

Dysplasia grading of colorectal polyps through convolutional neural network analysis of whole slide images ^{*}

Daniele Perlo¹, Enzo Tartaglione¹, Luca Bertero², Paola Cassoni², and Marco Grangetto¹

¹ University of Torino, Computer Science dept., Torino, Italy

² University of Torino, Pathology Unit, Dept. Medical Sciences, Torino, Italy
{daniele.perlo, luca.bertero}@unito.it

Abstract. Colorectal cancer is a leading cause of cancer death for both men and women. For this reason, histo-pathological characterization of colorectal polyps is the major instrument for the pathologist in order to infer the actual risk for cancer and to guide further follow-up. Colorectal polyps diagnosis includes the evaluation of the polyp type, and more importantly, the grade of dysplasia. This latter evaluation represents a critical step for the clinical follow-up. The proposed deep learning-based classification pipeline is based on state-of-the-art convolutional neural network, trained using proper countermeasures to tackle WSI high resolution and very imbalanced dataset. The experimental results show that one can successfully classify adenomas dysplasia grade with 70% accuracy, which is in line with the pathologists' concordance.

Keywords: Deep Learning, Multi Resolution, Colorectal Polyps, Colorectal Adenomas, Digital Pathology

1 Introduction

The cornerstone of conventional histo-pathological examination is the evaluation of hematoxylin & eosin slides by trained pathologists to detect and/or quantify specific features or patterns and provide a diagnostic evaluation. Based on this premise, whole slide image (WSI) analysis approaches based on Deep Learning (DL) are well suited to address the tasks posed by the histo-pathological evaluation [15]. During the last few years, many specific challenges have been tackled: from lymph node metastasis detection [3] to mitotic count [1]. The main aims of these approaches are multiple: i) improve pathologists' accuracy and thus diagnostic sensitivity; ii) speed-up the diagnostic workflow by addressing more menial, but time-consuming tasks; iii) improve diagnostic agreement by adopting standardized criteria.

Among the multiple fields of surgical pathology, gastrointestinal pathology is one

^{*} This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111, DeepHealth Project.

of the most represented [11], thus addressing this specific topic has the potential of significantly affecting the overall workflow of a pathology service. Colorectal polyps, pre-malignant lesions arising from the intestinal epithelium, are one of the most common gastrointestinal specimens submitted to histological examination. These lesions are usually collected during a colonoscopy, which represents the mainstay of colorectal cancer screening programs in many countries [4]. The development of these programs leads to a significant increase in this specific caseload of surgical pathology laboratories: the correct diagnostic assessment has far-reaching consequences both for the patient and the public health systems. Indeed, a correct diagnosis is obviously important for the management of the patient, but it is now well acknowledged that different types of polyps are associated with different risks of developing metachronous invasive carcinomas during the following years [14]. For this reason, specific algorithms have been established for tailoring patients' follow-up. Despite such clinical relevance, the concordance rates even among expert pathologists, in the diagnostic assessment of colorectal polyps, is far from optimal [9,10,20,24]. Although the distinction between non-adenomatous and adenomatous tissue is usually reliable, the inter-observer agreement between different histological types and dysplasia grades are sub-optimal. For instance, the concordance in assessing a tubulo-villous polyp or low grade dysplasia ranged around 70%.

In this work the main contributions are: i) the design of a deep learning pipeline to tackle the high dimensionality of WSI, working at single patches level; ii) the study on the physical resolutions suitable to deal automatically with the problem of classification of different colorectal polyps; iii) the study of different patch pre-processing approaches, where we find that, for the considered problem, the intensity of the dye present in the scans is the most informative feature of the tissue images.

2 Related work

Only a limited number of works explored histo-pathological examination through deep learning-based analysis of digital whole slide images [16,23,25]. Among these works, Korbar et al. [13] present a crop-based framework, developed using a ResNet architecture to classify different types of colorectal polyps from whole-slide images. This work provides empirical suggestions the residual network architecture achieves better performance than other models. Following their previous work, Korbar et al. introduce a revised version of Grad-CAM (gradient driven class activation mapping) [22] to visualize the attention map of the network for the annotated whole-slide [16]. Bychkov et al. [5] apply different architectures (convolutional and recurrent neural networks) in order to predict five-years disease survival probabilities for colorectal cancer and estimate the individual risk. This work explores the idea of using spatial information by feeding an LSTM network with the features extracted from image crops by a CNN. Recently, Wei et al. [25] propose an analysis model for annotated tissue and perform a study on the generalization of neural models with external medical

	HP	NORM	TA.HG	TA.LG	TVA.HG	TVA.LG	Total
Slides	62	30	34	232	44	55	457
R_t	158	112	145	777	264	245	1701
A_t [cm ²]	9.91	18.38	7.94	71.74	60.45	41.86	210.29

Table 1: Dataset composition.

institutions. In such work, a hierarchical evaluation mechanism is proposed to extend the classification of tissue fragments to the entire slide.

