University of Dundee

Discriminating dysplasia

Li, Wenqi; Coats, Maria; Zhang, Jianguo; McKenna, Stephen

Published in:
Medical Image Analysis

DOI:
[10.1016/j.media.2015.08.002](https://doi.org/10.1016/j.media.2015.08.002)

Publication date:
2015

Document Version
Peer reviewed version

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Li, W., Coats, M., Zhang, J., & McKenna, S. (2015). Discriminating dysplasia: optical tomographic texture analysis of colorectal polyps. *Medical Image Analysis*, 26(1), 57-69. <https://doi.org/10.1016/j.media.2015.08.002>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Copyright 2015 This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Discriminating dysplasia: optical tomographic texture analysis of colorectal polyps

Wenqi Li^a, Maria Coats^b, Jianguo Zhang^a, Stephen J. McKenna^a

^a*CVIP, School of Computing, University of Dundee, Dundee, UK, DD1 4HN*

^b*School of Medicine, University of Dundee, Dundee, UK, DD1 4HN*

Abstract

Optical projection tomography enables 3-D imaging of colorectal polyps at resolutions of 5 – 10 μm . This paper investigates the ability of image analysis based on 3-D texture features to discriminate diagnostic levels of dysplastic change from such images, specifically, low-grade dysplasia, high-grade dysplasia and invasive cancer. We build a patch-based recognition system and evaluate both multi-class classification and ordinal regression formulations on a 90 polyp dataset. 3-D texture representations computed with a hand-crafted feature extractor, random projection, and unsupervised image filter learning are compared using a bag-of-words framework. We measure performance in terms of error rates, F -measures, and ROC surfaces. Results demonstrate that randomly projected features are effective. Discrimination was improved by carefully manipulating various important aspects of the system, including class balancing, output calibration and approximation of non-linear kernels.

Keywords: Volumetric texture, Optical projection tomography, 3-D histopathology, Colorectal polypoid cancer

1. Introduction

Colorectal cancer is the third most common cancer in men (756k new cases per annum, 10.0%) and the second in women (614k new cases per annum, 9.2%) worldwide (Ferlay et al., 2013). Colonic polypoid cancers are the earliest detectable form of colorectal cancer and, if completely excised, can offer a potential cure to patients. Screening programmes have played a major role in keeping cancer mortality rates low. As of 2012, 65,535 polyps had been excised and recorded in the UK NHS bowel cancer screening programme

Figure 1: (a) An H&E-stained whole mount slide. (b) A virtual section from an OPT image of the same polyp with the section chosen to be close to the cut surface that resulted from removal of the physical section shown in (a).

9 (BCSP) database (Majumdar et al., 2012). Histological analysis of polyps is
 10 essential for accurate diagnosis. However, the current gold standard method,
 11 using Haematoxylin and Eosin (H&E) stained sections, has its limitations.
 12 This conventional technique involves taking a thin section of tissue from the
 13 centre of the polyp and this will not necessarily be representative of the
 14 whole specimen. Much ambiguity exists between experienced pathologists
 15 when making diagnoses using H&E sections due to features such as epithelial
 16 displacement (EPD) in which surface epithelial cells become misplaced into
 17 the stalk of the polyp mimicking true invasive cancer. Over-diagnosis of
 18 EPD as cancer has a confounding effect subjecting patients to unnecessary
 19 treatments and generating misleading epidemiology reports (Loughrey and
 20 Shepherd, 2015; Muto et al., 1973).

21 Optical projection tomography (OPT) is a relatively new 3-D imaging
 22 technique first developed to better our understanding of embryo develop-
 23 ment (Sharpe, 2009). It is a simple and affordable imaging technology that
 24 is well-suited for specimens up to 15 mm in size. OPT of colorectal polyps
 25 is non-destructive to the original tissue and enables virtual sectioning of the
 26 specimen at any orientation. Figure 1 shows a comparison between an H&E

section and a virtual OPT section of the same polyp. A significant advantage of tomography is the flexibility in viewing virtual sections and manipulating the image to gain more information. By contrast, once the cutting angle has been chosen for the H&E section, it cannot be changed. Histology sections are cut from the tissue once it has been embedded in paraffin wax and subsequently stained. These can be viewed at sub-micron resolutions whereas OPT provides a lower spatial resolution of about 5-10 μm (Sharpe et al., 2002). Near visible wave-length light is used to obtain OPT images and therefore polyps must be optically cleared in advance of scanning using Benzyl Alcohol Benzyl Benzoate (BABB).

Inter- and intra-observer variation exists when pathologists diagnose colorectal polyps, notably when grading dysplasia whether from H&E or OPT images (Coats et al., 2012). Reliable and repeatable automatic recognition systems are desirable. As a starting point, this paper proposes and evaluates methods for discrimination of dysplasia in OPT images of colorectal polyps. Specifically, the task addressed is that of differentiating between regions of low-grade dysplasia (LGD), high-grade dysplasia (HGD) and invasive cancer (ICA).

The use of automated image analysis for histopathological grading of colorectal polyps has been studied mainly with stained physical sections imaged using 2D microscopy (e.g., Atlamazoglou et al., 2001; Esgiar et al., 1998, 2002; Ficsor et al., 2008; Hamilton et al., 1997, 1995; Kalkan et al., 2012; Ozdemir and Gunduz-Demir, 2013; Shuttleworth et al., 2002). This body of work demonstrates that texture-based analysis is often an important component of lesion detection, segmentation and classification. 3-D texture features have been widely used in medical image analysis more broadly; a comprehensive review of 3-D texture analysis methods is available elsewhere (Depeursinge et al., 2014). Texture representation using local binary patterns (LBPs) and its variants forms an important component of many successful 3-D medical image classification systems e.g., for brain white matter lesion classification in MRI (Oppedal et al., 2012), analysis of lung CT (Sorensen et al., 2010), and retinal optical coherence tomography (Liu et al., 2011). Apart from hand-crafted features as exemplified by LBP, two fundamentally different directions for texture analysis have attracted increased attention in recent years. The first is domain-specific descriptors estimated from observations automatically. For example, deep belief networks have been used to learn features for 3-D brain image segmentation (Brosch and Tam, 2013), and independent subspace analysis (ISA) has been used for feature extraction from

H&E histology images of Glioblastoma Multiforme (Le et al., 2012), MR images (Liao et al., 2013) and video (Le et al., 2011). The second is random projection (RP), a non-adaptive dimensionality reduction tool motivated by compressive sensing theory, applied directly to image patches. RP compared favourably to state-of-the-art texture descriptors in 2D texture classification experiments (Liu and Fieguth, 2012) and it has been used for image classification (Bingham and Mannila, 2001) and for accelerating feature-based registration of 3-D neural ultrastructure (Akselrod-Ballin et al., 2011), for example. In this paper, for the purpose of discriminating dysplasia, these three contrasting approaches to texture representation are compared. Specifically, we compare RP, LBP (exemplifying hand-crafted features), and ISA (exemplifying unsupervised learning).

The main contribution of this paper is that it provides the first study in the literature on automatically discriminating between invasive cancer, high-grade dysplasia and low-grade dysplasia in optical projection tomography images. It builds on and extends earlier conference papers in which preliminary two-class classification experiments discriminating only between LGD and ICA were reported (Li et al., 2013a,b). While the task of discriminating between LGD, HGD and ICA can be cast as a three-class classification problem, this ignores the ordinal structure of these labels. Here, a classification model and an ordinal regression model, both based on margin maximisation, are compared and contrasted for this task. These raise issues of class imbalance and output calibration which are explored empirically. Two state-of-the-art strategies for fast approximation of non-linear kernels are also evaluated. After describing the methods used, results are reported on a data set of 90 polyps. Although the focus is on OPT images of colorectal polyps, the analysis and evaluation methods used should be applicable to other ordinal regression tasks in other image modalities.

2. Materials and Methods

2.1. Tissue Processing, Imaging, and Annotation

Ninety colorectal polyps were selected from the NHS Tayside Tissue Bank archive to be representative of the dysplasia subgroups: invasive cancer (ICA), high-grade dysplasia (HGD) and low-grade dysplasia (LGD). Thirty samples were selected for each of these three groups to give a balanced dataset. The H&E stained sections taken from each specimen were re-diagnosed by an experienced gastro-intestinal histopathologist according to

Figure 2: Direct renderings of OPT polyp images with polyp voxels rendered as opaque: (a), (b) a low-grade dysplasia (LGD) polyp, (c), (d) a high-grade dysplasia (HGD) polyp, (e), (f) an invasive cancer (ICA) polyp. Top row: viewing angles adjusted to view the natural surfaces of the polyps. Bottom row: viewing angles adjusted to view artefactual surfaces due to physical cuts.

