

Machine learning methods for histopathological image analysis

Daisuke Komura¹, Shumpei Ishikawa^{1*}

* Corresponding author

¹Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

© 2017. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract

Abundant accumulation of digital histopathological images has led to the increased demand for their analysis, such as computer-aided diagnosis using machine learning techniques. However, digital pathological images and related tasks have some issues to be considered. In this mini-review, we introduce the application of digital pathological image analysis using machine learning algorithms, address some problems specific to such analysis, and propose possible solutions.

Keywords: histopathology; deep learning; machine learning; whole slide images; computer assisted diagnosis; digital image analysis

1. Introduction

Pathology diagnosis has been performed by a human pathologist observing the stained specimen on the slide glass using a microscope. In recent years, attempts have been made to capture the entire slide with a scanner and save it as a digital image (Whole slide image, WSI) [1]. As a large number of WSI are being accumulated, attempts have been made to analyze WSIs using digital image analysis based on machine learning algorithms to assist tasks including diagnosis.

Digital pathological image analysis often uses general image recognition technology (e.g. facial recognition) as a basis. However, since digital pathological images and tasks have some unique characteristics, special processing techniques are often required. In this review, we describe the application of digital pathological image analysis using machine learning algorithms, and its problems specific to digital pathological image analysis and the possible solutions.

Since the overwhelming victory of the team using deep learning at ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 [2], most of the image recognition techniques have been replaced by deep learning. This is also true for pathological image analysis [3–5]. Therefore, even though many techniques introduced in this review are related to deep learning, most of them are also applicable for other machine learning algorithms.

2. Machine learning methods

Figure 1 shows typical steps for histopathological image analysis using machine learning. Prior to applying machine learning algorithms, some pre-processing should be performed. For example, when cancer regions are detected in WSI, local mini patches around 256×256 are sampled from large WSI. Then feature extraction and classification between cancer and non-cancer are performed in each local patch. The goal of feature extraction is to extract useful information for machine learning tasks. Various local features such as gray level co-occurrence Matrix (GLCM) and local binary pattern (LBP) have been used for histopathological image analysis, but deep learning algorithms such as convolutional neural network [3,4,6,7] starts the analysis from feature extraction. Features and classifiers are simultaneously optimized in deep learning and features learned in deep learning often outperforms other traditional features in histopathological image analysis.

Machine learning techniques often used in digital pathology image analysis are divided into supervised learning and unsupervised learning. The goal of supervised learning is to infer a function that can map the input images to their appropriate labels (e.g. cancer) well using training data. Labels are associated with a WSI or an object in WSIs. The algorithms for supervised learning include support vector machines, random forest and convolutional neural networks. On the other hand, the goal of unsupervised learning is to infer a function that can describe hidden structures from unlabeled images. The tasks include clustering, anomaly detection and dimensionality reduction. The algorithms for unsupervised learning include k-means, autoencoders and principal component analysis. For details of learning algorithms, please refer to books or reviews such as [8]. There are derivatives from these two learning such as semi-supervised learning and multiple instance learning, which are described in Section 4.2.2.

3. Machine learning application in digital pathology

3.1. Computer-assisted diagnosis

The most actively researched task in digital pathological image analysis is computer-assisted diagnosis (CAD), which is the basic task of the pathologist. Diagnosis is the task to map a WSI or multiple WSIs to one of the disease categories, meaning that it is essentially a supervised learning task. Since the errors made by a machine learning system reportedly differ from those made by a human pathologist [10], diagnostic accuracy could be improved using CAD system. CAD may also lead to the reduce variability in interpretations and prevent overlooking by investigating all pixels within WSIs.

Other diagnosis-related tasks include detection or segmentation of Region of Interest (ROI) such as tumor region in WSI, scoring of immunostaining [5,11], cancer staging [10,12], mitosis detection [13], gland segmentation [14], and detection and quantification of vascular invasion [15].

Most of these tasks analyze Formalin-fixed paraffin-embedded (FFPE) slides, and few research analyzed frozen sections to perform intraoperative diagnosis [16], which is another important application of CAD because the number of WSIs suitable for the analysis is less sufficient and the task is more challenging compared to the analysis of FFPE slides.

3.2. Content Based Image Retrieval

Content Based Image Retrieval (CBIR) retrieves similar images to a query image. In digital pathology, CBIR systems are useful in many situations, particularly in diagnosis, education, and research [17–21]. For example, CBIR systems can be used for educational purposes by students and beginner pathologists to retrieve relevant cases or histopathological images of tissues. In addition, such systems are also helpful to professional pathologists, particularly when diagnosing of rare cases.

Since CBIR does not necessarily require label information, unsupervised learning can be used. When label information is available, supervised learning approaches could learn better similarity measure than unsupervised learning approaches since the similarity between histopathological images may differ by definition. However, preparing sufficient number of labeled data can be a serious problem as will be described later.

In CBIR, not only accuracy but also high-speed search of similar images from numerous images are required. Therefore, various techniques for dimensionality reduction of image features such as principal component analysis and compact bilinear pooling [22], and fast approximate nearest neighbor search such as kd-tree and hashing [23] are utilized for high speed search.