These efforts show promising results, but the testing data size is small and, most importantly, they do not provide diagnosis based on both histological type and dysplasia grade. Our aim is thus to evaluate the efficacy of a deep neural network for the automatic histo-pathological classification of colorectal polyps employing a large training cohort and assessing both polyp histological type and dysplasia grade. Barbano et al. [2] shows how an hierarchical DL model for annotated tissue can take care of both colorectal polyps’ type and relative dysplasia degree.

3 Dataset

WSI composing the dataset are collected within the CE project *DeepHealth* [8]. This dataset contains all source WSI for *UniToPatho* [6] plus newer data. Here we analyze 457 WSI from colorectal cancer screening-undergoing patients. Slide scanning is obtained through a Hamamatsu Nanozoomer S210 scanner configured at $\times 20$ magnification ($0.4415 \mu\text{m}/\text{px}$) and stored as `.ndpi` file. Each WSI has been annotated by expert pathologists according to six classes chosen for our study: hyperplastic polyp (HP); normal tissue (NORM); tubular adenoma, high-grade dysplasia (TA.HG); tubular adenoma, low-grade dysplasia (TA.LG); tubulo-villous adenoma, high-grade dysplasia (TVA.HG) and tubulo-villous adenoma, low-grade dysplasia (TVA.LG).

Each slide is associated with some metadata (stored in NanoZoomer Digital Pathology Annotations `.ndpa` file format), including a collection of Region of Interests (RoIs) associated with the corresponding class. Each RoI is determined by the pathologist and is defined by a free-hand contour, identifying the tissue area exhibiting histological findings. The number and the size of RoIs is highly variable and depends on both the tissue availability and the histological analysis. Such heterogeneity unfortunately, leads to dataset unbalancing: the distribution of the data from T tissue classes in our dataset is shown in Tab. 1. In the table we read the number of WSIs, the number of ROIs R and total tissue area A_t for each t -th class, respectively.

4 Method

In this section we are going to describe and motivate the proposed method. In particular, the use of deep learning for classification already proved, in similar

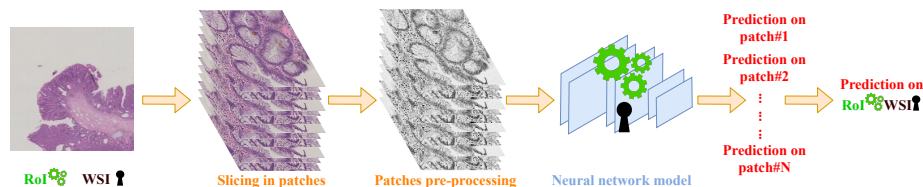


Fig. 1: The neural network is trained on RoI images (gears symbol) and tested on WSI (lock symbol).

	HP	NORM	TA.HG	TA.LG	TVA.HG	TVA.LG	Total
Train Slides	50	25	26	203	36	45	385
Test Slides	12	5	8	29	8	10	72
Train RoIs	133	98	113	695	240	208	1487
Validation RoIs	5	5	5	5	5	5	30
Test RoIs	20	9	27	77	19	32	184

Table 2: Dataset composition. Test RoIs are taken from a disjoint set of slides.

learning tasks, to be extremely effective and robust [16,25]. Direct classification on the (high resolution) whole slide, in our context, is unfeasible: the relevant features are local and can be detected at very low image scale. For this reason, the deep learning model is not trained on the full slides, but on some crops we refer to as *patches*. An high-level representation of our approach is depicted in Fig. 1. Once the model is trained on patches’ classification, in order to get the whole slide classification (at validation/test time), all the scores from the single patches are averaged on the whole slide. WSIs have large resolution and need to be cropped into patches. The first operation we perform on RoIs (even before slicing them into patches) is re-scaling them to some target resolution φ . using the Lancos-3 filter. Then, we slice the RoIs/WSIs into patches (224×224 pixels large) using sliding windows. These patches can be immediately normalized, using approaches like [18], or simply converting in gray-scale to reduce the expected color shift caused by hematoxylin and eosin.

