UNITOPATHO, A LABELED HISTOPATHOLOGICAL DATASET FOR COLORECTAL POLYPS CLASSIFICATION AND ADENOMA DYSPLASIA GRADING

Carlo Alberto Barbano* Daniele Perlo* Enzo Tartaglione*
Attilio Fiandrotti* Luca Bertero† Paola Cassoni† Marco Grangetto†

* Computer Science Department, University of Turin, 10149 Torino, Italy
† Medical Sciences Department, University of Turin, 10126, Torino, Italy

ABSTRACT
Histopathological characterization of colorectal polyps allows to tailor patients’ management and follow up with the ultimate aim of avoiding or promptly detecting an invasive carcinoma. Colorectal polyps characterization relies on the histological analysis of tissue samples to determine the polyps malignancy and dysplasia grade. Deep neural networks achieve outstanding accuracy in medical pattern recognition, however they require large sets of annotated training images. We introduce UniToPatho, an annotated dataset of 9536 hematoxylin and eosin (H&E) stained patches extracted from 292 whole-slide images, meant for training deep neural networks for colorectal polyps classification and adenomas grading. We present our dataset and provide insights on how to tackle the problem of automatic colorectal polyps characterization.

Index Terms— Deep Learning, Multi Resolution, Colorectal polyps, Colorectal Adenomas, Digital Pathology

1. INTRODUCTION
The demand for gastrointestinal histopathology is on the rise [1], fostered by the widespread adoption of cancer screening programs. Gastrointestinal histopathologists inspect tissue samples collected during colonoscopies, looking for hints that can predict the incidence of invasive carcinoma [2]. Colorectal polyps are pre-malignant lesions found in the intestinal mucosa that pathologists analyze to (i) ascertain the polyp type (hyperplastic, adenoma) and (ii) assess the dysplasia grade in case of adenomas. Examination of colorectal polyps represents a large share of histopathologists’ workload, thus methods for automating these tasks are highly sought. Despite such clinical relevance, the concordance rate even among expert pathologists, in the diagnostic assessment of colorectal polyps, is far from optimal [3][4]. Although the distinction between non-adenomatous and adenomatous tissue is usually reliable, the inter-observer agreement between different histological types and dysplasia grades is sub-optimal. For instance, the concordance in assessing a tubulovillous polyp or low grade dysplasia ranged around 70% [5].

Deep learning-based methods have shown promising results towards automating the pathologists’ work [5]. Korbar et al. [6] present a patch-based framework, developed using a ResNet architecture [7], to classify different types of colorectal polyps from whole-slide images. Their work provides empirical evidence that residual architectures are better suited at this task. Wei et al. [8] propose an analysis model for annotated tissue samples and perform a study on the generalization of neural models with external medical institutions. Their work describes a hierarchical evaluation mechanism to extend the classification of tissue fragments to the entire slide. Song et al. [9] propose a patch-based fully-convolutional approach for the classification and grading of adenomas, with a strong focus on model interpretability. They also highlight how different patch sizes should be used for adenomas classification and grading. However, the scarcity of datasets large enough and suitably labeled represents a major hurdle for training deep-learning based algorithms to predict polyp type and adenoma dysplasia grade.

This work provides the following contributions towards automatic colorectal polyps characterization, in the framework of the DeepHealth [10] project. First, we make available UniToPatho [1], a high-resolution annotated dataset of Hematoxylin and Eosin (H&E)-stained colorectal images [11]. UniToPatho enables training deep neural networks to classify different colorectal polyps types and adenomas grading. We make available our annotated dataset as a collection of high-resolution patches extracted at different scales. Second, we show that the direct application of a deep neural network fails to classify both the tissue type and adenoma dysplasia grade. Lastly, we propose a multi-resolution deep learning approach solving the previous issues, that achieves significant accuracy in the characterization of colorectal polyps.