3.3. Discovering new clinicopathological relationships

Historically, many important discoveries concerning diseases such as tumor and infectious diseases have been made by pathologists and researchers who have carefully and closely observed pathological specimens. For example, *H. pylori* was discovered by a pathologist who was examining the gastric mucosa of patients with gastritis [24].

Attempts have also been made to correlate the morphological features of cancers with their clinical behavior. For example, tumor grading is important in planning treatment and determining a patient's prognosis for certain types of cancer, such as soft tissue sarcoma, primary brain tumors, and breast and prostate cancer.

Meanwhile, thanks to the progress in digitization of medical information and advance in genome analysis technology in recent years, large amount of digital information such as genome information, digital pathological images, MRI and CT images has become available [9]. By analyzing the relationship between these data, new clinicopathological relationships, for example, the relationship between the morphological characteristic and the somatic mutation of the cancer, can be found [25,26]. However, since the amount of data is enormous, it is not realistic for pathologists and researchers to analyze all the relationships manually by looking at the specimens. This is where the machine learning technology comes in. For example, Beck et al. extracted texture information from pathological images of breast cancer and analyzed with L1 - regularized logistic regression, and indicated that the histology of stroma correlates with prognosis in breast cancer [27]. Other researches include prognosis predictions from histopathological image of cancer [28], prediction of somatic mutation [7], and discovery of new gene variants related to autoimmune thyroiditis based on image QTL [29].

4. Problems specific to histopathological image analysis

In this section, we describe unique characteristics of pathological image analysis and computational methods to treat them.

4.1. Very large image size

When images such as dogs or houses are classified using deep learning, small sized image such as 256×256 pixels is often used as an input. Images with large size often need to be resized into smaller size which is enough for sufficient distinction, as increase in the size of the input image results in the increase in the parameter to be estimated, the required computational power, and memory. In contrast, WSI contains many cells and the image could consist of as many as tens of billions of pixels, which is usually hard to analyzed as is. However, resizing the entire image to a smaller size such as 256×256 would lead to the loss of information at cellular level, resulting in marked decrease of the identification accuracy. Therefore, the entire WSI is commonly divided into partial

regions of about 256×256 pixels (“patches”), and each patch is analyzed independently, such as detection of ROIs. Thanks to the advances in computational power and memory, patch size is increasing (e.g. 960×960), which is expected to contribute to better accuracy. There is still a room for improvement in the method of integrating the result from each patch. For example, as the entire WSI could contain hundreds of thousands of patches, false positives are highly likely to appear even if individual patches are accurately classified. One possible solution for this is regional averaging of each decision, such that the regions is classified as ROI only when the ROI extends over multiple patches. However, this approach may suffer from false negatives, resulting in missing small ROIs such as isolated tumor cells [30].

In some applications such as staging of lymph node metastasis of specimens or patients, and staging of prostate cancer diagnosed by Glisson score of multiple regions within one slide, more sophisticated algorithms to integrate patch-level decisions are required. For example, for pN-staging of metastatic breast cancer, which was one of the tasks in Camelyon 17, multiple participating teams including us applied random forest classifiers of pixel or patch-level probabilities estimated by deep learning using various features such as estimated tumor size [30].

4.2. Insufficient labeled images

Probably the biggest problem in pathological image analysis using machine learning is that only a small number of training data is available. A key to the success of deep learning in general image recognition task is that training data is extremely abundant. Although label information at patch-level or pixel-level (e.g. inside/outside boundary of cancerous regions) is required in most tasks in digital pathology such as diagnosis, most labels of WSIs are at case-level (e.g. diagnosis) at most. Label information in general image analysis can be easily retrieved from the internet and it is also possible to use crowdsourcing labeling because anyone can identify objects and perform labeled work. However, only pathologists can label the pathological image accurately, and labeling at the regional level in a huge WSIs requires a lot of labor.

It is possible to reuse public ready-to-analyze data as training data in machine learning, such as ImageNet [31] in natural images and International Skin Imaging Collaboration [32] in macroscopic diagnosis of skin. In the field of digital pathology, The Cancer Genome Atlas (TCGA) [33] and Genotype-Tissue Expression (GTEx) [34,35] contain

many freely downloadable high-resolution WSIs, which may serve as potential training data. Furthermore, both of them also provide genomic profiles, which could be used to investigate relationships between genotype and morphology. The problem is that the WSIs in these repositories contain labels at the case-level, and some preprocessing or specialized machine learning algorithm for treating case-level labels is required to make them suitable for training data.

Many researches have attempted to solve the problem. Most of the approaches fall into one of the following categories: 1) efficient increase of label data, 2) utilization of weak label or unlabeled information, or 3) utilization of models/parameters for other tasks.