During training we augment data: we include vertical/horizontal flips and a random operation chosen between rotation, equalization, solarization, inversion and contrast enhancing, as proposed in [7].

In order to perform classification on the patches, we have used ResNet-18: it represents a good trade-off between complexity and performance and is one of the broadly-used to solve similar tasks [16,25]. Pre-trained deep neural networks (on the ImageNet classification task) can be effectively used as initialization for medical classification tasks, showing good performance [16].³

³ The pre-trained model used in all the experiments is available at <https://pytorch.org/docs/stable/torchvision/models.html>

5 Results

In this section we show and discuss the classification results obtained on the WSI biopsies dataset described in Sec. 3 with the method proposed in Sec. 4. We can easily expect high error rates, considering that the information about the adenoma type is a visually global information and requires features extracted at different scale than those for the dysplasia grade, which is a more local information. Here we are not interested in distinguishing different adenoma types, but their dysplasia grade. Towards this end, we will follow a hierarchical-like classification approach [26,27], grouping the adenoma classes into high grade (HG) and low grade (LG) dysplasia.

For all the experiments, we split the data at the whole slide level, in order to maintain the separation of tissues from different patients. For each class, 10% of total patients are considered as test set. We summarise the data split in Table 2. The validation set size is fixed to 5 RoIs for each class from the training set (likewise [25]). We train our model for 250 training epochs, and we choose the best one in terms of balanced accuracy (computed on the validation set). Adam has been used as optimizer, and all the hyper-parameters are tuned via grid-search: weight decay is set to 10^{-4} , learning rate $\eta = 10^{-4}$, exponential learning rate decay 0.99 per epoch, and minibatch size 16. Our algorithms are implemented in Python, using PyTorch 1.5, and training/inference runs over an NVIDIA GeForce GTX 1080 GPU.

5.1 Patches normalization

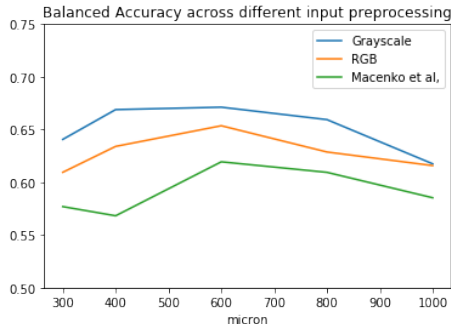


Fig. 2: Patches classification performance.

As a first step, we perform a study at different RoI resolutions: the goal here is to identify the best scale the deep model is able to extract the features. Towards this end, we consider 8 possible patches resolutions $\varphi \in [300; 1000]$ μm , and 3 possible input preprocessing strategies: use of the original patches (RGB), conversion to gray-scale (gray) and the use of a standard slide normalization strategy (Macenko et al. [18]), resulting in 24 training possibilities, which are

reported in Fig. 2. For our classification task, the use of gray-scale images does not remove useful information (which might be embedded in the color) and, on the contrary, helps in removing the expected color bias [19,21]. From our results we learn that, for the particular classification task we aim at solving, the relevant features are embed in the image texture and the signal strength, while the direct use of the RGB image does not compensate the color bias, or even standard slide normalization strategies like [18] destroy some useful information which is not embed in the color feature. For these reasons, we will focus our analysis using gray-scale patch images as input for our model.