2. THE UNITOPATHO DATASET

UniToPatho is a dataset of annotated high-resolution H&E-stained images, comprising different histological samples of colorectal polyps, collected from patients undergoing cancer screening. The dataset is a collection of the most relevant patch images extracted

<p>|</p>
<table>
<thead>
<tr>
<th>HP</th>
<th>NORM</th>
<th>TA</th>
<th>LG</th>
<th>TVA</th>
<th>LG</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slides</td>
<td>41</td>
<td>21</td>
<td>26</td>
<td>146</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>σ = 7000</td>
<td>59</td>
<td>74</td>
<td>98</td>
<td>411</td>
<td>93</td>
<td>132</td>
</tr>
<tr>
<td>σ = 800</td>
<td>545</td>
<td>950</td>
<td>454</td>
<td>3618</td>
<td>916</td>
<td>2186</td>
</tr>
<tr>
<td>Total</td>
<td>604</td>
<td>1024</td>
<td>552</td>
<td>4029</td>
<td>1009</td>
<td>2318</td>
</tr>
</tbody>
</table>

Table 1. UniToPatho class distribution for whole image slides (top) and the two patch scales made available (bottom).

The dataset is available at https://ieee-dataport.org/open-access/unitopatho

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825111, DeepHealth Project.
Fig. 1. Example of 800 × 800µm patches for the six UniToPatho colorectal polyps classes.

from 292 whole-slide images (simply slides in the following), in accordance with UniTo pathologists’ evaluation. The slides are acquired through a Hamamatsu Nanozoomer S210 scanner at 20× magnification (0.4415 µm/px), as exemplified in Fig. 1. Each slide belongs to a different patient and is annotated by expert UniTo pathologists, according to six classes as follows:

NORM - Normal tissue
HP - Hyperplastic Polyp
TA.HG - Tubular Adenoma, High-Grade dysplasia
TA.LG - Tubular Adenoma, Low-Grade dysplasia
TVA.HG - Tubulo-Villous Adenoma, High-Grade dysplasia
TVA.LG - Tubulo-Villous Adenoma, Low-Grade dysplasia

Hyperplastic polyps usually exhibit no malignant potential [12], while adenomas are more likely to progress into invasive carcinomas. Tubular and tubulo-villous are common colorectal adenomas, with villous adenomas generally presenting higher malignant potential given the larger surface [12]. Adenomas are associated with a grade of dysplasia, which measures the abnormality in cellular growth and differentiation [13]. Higher grade dysplasia indicates higher malignant potential.

We split the slides into a train set and a test set with a 70% to 30% ratio, resulting in 204 slides used for training and 88 for testing. Therefore, each slide is represented either in the train set or in the test set, but not in both. Following the approach in [9] [14], we release square patches cropped from the 292 slides. From each slide, we crop multiple non-overlapping square patches at different scales. We denote the side of the underlying physical area with σ, measured in µm, which we vary from 100 to 8000. The number of patches obtained for each slide hence depends on multiple factors including σ, the slide size and the polyp type. As common in similar datasets, the different classes are highly unbalanced in the dataset. We make publicly available a total of 9536 patches, 8669 of which extracted at σ = 800 (1812 × 1812 pixels patches) and 867 at σ = 7000 (15,855 × 15,855 pixels patches). These are the sizes that will be used in Sec. 3. Tab. 1 provides a summary of the class distribution for the whole slides and the released patches.

<table>
<thead>
<tr>
<th>Patch scale σ [µm]</th>
<th>100</th>
<th>800</th>
<th>1500</th>
<th>4000</th>
<th>7000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA (6-class)</td>
<td>0.40</td>
<td>0.45</td>
<td><strong>0.46</strong></td>
<td>0.41</td>
<td>0.37</td>
<td>0.38</td>
</tr>
<tr>
<td>NORM</td>
<td>0.70</td>
<td>0.66</td>
<td>0.72</td>
<td>0.76</td>
<td>0.78</td>
<td>0.71</td>
</tr>
<tr>
<td>HP</td>
<td>0.81</td>
<td><strong>0.92</strong></td>
<td>0.85</td>
<td>0.70</td>
<td>0.60</td>
<td>0.69</td>
</tr>
<tr>
<td>TA (HG+LG)</td>
<td>0.65</td>
<td>0.66</td>
<td>0.65</td>
<td>0.71</td>
<td><strong>0.76</strong></td>
<td>0.70</td>
</tr>
<tr>
<td>TVA (HG+LG)</td>
<td>0.64</td>
<td>0.67</td>
<td>0.68</td>
<td>0.74</td>
<td><strong>0.84</strong></td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 2. Preliminary experiments: overall BA for all of the six classes (first row) and BA for each polyp type, plus normal tissue.