Figure 3: Images showing slices with regions annotated as (a) LGD, (b) HGD, and (c) ICA

101 the NHS BCSP and WHO guidelines to reduce intra-observer bias (Hamilton
 102 et al., 2000; NHS BCSP, 2007). Images were acquired using OPT in emission
 103 mode under ultraviolet light and Cy3 dye at a voxel resolution of $6.7 \mu\text{m}^3$.
 104 Each image was of one colorectal polyp and had 1024^3 voxels. Figure 2 shows
 105 renderings of three of these colorectal polyp OPT images.

106 Each 3-D image was manually annotated with 3-D regions by an indi-
 107 vidual experienced in interpreting OPT images. Characteristic regions were
 108 annotated in each polyp, i.e., regions of ICA were annotated in polyps la-
 109 belled as ICA, regions of HGD were annotated in polyps labelled as HGD,
 110 and regions of LGD were annotated in polyps labelled as LGD. Each region's
 111 boundary was delineated such that the annotator had high confidence that all
 112 tissue within the region was correctly labelled. The H&E slide corresponding

Figure 4: Statistics of annotated slices per class.

113 to the cut surface of each polyp was used as guidance for this annotation.
 114 Annotations were performed using the software tool ITK-SNAP (Yushkevich
 115 et al., 2006) by delineating 2D regions every 4 or 5 slices and then inter-
 116 polating between them. Figure 3 shows some examples of annotated slices.
 117 Figure 4 summarises the quantities of voxels annotated per slice.

118 2.2. Patch Encoding

119 In the following, *patch* denotes a cube-shaped image region to be classi-
 120 fied; *window* denotes a smaller cube-shaped region of d^3 voxels from which
 121 local texture features are extracted. Figure 5 shows example patches from
 122 LGD, HGD and ICA. LGD tends to have somewhat regular texture with
 123 low spatial frequency as shown in Figure 5 (a); LGD normally has tubular
 124 morphological structure. ICA shown in Figure 5 (c) contains more homoge-
 125 neous micro-texture patterns corresponding to more dense tissue with villous
 126 structure. HGD is intermediate in appearance as shown in Figure 5 (b). The
 127 bag-of-words framework is adopted to encode patches with a pre-learned vi-
 128 sual dictionary (Zhang et al., 2007). The procedure of encoding a patch is
 129 illustrated in Figure 6. A window slides through the patch and at each lo-
 130 cation, a texture feature vector is extracted from the window. The feature
 131 vectors are quantised into visual words by matching with the most similar
 132 visual word in a pre-learned dictionary. Finally, an L1-normalised histogram
 133 of visual word frequencies constitutes the patch representation. Window step
 134 size was set to half the window width. Dictionaries of 200 visual words were
 135 obtained with *k*-means++ (Arthur and Vassilvitskii, 2007). Bag-of-words
 136 provides a compact representation of fixed dimensionality regardless of the
 137 number of local windows used. It offers a uniform approach to comparative

Figure 5: OPT image patches from regions labelled as (a) low-grade dysplasia (LGD), (b) high-grade dysplasia (HGD), and (c) invasive cancer (ICA)

138 evaluation of local feature extraction methods. The next section elaborates
 139 the feature extraction methods used.

140 *2.3. Extracting Local 3-D Texture Features*

141 Random projection (RP), local binary patterns (LBP) and independent
 142 subspace analysis (ISA) techniques were adopted to compute 3-D texture
 143 representations. These methods were selected from the vast literature on
 144 texture features mainly because they are among the most advanced and po-
 145 tentially effective for OPT classification. Computational efficiency was also
 146 considered when choosing these methods since the OPT datasets in our exper-
 147 iments are large. Random projection was reported effective in representing
 148 general 2-D textures (Liu and Fieguth, 2012). It can be extended to 3-D and
 149 computed efficiently. Local binary patterns are invariant to local contrast
 150 changes and local rotations. These properties are potentially capable of rep-

Figure 6: The procedure of encoding a patch as a bag-of-words histogram.

151 resending micro-textures in OPT images. An approximation of 3-D LBP was
 152 employed which is not fully invariant to arbitrary 3-D rotations but com-
 153 putationally feasible. Independent subspace analysis can be used to learn a
 154 set of phase- and shift-invariant filters that have similarities to features com-
 155 puted by complex cells in the V1 area of primate visual cortex. ISA learning
 156 scales well to large training sets (e.g., Le et al., 2011). The following sections
 157 present technical details of the feature extraction procedures.

158 2.3.1. Random Projection

Let $\mathbf{X}$ be a $d^3 \times N$ matrix in which the elements in each column are the voxel values of one of N windows. The random projection method maps windows onto a k -dimensional subspace using a suitably generated $k \times d^3$ random projection matrix $\mathbf{R}$ (Equation (1)).

$$\hat{\mathbf{X}} = \mathbf{R}\mathbf{X}. \quad (1)$$

159 Each element in the matrix $\mathbf{R}$ is an independent sample from a standard nor-
 160 mal distribution, i.e., a Gaussian with zero mean and unit variance. After
 161 projection, the columns in $\hat{\mathbf{X}}$ are considered as the window descriptors. Ac-
 162 cording to the Johnson-Lindenstrauss lemma (Dasgupta and Gupta, 2003),
 163 data points in $\mathbb{R}^{d^3}$ are embedded into the lower dimensional Euclidean space
 164 $\mathbb{R}^k$ such that pair-wise distances between columns in $\mathbf{X}$ are approximately
 165 preserved. The computational complexity of RP is only $O(d^3 k N)$. In ex-
 166 periments reported in this paper, k was set to 200 unless $d^3 < 200$ in which
 167 case the window was transformed with a square ($d^3 \times d^3$) random projection
 168 matrix.

169 2.3.2. Local Binary Patterns

170 Computing *local binary patterns* (LBP) from 2D images involves thresh-
 171 olding each 3×3 -pixel neighbourhood at the value of its central pixel thus

172 obtaining an 8-bit binary code. A histogram of these codes over an image
 173 window can then be used as a local descriptor. This representation, known
 174 as $\text{LBP}_{8,1}$, (i.e., 8-bit LBP with radius 1 neighbourhood), is not rotation in-
 175 variant. Ojala et al. (2002) found that the vast majority of binary codes in
 176 a local neighbourhood are so called *uniform patterns*. To achieve rotational
 177 invariance (around the central pixel) using uniform patterns, all non-uniform
 178 LBP patterns are stored in a single bin in the histogram computation. The
 179 length of the uniform $\text{LBP}_{8,1}$ descriptor is 59 bins, which is smaller than the
 180 permutations (2^8 bins). In a 3-D volumetric image, designing LBPs that
 181 are invariant to arbitrary rotations is not straight-forward as the ordering of
 182 3-D neighbourhood is undefined. Fehr and Burkhardt (2008) addressed the
 183 problem by computing spherical correlations in the frequency domain. The
 184 approach is robust to 3-D rotations however it is computationally expensive.
 185 We choose to approximate the 3-D LBP with uniform $\text{LBP}_{8,1}$ descriptors
 186 computed in each of three orthogonal planes, taken to be aligned with the
 187 image axes for convenience (Zhao and Pietikainen, 2007). This is compu-
 188 tationally feasible in our experiments. In addition, this approximated 3-D
 189 LBP has been demonstrated to be competitive with a full volumetric LBP
 190 for encoding texture features in microscopic images (Majtner and Svoboda,
 191 2014). The 3-D uniform $\text{LBP}_{8,1}$ operator encodes a window as a histogram
 192 with 177 bins.