5.2 Study on patches resolution for WSI classification

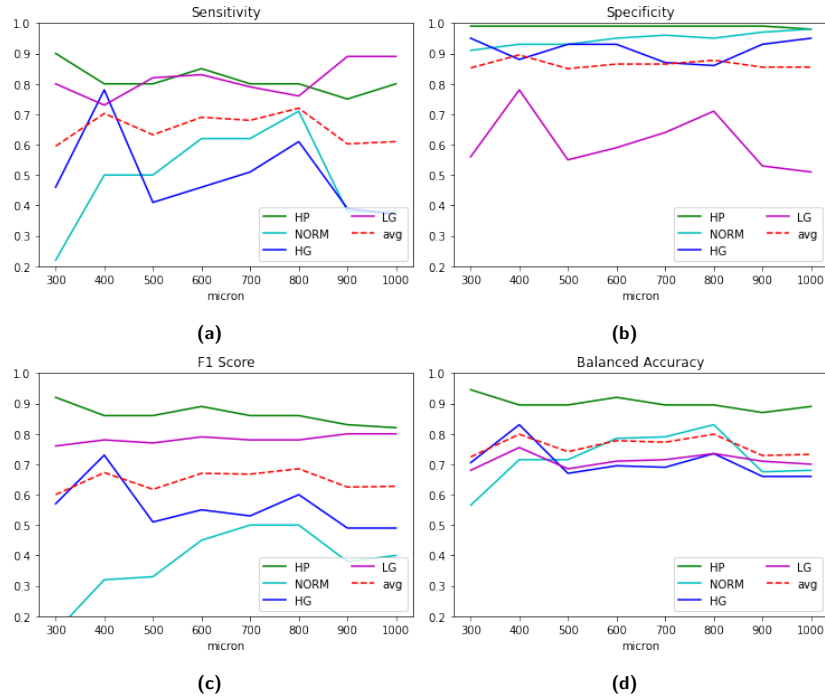


Fig. 3: WSI inference performance comparison between different tissue categories at different patches resolutions: sensitivity (a), specificity (b), F1-score (c) and balanced accuracy (d). Red dashed line is the average performance (avg).

Here we will inspect more in depth the study on WSI classification performance using gray-scaled input. Fig. 3 provides a general overview of some metrics evaluated. There is not a clear choice regarding the optimal scale features have to be extracted. If our goal is to maximize the sensitivity for the HG class, we should choose 400 μm : inspecting the HP’s specificity for the same scale, we observe a drop which, however, is overall tolerable. F1-score gives us a more global information: indeed, for the HG class, 400 μm is the best one. However, if we look

		Accuracy	Sensitivity	Specificity
Hyperplastic	Our (400 μm)	0.90	0.80	0.99
	Our (600 μm)	0.92	0.85	0.99
	Pathologist [9]	0.79	0.30	0.97
Low Grade	Our (400 μm)	0.76	0.73	0.78
	Our (600 μm)	0.71	0.83	0.59
	Pathologist [9]	0.66	0.57	0.69
High Grade	Our (400 μm)	0.83	0.78	0.88
	Our (600 μm)	0.70	0.46	0.93
	Pathologist [9]	0.83	0.81	0.84

Table 3: Human dysplasia diagnostic performance comparison

at average performance on all classes (avg), focusing on F1-score and balanced accuracy, we can observe similar performance for 400 μm and 600-800 μm .

It is important to compare the model performance with the results obtained by human pathologists. Table 3 reports performance comparison for HP, LG and HG in terms of balanced accuracy, sensitivity and specificity. Here, human pathologist’s average performance is taken from Denis et al.’s work [9], evaluated on qualitatively similar data. As we observe, our performance is very close to the pathologists’. In particular, HP classification increases of more than 10% in accuracy, showing a quite significant improvement in terms of sensitivity. LG classification improves as well up to 10% in balanced accuracy, yielding a significant improvement both in terms of sensitivity and specificity. HG classification score is in the same order than human pathologists (this finding is likely to be due to HG features that are known to be visually easier to detect).

5.3 WSI classification with 600 μm patches

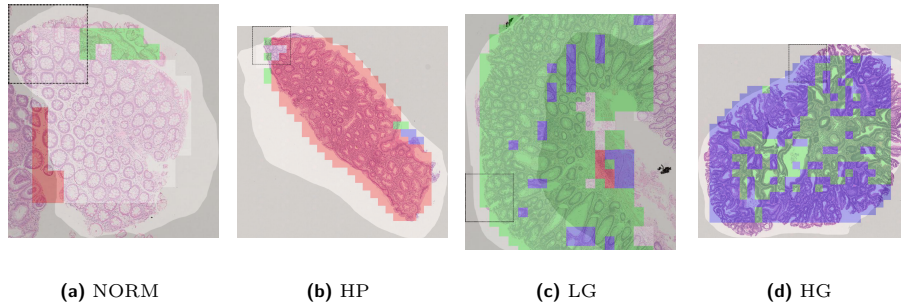


Fig. 4: Patch classification: each box is located at the center of the corresponding patch with a color representing the predicted class: HP (red), NORM (white), LG (green), HG (blue). The black dashed square visually represents the patch scale ($\varphi = 600 \mu\text{m}$).

