches at magnification m with maximal score.

Once the latent variables are estimated, we infer the class label $\hat{y}$ that maximizes the score of prediction, given the model's parameter w:

$$\hat{y} = \arg\max_{y} f_w(X^{(n)}, y) \tag{3}$$

3.2.3. Training objective function

During training, we learn the model parameters w that maximize the classifier's score of prediction given the ground truth labels. Given N training instances X, we use the standard multiclass latent SVM (LSVM) objective function (Felzenszwalb et al., 2008) to optimize the weight parameters in eq.(4). In our optimization, we use a zero-one loss Δ to penalize wrongly predicted labels $\hat{y}$ given the ground truth label $y^{(n)}$. A coefficient α is used as an additional cost on the model for making classification mistakes on the least represented classes during training. This coefficient is particularly useful in the case of highly imbalanced datasets such as ovarian carcinomas subtypes. α is defined as a vector corresponding to the prevalence of each class in the training set and is used to bias the model to pay more attention to the minority class.

$$\begin{aligned} \mathcal{L}_w = \min_w \frac{1}{2}\|w\|^2 \quad &+ \quad C \sum_{n=1}^{N} \max_{\hat{y} \in \mathcal{Y}} f_w(X^{(n)}, \hat{y}) + \alpha \Delta(y^{(n)}, \hat{y}) \\ &- \quad C \sum_{n=1}^{N} f_w(X^{(n)}, y^{(n)}), \end{aligned} \tag{4}$$

Algorithm 1: Training Context-Aware LSVM Model

Input: : Labeled training instances and hyper-parameters $\mathcal{X}, y, V, C$ (described in Table 1)
T: number of training iterations, $X^{(n)}$: n^{th} WSI $\in \mathcal{X}$

Output: Optimal set of parameters w and latent variables ν (described in Table 1)

Initialize w_1 (w_i: value of w at iteration i);
for $t \leftarrow 1$ ***to*** T **do**
  for $n \leftarrow 1$ ***to*** $|\mathcal{X}|$ **do**
    for $p \in X^{(n)} = [x_1^1, x_2^1, \ldots, x_1^2, x_2^2, \ldots, x_i^j, \ldots, x_P^M]^{(n)}$ **do**
      Compute magnification-specific scores using eq.(1);
      Set the latent variables ν using eq.(2);
      Infer the predicted class labels y given current w and ν via eq.(3);
    end
  end
  Compute the loss function using predicted labels and eq.(4) ;
  Compute the gradient of the loss function as in eq.(5);
  Compute $[w_{t+1}, w_t^*, \text{gap}]$ using (Do and Artières, 2009), alg.1;
  if $gap \leq \epsilon$ *or* $t == T$ **then**
    return w_t^*;
  end
end

where C is the usual LSVM slack tradeoff constant.

Equation (4) is a non-convex optimization problem. However, the learning problem becomes convex once the latent variable ν is fixed for positive instances. Therefore, we train the LSVM by an iterative algorithm that alternates between inferring ν on positive instances and optimizing the model parameters w. To solve eq.(4), we use the non-convex regularized bundle optimization (NRBM) introduced by Do et al. (Do and Artières, 2009). This iterative optimization method is an extension of the popular cutting plane technique to non-convex functions. Briefly, at each iteration, this method finds a new linear cutting plane using the sub-gradient of the objective function. The sub-gradient of eq.(4) corresponds to the following linear equation:

$$\frac{\partial \mathcal{L}_w}{\partial w} = w + \frac{C}{N}(\phi(X^{(n)}, y^*, \nu^*) - \phi(X^{(n)}, \hat{y}, \hat{\nu})) \tag{5}$$

where $\hat{\nu} = \arg\max_{\nu} w^T \phi(X^{(n)}, \hat{y}, \nu)$ is the latent variable inferred for the predicted label $\hat{y}$. Each cutting plane is then added to a piecewise quadratic approximation of the objective function, thus leading to an increasingly accurate approximation. The different steps to train the proposed classification model are described in algorithm 1.

3.2.4. Applying the trained classifier to unseen tissue images

Given the model parameters w learned using the above procedure (eq.(4)); we perform the inference on test images. This inference will score all given WSI-class label pairs and provide a discriminative set of latent ROIs for the unseen test WSI. We label a new WSI X with class label

y^* using the following equation:

$$(y^*, \nu^*) = \arg\max_{y,\nu} w^T \phi(X, y, \nu). \tag{6}$$

Note that the inference, which involves the enumeration over possible values for y and ν, is feasible since the set of possible class labels (five) and discrete latent variables (hundreds) is limited.

3.3. Feature Representation

3.3.1. Dictionary Learning

We adopt a feature learning strategy to find a feature representation ψ for each patch from a WSI. We use K-means clustering in a bag-of-words framework to learn an over-complete dictionary (the size of the dictionary is greater than the dimensionality of the input images) of visual words that can best reconstruct an input image.

We first randomly collect local RGB image descriptors $d \in \mathcal{R}^{w_d \times h_d \times 3}$ from our set of whole slide images. These local descriptors form the input to our K-means based feature learning strategy. To reduce the chances of obtaining highly correlated visual words when learning the dictionary, we use ZCA whitening (Krizhevsky and Hinton, 2010) to re-scale the input local descriptors and remove the correlation. This operation reduces the redundancies in the data by removing the covariance between local descriptors and normalizing the variance while keeping the re-scaled data as close as possible to the original data.

After whitening the input, we use K-means clustering to find a set $K = \{c_1, c_2, \ldots, c_{|K|}\}$ of centroids. These centroids form our dictionary $D \in \mathcal{R}^{w_d \times h_d \times 3 \times K}$ of "visual words". We use a common heuristic to initialize the K-means algorithm, which is to randomly initialize the centroids from a normal distribution then normalize them to unit length (Coates et al., 2011).