193 2.3.3. Independent Subspace Analysis

Independent subspace analysis (ISA) is an unsupervised feature learning
 method based on natural image statistics (Hyvärinen and Hoyer, 2000). In
 the ISA method, image filters $\mathbf{W} = \{\mathbf{w}_t\}_{t=1}^T$ can be learned from a set of
 image windows $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$. The invariances are achieved by grouping the
 image filters into subspaces: the filter responses within the same subspace are
 pooled together, while among different groups the filter responses are treated
 independently. The model is estimated by maximising the log-likelihood
 function (Hyvärinen et al., 2009):

$$\log \mathcal{L}(\mathbf{X}|\mathbf{W}) = - \sum_{n=1}^N \sum_{l=1}^L \sqrt{\sum_{t \in g(l)} (\mathbf{w}_t^\top \mathbf{x}_n)^2}, \quad (2)$$

$$\text{subject to : } \mathbf{W}\mathbf{W}^\top = \mathbf{I}, \quad (3)$$

194 where each image filter $\mathbf{w}_t$ is a d^3 -dimensional vector which is applied to the
 195 image window; $g(l)$ is the set of indices of l -th group in the total L groups; $\mathbf{I}$

196 is an identity matrix. The constraint (3) is introduced to reduce the number
 197 of free parameters and leads to more stable solutions (Hyvärinen, 2013).

Figure 7: (a-f) Six groups of 4 filters learned from 2700 windows (filter size: $13 \times 13 \times 13$). Filters within the same subspace (group) shared some similarities while filters in different subspaces had different orientations and frequencies.

198 This optimisation was performed using stochastic gradient descent after
 199 applying whitening transformations to remove correlations between voxels
 200 and then PCA to reduce the dimensionality to 300 (whenever the number of
 201 voxels in the window was greater than 300). A set of 300 filters forming 75 in-
 202 dependent subspaces (4 filters per subspace) were *simultaneously* learned by
 203 optimising Formula (2). For each optimisation process 200 passes of stochas-
 204 tic gradient descent with adaptive learning rate were applied through the
 205 entire set of training windows. After training, windows are represented as
 206 vectors containing 300 filter responses. Some filters learned from the train-
 207 ing window set are visualised in Figure 7. These 300 responses are further
 208 pooled into a 150-dimensional vector as the final window representation. The
 209 pooling process takes the square root of squared sum of two responses in the
 210 same subspace (illustrated in Figure 8). Filters within a given subspace are
 211 similar.

212 2.4. Multi-Class Patch Classification

213 When formulated as a three-class classification problem, the task of dis-
 214 criminating between LGD, HGD and ICA was addressed using a set of three

Figure 8: The procedure of encoding an image window with ISA model.

support vector machine (SVM) binary classifiers (Allwein et al., 2001). Each classifier was trained to discriminate one class from the others. This one-vs-rest approach leads to unbalanced data sets for each of the classifiers. Furthermore, it is important to calibrate the outputs of the classifiers before comparing them in order to infer the class label. A further consideration is the use of non-linear kernels in the classifiers which, given large data sets, can be approximated for practical reasons. This section describes the classification method used and considers available solutions to each of these issues.

2.4.1. Binary Subproblem

Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^l$ where $\mathbf{x}_i$ is the i^{th} bag-of-words feature vector and $y_i \in \{-1, 1\}$ is its corresponding class label, a classification function f that maps the feature vector to the label set $\{-1, 1\}$ can be found using the widely used SVM formulation by minimising

$$L(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i, \quad (4)$$

$$\text{subject to : } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (5)$$

where C is a parameter controlling the trade-off between model complexity and training errors; ξ_i , $i = 1, \dots, l$ are slack variables; $\mathbf{w}$ is the weight vector, b is a bias weight, and $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ is the SVM hyperplane to be optimised. This optimisation was performed using the primal form solver provided in the LIBLINEAR package (Fan et al., 2008). C was searched over the set $\{2^\lambda \mid \lambda \in \mathbf{Z} \wedge [-15, 15]\}$.

230 *2.4.2. Handling Class Imbalance*

The binary classification sub-problems do not have balanced data sets because in each case one class is being discriminated against all other classes. Class imbalance also arises because HGD is a less commonly assigned label than LGD or ICA (see Figure 4). This problem can be addressed by replacing the free parameter C with C_p and C_n for positive and negative classes respectively. Formula (4) then becomes:

$$\text{Minimise}_{\mathbf{w}, \xi^+, \xi^-} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C_p \sum_{i=1}^{l^+} \xi_i^+ + C_n \sum_{i=1}^{l^-} \xi_i^-, \quad (6)$$

$$\text{subject to : } \mathbf{w}^\top \mathbf{x}_i^+ + b \geq 1 - \xi_i^+, \quad (7)$$

$$- \mathbf{w}^\top \mathbf{x}_i^- - b \geq 1 - \xi_i^-, \quad (8)$$

$$\xi_i^+ \geq 0, \quad \xi_i^- \geq 0, \quad (9)$$

231 where $\mathbf{x}_i^+$, $\mathbf{x}_i^-$ are positive and negative training examples in the one-vs-rest
 232 setting; l^+ and l^- are the numbers of such examples; ξ_i^+ , $i = 1, \dots, l^+$ and
 233 ξ_i^- , $i = 1, \dots, l^-$ are slack variables; $\mathbf{w}$ is the weight vector, b is a bias
 234 weight, and $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ is the SVM hyperplane to be optimised; C_p and
 235 C_n are set as $C_p = \frac{l^+ + l^-}{2l^+} C_0$ and $C_n = \frac{l^+ + l^-}{2l^-} C_0$. C_0 was searched over the set
 236 $\{2^\lambda \mid \lambda \in \mathbf{Z} \wedge [-15, 15]\}$.

237 *2.4.3. Output Calibration*

Three binary classifiers are trained for the three-class classification problem. Traditionally in the one-vs-rest scheme the final output score for a test example is the highest among the scores given by the three classifiers. However, since the three classifiers are trained independently, the scores are not necessarily comparable. This situation is helped by calibrating the scores prior to making this comparison by using Platt's scaling method to obtain values that can be treated as class probability estimates (Lin et al., 2007). Calibration maps the binary classifier output ($\mathbf{w}^\top \mathbf{x}$) onto values that can be treated as probabilities with a parameterised sigmoid function:

$$P(y = 1 | \mathbf{x}; \mathbf{w}, A, B) = \frac{1}{1 + \exp(A\mathbf{w}^\top \mathbf{x} + B)}, \quad (10)$$

where A and B are learned from a validation set. For k -class classification, k such sigmoid functions are estimated, one per binary classifier. Unbalanced datasets mean that the fitting of these sigmoid functions is more heavily

effected by over-represented classes. This problem was recently addressed by Wallace and Dahabreh (2014) who introduced bagging of under-sampling data estimators to refine the calibration procedure. Similarly, we apply the bagging method to one-vs-rest classifier calibration. More specifically, T balanced patch sets are formed by randomly discarding patches from over-represented classes and T calibration models are learned from these sets. The final probability estimate from a binary classifier is an average of the T models, i.e.,

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \frac{1}{T} \sum_{t=1}^T P^{(t)}(y = 1|\mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)}), \quad (11)$$

where $P^{(t)}(y = 1|\mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)})$ is obtained by applying Equation (10) with $A^{(t)}$ and $B^{(t)}$ estimated from the t -th balanced patch set.

2.4.4. Non-Linear Kernel Approximation

Using appropriate non-linear kernels that map feature vectors into a high dimensional space can improve the classification performance of bag-of-words encoding. For example, the χ^2 kernel has been used for classification of endoscopic images (Tamaki et al., 2013). However, this is computationally prohibitive for large-scale problems due to the expensive operation of constructing a Gram matrix over all training data. Recently the method of approximating kernels with explicit feature maps enabled the use of non-linear kernels with relatively low computational cost on large scale datasets (Maji et al., 2013; Vedaldi and Zisserman, 2012). The main idea is that for a homogeneous kernel $k(\mathbf{x}, \mathbf{y}) : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$, it is possible to use an approximation function $\Phi(\cdot) : \mathbb{R}^M \mapsto \mathbb{R}^{M'}$ so that $k(\mathbf{x}, \mathbf{y}) \approx \Phi(\mathbf{x})^\top \Phi(\mathbf{y})$. The $\Phi(\cdot)$ can be data-independent. A linear SVM can then be used directly with $\Phi(x)$ for classification. This leverages the high performance of non-linear kernels while maintaining scalability to large scale problems.

Given two OPT patches encoded with bag-of-words, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, the χ^2 kernel is computed as $k_{\text{chi}}(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^M (2\mathbf{x}_m \mathbf{y}_m) / (\mathbf{x}_m + \mathbf{y}_m)$, and the histogram intersection kernel is computed as $k_{\text{hist}}(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^M \min(\mathbf{x}_m, \mathbf{y}_m)$. These two kernels were approximated with methods proposed by Vedaldi and Zisserman (2012) and Maji et al. (2013) respectively.