Considering that the overall performance shown by 400, 600, 700 and 800 μm is similar, we decided here on to focus on $\varphi = 600 \mu\text{m}$. Such a scale is a fair compromise, considering that other works in the literature focus on similar scales [16,25]. Fig. 4 reports a patch-level classification result for the four possible WSI

		(a) $\varphi = 600 \mu\text{m}$, gray-scale						(b) $\varphi = 600 \mu\text{m}$, RGB			
		Predicted						Predicted			
		HP	NORM	HG	LG			HP	NORM	HG	LG
Gr. truth	HP	0.85	0	0.05	0.1	Gr. truth	HP	0.75	0.05	0	0.2
	NORM	0.12	0.75	0	0.12		NORM	0	0.62	0	0.38
	HG	0.02	0	0.63	0.35		HG	0	0.02	0.61	0.37
	LG	0.03	0.09	0.18	0.7		LG	0.03	0.06	0.15	0.76

Table 4: WSI inferences: confusion matrices.

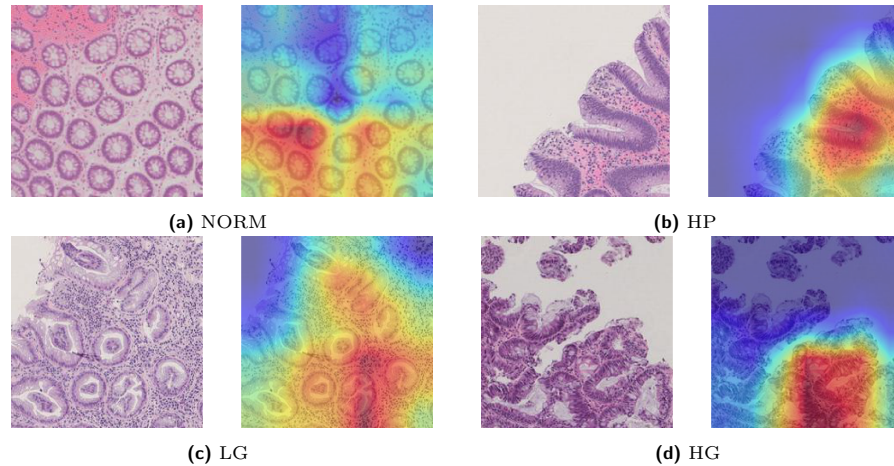


Fig. 5: Regions where the trained neural network model focuses on $600 \mu\text{m}$ patches.

classes. In particular, we observe that the model finds some HG patches within the LG WSI (Fig.4c), and viceversa (Fig.4d). This is an expected behavior, given that the dysplasia grade is provided by the pathologists according to the quantity of tissue (in our case, the number of patches) with high-grade dysplasia.

At $\varphi = 600 \mu\text{m}$, the classification between TA and TVA classes in general is poor: this is due to the larger scale required to extract proper features for adenoma classification. This, however, is not our goal, since we are here interested in classifying the dysplasia grade. Hence, we group HG and LG and we obtain the confusion matrix shown in Table 4 on WSI: the score is competitive to the human classification, as described in Sec. 5.2. We also report the confusion matrix for the equivalent model, using RGB images: as also observed in Sec. 5.1, the use of gray-scale images positively impacts on the WSI inference task. Additionally, we inspect the areas our deep model focuses in order to perform classification by using Grad-CAM. Fig. 5 shows that areas of focus are consistent with the most relevant features of each histo-pathological category. For example, the hot spot of the HP sample is on a serrated gland which is a characteristic finding of this entity.

6 Conclusion

In this work we have designed a neural network-based pipeline for the classification of colorectal polyps in histopathological slides. We found performance benefits by applying grayscale Luma transformation [17] to input tissue patches. We focused on four tissue classes: normal, hyperplastic, high-dysplasia and low-dysplasia adenoma. The dysplasia degree of adenomas is a very important evaluation element for the histopathologist because it leads to different post-polypectomy surveillance protocols [12]. The collected data enable a classification on the dysplasia degree in adenomas. The classification is performed by ResNet-18, inspecting WSI in single patches, and then classified averaging scores on all the patches. Our experiments show a performance which is very close to human pathologists [9]. Future work includes the design of a neural network model able to learn to extract relevant tissue RoIs from the whole slide, evaluated by pathologists' agreement.