4.2.1. Efficient labeling

One way to increase training data is to reduce the working time of pathologists to specify ROIs in the WSI. Easy-to-use GUI tools help pathologists efficiently label more samples in shorter periods of time. For example, Cytomine [36] not only allows pathologists to surround ROIs in WSIs with ellipses, rectangles, polygons or freehand drawings, but also applies content-based image retrieval algorithms to speed up annotation. Another interesting idea to reduce working time is to automatically localize ROIs during diagnosis, which uses the usual working time for diagnosis as labeling by tracking pathologists' behavior. This approach tracks pathologists' eye movement [37], mouse cursor positions [38] and change in viewport [39]. However, localizing ROIs accurately from these tracking data is not always easy since pathologists do not always spend time looking at ROIs, and boundary information obtained by these approaches tends to be less clear.

Another approach that utilizes a machine learning method is active learning [40–42]. This is generally effective when the acquisition cost of label data is large (i.e. pathological images). Active learning is a method used in supervised learning, and it automatically chooses the most valuable unlabeled sample (i.e. the one that is expected to improve the identification performance of classifiers when labeled correctly and used as a training data) and display it for labeling by pathologists. Since this approach is likely to increase discrimination performance with smaller number of labeled images, the total labeling time to obtain the same discrimination performance will be shortened [40]. Many criteria such as uncertainty sampling [43], expected error reduction [44] have been proposed for selecting valuable unlabeled samples.

4.2.2. Incorporating insufficient label

Even if the exact position of the ROI in a WSI is not known, it is possible that the information regarding the presence/absence of the ROI in the WSI is available from the pathological diagnosis assigned to the WSI or WSI-level labels. These so-called weak labels are easy to obtain compared to patch-level labels even when the WSIs have no further information, and in this regard, WSIs are considered as a “bag” made with many patches (instances) in machine learning settings. When diagnosing cancer, WSI is labeled as cancer if at least one patch contains cancerous tissue, or normal if none of the patches contain cancerous tissue. This setting is a problem of multiple instance learning [45] or weakly-supervised learning [46,47]. In a typical multiple instance learning problem, positive bags contain at least one positive instance and negative bags do not contain any positive instances. The aim of multiple instance learning is to predict bag or instance label based on training data that contains only bag labels. Various methods in multiple instance learning have been applied to histopathological image analysis including boosting-based approach [46], support vector machine-based approach [48] and deep learning-based approach [47].

In contrast, semi-supervised learning [49–51] utilizes both labeled and unlabeled data. Unlabeled data is used to estimate the true distribution of labeled data. For example, as shown in Figure 1, decision boundary which takes only the labeled samples into account would form a vertical line, but that considering both labeled and unlabeled samples would form a slanting line, which could be more accurate. Since semi-supervised learning is considered particularly effective when samples in the same class form a well-discriminative cluster, relatively easy problem could be a good target.

4.2.3. Reusing parameters from another task

Performing supervised learning using too few training data would only result in insufficient generalization performance. This is true especially in deep learning, where the number of parameters to be learned is very large. In such a case, instead of learning the entire model from scratch, learning often starts by using (a part of) parameters of a pre-trained model optimized in another similar task. Such a learning method is called transfer learning. In CNN, layers before the last (typically three) fully-connected layers are regarded as feature extractors. The fully-connected layers are often replaced by a new network suitable for the target task. The parameters in earlier layers can be used as is, or

as initial parameters and then the network is learned partially or fully from the training data of the target task (so-called fine-tuning). In pathological images, no network learned from tasks using other pathological images are available, and thus networks learned using ImageNet, which is a database containing vast number of general images, are often used. For example, Xu et al., performed classification and segmentation tasks on brain and colon pathological images using features extracted from CNN trained on ImageNet, and achieved state-of-the-art performance [52]. Although the pathological image itself looks very different to the general images (e.g. cats and dogs), they share common basic image structures such as lines and arcs. Since earlier layers in deep learning capture these basic image structures, such pre-trained models using general images work well in histopathological image analysis. Nevertheless, if models pre-trained on sufficient number of diverse tissue pathology images are available, they may outperform the ImageNet pre-trained models.

4.3. Different levels of magnification result in different levels of information

Tissues are usually composed of cells, and different tissues show distinct cellular features. Information regarding cell shape is well captured in high-power field microscopic images, but structural information such as a glandular structure made of many cells are better captured in a lower-power field (Figure 2). Because cancerous tissues have both cellular and structural atypia, images taken at multiple magnifications would each contain important information. Pathologists diagnose diseases by acquiring different kinds of information from the cellular level to the tissue level by changing magnifications of a microscope. Even in machine learning, researches utilizing images at different magnifications exist. As mentioned above, it is difficult to handle the images at its original resolution directly, images are often resized to correspond to various magnifications and used as input for analysis. Regarding diagnosis, the most informative magnification is still controversial [10,53], but improvement in accuracy is sometimes achieved by inputting both high and low magnification images simultaneously, probably depending on the types of diseases and tissues, and machine learning algorithms.