3. PRELIMINARY ANALYSIS

We perform our preliminary experiments on the UniToPatho dataset with a baseline strategy owing to state-of-the-art methods [6,9]. The method which will be described in the next section builds upon the lessons learned during these experiments.

First, we randomly color-jitter the patches described in the previous section to augment the dataset, as proposed in [8,5]. Next, as in [6,8], we train a deep convolutional neural network for image classification on the patches belonging to the train slides. Namely, we train an ImageNet-pretrained ResNet-18 [4] with SGD for 50 epochs, with an initial learning rate of 0.01 decayed by a factor of 0.1 every 20 epochs. The output layer is reshaped to match the UniToPatho 6-class classification problem. Each patch is downsampled to the standard 224 × 224 pixels ImageNet size prior being fed into the ResNet. We repeat the whole training process for different scales, choosing σ in the [100, 8000] range. For each σ, the trained network is eventually tested on the patches extracted at the same scale from the test slides.

We analyze the the results of our preliminary experiments first in terms of polyp type classification accuracy, next in terms of adenoma grade prediction for adenoma samples only. In the following, the classification accuracy will be defined in terms of Balanced Accuracy (BA) to cope with the unbalanced samples in the dataset. In Tab. 2 (first row) we show the BA achieved when attempting to discriminate all 6 classes with the baseline approach as function of the patch scale σ: it can be noted that even the best accuracy achieved at σ = 1500 is quite low. Nonetheless, we conjecture that different polyp types can be better recognized at different scales: as a consequence, in Tab. 2 we also show the BA concerning the classification of each single polyp type HP, TA, TVA plus NORM. The TA and TVA classes encompass the respective low grade (LG) and high grade (HG) subclasses. Indeed, breaking down the accuracy on a per-class basis reveals that different types of polyps achieve top-classification accuracy at different scales.

Hyperplastic Polyps (HP) are best classified at a finer 800µm scale: we hypothesize these types of benign polyps are best discriminated by looking at smaller-scale details such as gland edges [15]. Conversely, Tubuloadenomas (TA) are best classified at a coarser 7000µm scale: we hypothesize this type of polyps is best discriminated by looking at large-scale macro structures such as entire glands shapes [16].

Coming to the problem of predicting the grade (LG or HG) of TA and TVA adenomas, we investigate whether that could be best predicted at some particular scale σ. Our experiments proved however inconclusive, i.e. the adenoma grade classification accuracy does not appear being a function of the patch scale. In fact, visual inspection of the 224 × 224px downsampled patches reveals that the downsam-
4. PROPOSED METHOD

This section details our proposed approach towards classification of UniToPatho images: a multi-resolution ensemble of cascaded classifiers. The method relies on three cascaded ResNet-18 classifiers, having the output layer specifically adapted for each classification task. The classifiers are trained on patches extracted at either \( \sigma = 800 \) or \( \sigma = 7000 \), following the procedure described in the previous section.

The overall process of inference is depicted in Fig. 2 with the input being a single \( 7000 \times 7000 \mu m \) patch which is used by the three ResNet-18 classifiers mentioned above. For classifiers working at \( \sigma = 800 \), we crop the input image into smaller \( 800 \times 800\mu m \) sub-patches which are used to generate a prediction for the entire image. Also, all of the patches are downsampled to \( 224 \times 224 \) pixels, unless stated otherwise.

4.1. HP polyps detection

First, HP polyps are discriminated from adenomas and normal images via a binary classifier that takes as input sub-patches of size \( 800 \times 800\mu m \) extracted from the larger input image. In fact, Tab. 3 shows that HP polyps are identified with top accuracy at this scale. In the case the patch content is classified as being HP, the second classifier in the cascade is invoked.

4.2. Adenoma detection

Second, TA adenomas are discriminated from TVA adenomas and from normal images via a ternary classifier taking as input the entire \( 7000 \times 7000 \mu m \) patch. In fact, Tab. 3 shows that TA and TVA adenomas are identified with top accuracy at this scale. In the case the patch content is classified as a TA or TVA polyp, the third and last classifier is invoked to infer its grade, otherwise the tissue will be labeled as NORM.