2.4.5. Inherently Multi-class Classifiers

The one-vs-rest SVM classifications were compared with three inherently multi-class classifiers: random forest (RF) (Breiman, 2001), k -nearest neigh-

263 bours (KNN), and multi-layer perceptrons (MLP). A non-linear classification
 264 function $\hat{f}$ that maps the feature vector to the label set $\{\text{LGD}, \text{HGD}, \text{ICA}\}$
 265 was trained for each method to discriminate three classes. For RF classi-
 266 fier we report performances of 2,000 randomised classification trees. k was
 267 searched over the set $\{5, 10, 20\}$ in the KNN classifier. We set one hidden
 268 layer with 20 neurons for the MLP¹. For the one-vs-rest SVM in the compari-
 269 son, the procedures described in Section 2.4.2, Section 2.4.3 and Section 2.4.4
 270 were used (denoted as KSVM). The non-linear kernel in KSVM was computed
 271 as the 1400-dimensional χ^2 approximation (the *Chi2* method in Figure 15)
 272 described in Section 2.4.4.

273 2.5. Ordinal Patch Regression

274 The labels of dysplastic change are qualitative measurements that reflect
 275 increasing severity in diagnosis from LGD to HGD to ICA. This suggests
 276 that a patch label $r \in \{\text{LGD}, \text{HGD}, \text{ICA}\}$ should be an ordinal variable rather
 277 than a nominal one; the order is $\text{LGD} \prec \text{HGD} \prec \text{ICA}$.

278 2.5.1. Large Margin Formulation

Given a training set $\{\mathbf{x}_i, r_i\}_{i=1}^l$ where $\mathbf{x}_i$ is the i^{th} bag-of-words feature
 vector and $r_i \in \{\text{LGD}, \text{HGD}, \text{ICA}\}$ is its corresponding ordinal label, a
 function g that maps the feature vector to the label set $\{\text{LGD}, \text{HGD}, \text{ICA}\}$
 can be found using a structural risk minimisation formulation of ordinal
 regression. Herbrich et al. (1999) show how to estimate g by learning a
 ranking function $f(\mathbf{x}) \in \mathbb{R}$ so that the pair-wise orders are preserved, i.e.,
 $f(\mathbf{x}_i) \prec f(\mathbf{x}_j) \iff r_i \prec r_j$. The score $f(\mathbf{x}_i)$ is thresholded to determine
 the ordinal category of $\mathbf{x}_i$, i.e., $g(\mathbf{x}) = k \iff f(\mathbf{x}) \in [\theta_k, \theta_{k+1}]$ where θ_k and
 θ_{k+1} are the learned thresholds of the k -th category. In the case of a linear
 mapping $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, the weight vector $\mathbf{w}$ can be learned by minimising
 empirical risk on the pair-wise order set $\mathcal{P} = \{(i, j) | r_i \prec r_j\}$:

$$L(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{(i,j) \in \mathcal{P}} \xi_{ij}, \quad (12)$$

$$\text{subject to : } t_{ij}(\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \mathbf{x}_i) \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0, \quad \forall (i, j) \in \mathcal{P},$$

¹We used the `newpr` function from the Matlab Neural Network Toolbox. (URL: <http://uk.mathworks.com/products/neural-network/>)

where t_{ij} is the order label, i.e., $t_{ij} = 1$ iff $r_i \prec r_j$; $t_{ij} = -1$ otherwise; $\xi_{ij}, \forall (i, j) \in \mathcal{P}$ are slack variables.

There are three types of pair-wise ordering in $\mathcal{P}$, i.e., LGD $\prec$ HGD, LGD $\prec$ ICA and HGD $\prec$ ICA (The number of pairs of each type is denoted as N_1 , N_2 , and N_3 respectively). Similar to the balancing of multi-class classification in Section 2.4.2, the balance problem is also considered on the pair-wise order set. Specifically, we replace C with three cost parameters to be proportional to the number of pair-wise preferences of each type respectively, i.e., $C_{\text{LGD} \prec \text{HGD}} = C_0(N_1 + N_2 + N_3)/N_1$, $C_{\text{LGD} \prec \text{ICA}} = C_0(N_1 + N_2 + N_3)/N_2$, and $C_{\text{HGD} \prec \text{ICA}} = C_0(N_1 + N_2 + N_3)/N_3$, where C_0 is a free parameter; the value is searched within the set $\{2^\lambda \mid \lambda \in \mathbf{Z} \wedge [-15, 15]\}$.

2.5.2. Solving the Optimisation Problem

The objective function in Formula (12) is a standard quadratic programming problem that can be solved by many existing convex optimisation packages. However, the size of the pair-wise order set grows quadratically with the number of samples, e.g., 100 LGD and 100 ICA patches gives 10,000 pair-wise orders. Most solvers are not feasible due to the scale of our problem. Here we choose the fast rank SVM solver proposed by Chapelle and Keerthi (2010) which tackles the primal form of rank SVM with Newton's method. Instead of computing the inverse Hessian matrix in the Newton step, the fast rank SVM approximates the inverse with a conjugate gradient method. This approximation is both fast and memory-efficient.

With the estimated $\mathbf{w}$, the optimal threshold θ_k is set to be in the middle of the closest correctly separated training pair in the k -th and $(k + 1)$ -th category, i.e.,

$$\theta_k = \frac{\mathbf{w}^\top \mathbf{x}_{i_k} + \mathbf{w}^\top \mathbf{x}_{j_k}}{2}, \quad (13)$$

$$\text{where: } (i_k, j_k) = \underset{(i,j) \in \mathcal{P}_k}{\operatorname{argmin}} (\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \mathbf{x}_i), \quad (14)$$

$$\mathcal{P}_k = \{(i, j) \mid r_i = k \wedge r_j = k + 1 \wedge (\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \mathbf{x}_i) \geq 1\}. \quad (15)$$

2.6. Patch Sampling and Cross-validation

Empirical evaluations used patches sampled with a systematic uniform random sampling (SURS) strategy. The major advantage of SURS over repeated uniform random sampling is its statistical efficiency. The sampling method is illustrated in Figure 9. The displacement between points

Figure 9: Patch sampling with systematic uniform random sampling (SURS) applied to an annotated region. A regular sampling grid with fixed displacement between sampling points is positioned at random. Patches are sampled with their centres at the grid points.

on the sampling grid was set to 31 voxels along each of the three image axes. Patches were sampled with their centres at the grid points. A set of 20,000 patches was sampled from the 90 polyps at each of ten patch sizes, $d \in \{11, 21, 31, \dots, 111\}$, giving 200,000 patches in total. This enabled exploration of the effect of patch size as a system parameter.

Experiments based on 10-fold cross-validation with stratified splitting were carried out to estimate generalisation capability of the systems as follows. (1) All patches from the same image are in the same fold; this ensures that no part of the polyp participating in the test is in the training set (cross-polyp generalisation); (2) The testing fold is generated so that it always contains the same proportions of polyp images in the three classes as the original dataset to ensure that there is no dominating class in the test dataset. Within the training folds, we further applied 3-fold cross-validation to search the appropriate system parameters (e.g., SVM calibration parameters, cost parameters). This experimental design is illustrated in Figure 10. The results in the following section are based on this cross-validation scheme unless otherwise specified.

2.7. Performance Metrics

There are six types of mis-classification, i.e., LGD as ICA, LGD as HGD, ICA as HGD, ICA as LGD, HGD as LGD, HGD as ICA. Error rates were calculated for each error type respectively. Performance was also compared with measures of overall mis-classification rate, absolute error rate and F -measure.

Figure 10: Cross-validation scheme.