3.3.2. Feature Encoding

The learned centroids are used to map any RGB input data d to a code vector that minimizes the reconstruction error. This code vector is a parsimonious and simpler representation than the original data that ends up being more suitable for classification tasks.

We use a non-linear mapping function $f_k(d)$ for every centroid, as a soft-quantization method to map each input d to a $|K|$-dimensional code vector. $f_k(d)$ selects the set of visual words c_k with highest activation. For a given input d, $f_k(d)$ is defined as follows:

$$f_k(d) = \max\{0, \mu(z) - z_k\} \tag{7}$$

$$z_k = \|d - c_k\|_2^2, \tag{8}$$

where $\mu(z)$ is the mean of elements in cluster z. For any given local descriptor d, the mapping function f returns a feature vector of size equal to the number of centroids by introducing a form of competition between different centroids. By applying the mapping function f to many local descriptors of an image, we obtain a feature representation of the entire image. More specifically, we densely apply the function f to all local descriptors d of size $w_d \times h_d$ and obtain a $|K|$-dimensional feature representation at every location i, j of a local descriptor. We will refer to $w_d \times h_d$ as the receptive field of the feature encoding step. The distance separating two consecutive local descriptors on the image is usually referred to as stride s. After applying the mapping

function f over all local descriptors, we obtain a feature representation ψ for a patch x of size $W_p \times H_p \times 3$ from a WSI. ψ is of size $(\frac{W_p - w_d}{s} + 1) \times (\frac{H_p - h_d}{s} + 1) \times K$.

Once the new image representation ψ computed for all image descriptors d, we use max pooling (Coates et al., 2011) to reduce the final features dimensionality before classification. Pooling induces translation invariance by aggregating feature responses from a small spatial region of the input image. We use max-pooling over square sub-regions of the new image representation. Specifically, we split ψ into four equal-sized quadrants and compute the max over all local descriptors' feature representation in each quadrant. We obtain a final pooled feature vector of length $4|K|$.

3.4. Implementation details

To create the dictionary of visual words D, we sample visual words from each training set multi-magnification patches. In order to build an over-complete dictionary, we ensure that the number of local descriptors extracted is reasonably large. In practice, training a K-means dictionary requires a larger number of input local descriptors than is necessary for other algorithms (e.g. sparse feature learning). We collected a total of 400 000 local descriptors randomly extracted from the total set of patches from whole slide images. We used different receptive field (size of local descriptors) sizes for local descriptors extracted from different patch magnifications. We used receptive fields of size 12×12 for local descriptors from 4x patches, 8×8 for local descriptors from 10x patches and 4×4 for local descriptors from patches at magnifications 20x and 40x. The receptive field sizes were set via cross-validation. In practice we observed that smaller receptive fields tend to give better performance. We learned a dictionary of $|K|= 4000$ centroids from these local descriptors. To encode features from the learned dictionary, we densely applied the mapping function f to all image patches with a stride $s = 1$ which was the most computationally efficient given our dataset size and hardware equipment. This resulted in a 4000-dimensional feature vector per patch. The dictionary size was also defined via cross-validation over the training set but the classification error was relatively insensitive to this parameter varying only about 1% when changing $|K|$ by $\pm 5\%$.

Linear SVM is used to initialize the model parameters w. Latent variables v can then be inferred for positive samples using eq.(2). Subsequently, we assign a tissue sample to the cell-type maximizing the score of prediction given all latent variables following eq.(3). At test time, multi-magnification patches extracted in a pyramidal manner are used to detect patterns learnt during training. We then find the carcinoma type y that maximizes the score of prediction given the observed latent variables (patches) and model's parameters for the unseen tissue section. Selecting V salient patches from the set of P total patches in X can be done in $O(Plog(V))$ time. In our experiments, this inference takes 0.02 seconds for a WSI represented by 120 patches on an Intel E8400 CPU @3.00GHz using unoptimized MATLAB R2014b code.

4. Experiments and results

4.1. Experimental settings

We evaluated our method on a dataset [1] of 133 whole slide H&E tissue sections from 133 different patients. The dataset was digitized using Aperio ScanScope™ digital slide scanner (Leica

[1] In order to facilitate direct comparisons to our work, we make this dataset available along with the Matlab code at the following URL: http://tinyurl.com/hn83mvf.

Class	HGSC	EN	MC	LGSC	CC
Training set	25	14	12	11	6
Test set	24	14	11	10	6

Table 2: Dataset representation: number of cases per class.

Biosystems, Nussloch, Germany) with a highest magnification of 40x. Three expert pathologists were provided with patient's WSI and other immunologic and cytologic tests. Each patient was labelled with a carcinoma type after common agreement of the experts. Tissues in this dataset were carefully chosen to represent and gather different challenging aspects of ovarian carcinoma diagnosis. These challenges are mainly caused by genetic tissue variability, staining incoherence and scanning heterogeneity. For example, samples of serous carcinomas include different grades of HGSC (e.g. malignant, borderline, mixed) often leading to confusions with EN and LGSC. Second, this dataset also reflects real-world difficulties by being imbalanced in favor of serous carcinomas, which are the most frequently diagnosed subtypes. For the following set of experiments, we randomly sampled 68 patients for training and 65 for test. The class distribution was kept similar between train and test sets as shown in table 2.

We extracted non-overlapping patches at multiple magnification levels and created the image pyramid for each tissue image (figure 2). To create the pyramid, each WSI was partitioned into tiles corresponding to different magnifications. First, we extracted low magnification patches (4x) that correspond to patches of size $10,000 \times 10,000$ pixels from the original WSI. Note that on average, WSIs in our dataset are of size $50,000 \times 50,000$ pixels, thus we used 3 patches at 4x magnification to cover as much tissue as possible. From these low magnification patches, we created a pyramid by partitioning the image into tiles. The four quadrants of a 4x patch as well as the middle area was used to form the set of 10x patches (5 patches at 10x). This procedure was carried out to collect 20x patches (corresponding to the 4 quadrants of each 10x patch) and 40x patches (4 quadrants of the 40x patch). All patches were re-sized to 500×500 pixels and used for feature extraction.