4.4. WSI as orderless texture-like image

Pathological image is different from cats and dogs in nature, in a sense that it shows repetitive pattern of minimum components (usually cells). Therefore, it is rather closer to

texture than object. CNN acquires shift invariance to a certain extent by pooling operations. In addition, even normal CNN can learn texture-like structure by data augmentation by shifting the tissue image with a small stride. Meanwhile, there has been methods which utilize texture structure more intensively, such as gray level co-occurrence matrix [54], local binary pattern [55], Gabor filter bank, and recently developed deep texture representations using a CNN [56]. Deep texture representations are computed using a correlation matrix of feature maps in a CNN layer. Converting the CNN features to texture representations would lead to the acquisition of invariance regarding cell position, while utilizing good representations learned by CNN. Another advantage of deep texture representation is that there are no constraints on the size of input image, which is very suitable for large image size of WSI. The boundary between texture and non-texture is unclear, but a single cell or a single structure is obviously not a texture. Better approach would thus depend on the object to be analyzed.

4.5. Color variation and artifacts

WSIs are created through multiple processes: pathology specimens are sliced and placed on a slide glass, stained with hematoxylin and eosin, and then scanned. At each step undesirable effects, which are unrelated to the underlying biological factors, could be introduced. For example, when tissue slices are being placed onto the slides, they may be bent and wrinkled; dust may contaminate the slides during scanning; blur attributable to different thickness of tissue sections may occur (Figure 3); and sometimes tissue regions are marked by color markers. Since these artifacts could adversely affect the interpretation, specific algorithms to detect artifacts such as blur [57] and tissue-folds [58] have been proposed. Such algorithms may be used for preprocessing WSIs.

Another serious artifact is color variation as shown in Figure 4. The sources of variation include different lots or manufacturers of staining reagents, thickness of tissue sections, staining conditions and scanner models. Learning without considering the color variation could worsen the performance of machine learning algorithm. If sufficient data on every stained tissue acquired by every scanner can be incorporated, the influence of color variation on classification accuracy may become negligible; however, that seems very unlikely at the moment.

To address this issue, various methods have been proposed so far including conversion to gray scale, color normalization [59–61], and color augmentation [57]. Conversion to

grayscale is the easiest way, but it ignores the important information regarding the color representation used routinely by pathologists. In contrast, color normalization tries to adjust the color values of an image on a pixel-by-pixel basis so that the color distribution of the source image matches to that of a reference image. However, as the components and composition ratios of cells or tissues in target and reference images differ in general, preprocessing such as nuclear detection using a dedicated algorithm to adjust the component is often required. For this reason, color normalization seems to be suitable when WSIs analyzed in the tasks contain, at least partially, similar compositions of cells or tissues.

On the other hand, color augmentation is a kind of data augmentation performed by applying random hue, saturation, brightness, and contrast. The advantage of color augmentation lies in the easy implementation regardless of the object being analyzed. Color augmentation seems to be suitable for WSIs with smaller color variation, since excessive color change in color augmentation could lead to the loss of color information in the final classifier. As color normalization and color augmentation could be complementary, combination of both approaches may be better.

5. Summary and Outlook

Digital histopathological image recognition is a very suitable problem for machine learning since the images themselves contain information sufficient for diagnosis. In this review, we brought up problems in digital histopathological image analysis using machine learning. Due to great efforts made so far, these problems are becoming tractable, but there is still room for improvement. Most of these problems are likely to be solved once a large number of well-annotated WSIs become available. Most WSIs are currently owned by each institute and kept private: making these data public, or sharing them among institutes, will alone boost the development of more sophisticated digital histopathological image analysis.

Acknowledgement

This study was supported by JSPS Grant-in-Aid for Scientific Research (A), No. 25710020 (SI).

- [1] Pantanowitz L. Digital images and the future of digital pathology. J Pathol

Inform 2010;1. doi:10.4103/2153-3539.68332.

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Adv. Neural Inf. Process. Syst. 25, Curran Associates, Inc.; 2012, p. 1097–1105.

[3] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. ArXiv150407947 Cs 2015.

[4] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214–23. doi:10.1016/j.neucom.2016.01.034.

[5] Sheikhzadeh F, Guillaud M, Ward RK. Automatic labeling of molecular biomarkers of whole slide immunohistochemistry images using fully convolutional networks. ArXiv161209420 Cs Q-Bio 2016.

[6] Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286. doi:10.1038/srep26286.

[7] Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer. bioRxiv 2017:64279. doi:10.1101/064279.

[8] Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press; 2012.

[9] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113–1120.

[10] Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. ArXiv160605718 Cs Q-Bio 2016.

[11] Mungle T, Tewary S, Das DK, Arun I, Basak B, Agarwal S, et al. MRF-ANN: a machine learning approach for automated ER scoring of breast cancer immunohistochemical images. J Microsc 2017;267:117–29. doi:10.1111/jmi.12552.

[12] Wang D, Foran DJ, Ren J, Zhong H, Kim IY, Qi X. Exploring automatic prostate histopathology image gleason grading via local structure modeling. 2015 37th

Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC, 2015, p. 2649–52.

doi:10.1109/EMBC.2015.7318936.

[13] Shah M, Rubadue C, Suster D, Wang D. Deep Learning Assessment of Tumor Proliferation in Breast Cancer Histological Images. ArXiv161003467 Cs 2016.

[14] Chen H, Qi X, Yu L, Heng PA. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, 2016, p. 2487–96. doi:10.1109/CVPR.2016.273.