4.3. Dysplasia grading

Finally, a binary classifier is used to determine the dysplasia grade for TA and TVA adenomas. Sec. 3 suggested that downsampling the patches to \( 224 \times 224 \) might be detrimental towards inferring the dysplasia grade due to the loss of important features such as cells nuclei, as also observed by [9]. Thus, only for this specific classification task, we skip downsampling to prevent the loss of fine grained details. To account for the increased size of the feature space, we add an adaptive average pooling layer [17, 18] before the fully connected layer of the ResNet-18 network. We repeat the experiment of Sec. 3 sweeping the patch scale \( \sigma \) this time without downsampling and we find that adenoma classification peaks \( (0.81 \text{ BA}) \) for \( \sigma = 800 \). As a consequence, we apply the grade classifier on sub-patches extracted from the input image at the scale \( \sigma = 800 \). Finally, we infer the dysplasia grade for the entire input image according to a threshold \( T_d \): if the ratio of high grade predictions on the sub-patches is higher than \( T_d \), the image is labeled as HG, otherwise as LG. This choice is motivated by pathologists grading guidelines, where HG can be decided on the base of small portions of tissue; clearly, setting smaller values for \( T_d \) can mimic this behaviour [12].

5. EXPERIMENTAL RESULTS

We evaluate our proposed method on the \( 7000 \times 7000 \mu m \) test patches of UniToPatho. As a preliminary step, we tune the classifier responsible for the dysplasia grade prediction to find a suitable \( T_d \) threshold. In the following experiments we choose \( T_d = 0.2 \), as it strikes a favorable balance between false positives and sensitivity. Lower thresholds may be preferred, for example, in clinical applications where minimizing the false negatives rate is more important.

Fig. 3 shows the 6-class confusion matrix of the proposed method compared to the baseline approach described in Sec. 3 in this latter case we employ \( \sigma = 1500 \) which, as already discussed, yields the best overall BA. Looking at the diagonal, it is quite evident that the proposed method significantly improves in average accuracy, that leaps from 0.46 to 0.67 (50% relative increase). We notice how the baseline model is biased towards the lower grade classes: this represents further proof that subsampling the images results in the lack of useful features to distinguish high grade from lower grade and normal tissue. On the other hand, the multi-resolution approach shows remarkable improvements in assessing the correct grade.

Tab. 4 reports other metrics that are common in the related literature. Our proposed method achieves quite high specificity for all classes, and we also observe promising sensitivity values especially for the higher-risk TA.HG and TVA.HG classes.

Finally, we also analyze the per-type performance, as done in Sec. 3. The results are shown Tab. 4. Despite the small HP class accuracy drop - which could be due to the simple inference method - we observe an increase for all of the other tissue classes. Notably, we obtain a great reduction in false positive adenoma predictions,
Fig. 3. Confusion matrices for the (a) baseline at $\sigma = 1500$ and (b) multi-resolution ensemble, reaching a BA of 0.46 and 0.67 respectively.

and, most importantly, a more precise distinction between TA and TVA adenomas.

6. CONCLUSIONS

This work presents UniToPatho, a histopathological dataset of colorectal polyps obtained from 292 high-resolution annotated images. UniToPatho provides annotation for hyperplastic polyps, adenomas (tubular or tubulo-villous) and their dysplasia grade (low or high). We show that a single deep neural network fails at correctly classifying the tissue type. We highlight that each of the classes discriminant features are extracted at different resolutions, making a direct classification of the tissue at a fixed scale a sub-optimal approach. This observation allowed us to design a multi-resolution deep learning strategy that, employing an ensemble of classifiers, achieves 67% accuracy. From a clinical perspective, the most relevant results are the differentiation capability between tubular and tubulo-villous adenomas and the dysplasia grading, which are the most difficult tasks for pathologists [19, 20, 21, 22]. The challenge to improve the automatic diagnostic accuracy is however still open, as UniToPatho is publicly available and still growing. Our future work will focus on collecting samples from other institutions to assess the cross-laboratory generalization capability.

7. REFERENCES

[3] Bernard Denis, Carol Peters, Catherine Chapelain, Isabelle Kleinclauss, Anne Fricker, Richard Wild, Bernard Auge, Is-


[11] Luca Bertero; Carlo Alberto Barbano; Daniele Perlo; Enzo Tartaglione; Paola Cassoni; Marco Grangetto; Attilio Fiandrotti; Alessandro Gambella; Luca Cavallo, “Unitopatho,” https://dx.doi.org/10.21227/9fsv-tm25, 2021.