In all experiments, patches were extracted in a hierarchical manner (figure 2) where higher magnification patches were contained in a given lower magnification patch. After patch extraction, each WSI was represented by a total set of 318 patches ($3+3\times5+3\times5\times4+3\times5\times4\times4$) from 4 different magnifications. Note that only 4x patches where selected randomly while others were selected by the model and are qualified as salient regions/patches. Although the selection of 4x patches was random, we used large non-overlapping patches to cover at least $\frac{2}{3}$ of each a tissue slide where the ratio of tissue to background is also approximately $\frac{2}{3}$. Regarding the construction of the pyramid of patches from different magnifications, the choice of number of patches extracted per magnification level mimics pathologists analysis of tissue slides which consists in a first global assessment of the tissue slide at 4x then random selection of certain regions for more detailed analysis at higher microscope resolutions. In fact, as discriminative patterns, as well as non-discriminative ones, generally reoccur uniformly in ovarian carcinomas tissue slides randomly sampling patches at 4x (with a reasonably large field of view) does not result in omitting discriminative information. Note that the random selection at 4x was purely an experimental design choice suitable for ovarian carcinomas and should be refined for other applications, however, given a pyramid of tissue patches from a WSI, our salient region detection is automatic and transferable to any histopathology image classification task.

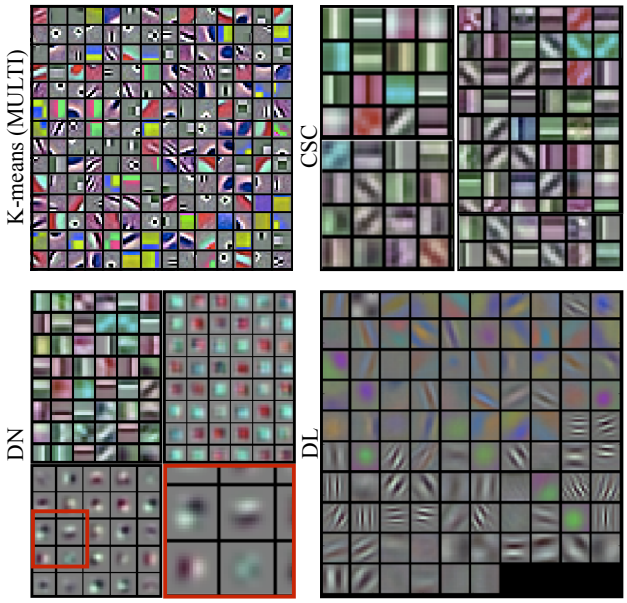

Figure 4: Filters learned using different feature learning techniques. CSC: convolutional sparse coding, DN: 3-layer deconvolution network (BenTaieb et al., 2015), DL: deep learning CNN model (Krizhevsky et al., 2012).

To assess the sensitivity of our model to the training dataset, we shuffled the total dataset and performed 5 rounds of training on half of the dataset, testing on the remaining half. The classifier hyper-parameters: V (number of ROI to select) and C (slack variable in SVM) were set via leave-one-out cross validation on training data. These settings were kept constant in each of the following experiments.

4.2. Feature construction

We tested multiple configurations (supervised and unsupervised) for learning the dictionary of visual words from the training set.

1. *MIX* dictionary: In this configuration, we used visual words extracted from different magnification patches (4x, 10x, 20x and 40x) and learned a dictionary of 4000 words. This dictionary combines multiple magnifications.
2. *MULTI* dictionary: We learned a dictionary separately for each different magnification to insure there is an equal number of centroids or visual words selected from each level of the magnification pyramid. The total number of visual words was kept fixed with 1000 words per magnification level forming a final dictionary of 4000 (1000×4) visual words.
3. *CSP* dictionary: We tested a class-specific dictionary using a supervised version of the proposed feature learning method. With our dataset being highly imbalanced, visual words learned with K-means can be biased toward the most represented classes. In this experiment, we combined five different dictionaries (one per ovarian carcinoma cell-type) from four different magnification visual words extracted individually for each magnification and

for each class. We used 800 words per class for a total of 4000 visual words. The 800 visual words were extracted from all magnification levels using the MULTI scheme.

We compared our learning-based features with popular hand designed features as well as recent convolution-based feature learning techniques.

To extract hand designed features, each multi-magnification patch was represented with a feature vector composed of color (RGB color histograms) and texture features (SIFT and Local Binary Pattern). Our latent SVM model was used to learn the set of weights w that selects the most discriminative features per patch at a single magnification. Note that we did not use the structured relation between magnifications in this experiment.

Given the recent success of deep learning and convolution-based feature learning techniques for pattern recognition applications, we also extracted convolution-based features using convolution sparse coding (CSC) (Zhou et al., 2014), a 3-layer deconvolution network (DN) proposed in our earlier work (BenTaieb et al., 2015) and the popular deep learning convolution neural network (CNN) proposed by Krizhevsky et al. (Krizhevsky et al., 2012). Parameters (e.g. dictionary and receptive field sizes for each image magnification) of CSC and DN were defined using cross-validation as described in section 3.4. Examples of the learned dictionaries are presented in figure 4.