[15] Caie PD, Turnbull AK, Farrington SM, Oniscu A, Harrison DJ. Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer. J Transl Med 2014;12:156. doi:10.1186/1479-5876-12-156.

[16] Abas FS, Gokozan HN, Goksel B, Otero JJ, Gurcan MN. Intraoperative neuropathology of glioma recurrence: cell detection and classification. vol. 9791, 2016, p. 979109-979109–10. doi:10.1117/12.2216448.

[17] Caicedo JC, González FA, Romero E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. J Biomed Inform 2011;44:519–28. doi:10.1016/j.jbi.2011.01.011.

[18] Mehta N, Raja'S A, Chaudhary V. Content based sub-image retrieval system for high resolution pathology images using salient interest points. Eng. Med. Biol. Soc. 2009 EMBC 2009 Annu. Int. Conf. IEEE, IEEE; 2009, p. 3719–3722.

[19] Qi X, Wang D, Roderio I, Diaz-Montes J, Gensure RH, Xing F, et al. Content-based histopathology image retrieval using CometCloud. BMC Bioinformatics 2014;15:287. doi:10.1186/1471-2105-15-287.

[20] Sridhar A, Doyle S, Madabhushi A. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. J Pathol Inform 2015;6:41. doi:10.4103/2153-3539.159441.

[21] Vanegas JA, Arevalo J, González FA. Unsupervised feature learning for content-based histopathology image retrieval. 2014 12th Int. Workshop Content-Based Multimed. Index. CBMI, 2014, p. 1–6. doi:10.1109/CBMI.2014.6849815.

[22] Gao Y, Beijbom O, Zhang N, Darrell T. Compact Bilinear Pooling. ArXiv151106062 Cs 2015.

[23] Zhang X, Liu W, Dundar M, Badve S, Zhang S. Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval. IEEE Trans Med

Imaging 2015;34:496–506. doi:10.1109/TMI.2014.2361481.

- [24] Marshall B. A Brief History of the Discovery of *Helicobacter pylori*. *Helicobacter Pylori*, Springer, Tokyo; 2016, p. 3–15. doi:10.1007/978-4-431-55705-0_1.
- [25] Molin MD, Matthaei H, Wu J, Blackford A, Debeljak M, Rezaee N, et al. Clinicopathological correlates of activating GNAS mutations in intraductal papillary mucinous neoplasm (IPMN) of the pancreas. *Ann Surg Oncol* 2013;20:3802–8. doi:10.1245/s10434-013-3096-1.
- [26] Yoshida A, Tsuta K, Nakamura H, Kohno T, Takahashi F, Asamura H, et al. Comprehensive histologic analysis of ALK-rearranged lung carcinomas. *Am J Surg Pathol* 2011;35:1226–34. doi:10.1097/PAS.0b013e3182233e06.
- [27] Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, Vijver MJ van de, et al. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Sci Transl Med* 2011;3:108ra113-108ra113. doi:10.1126/scitranslmed.3002564.
- [28] Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* 2016;7:12474. doi:10.1038/ncomms12474.
- [29] Barry JD, Fagny M, Paulson JN, Aerts H, Platig J, Quackenbush J. Histopathological image QTL discovery of thyroid autoimmune disease variants. *bioRxiv* 2017:126730. doi:10.1101/126730.
- [30] CAMELYON17 n.d. <https://camelyon17.grand-challenge.org/> (accessed August 21, 2017).
- [31] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 2015;115:211–52. doi:10.1007/s11263-015-0816-y.
- [32] Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *ArXiv160501397 Cs* 2016.
- [33] Genomic Data Commons Data Portal (Legacy Archive) n.d.
- [34] The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;45:580–5. doi:10.1038/ng.2653.

- [35] Biospecimen Research Database n.d. <https://brd.nci.nih.gov/brd/image-search/searchhome> (accessed August 30, 2017).
- [36] Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinforma Oxf Engl* 2016;32:1395–401. doi:10.1093/bioinformatics/btw013.
- [37] Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study n.d. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0103447> (accessed July 20, 2017).
- [38] Raghunath V, Braxton MO, Gagnon SA, Brunyé TT, Allison KH, Reisch LM, et al. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. *J Pathol Inform* 2012;3:43. doi:10.4103/2153-3539.104905.
- [39] Mercan E, Aksoy S, Shapiro LG, Weaver DL, Brunyé TT, Elmore JG. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images: a Comparative Study. *J Digit Imaging* 2016;29:496–506. doi:10.1007/s10278-016-9873-1.
- [40] Doyle S, Monaco J, Feldman M, Tomaszewski J, Madabhushi A. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinformatics* 2011;12:424. doi:10.1186/1471-2105-12-424.
- [41] Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification. *J Mach Learn Res* 2001;2:45–66.
- [42] Lewis DD, Gale WA. A Sequential Algorithm for Training Text Classifiers. In: Croft BW, van Rijsbergen CJ, editors. *SIGIR '94 Proc. Seventeenth Annu. Int. ACM-SIGIR Conf. Res. Dev. Inf. Retr. Organised Dublin City Univ., London: Springer London; 1994*, p. 3–12. doi:10.1007/978-1-4471-2099-5_1.
- [43] Homeyer A, Schenk A, Dahmen U, Dirsch O, Huang H, Hahn HK. A comparison of sampling strategies for histological image analysis. *J Pathol Inform* 2012;2. doi:10.4103/2153-3539.92034.
- [44] Padmanabhan RK, Somasundar VH, Griffith SD, Zhu J, Samoyedny D, Tan KS, et al. An Active Learning Approach for Rapid Characterization of Endothelial Cells in Human Tumors. *PLoS ONE* 2014;9. doi:10.1371/journal.pone.0090495.
- [45] Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance

problem with axis-parallel rectangles. *Artif Intell* 1997;89:31–71. doi:10.1016/S0004-3702(96)00034-3.

[46] Xu Y, Zhu J-Y, Chang EI-C, Lai M, Tu Z. Weakly supervised histopathology cancer image segmentation and classification. *Med Image Anal* 2014;18:591–604. doi:10.1016/j.media.2014.01.010.

[47] Jia Z, Huang X, Chang EI-C, Xu Y. Constrained Deep Weak Supervision for Histopathology Image Segmentation. *ArXiv170100794 Cs* 2017.

[48] BenTaieb A, Li-Chang H, Huntsman D, Hamarneh G. A structured latent model for ovarian carcinoma subtyping from histopathology slides. *Med Image Anal* 2017;39:194–205. doi:10.1016/j.media.2017.04.008.

[49] Peikari M, Zubovits J, Clarke G, Martel AL. Clustering Analysis for Semi-supervised Learning Improves Classification Performance of Digital Pathology. *Mach. Learn. Med. Imaging*, Springer, Cham; 2015, p. 263–70. doi:10.1007/978-3-319-24888-2_32.

[50] Miyato T, Maeda S, Koyama M, Ishii S. Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning. *ArXiv170403976 Cs Stat* 2017.

[51] Rasmus A, Valpola H, Honkala M, Berglund M, Raiko T. Semi-Supervised Learning with Ladder Networks. *ArXiv150702672 Cs Stat* 2015.

[52] Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics* 2017;18:281. doi:10.1186/s12859-017-1685-x.

[53] Gupta V, Bhavsar, Arnav. Breast Cancer Histopathological Image Classification: Is Magnification Important?, n.d.

[54] Saito A, Numata Y, Hamada T, Horisawa T, Cosatto E, Graf H-P, et al. A novel method for morphological pleomorphism and heterogeneity quantitative measurement: Named cell feature level co-occurrence matrix. *J Pathol Inform* 2016;7. doi:10.4103/2153-3539.189699.

[55] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit* 1996;29:51–9. doi:10.1016/0031-3203(95)00067-4.

[56] Lin T-Y, RoyChowdhury A, Maji S. Bilinear CNN Models for Fine-grained

Visual Recognition. ArXiv150407889 Cs 2015.

[57] Wu H, Phan JH, Bhatia AK, Shehata B, Wang MD. Detection of Blur Artifacts in Histopathological Whole-Slide Images of Endomyocardial Biopsies. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2015;2015:727–30. doi:10.1109/EMBC.2015.7318465.

[58] Kothari S, Phan JH, Wang MD. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inform 2013;4. doi:10.4103/2153-3539.117448.

[59] Bejnordi BE, Litjens G, Timofeeva N, Otte-Höller I, Homeyer A, Karssemeijer N, et al. Stain Specific Standardization of Whole-Slide Histopathological Images. IEEE Trans Med Imaging 2016;35:404–15. doi:10.1109/TMI.2015.2476509.

[60] Ciompi F, Geessink O, Bejnordi BE, de Souza GS, Baidoshvili A, Litjens G, et al. The importance of stain normalization in colorectal tissue classification with convolutional networks. ArXiv170205931 Cs 2017.

[61] Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014;61:1729–38. doi:10.1109/TBME.2014.2303294.

[62] Lafarge MW, Pluim JPW, Eppenhof KAJ, Moeskops P, Veta M. Domain-adversarial neural networks to address the appearance variability of histopathology images. ArXiv170706183 Cs 2017.