References

1. Balkenhol, M.C., Tellez, D., Vreuls, W., Clahsen, P.C., Pinckaers, H., Ciompi, F., Bult, P., van der Laak, J.A.: Deep learning assisted mitotic counting for breast cancer. *Laboratory Investigation* 99(11), 1596–1606 (2019)
2. Barbano, C.A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., Grangetto, M.: Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading (2021)
3. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318(22), 2199–2210 (2017)
4. Bevan, R., Rutter, M.D.: Colorectal cancer screening—who, how, and when? *Clinical endoscopy* 51(1), 37 (2018)
5. Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Wallander, M., Lundin, M., Haglund, C., Lundin, J.: Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports - Nature* 8 (2018)
6. Cavallo, L.B.C.A.B.D.P.E.T.P.C.M.G.A.F.A.G.L.: Unitopatho (2021), <https://dx.doi.org/10.21227/9fsv-tm25>
7. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 113–123 (2019)
8. DeepHealth: Deep-learning and hpc to boost biomedical applications for health (2019), <https://deephealth-project.eu/>
9. Denis, B., Peters, C., Chapelain, C., Kleinclaus, I., Fricker, A., Wild, R., Auge, B., Gendre, I., Perrin, P., Chatelain, D., et al.: Diagnostic accuracy of community pathologists in the interpretation of colorectal polyps. *European journal of gastroenterology & hepatology* 21(10), 1153–1160 (2009)
10. Foss, F.A., Milkins, S., McGregor, A.H.: Inter-observer variability in the histological assessment of colorectal polyps detected through the nhs bowel cancer screening programme. *Histopathology* 61(1), 47–52 (2012)

11. Gonzalez, R.S.: Updates and challenges in gastrointestinal pathology. *Surgical Pathology Clinics* 13(3), ix (2020)
12. Hassan, C., Antonelli, G., Dumonceau, J.M., Regula, J., Bretthauer, M., Chaussade, S., Dekker, E., Ferlitsch, M., Gimeno-Garcia, A., Jover, R., et al.: Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy (esge) guideline–update 2020. *Endoscopy* 52(08), 687–700 (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778. IEEE Computer Society (2016)
14. He, X., Hang, D., Wu, K., Nayor, J., Drew, D.A., Giovannucci, E.L., Ogino, S., Chan, A.T., Song, M.: Long-term risk of colorectal cancer after removal of conventional adenomas and serrated polyps. *Gastroenterology* 158(4), 852–861 (2020)
15. Janowczyk A, M.A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics* pp. 7–29 (2016)
16. Korbar, B., Olofson, A.M., Mirafior, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S.: Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics* 8 (2017)
17. Luma: [https://en.wikipedia.org/wiki/Luma_\(video\)](https://en.wikipedia.org/wiki/Luma_(video))
18. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. pp. 1107–1110. IEEE (2009)
19. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. pp. 309–319. Springer International Publishing, Cham (2020)
20. Mollasharifi, T., Ahadi, M., Jamali, E., Moradi, A., Asghari, P., Maroufizadeh, S., Kazeminezhad, B.: Interobserver agreement in assessing dysplasia in colorectal adenomatous polyps: A multicentric iranian study. *Iranian Journal of Pathology* pp. 167–174 (2020)
21. Roy, S., kumar Jain, A., Lal, S., Kini, J.: A study about color normalization methods for histopathology images. *Micron* 114, 42 – 61 (2018)
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *ICCV*. pp. 618–626. IEEE Computer Society (2017)
23. Song, Z., Yu, C., Zou, S., Wang, W., Huang, Y., Ding, X., Liu, J., Shao, L., Yuan, J., Gou, X., et al.: Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. *BMJ open* 10(9), e036423 (2020)
24. Van Putten, P.G., Hol, L., Van Dekken, H., Han van Krieken, J., Van Ballegooijen, M., Kuipers, E.J., Van Leerdam, M.E.: Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening. *Histopathology* (2011)
25. Wei, J.W., Suriawinata, A.A., Vaickus, L.J., Ren, B., Liu, X., Lisovsky, M., Tomita, N., Abdollahi, B., Kim, A.S., Snover, D.C., et al.: Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Network Open* 3(4), e203398–e203398 (2020)
26. Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., Yu, Y.: Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2740–2748 (2015)
27. Zhu, X., Bain, M.: B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890* (2017)