328 Overall mis-classification rate is the number of mis-classified cases divided
 329 by the total number of test cases, N , without considering ordinal informa-
 330 tion. The absolute error rate is $(\sum_{n=1}^N |e_n|)/N$, where e_n is a scalar error
 331 value. In the case of correct prediction, $e_n = 0$. In the case of an out-by-one
 332 error (LGD confused with HGD or HGD confused with ICA), $e_n = 1$. In
 333 the case of an out-by-two error (LGD confused with ICA) $e_n = 2$. For ex-
 334 periments on handling class imbalance, averaged F -measure was used. The
 335 averaged F -measure is an average over F -measures with respect to each class.
 336 The 95% confidence intervals of each type of measurement were obtained by
 337 bootstrapping (DiCiccio and Efron, 1996) with $n = 10,000$; specifically we
 338 generated n bootstrap replicates of the classifier outputs, and calculated av-
 339 erage F -measures of each bootstrap replicate. The confidence intervals were
 340 computed with n average F -measures using the `boot.ci` function from the
 341 `boot` package (R implementation). When evaluating window and patch size
 342 parameters, and kernel approximations, performance was measured with Av-
 343 eraged Area under the ROC curve (AAUC) which is not affected by specific
 344 choices of thresholds on the raw SVM outputs. Error bars indicating 95%
 345 confidence intervals were estimated with the method proposed by DeLong
 346 et al. (1988). When comparing features with ordinal regression formulations,
 347 ROC surfaces were constructed with true positive rates. Volumes under the

Figure 11: Cobweb diagrams showing number of mis-classifications for multi-class classification and ordinal regression formulations.

surfaces were further calculated using the algorithm proposed by Luo and Xiong (2012).

3. Results

3.1. Overall Comparison of Formulations and Feature Types

Figure 11 shows cobweb diagrams for multi-class classification and ordinal regression for each of the three feature types. Table 1 reports confusion matrices. Table 2 reports the misclassification rates, absolute error rates, and average F-measures. These results were obtained using a patch size of $81 \times 81 \times 81$ and a window size of $13 \times 13 \times 13$ using the procedures described in Sections 2.4.2 and 2.4.3. (Experiments exploring the effect of varying such parameters are reported in the following section). The values in Table 2 suggest that RP outperforms ISA and LBP; the performance rank of RP is consistent for all measures considered for both problem formulations. Table 3 compares classification performance of the four types of multi-class classifier described in Section 2.4.5. In most cases, RF, KNN and MLP performed worse than one-vs-rest KSVM. MLP gave the most competitive performance among the three inherently multi-class classifiers.

Table 1: Multi-class classification (left) and ordinal regression (right) confusion matrices.

(a) Random projection features.

		Actual labels					Actual labels		
		ICA	HGD	LGD			ICA	HGD	LGD
Pred.	ICA	5397	1356	702	Pred.	ICA	4048	1038	196
	HGD	1604	2406	1041		HGD	3162	2533	2197
	LGD	569	794	6257		LGD	360	985	5607

(b) Local binary patterns features.

		Actual labels					Actual labels		
		ICA	HGD	LGD			ICA	HGD	LGD
Pred.	ICA	5465	1547	644	Pred.	ICA	3154	759	250
	HGD	1271	1833	954		HGD	3715	2588	1949
	LGD	834	1176	6402		LGD	701	1209	5801

(c) Independent subspace analysis features.

		Actual labels					Actual labels		
		ICA	HGD	LGD			ICA	HGD	LGD
Pred.	ICA	4752	1242	669	Pred.	ICA	4093	1074	345
	HGD	1643	2189	1537		HGD	2742	2416	2708
	LGD	1175	1125	5794		LGD	735	1066	4947

3.2. One-Vs-Rest Classification

Figure 12 plots AAUCs averaged over 10 folds for each one-vs-rest classifier for various window and patch sizes. These plots suggest that patch size has greater impact than window size. At patch size $81 \times 81 \times 81$ all features reached high performance. At large window size ($13 \times 13 \times 13$) we generated a smaller number of windows than with smaller window sizes; this was computationally more efficient without significantly decreasing the AAUC. These parameter values were used in other experiments unless otherwise specified. RP gave generally good results compared to the other two types of feature.

3.2.1. Classifier Calibration

Figure 13 shows calibration of an HGD-vs-rest SVM classifier. Thresholding the raw output scores at 0 gave an F -measure of 0.2126 with 95% confidence interval (0.1992, 0.2259). After calibration using Equation (10) (see

(a) Random projection features.

(b) Local binary patterns features.

(c) Independent subspace analysis features.

Figure 12: Averaged Area under the ROC curve (AAUC) using different types of feature with varied window size and patch size. Error bars show 95% confidence intervals. (Best viewed in colour.)

Figure 13: Calibrating the HGD-vs-rest classifier output of all testing folds. Figure best viewed in colour. (a) Histogram and (b) Gaussian kernel density estimates of raw SVM scores. (c–e) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (10). (f–h) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (11). The diagonal solid lines in the reliability diagrams indicate perfectly calibrated classifiers.

Table 2: Multi-class classification (Multi.) and ordinal regression (Ordi.) results for different features (with 95% confidence intervals).

Feature type	Mis-classification rate	Absolute error rate	Average F -measure
RP - Mult.	0.3038 (0.2974, 0.3102)	0.3659 (0.3578, 0.3742)	0.6675 (0.6610, 0.6740)
ISA - Mult.	0.3522 (0.3457, 0.3589)	0.4312 (0.4223, 0.4399)	0.6172 (0.6105, 0.6238)
LBP - Mult.	0.3457 (0.3391, 0.3523)	0.4227 (0.4141, 0.4313)	0.6216 (0.6149, 0.6283)
RP - Ordi.	0.4253 (0.4185, 0.4321)	0.4481 (0.4406, 0.4556)	0.5735 (0.5667, 0.5802)
ISA - Ordi.	0.4397 (0.4328, 0.4466)	0.4965 (0.4882, 0.5047)	0.5542 (0.5475, 0.5608)
LBP - Ordi.	0.4280 (0.4211, 0.4348)	0.4729 (0.4648, 0.4809)	0.5595 (0.5530, 0.5662)

Table 3: Multi-class classification results for different classifiers and features (with 95% confidence intervals).

Feature type	Mis-classification rate	Absolute error rate	Average F -measure
RP - RF.	0.3110 (0.3046, 0.3174)	0.4024 (0.3935, 0.4113)	0.6238 (0.6171, 0.6308)
ISA - RF.	0.3373 (0.3308, 0.3439)	0.4177 (0.4089, 0.4265)	0.5994 (0.5927, 0.6060)
LBP - RF.	0.3234 (0.3168, 0.3300)	0.4199 (0.4109, 0.4291)	0.5913 (0.5846, 0.5981)
RP - KNN.	0.3572 (0.3507, 0.3638)	0.4671 (0.4577, 0.4764)	0.6026 (0.5958, 0.6094)
ISA - KNN.	0.3621 (0.3554, 0.3686)	0.4669 (0.4574, 0.4763)	0.5832 (0.5764, 0.5900)
LBP - KNN.	0.3650 (0.3583, 0.3717)	0.4768 (0.4673, 0.4863)	0.5772 (0.5704, 0.5840)
RP - MLP.	0.3031 (0.2968, 0.3095)	0.3716 (0.3633, 0.3799)	0.6558 (0.6491, 0.6625)
ISA - MLP.	0.3373 (0.3308, 0.3437)	0.4135 (0.4048, 0.4222)	0.6117 (0.6050, 0.6184)
LBP - MLP.	0.3462 (0.3395, 0.3527)	0.4619 (0.4524, 0.4714)	0.5976 (0.5906, 0.6046)
RP - KSVM.	0.2987 (0.2924, 0.3051)	0.3641 (0.3558, 0.3742)	0.6746 (0.6680, 0.6812)
ISA - KSVM.	0.3227 (0.3162, 0.3292)	0.3851 (0.3767, 0.3933)	0.6484 (0.6418, 0.6550)
LBP - KSVM.	0.2996 (0.2932, 0.3059)	0.3655 (0.3573, 0.3739)	0.6636 (0.6570, 0.6703)

Figure 13(c-e)), thresholding probabilities at 0.5 gave an improved F -measure of 0.3286 with 95% confidence interval (0.3090, 0.3368). However, the probability of HGD is systematically underestimated by the sigmoid function in Equation (10) because the model is biased towards the “rest” class due to the imbalanced training set. This can be seen from Figures 13(a-e). In Figures 13(f-h) we visualise the probability outputs of the bagging method. The under-sampling of the dominating class and the ensemble strategy mitigated the bias problem. In the reliability diagrams (Figures 13(e, h)), the observed probabilities are grouped into 25 bins and the observed frequency of positives are plotted against mean probability in each bin. For a perfect reliability, the observed frequency and the predicted probability should be equal (shown as dotted line in diagonal direction). By using the bagging method, the probability outputs from 0.8 to 1.0 are more reliable compared to the original Platt’s method. We are able to further improve the F -measure

Figure 14: Box plots summarising the distribution of F -measures over cross-validation folds with ISA features (patch size: $81 \times 81 \times 81$, window size: $13 \times 13 \times 13$). *Balanced* denotes use of the balanced cost function. *Scaled* denotes use of output calibration. *Scaled and Balanced* denotes use of both a balanced cost function and output calibration.