Different encoding strategies were used along with each feature learning technique: we used standard hard quantization with CSC, Fisher vector encoding with DN and the soft quantization function f described above in eq.(2) with our K-means dictionaries. As for the CNN, given the limited size of our dataset we used a pre-trained model that we fine-tuned on our dataset, as recent studies have shown that transfer learning using pre-trained networks generally results in better performance (Shin et al., 2016). We fine-tuned the CNN model using all patches extracted at multiple magnifications from our dataset as explained in section 4.1. We used cross-validation on the training set to fix the hyper-parameters of the stochastic gradient descent optimization during finetuning. At test time, we experimented with features extracted at the different layers of the network (i.e conv4, conv5, fc6, fc7). Best results were obtained using features from the last layer of the CNN (fc7). Our cross-validation experiments showed that using patches from multiple magnifications resulted in better performance than using a single magnification, regardless of the magnification level.

All extracted features (hand designed, CSC, DN, CNN, proposed K-means features) were then fed to a multiclass linear SVM and to our proposed classifier to infer a class label for each WSI. When using linear SVM, we used max-voting over the predicted scores for all patches from a WSI to obtain a final class label per patient. This step was not needed when using our proposed classifier as it handles multiple magnification patches and infers a class-label per WSI.

Table 3 shows the multiclass classification accuracy using all three different configurations of dictionaries (MIX, MULTI, CSP) and other feature learning techniques. Our results show a significant accuracy gain (on average +17%) using learnt features compared to hand designed features and a clear advantage when learning the dictionary from multiple magnification visual words (MULTI and CSP).

Our experiments also showed that K-means feature learning, when designed for multiple classes and multiple magnifications, can outperform other highly non-linear feature learning techniques (e.g. 26% better than CNN, 15% better than DN). Note that even though we have used pretrained architectures (AlexNet) we then fine-tuned them for our problem. Designing novel deep architectures especially crafted for ovarian carcinomas subtyping may be an interesting future work. Such deep architectures may become particularly useful as bigger datasets

Features	H-D	CSC	DN	CNN	K-means		
					MIX	MULTI	CSP
Linear SVM	37.5 ± 0.1	45.0 ± 0.1	**68.5±0.1**	44.0 ± 0.5	58.5 ± 0.1	62.2 ± 0.1	66.2 ± 0.1
Proposed	43.5 ± 0.1	50.0 ± 0.1	68.5 ± 0.1	50.0 ± 0.4	62.0 ± 0.1	62.3 ± 0.1	**76.2±0.1**
T(Dictionary)	–	~ 24h	~ 36h	~ 2h	~ **0.2h**	~**0.2h**	~ **0.2h**
T(Encoding)	~ 1h	~ 14h	~ 24h	~ 1h	~ **0.5h**	~**0.5h**	~ **0.5h**

Table 3: Classification accuracy using different feature types. Accuracy is reported in percent (% ± std). H-D refers to hand designed features. T(Dictionary) corresponds to the average computation time necessary to learn the dictionary or set of filters for CNN and DN. T(Encoding) corresponds to the computation time necessary to encode features for the entire test set using the learned dictionary/filters. Note that CNN (Krizhevsky et al., 2012) and DN (BenTaieb et al., 2015) features were extracted on a 12GB NVIDIA GPU, others were on a single CPU.

become available.

When comparing the dictionaries obtained using convolution-based features vs K-means feature learning, we observed both producing similar patterns including edge-like and cell-like shapes (figure 4). This confirms that the adequate design of the dictionary, specifically adapted to histopathology images, can result in learning complex visual patterns without requiring extensive computational times (e.g. K-means feature learning was 10 to 120 times faster than other feature learning techniques).

An important factor of the success of K-means as feature learning technique was the total number and the receptive field size of the local-descriptors extracted to train the dictionary. Our experiments showed that a larger number of sampled local descriptors and smaller receptive field sizes usually resulted in better performance. While the aforementioned meta-parameters are critical, setting such parameters is usually simpler and more intuitive than setting meta-parameters for deep learning (e.g. number of layers, number of feature maps per layer) or other convolution-based techniques (e.g optimization parameters and sparsity coefficient for DN). Using different feature learning strategies, we also observed the importance of the encoding strategy used. In fact, DN features were combined with Fisher vector encoding, as proposed in (BenTaieb et al., 2015) which maps features into a very high-dimensional space, facilitating their linear separability. Unsurprisingly, using latent SVM with DN features encoded with Fisher vectors did not improve the classification performance when compared to linear SVM as features are already highly separable. This may also signal a tradeoff between classification model complexity and features representation capacity.

Finally, when using a class-specific dictionary (CSP), we were able to reach an average classification accuracy of up to 76%, outperforming all other techniques. It is worth noting that applying a random classifier, on our highly imbalanced and relatively small dataset, would achieve only 20% average classification accuracy. Also, our results demonstrate the contrast between adding supervision to the K-means feature learning method when designing CSP dictionary and using a supervised deep learning model such as CNN. In fact, while the performance of deep learning models is extremely bound to the availability of large training datasets, K-means clustering shows to be more robust to the dataset size.

4.3. Contextual representation

To demonstrate the benefit of using multiple magnifications in a structured formulation, we used 4x, 10x, 20x and 40x patches individually or combined to form one, two, three or four

Magnification Levels	1	2	3	4
DN (BenTaieb et al., 2015)	68.5%	75.0%	78.2%	78.0%
K-means (CSP)	76.2%	78.0%	**80.0%**	62.5%

Table 4: Classification accuracy using different magnification levels. We report the best accuracy achieved at each level. At level 1, the best accuracy was achieved using patches from 10x magnification, and levels 2 to 4 accuracies were achieved using: {10x-20x} (level 2), {4x-10x-20x} (level 3), {4x-10x-20x-40x} (level 4).

pyramidal layers, e.g. {4x} ,{10x}, {20x}, {40x}, {4x,10x}, {10x,20x}, {20x,40x}, {4x,10x,20x}, etc. We used the best performing features obtained in the previous experiment (table 3): our learnt CSP dictionary and DN (BenTaieb et al., 2015) features. The proposed latent SVM with structured latent variables was used as classifier. Table 4 shows the top classification accuracy using one, two, three or four magnification levels to represent a WSI.