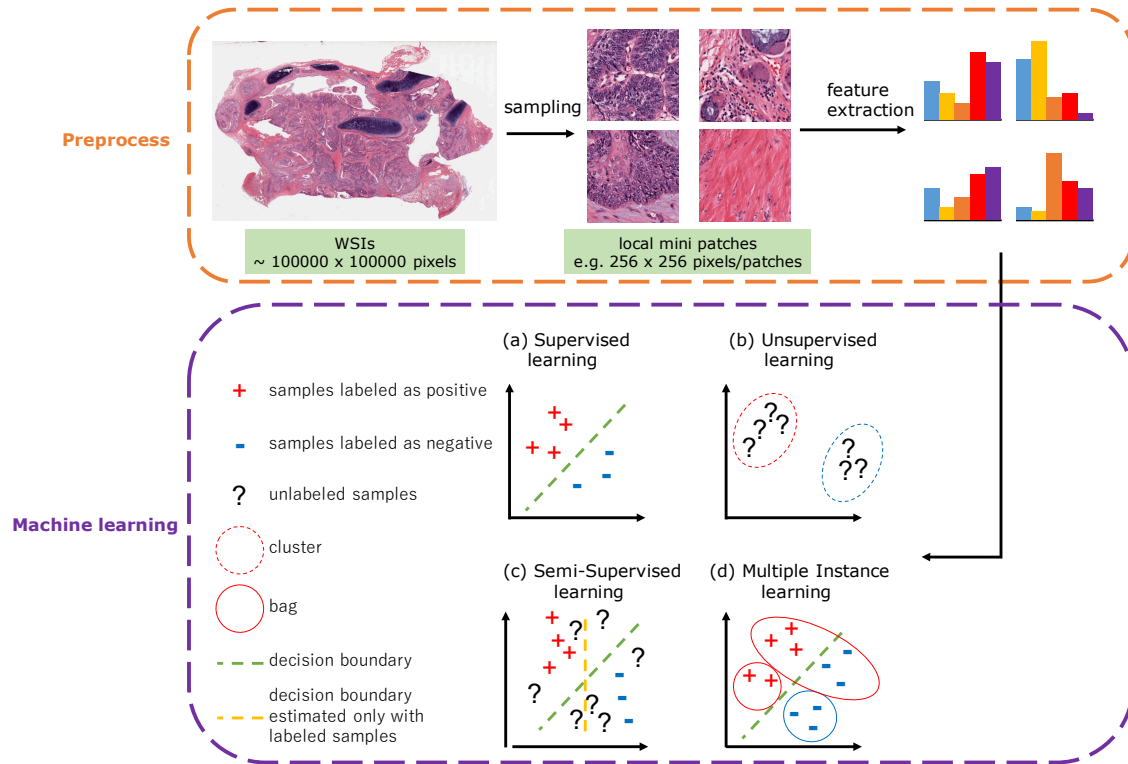


Figure 1. Typical steps for machine learning in digital pathological image analysis. After preprocessing whole slide images, various types of machine learning algorithms could be applied including (a) supervised learning (see Section 2), (b) unsupervised learning (see Section 2), (c) semi-supervised learning (see Section 4.2.2), and (d) multiple instance learning (see Section 4.2.2). The histopathological images are adopted from The Cancer Genome Atlas (TCGA)[9]

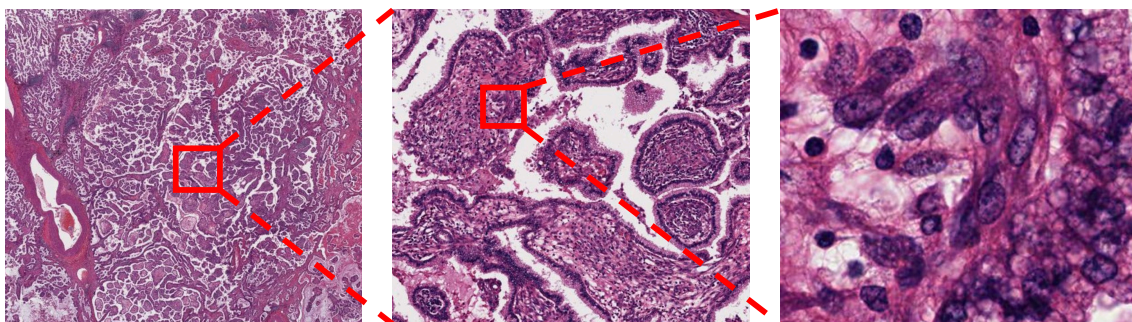


Figure 2. Multiple magnification levels of the same histopathological image. Right images show the magnified region indicated by red box on the left images. Leftmost image clearly shows papillary structure, and rightmost image clearly shows nucleus of each cell. The histopathological images are adopted from TCGA[9]