392 from 0.3286 to 0.4637 with 95% confidence interval (0.4526, 0.4748). In the
 393 other experiments reported here we applied Equation (11) for the calibration
 394 procedure.

395 3.2.2. Effect of Class Balancing and Calibration

396 We conducted a group of experiments to investigate the effect of balanc-
 397 ing the cost function (Section 2.4.2) and output calibration (Section 2.4.3)
 398 on classification performance. As a *baseline* we use the standard SVM (For-
 399 mula (4)). To obtain decisions from one-vs-rest classifiers, raw SVM scores
 400 were thresholded at 0 and calibrated outputs at 0.5. Multi-class classification
 401 decisions were obtained by applying the max-win strategy described in Sec-
 402 tion 2.4.1. Figure 14 shows comparisons with the baseline SVM in terms of
 403 F -measures of each fold. Classification results after class balancing and cali-
 404 bration were significantly different from the baseline ($p < 0.0001$; McNemar’s
 405 test (Dietterich, 1998)). Use of the balanced SVM cost function generally
 406 improved performance, especially for HGD-vs-rest classification where the
 407 baseline suffers from the very unbalanced dataset (see Figure 4). SVM out-
 408 put calibration helped address the data imbalance problem and improved
 409 performance compared to the baseline. Combining the balanced cost func-

tion and output calibration does not lead to further improvement in terms of F -measure as compared to each applied individually.

3.2.3. Effect of Kernel Approximation

The effect of approximating the histogram intersection and χ^2 kernels as discussed in Section 2.4.4 was evaluated. Two approximation strategies were compared: 1) B-spline approximations² (Maji et al., 2013) of the χ^2 kernel (denoted as *Chi_1*) and the histogram intersection kernel (denoted as *Min_1*); and 2) the homogeneous feature map³ (Vedaldi and Zisserman, 2012) for the χ^2 kernel (denoted as *Chi_2*) and the histogram intersection kernel (denoted as *Min_2*). Figure 15 gives AUC results using those approximations compared with the baseline result of the plain linear SVM (error bars indicate 95% confidence intervals). It can be seen that, generally, the approximations of non-linear kernels can improve dysplasia classification over the baseline for all three feature types if the feature dimensionality is carefully tuned; the exception was that the performance improvement of B-spline approximations of the histogram intersection kernel was not stable and conclusive. The homogeneous feature map consistently outperformed the B-spline approximation although the improvement was marginal. The χ^2 kernel worked better than the histogram intersection kernel in most cases.

3.3. Ordinal Regression

Figure 16 shows distributions of rank SVM scores obtained along with the two thresholds for three-class ordinal regression. By varying the two thresholds ROC surfaces can be created. Each point on an ROC surface represents a tuple of three true positive rates (TPR). Figure 17(a) shows ROC surfaces for the three feature types.

For visual comparison, each surface was mapped onto the plane $\text{TPR}_{\text{LGD}} + \text{TPR}_{\text{ICA}} + \text{TPR}_{\text{HGD}} = 1$ and the signed distances between points on the surface and the plane were colour-coded (Figure 17(b)). The plane represents the ROC surface for random classification. The signed distances give an indication of how much better the classifier is than random guessing. The

²We used the Matlab implementation by Maji et al. (2013). (*URL*: <http://ttic.uchicago.edu/~smaji/projects/libbspline-release1.0.tar.gz>)

³We used the Matlab implementation provided in VLFeat package (Vedaldi and Fulkerson, 2008).

Figure 15: Averaged AUCs for different kernel types (with feature dimensionality given in parentheses). Features were extracted with patch size $81 \times 81 \times 81$ and window size $13 \times 13 \times 13$. “Chi_1” and “Min_1” denote B-spline based kernel approximation of χ^2 and histogram intersection kernels respectively. “Chi_2” and “Min_2” denote homogeneous feature map of χ^2 and histogram intersection kernels respectively. Error bars indicate 95% confidence intervals.

signed distance maps for different feature types were further compared (Figure 17(c)). Compared to ISA features, RP and LBP give better performances in high TPR range of LGD (high LGD specificity range), whereas ISA is better in high TPR range of ICA (high ICA specificity range). In the comparison of RP and LBP, RP performs better when true positive rate of HGD is in low range (in low HGD specificity).

The volume under an ROC surface (VUS) is the expected proportion of correctly ranked triplets uniformly drawn from all possible samples of triplets (He and Frey, 2008). It is an extension of area under the ROC curve

Figure 16: Distributions of training set ranking scores. Dashed and dotted vertical bars are thresholds for the three ordinal categories, obtained using Equation (13).

Figure 17: (a) ROC surfaces for three-class ordinal regression (patch size: $81 \times 81 \times 81$, window size: $13 \times 13 \times 13$). Colour on the $TPR_{LGD} - TPR_{ICA}$ plane encodes TPR_{HGD} value. (b) Signed distances from points on the ROC surfaces to the plane $TPR_{LGD} + TPR_{ICA} + TPR_{HGD} = 1$. (c) Differences between the maps in (b). (Best viewed in colour.)

449 to the three-class case. The VUS and its 95% confidence interval for the RP,
 450 ISA and LBP methods were 0.5901 (0.5821, 0.5981), 0.5215 (0.5136, 0.5294)
 451 and 0.5325 (0.5246, 0.5405) respectively. In this experiment RP features
 452 showed significantly better performance than ISA and LBP in terms of VUS.

453 4. Concluding Remarks

454 Large margin multi-class classification and ordinal regression techniques
 455 were adopted with a focus on practical considerations. Their performance
 456 with three types of texture descriptor was demonstrated. The results showed
 457 that random projection features and bag-of-words framework was the most
 458 effective. Although the RP method is not invariant to rotations, it outper-
 459 formed carefully designed LBP operators as well as automatically learned ISA
 460 filters for all parameter settings. Similar observations were also reported in
 461 several general texture classification systems (Liu and Fieguth, 2012; Varma
 462 and Zisserman, 2003) where classification of filter-based features did not out-
 463 perform raw pixels as patch representation. The RP method generates a
 464 low-dimensional representation without making strong assumption about the
 465 nature of the texture being analysed. These attributes may partially explain
 466 its relatively good performance in OPT analysis.

467 The class balancing problem is important and the performance can be
 468 further improved if it is carefully managed. Using the approximation of
 469 non-linear kernels, especially χ^2 kernel, also improves the classification per-
 470 formance slightly.

471 In terms of computational complexity, encoding a window with RP and
 472 ISA methods is similar: the output window feature is a simple multiplication
 473 of the vectorised window with a matrix (either a random projection matrix⁴
 474 or a set of filters learned with ISA). The LBP operator⁵ is more expensive as it
 475 involves a neighbourhood-pixel thresholding and a uniform pattern matching
 476 procedure. ISA requires a training phase in order to estimate filters from

⁴The complexity of RP method can be further reduced by constructing sparse random projections with a simpler sampling distribution than a standard normal distribution (e.g., Achlioptas, 2003; Li et al., 2006). We use the standard normal distribution for RP matrix because both methods achieve similar performance and a standard normal distribution is easier to implement.

⁵For LBP operator we use the C++ implementation by Zhao and Pietikainen (2007) (URL: <http://www.cse.oulu.fi/CMV/Downloads/LBPSoftware>).

477 training patches. In our experiments learning a set of 300 filters⁶ from a $81 \times$
 478 81×81 patch set required approximately 4 hours; Encoding and classifying
 479 200 test patches with size $81 \times 81 \times 81$ required approximately 30 seconds for
 480 the RP and ISA methods, and 4 minutes for the LBP method on a 3.4GHz
 481 Intel i7 CPU with 16Gb memory.