Our experiments confirm that despite the type of features employed, the model learns more discriminative information and is able to generalize better when using a composition of patches from multiple magnifications (e.g. three magnifications: {4x-10x-20x}, achieves the optimum accuracy of 95.0%). Surprisingly, using the highest magnification 40x did not help the classification accuracy, a possible explanation maybe that the highest magnification does not transmit the structural appearance of the tissue, and, hence, using lower magnifications is advantageous. In practice, clinicians often use 10x to 20x magnifications and inspect the highest magnification level when uncertain on the diagnosis. We reach a plateau (or a slight dip) when using all four magnification levels (4x to 40x) available in our dataset, which may signal a tradeoff between utilizing information from more magnification levels and the resulting increased model complexity.

4.4. Sensitivity of the model

We evaluated our model against other "baseline" works using different training set sizes for a fixed test set of 20 patients (4 patients per class) and using our K-means CSP features. First, linear SVM was compared against as this allows us to assess the utility of using structured latent variables to model our problem. Along the line of weakly-supervised approaches adopted for cancer subtypes classification, we also compared our method to the recent work of Xu et al. (Xu et al., 2012, 2014) which is based on a multiple instance learning (MIL) framework. Finally, we show the performance of our full pipeline (K-means CSP features with latent structured SVM model) compared to our latest work using DN features with Fisher encoding and linear SVM as classifier (BenTaieb et al., 2015). All baselines were tested in similar experimental settings where we used two magnification levels (10x and 20x) to form a composition of patches for each WSI. Table 5 shows the multiclass classification accuracy obtained with each method. We also report the average training classification accuracy to be contrasted with test accuracy in order to estimate each model's generalization ability.

Our experiments showed that on average, the proposed method outperforms other baselines when using a 3-layer pyramid of magnifications (90% accuracy when using 60 training patients and testing on 20). We also observed that the proposed structured latent SVM model outperforms linear SVM with a large margin (up to 35%) which confirms our hypothesis that identifying salient regions through the use of latent variables helps training more accurate classifiers. Furthermore, we observed a clear gain (25% better) over MIL (Xu et al., 2012, 2014). This can

Train vs Test	20 vs 20		40 vs 20		60 vs 20		80 vs 20	
	Train	Test	Train	Test	Train	Test	Train	Test
SVM	100.0%	40.0%	100%	50.0%	92.5%	55.0%	91.6%	50.0%
MIL (Xu et al., 2014)	100.0%	55.0%	95.0%	65.0%	92.5%	65.0%	92.5%	60.0%
DN (BenTaieb et al., 2015)	100.0%	45.0%	98.5%	65.0%	86.7%	75.0%	91.6%	90.0%
Proposed - 2 levels	91.6%	60.0%	92.5%	65.0%	78.3%	75.0%	85.0%	70.0%
Proposed - 3 levels	**86.3%**	**75.0%**	**88.3%**	**85.0%**	**92.5%**	**90.0%**	**92.5%**	**85.0%**

Table 5: Classifier performance compared with baselines on different training set sizes. Train vs Test refers to the total number of training and test samples used.

be explained by the fact that MIL only selects discriminative patches without considering any hierarchy. Note that the performance of our latent SVM model using only one magnification (no structured representation, thus only acting as patch selection without contextual representation) is similar to MIL ($\sim$ 62%, see table 3).

The proposed method also achieves competing results with the highly non-linear Fisher features used in (BenTaieb et al., 2015). Despite its ability to generate linearly discriminative features, DN required long hours of training (>18 hours on 4800 patches randomly extracted at two magnification levels on an Intel CPU E8400 @3.00 GHz vs. 30 minutes using our CSP features with latent structured SVM model) which limits its applicability to higher magnification levels. Using our proposed method with 3 magnification levels allowed us to further improve the classification accuracy outperforming our earlier results by 15% using 60 training samples.

We also tested the sensitivity of the proposed method to different training set sizes. Generally, it is expected that larger training set sizes results in more accurate models which is what we observe up to 60 training samples. However, we also observed a slight drop in accuracy when using 80 samples. This is most likely due to the high variability between different training samples. In fact, tissue images used in this study were gathered from different centers and show high variability in staining and appearance.

To assess the generalization ability of the method, we report in table 3 the average training classification accuracy. Generally, linear models, such as SVM, tend to overfit to the training data. In contrast, weakly-supervised models (e.g. LSVM and MIL) are generally able to effectively avoid overfitting as they enforce the predictions to be based on the most discriminative subset of the data.

A final important factor to estimate is the sensitivity of the model to the class imbalance. In the case of ovarian carcinomas, the imbalance in our data is a direct consequence of the corresponding prevalence of each subtype in practice. Hence, the purpose of this experiment is to assess the model's robustness towards highly imbalanced training sets. Recall that in the proposed pipeline we addressed the class imbalance using class-specific K-means dictionary learning (described in subsection 4.2) as well as a weighted zero-one loss (described in eq.(4)). For this purpose, we tested the model on a fixed test set of 5 patients (one per class) and created different training sets with increasing levels of imbalance ratios between the most represented and least represented classes: 1:1 (perfectly balanced), 2:1, 3:1, and 4:1 (highly imbalanced). We used the mean F1 score to evaluate the impact of the imbalance on the classifier's prediction per class when each class has equal contribution to the F1 score. We observed a constant F1 score of 0.73 when using training sets with imbalance ratios of 1:1, 2:1 and 3:1 and an F1 score of 0.53 for a ratio of 4:1. Generally, the proposed model accurately recognized HGSC, CC and MC cases

Clinician	# 1	# 2	# 3	# 4	# 5	# 6	AUTO
Kappa	0.90	0.93	0.89	0.84	0.90	0.89	0.89

Table 6: Performance of our proposed automatic system compared to clinicians.

but misclassified EN and LGSC as HGSC (most prevalent class). Despite the 20% decrease in F1 score, the proposed model showed to be more robust to class imbalance than linear SVM for which we observed a significant drop in F1 score (from 0.66 to 0.43, i.e a 23% drop, when going from 1:1 to 2:1 only) for imbalanced training sets.