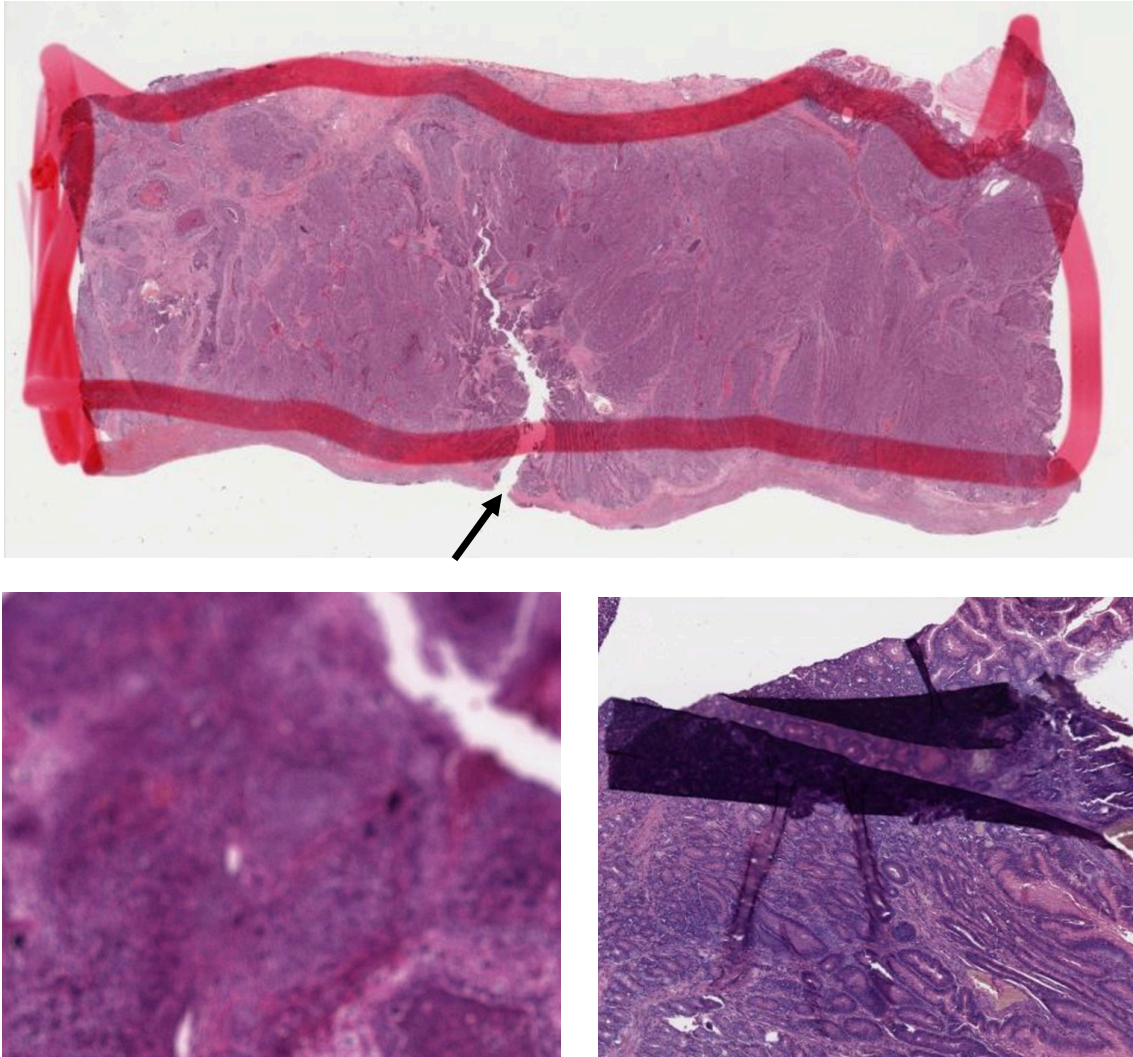


Figure 3. Artifacts in WSIs. Top: tumor region is outlined with red marker. The arrow indicates a tear possibly formed during the tissue preparation process. Left bottom: blurred image. Right bottom: folded tissue section. The histopathological images are adopted from TCGA[9]

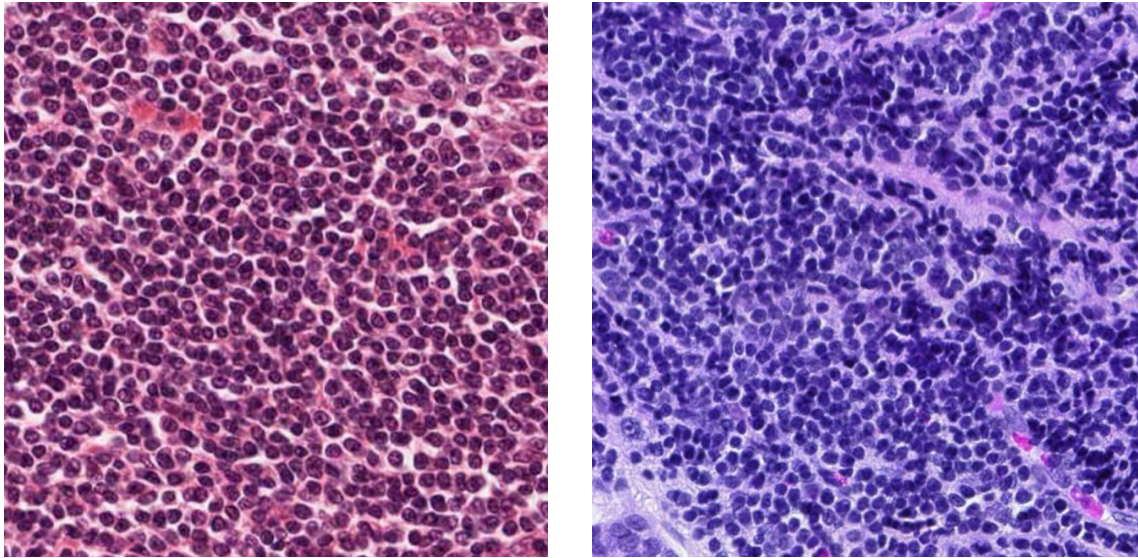


Figure 4. Color variation of histopathological images. Both of these two images show lymphocytes. The histopathological images are adopted from TCGA[9]