482 LGD was the class most readily distinguished. Unsurprisingly, being
 483 the intermediate class, HGD was the most often confused class. In terms
 484 of problem formulations, ordinal regression performance was slightly worse
 485 than multi-class classification in terms of mis-classification rate, absolute
 486 error rate and averaged F -measure (Table 2). However ordinal regression
 487 makes less confusion between the ICA and LGD classes (Figure 11). The
 488 ordinal regression formulation is a better choice over one-vs-rest classifica-
 489 tion when the focus concentrates on minimising the risk of mis-classification
 490 between LGD and ICA. Ordinal regression is also simpler than multi-class
 491 classification in the sense that only one model is trained while in multi-class
 492 classification three models are trained. Dysplastic change is naturally a con-
 493 tinuous phenomenon on which ordinal grading imposes artificial categories.
 494 The ordinal regression method works by first mapping samples onto the real
 495 line and it would be interesting to investigate using this continuous map in
 496 diagnostic histopathology.

497 Compared to the emerging study of using OPT for colorectal polyps anal-
 498 ysis, extensive literature exists on automatic analysis of more traditional
 499 histopathology images for a range of analysis tasks, clinical settings, and dis-
 500 ease types. However, we are not aware of a suitable study with which to make
 501 direct comparisons (i.e., patch discrimination between LGD, HGD and ICA
 502 in colorectal polyps). The proposed system achieved promising performance
 503 in discriminating OPT image patches. Further improvements and incorpor-
 504 ation into software tools would be needed to enable translation or adoption
 505 for clinical research. Larger data sets of annotated images would be needed
 506 to train and validate such tools because the visual appearance of dysplastic
 507 change is complex and the variations across polyps are large.

508 It is worth noting that inter-observer variation exists in polyp diagnosis
 509 with OPT (Coats et al., 2013). In order to minimise the effect of such vari-
 510 ation and uncertainty in our study, the ground truth was annotated within

⁶For ISA model estimation we use the Matlab implementation by Hyvärinen et al. (2009) (*URL*: <http://www.naturalimagestatistics.net>).

high confidence regions by an experienced pathologist rather than trying to delineate accurate region boundaries. The annotated regions were meanwhile cross-checked and calibrated with the corresponding H&E slices to ensure the confidence of the obtained ground truth. Nevertheless, there may still exist some uncertainty in the ground truth. Quantifying this using multiple pathologists would be interesting for a future study.

OPT enables access to polyps' surface morphology and internal structure at the same time. In the future one interesting direction is to combine morphological analysis of colorectal polyp surfaces (e.g., Zhang et al., 2014) with 3-D texture analysis. Recently, deep learning techniques have yielded promising results in learning image representations from large datasets. Applying such techniques to OPT datasets would be an interesting direction. Finally, building accurate classification and regression systems usually requires large amounts of manually annotated data; it is expensive to obtain such manual annotations. Another research direction is to develop training algorithms that efficiently utilise limited amounts of annotation (e.g., Li et al., 2013c).

References

- Achlioptas, D., 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences* 66, 671–687.
- Akselrod-Ballin, A., Bock, D., Reid, R.C., Warfield, S.K., 2011. Accelerating image registration with the Johnson–Lindenstrauss lemma: Application to imaging 3-D neural ultrastructure with electron microscopy. *IEEE Transactions on Medical Imaging* 30, 1427–1438.
- Allwein, E.L., Schapire, R.E., Singer, Y., 2001. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research* 1, 113–141.
- Arthur, D., Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding, in: *ACM-SIAM symposium on Discrete Algorithms*, pp. 1027–1035.
- Atlamazoglou, V., Yova, D., Kavantzias, N., Loukas, S., 2001. Texture analysis of fluorescence microscopic images of colonic tissue sections. *Medical and Biological Engineering and Computing* 39, 145–151.

544 Bingham, E., Mannila, H., 2001. Random projection in dimensionality re-
545 duction: Applications to image and text data, in: ACM International
546 Conference on Knowledge Discovery and Data Mining, pp. 245–250.

547 Breiman, L., 2001. Random forests. *Machine learning* 45, 5–32.

548 Brosch, T., Tam, R., 2013. Manifold learning of brain MRIs by deep learn-
549 ing, in: *Medical Image Computing and Computer-Assisted Intervention*,
550 Springer. pp. 633–640.

551 Chapelle, O., Keerthi, S.S., 2010. Efficient algorithms for ranking with SVMs.
552 *Information Retrieval* 13, 201–215.

553 Coats, M., Wedden, S., Farrell, J., Cranston, G., Mitchell, L., Wilson, J.,
554 Steele, R., Carey, F., 2012. Optical projection tomography: Can it help di-
555 agnose the colorectal polypoid cancer? *Gastroenterology* 142, S178–S179.

556 Coats, M., Wedden, S., Keogh, R., Steele, R., Carey, F., 2013. Diagnosis of
557 colorectal polyps using optical projection tomography how well do pathol-
558 ogists agree? *United European Gastroenterology Journal* 1, (Supplement
559 1) A233.

560 Dasgupta, S., Gupta, A., 2003. An elementary proof of a theorem of Johnson
561 and Lindenstrauss. *Random Structures & Algorithms* 22, 60–65.

562 DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the
563 areas under two or more correlated receiver operating characteristic curves:
564 a nonparametric approach. *Biometrics* 40, 837–845.

565 Depeursinge, A., Foncubierta-Rodriguez, A., Van De Ville, D., Müller, H.,
566 2014. Three-dimensional solid texture analysis in biomedical imaging: Re-
567 view and opportunities. *Medical Image Analysis* 18, 176–196.

568 DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. *Statistical*
569 *Science* 11, 189–212.

570 Dietterich, T.G., 1998. Approximate statistical tests for comparing su-
571 pervised classification learning algorithms. *Neural Computation* 10,
572 1895–1923.

573 Esgiar, A.N., Naguib, R., Bennett, M.K., Murray, A., 1998. Automated
574 feature extraction and identification of colon carcinoma. *Analytical and*
575 *quantitative cytology and histology* 20, 297–301.

576 Esgiar, A.N., Naguib, R.N., Sharif, B.S., Bennett, M.K., Murray, A., 2002.
577 Fractal analysis in the detection of colonic cancer images. *IEEE Transactions*
578 *on Information Technology in Biomedicine* 6, 54–58.

579 Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. Liblinear:
580 A library for large linear classification. *The Journal of Machine Learning*
581 *Research* 9, 1871–1874.

582 Fehr, J., Burkhardt, H., 2008. 3D rotation invariant local binary patterns,
583 in: *International Conference on Pattern Recognition*, pp. 1–4.

584 Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers,
585 C., Rebelo, M., Parkin, D., Forman, D., Bray, F., 2013. Globocan 2012
586 v1.0, cancer incidence and mortality worldwide: IARC CancerBase no.
587 11 [internet]. Lyon, France: International agency for research on cancer.
588 Available from: <http://globocan.iarc.fr>, accessed on 04/march/2014.

589 Ficsor, L., Varga, V.S., Tagscherer, A., Tulassay, Z., Molnar, B., 2008. Auto-
590 mated classification of inflammation in colon histological sections based on
591 digital microscopy and advanced image analysis. *Cytometry Part A* 73A,
592 230–237.

593 Hamilton, P.W., Bartels, P.H., Thompson, D., Anderson, N.H., Montironi,
594 R., Sloan, J.M., 1997. Automated location of dysplastic fields in colorectal
595 histology using image texture analysis. *The Journal of Pathology* 182,
596 68–75.

597 Hamilton, P.W., Thompson, D., Sloan, J.M., Bartels, P.H., 1995. Knowledge-
598 guided segmentation and morphometric analysis of colorectal dysplasia.
599 *Analytical and Quantitative Cytology and Histology* 17, 172–182.

600 Hamilton, S.R., Aaltonen, L.A., International Agency for Research on Cancer
601 and World Health Organization and others, 2000. *Pathology and genetics*
602 *of tumours of the digestive system*. IARC press Lyon.