4.5. Agreement with clinicians

In a final set of experiments, we show the agreement between our automatic classifier and six[2] clinicians trained and tested on the same dataset and in similar conditions. Clinicians were provided with 40 WSI for training and were tested on 40 unseen WSI. These WSI were selected by expert pathologists to contain the largest amount of tumour in order to assess if in ideal conditions, clinicians diagnosis could be more reproducible. After training, clinicians were asked to predict a carcinoma type for each patient of the test set and were provided with immunostaining results to confirm or modify their predictions.

We used our feature learning technique with a class-specific dictionary (CSP) and patches from 3 magnification levels to describe a WSI. We did not use immunostaining results as features to our automatic system. We report the average Cohen's Kappa score κ in eq. (9) for each clinician with all other clinicians as well as for our automatic system with all clinicians.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}, \tag{9}$$

where $P(a)$ is the observed probability of agreement and $P(e)$ is the expected probability of agreement by chance [3]. Cohen's Kappa score is a comparison between two observers or raters who are examining the same set of categorical data.

The average κ for each observer (i.e average agreement of an observer with all remaining ones) is reported in table 6. On average, in ideal conditions, clinicians' agreement with each other ranges from 0.84 to 0.90 which is relatively good but still imperfect. The Kappa achieved by the automatic system reaches 0.89, which indicates substantial agreement with clinicians. On average, our automatic system showed an equivalent agreement with clinicians than the average of clinicians' agreement with each other ($\kappa = 0.89$).

4.6. Automatically detected salient patches

We show in figure 5, saliency maps obtained after the automatic selection of ROI by our model on test images. While there is no guarantee that salient regions automatically detected by our model using the feature representation of patches will (or should) always be interpretable, correlate with clinician's diagnosis or highlight specific morphological patterns, we systematically observed meaningful correspondences. In fact, the classifier discards apoptotic-looking

[2] Jocelyne Arseneau, Patricia M. Baker, Carol A. Ewanowich, Dan Fontaine, Robin Parker, and Martin Köbel.

[3] $P(a)$ is the number of times two observers agreed on cases, normalized by the total number of cases. $P(e)$ is the probability of each observer randomly predicting each class, assuming independent observers' predictions.

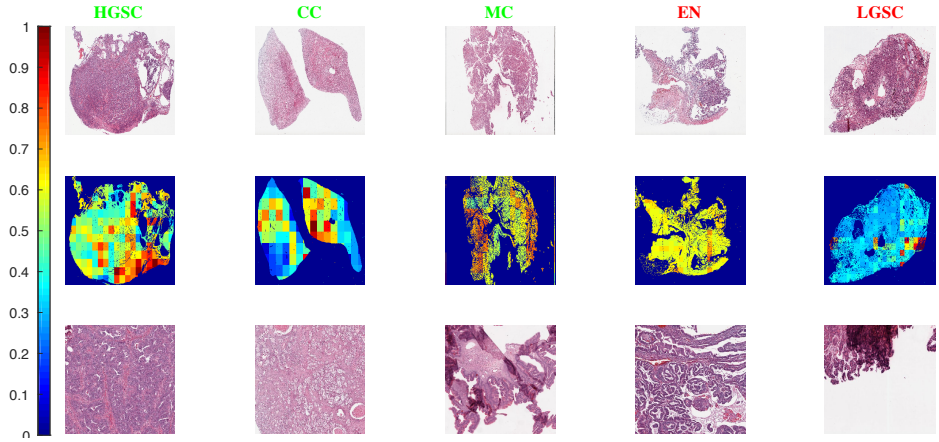

Figure 5: Automatic selection of salient ROI. Rows correspond to: 1) original WSI, 2) saliency map of the scores predicted by the classifier from a hierarchy of patches extracted at magnifications {10x, 20, 40x}, 3) 10x patch with highest classification score. The two last columns (EN and LGSC) correspond to cases mis-labelled as HGSC. Note how the selected regions (last row) for the mis-labelled cases (two last columns) are over-stained and visually appear highly similar to HGSC with very dark and abundant nuclei. Blue arrows show tissue foldings on MC carcinoma and papillary areas on CC carcinoma.

areas corresponding to dying cells. These apoptotic regions are common to all cancer subtypes hence do not contain any discriminative information but rather mislead the classification. Also, we observed that salient patches from CC cases often contain papillary-looking areas of tissue while salient regions on MC cases show prominent tissue foldings (as seen in the last row of figure 5 with blue arrows). Both these characteristics are generally used in clinical practice to diagnose MC and CC tumours. Finally, when visualizing salient patches chosen as examples of mislabeled cases for EN and LGSC (figure 5) we observed that over-stained samples visually appear very similar to HGSC with consistent dark and prominent nuclei (as seen in figure 5, last row). These cases reflect the difficulty of the classification task for an automatic system but also the benefit of visualizing regions selected by the trained classifier. In fact, in the context of a computer-aided diagnosis, visualizing salient patches reveals critical information about the automatic system, i.e which regions led it to make a particular prediction. Clinicians may find this information insightful.

In figure 6, we show the most discriminative features per class. These features correspond to the dictionary centroids for which the classifier has the highest weights w. Generally, selected centroids capture different texture patterns with a variety of directed edges that are not always semantically interpretable but we also observe different nuclei shapes specific to each carcinoma subtype. For instance, circular and uniform nuclei are representative of CC carcinomas while HGSC show more heterogeneous shapes. This correlates with what pathologists describe as biological markers for these subtypes (Prat, 2012).

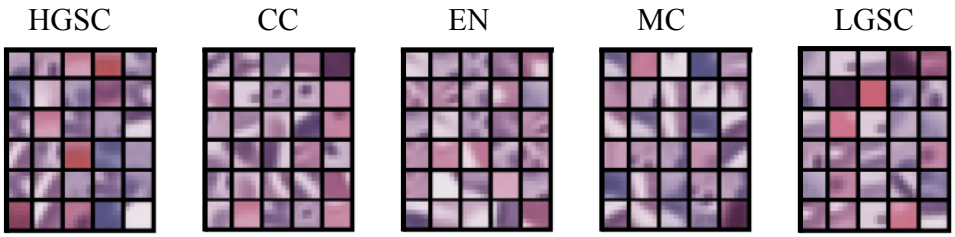

Figure 6: Top-10 centroids obtained after training the classifier.

5. Conclusions

Agreement on the subclassification of ovarian carcinoma subtypes can be resource intensive and time consuming (Cramer et al., 1987; Lund et al., 1991), and can be very difficult in some cases. Accurate subclassification is necessary to fully exploit pathologists' understanding of these subtypes and yield improvements in the management of these malignancies. Diagnostic reproducibility is subject to many parameters such as human variability, carcinomas level of differentiation but also staining, WSI acquisition or microscope type. These different levels of variability constitute limitations to advances in ovarian carcinoma subtypes understanding, treatment and prognosis.

In this paper, we proposed an automatic classification system for ovarian carcinoma subtypes diagnosis. Our model was inspired by clinicians' approach to the analysis of tissues and their training procedure. We proposed a multi-magnification representation of tissues that uses the contextual information (multiple fields of view) to identify salient regions on a WSI and use them to infer a diagnosis or carcinoma cell-type. We show in this work, the generality of our classifier to different feature types while out-performing state-of-the-art techniques proposed for histopathology image classification. Our learning framework, achieves a classification accuracy of 90% while trained on a challenging dataset of 60 whole slide images and shows a strong agreement with six clinicians trained and tested on the same dataset.

To fully evaluate our system's robustness to batch effect (e.g. human variability, staining and operating conditions, etc.), we will need to test our pipeline on a larger dataset covering data from different centers, staining techniques and larger cohorts. Also, further quantification of the classifier's ability to handle a variety of feature types may involve the use of other feature learning strategies proposed for histopathology such as wavelet-based Qureshi et al. (2008, 2009) or class-specific sparse coding approaches Sirinukunwattana et al. (2015).

It is important to note that the semantic gap associated with computer aided diagnosis systems (like the proposed method), which use feature learning approaches and, in general, black-box machine learning systems, may hinder their applicability in clinical practice. Nonetheless, while this was not the entire focus of this work, we believe the visualization of the proposed automatically detected salient regions (i.e Figure 5) may provide some insight to the user and may constitute the first steps towards improved and more interpretable machine learning systems for histopathology.

Acknowledgements: Authors gratefully thank the Natural Sciences and Engineering Research Council of Canada for funding and the support of NVIDIA Corporation with the donation of the

Titan X Pascal GPU used in this research.

References

Artan, Y., Haider, M. A., Langer, D. L., van der Kwast, T. H., Evans, A. J., Yang, Y., Wernick, M. N., Trachtenberg, J. and Yetik, I. S. (2010), 'Prostate cancer localization with multispectral mri using cost-sensitive support vector machines and conditional random fields', *IEEE Transactions on Image Processing* **19**(9), 2444–2455.

Barker, J., Hoogi, A., Depeursinge, A. and Rubin, D. L. (2016), 'Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles', *Medical image analysis* **30**, 60–71.

Basavanhally, A., Ganesan, S., Feldman, M., Shih, N., Mies, C., Tomaszewski, J. and Madabhushi, A. (2013), 'Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides.', *IEEE Transactions on Biomedical Engineering* **60**(8), 2089–2099.

BenTaieb, A., Li-Chang, H., Huntsman, D. and Hamarneh, G. (2015), Automatic diagnosis of ovarian carcinomas via sparse multiresolution tissue representation, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 629–636.

BenTaieb, A., Nosrati, M. S., Li-Chang, H., Huntsman, D. and Hamarneh, G. (2016), 'Clinically-inspired automatic classification of ovarian carcinoma subtypes', *Journal of pathology informatics* **7**.

Chang, H., Nayak, N., Spellman, P. T. and Parvin, B. (2013), Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching, *in* 'Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 91–98.

Coates, A., Ng, A. Y. and Lee, H. (2011), An analysis of single-layer networks in unsupervised feature learning, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 215–223.

Cramer, S., Roth, L., Ulbright, T. M., Mazur, M., Nunez, C., Gersell, D., Mills, S. and Kraus, F. (1987), 'Evaluation of the reproducibility of the world health organization classification of common ovarian cancers. with emphasis on methodology.', *Archives of pathology & laboratory medicine* **111**(9), 819–829.

DiFranco, M. D., O'Hurley, G., Kay, E. W., Watson, R. W. G. and Cunningham, P. (2011), 'Ensemble based system for whole-slide prostate cancer probability mapping using color texture features', *Computerized medical imaging and graphics* **35**(7), 629–645.

Do, T. and Artières, T. (2009), Large margin training for hidden markov models with partially observed states, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', pp. 265–272.

Doyle, S., Feldman, M., Tomaszewski, J. and Madabhushi, A. (2012), 'A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies', *IEEE Transactions on Biomedical Engineering* **59**(5), 1205–1218.

Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008), A discriminatively trained, multiscale, deformable part model, *in* 'Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1–8.

Frable, W. J. (2006), 'Surgical pathology-second reviews, institutional reviews, audits, and correlations: What's out there? error or diagnostic variation?', *Archives of pathology & laboratory medicine* **130**(5), 620.

Gavrielides, M. A., Gallas, B. D. and Hewitt, S. M. (2015), Uncertainty in the assessment of immunohistochemical staining with optical and digital microscopy: lessons from a reader study, *in* 'SPIE Medical Imaging', International Society for Optics and Photonics, pp. 94200V–94200V.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M. and Yener, B. (2009), 'Histopathological image analysis: A review', *Reviews in Biomedical Engineering* **2**, 147–171.

Gutiérrez, R., Rueda, A. and Romero, E. (2013), Learning semantic histopathological representation for basal cell carcinoma classification, *in* 'SPIE Medical Imaging', International Society for Optics and Photonics, pp. 86760U–86760U.

Hipp, J., Flotte, T., Monaco, J., Cheng, J., Madabhushi, A., Yagi, Y., Rodriguez-Canales, J., Emmert-Buck, M., Dugan, M. C. and Hewitt, S. (2011), 'Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess', *Journal of pathology informatics* **2**(1), 25.

Irshad, H., Veillard, A., Roux, L. and Racoceanu, D. (2014), 'Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential', *IEEE reviews in biomedical engineering* **7**, 97–114.

Kothari, S., Phan, J. H., Osunkoya, A. O. and Wang, M. D. (2012), Biological interpretation of morphological patterns in histopathological whole-slide images, *in* 'Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine', ACM, pp. 218–225.

Kothari, S., Phan, J. H., Young, A. N. and Wang, M. D. (2013), 'Histological image classification using biologically interpretable shape-based features', *BMC medical imaging* **13**(1), 9.

Krizhevsky, A. and Hinton, G. E. (2010), Factored 3-way restricted boltzmann machines for modeling natural images, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 621–628.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., Graham, A. R., Descour, M. R., Davis, J. R. and Weinstein, R. S. (2006), 'Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience', *Human pathology* **37**(12), 1543–1556.

Lalwani, N., Prasad, S. R., Vikram, R., Shanbhogue, A. K., Huettner, P. C. and Fasih, N. (2011), 'Histologic, molecular, and cytogenetic features of ovarian cancers: implications for diagnosis and treatment', *Radiographics* **31**(3), 625–646.

Lund, B., Thomsen, H. and Olsen, J. (1991), 'Reproducibility of histopathological evaluation in epithelial ovarian carcinoma. clinical implications', *Apmis* **99**(1-6), 353–358.

Petushi, S., Garcia, F. U., Haber, M. M., Katsinis, C. and Tozeren, A. (2006), 'Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer', *BMC Medical Imaging* **6**(1), 14.

Prat, J. (2012), 'New insights into ovarian cancer pathology', *Annals of Oncology* **23**(suppl 10), x111–x117.

Qureshi, H. A., Rajpoot, N. M., N, T. W. and Hans, V. (2009), 'A robust adaptive wavelet-based method for classification of meningioma histology images'.

Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R. and Gurcan, M. (2008), Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 196–204.

Racoceanu, D. and Capron, F. (2016), 'Semantic integrative digital pathology: Insights into microsemiological semantics and image analysis scalability', *Pathobiology* **83**(2-3), 148–155.

Roa-Peña, L., Gómez, F. and Romero, E. (2010), 'An experimental study of pathologist's navigation patterns in virtual microscopy', *Diagnostic Pathology* **5**, 71.

Romo, D., García-Arteaga, J. D., Arbeláez, P. and Romero, E. (2014), A discriminant multi-scale histopathology descriptor using dictionary learning, *in* 'SPIE Medical Imaging', International Society for Optics and Photonics, pp. 90410Q–90410Q.

Roux, L., Racoceanu, D., Lomenie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G. and Gurcan, M. N. (2013), Mitosis detection in breast cancer histological images: An icpr 2012 contest, *in* 'Journal of Pathology Informatics', Vol. 4, pp. 2–8.

Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and S, R. M. (2016), 'Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning', *IEEE transactions on medical imaging* **35**(5), 1285–1298.

Sirinukunwattana, K., Khan, A. M. and Rajpoot, N. (2015), 'Cell words: Modelling the visual appearance of cells in histopathology images', *Computerized Medical Imaging and Graphics* **42**, 16–24.

Soslow, R. A. (2008), 'Histologic subtypes of ovarian carcinoma: an overview', *International Journal of Gynecologic Pathology* **27**(2), 161–174.

Veta, M., Pluim, J. P., Van Diest, P. J. and Viergever, M. A. (2014), 'Breast cancer histopathology image analysis: a review', *transactions on bio-medical engineering* **61**(5), 1400–1411.

Wang, C. and Yu, C. (2013), 'Automated morphological classification of lung cancer subtypes using h&e tissue images', *Machine vision and applications* **24**(7), 1383–1391.

Xu, Y., Zhang, J., Eric, I., Chang, C., Lai, M. and Tu, Z. (2012), Context-constrained multiple instance learning for histopathology image segmentation, *in* 'Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 623–630.

Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M. and Tu, Z. (2014), 'Weakly supervised histopathology cancer image segmentation and classification', *Medical image analysis* **18**(3), 591–604.

Zhang, Y., Zhang, B., Coenen, F. and Lu, W. (2013), 'Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles', *Machine vision and applications* **24**(7), 1405–1420.

Zhou, Y., Chang, H., Barner, K., Spellman, P. and Parvin, B. (2014), Classification of histology sections via multispectral convolutional sparse coding, *in* 'IEEE Computer Vision and Pattern Recognition (CVPR)', pp. 3081–3088.