- 603 He, X., Frey, E.C., 2008. The meaning and use of the volume under a
604 three-class ROC surface (VUS). *IEEE Transactions on Medical Imaging*
605 27, 577–588.
- 606 Herbrich, R., Graepel, T., Obermayer, K., 1999. Large margin rank bound-
607 aries for ordinal regression, in: *Advances in Neural Information Processing*
608 *Systems*, MIT. pp. 115–132.
- 609 Hyvärinen, A., 2013. Independent component analysis: recent advances.
610 *Philosophical Transactions of the Royal Society A: Mathematical, Physical*
611 *and Engineering Sciences* 371, 20110534.
- 612 Hyvärinen, A., Hoyer, P., 2000. Emergence of phase- and shift-invariant
613 features by decomposition of natural images into independent feature sub-
614 spaces. *Neural Computation* 12, 1705–1720.
- 615 Hyvärinen, A., Hurri, J., Hoyer, P., 2009. *Natural Image Statistics: A Proba-
616 bilistic Approach to Early Computational Vision*. Computational Imaging
617 and Vision, Springer.
- 618 Kalkan, H., Nap, M., Duin, R.P., Loog, M., 2012. Automated classification
619 of local patches in colon histopathology, in: *International Conference on*
620 *Pattern Recognition*, pp. 61–64.
- 621 Le, Q.V., Han, J., Gray, J.W., Spellman, P.T., Borowsky, A., Parvin, B.,
622 2012. Learning invariant features of tumor signatures, in: *IEEE Interna-
623 tional Symposium on Biomedical Imaging*, pp. 302–305.
- 624 Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y., 2011. Learning hierarchical
625 invariant spatio-temporal features for action recognition with independent
626 subspace analysis, in: *IEEE Conference on Computer Vision and Pattern*
627 *Recognition*, pp. 3361–3368.
- 628 Li, P., Hastie, T.J., Church, K.W., 2006. Very sparse random projections, in:
629 *ACM International Conference on Knowledge Discovery and Data Mining*,
630 pp. 287–296.
- 631 Li, W., Coats, M., Zhang, J., McKenna, S.J., 2013a. Comparative analysis of
632 feature extraction methods for colorectal polyp images in optical projection
633 tomography, in: *Medical Image Understanding and Analysis*, pp. 67–72.

- 634 Li, W., Zhang, J., McKenna, S., Coats, M., Carey, F., 2013b. Classification
635 of colorectal polyp regions in optical projection tomography, in: IEEE
636 International Symposium on Biomedical Imaging, pp. 736–739.
- 637 Li, W., Zhang, J., Zheng, W.S., Coats, M., Carey, F.A., McKenna, S.J.,
638 2013c. Learning from partially annotated opt images by contextual rel-
639 evance ranking, in: Medical Image Computing and Computer-Assisted
640 Intervention, Springer. pp. 429–436.
- 641 Liao, S., Gao, Y., Oto, A., Shen, D., 2013. Representation learning: A uni-
642 fied deep learning framework for automatic prostate mr segmentation, in:
643 Medical Image Computing and Computer-Assisted Intervention, Springer.
644 pp. 254–261.
- 645 Lin, H.T., Lin, C.J., Weng, R.C., 2007. A note on Platt’s probabilistic
646 outputs for support vector machines. *Machine Learning* 68, 267–276.
- 647 Liu, L., Fieguth, P., 2012. Texture classification from random features. *IEEE*
648 *Transactions on Pattern Analysis and Machine Intelligence* 34, 574–586.
- 649 Liu, Y.Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J.S., Rehg, J.M.,
650 2011. Automated macular pathology diagnosis in retinal OCT images using
651 multi-scale spatial pyramid and local binary patterns in texture and shape
652 encoding. *Medical Image Analysis* 15, 748–759.
- 653 Loughrey, M.B., Shepherd, N.A., 2015. The pathology of bowel cancer screen-
654 ing. *Histopathology* 66, 66–77.
- 655 Luo, J., Xiong, C., 2012. DiagTest3Grp: An R package for analyzing diag-
656 nostic tests with three ordinal groups. *Journal of statistical software* 51,
657 1.
- 658 Maji, S., Berg, A.C., Malik, J., 2013. Efficient classification for additive
659 kernel SVMs. *IEEE Transactions on Pattern Analysis and Machine Intel-*
660 *ligence* 35, 66–77.
- 661 Majtner, T., Svoboda, D., 2014. Comparison of 3D texture-based image
662 descriptors in fluorescence microscopy, in: *Combinatorial Image Analysis*.
663 Springer, pp. 186–195.

- 664 Majumdar, D., Patnick, J., Nickerson, C., Rutter, M., 2012. Analysis of
665 colorectal polyps detected in the English NHS bowel cancer screening pro-
666 gramme with emphasis on advanced adenoma and polyp cancer detected.
667 Gut 61, A67–A67.
- 668 Muto, T., Bussey, H., Morson, B., 1973. Pseudo-carcinomatous invasion in
669 adenomatous polyps of the colon and rectum. Journal of clinical pathology
670 26, 25.
- 671 NHS BCSP, 2007. Reporting lesions in the NHS bowel cancer screening pro-
672 gramme - guidelines from the bowel cancer screening programme pathology
673 group. [http://www.cancerscreening.nhs.uk/bowel/publications/](http://www.cancerscreening.nhs.uk/bowel/publications/nhsbcsp01.pdf)
674 [nhsbcsp01.pdf](http://www.cancerscreening.nhs.uk/bowel/publications/nhsbcsp01.pdf). Accessed on 13/07/2014.
- 675 Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and
676 rotation invariant texture classification with local binary patterns. IEEE
677 Transactions on Pattern Analysis and Machine Intelligence 24, 971–987.
- 678 Oppedal, K., Engan, K., Aarsland, D., Beyer, M., Tysnes, O.B., Eftestol, T.,
679 2012. Using local binary pattern to classify dementia in MRI, in: IEEE
680 International Symposium on Biomedical Imaging, pp. 594–597.
- 681 Ozdemir, E., Gunduz-Demir, C., 2013. A hybrid classification model for
682 digital pathology using structural and statistical pattern recognition. IEEE
683 Transactions on Medical Imaging 32, 474–483.
- 684 Sharpe, J., 2009. Optical projection tomography, in: Advanced Imaging in
685 Biology and Medicine. Springer, pp. 199–224.
- 686 Sharpe, J., Ahlgren, U., Perry, P., Hill, B., Ross, A., Hecksher-Sørensen, J.,
687 Baldock, R., Davidson, D., 2002. Optical projection tomography as a tool
688 for 3D microscopy and gene expression studies. Science 296, 541–545.
- 689 Shuttleworth, J., Todman, A., Naguib, R., Newman, B., Bennett, M., 2002.
690 Multiresolution colour texture analysis for classifying colon cancer images,
691 in: Annual Conference and the Annual Fall Meeting of the Biomedical
692 Engineering Society EMBS/BMES Conference, 2002. Proceedings of the
693 Second Joint, pp. 1118–1119.

694 Sorensen, L., Shaker, S.B., De Bruijne, M., 2010. Quantitative analysis of
695 pulmonary emphysema using local binary patterns. *IEEE Transactions on*
696 *Medical Imaging* 29, 559–569.

697 Tamaki, T., Yoshimuta, J., Kawakami, M., Raytchev, B., Kaneda, K.,
698 Yoshida, S., Takemura, Y., Onji, K., Miyaki, R., Tanaka, S., 2013.
699 Computer-aided colorectal tumor classification in NBI endoscopy using
700 local features. *Medical Image Analysis* 17, 78–100.

701 Varma, M., Zisserman, A., 2003. Texture classification: Are filter banks
702 necessary?, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. II–691.

704 Vedaldi, A., Fulkerson, B., 2008. VLFeat: An open and portable library of
705 computer vision algorithms. <http://www.vlfeat.org/>.

706 Vedaldi, A., Zisserman, A., 2012. Efficient additive kernels via explicit feature
707 maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
708 34, 480–492.

709 Wallace, B.C., Dahabreh, I.J., 2014. Improving class probability estimates
710 for imbalanced data. *Knowledge and Information Systems* 41, 33–52.

711 Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C.,
712 Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical
713 structures: Significantly improved efficiency and reliability. *NeuroImage*
714 31, 1116–1128.

715 Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C., 2007. Local features and
716 kernels for classification of texture and object categories: A comprehensive
717 study. *International Journal of Computer Vision* 73, 213–238.

718 Zhang, J., McKenna, S.J., Zhang, J., Coats, M., Carey, F., 2014. Analysing
719 the surface morphology of colorectal polyps: Differential geometry and pit
720 pattern prediction, in: *Medical Image Understanding and Analysis*, pp.
721 67–72.

722 Zhao, G., Pietikainen, M., 2007. Dynamic texture recognition using local bi-
723 nary patterns with an application to facial expressions. *IEEE Transactions*
724 *on Pattern Analysis and Machine Intelligence* 29, 915–928.