The American Journal of Pathology, Vol. 191, No. 10, October 2021

The American Journal of
PATHOLOGY

ajp.amjpathol.org

Artificial Intelligence and Deep Learning in Pathology Theme Issue

MINI-REVIEW

Searching Images for Consensus

Can AI Remove Observer Variability in Pathology?

Hamid R. Tizhoosh,* Phedias Diamandis,† Clinton J.V. Campbell,‡ Amir Safarpoor,* Shivam Kalra,* Danial Maleki,* Abtin Riasatian,* and Morteza Babaie*

From the Kimia Laboratory, University of Waterloo, Waterloo; the Department of Laboratory Medicine and Pathobiology,† University of Toronto, Toronto; and the Department of Pathology and Molecular Medicine,‡ Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada*

Accepted for publication January 25, 2021.

Address correspondence to Hamid R. Tizhoosh, Ph.D., University of Waterloo, Waterloo, Ontario, Canada. E-mail: tzhoosh@uwaterloo.ca.

One of the major obstacles in reaching diagnostic consensus is observer variability. With the recent success of artificial intelligence, particularly the deep networks, the question emerges as to whether the fundamental challenge of diagnostic imaging can now be resolved. This article briefly reviews the problem and how eventually both supervised and unsupervised AI technologies could help to overcome it. *(Am J Pathol 2021, 191: 1702—1708; https://doi.org/10.1016/j.ajpath.2021.01.015)*

Observer variability in pathology is described as the degree of variation between the diagnostic interpretations when a set of cases are examined by two or more independent clinicians. While there are many ways to estimate these errors, in most circumstances, this estimation requires large-scale case reviews by two or more blinded pathologists examining the same cases independently. While straightforward in theory, this quantification has proven difficult, as it is a costly and time-consuming process and appears to vary across different sample types and pathologists. Overall, cancer diagnoses tend to be more highly variable than non-neoplastic cases, especially as the number of diagnostic criteria continues to evolve in the era of molecular diagnostics.[1,2] Variability also appears to depend on specific tumor type and tumor frequency, with breast and gynecologic malignancies and less common atypical tumor types showing higher observer variability than skin and gastrointestinal lesions and more common tumor types, respectively.[2–5] Similarly, observer factors are an important consideration. Specific training and disease distribution may substantially change from a pathologist's original training or working environment, leading to specific diagnostic biases.[6] In some studies in which pathologists shared difficult cases with intradepartmental colleagues, the diagnosis was changed in as many as 13% of cases.[5,7]

The overall reported error rates in pathology are relatively low and are currently of little clinical significance at a population level.[8,9] However, the transition to more personalized care and the reliance on more precise diagnostics for guiding patient management will likely make previously inconsequential errors more crucial, especially for individual patients.[10] Therefore, subtle differences between diagnostic considerations currently too subjective for consistent agreement between independent pathologists may soon lead to life-saving decisions in personalized medicine. A false positive diagnosis could result in a patient receiving inappropriate or unnecessary toxic chemotherapy in oncology. Conversely, false negative results may delay or deny the benefit of effective therapy in another patient. Unfortunately, many errors appear to be the result of pathologists missing a pathologic finding on a slide or not

No Funding.

Disclosures: S.K. is a Ph.D. student at the University of Waterloo and had a Mathematics of Information Technology and Complex Systems internship at Huron Digital Pathology (St. Jacobs, ON, Canada). H.T. has received consultant's fees from Huron Digital Pathology in 2020.

This article is part of a mini-review series on the applications of artificial intelligence and deep learning in advancing research and diagnosis in pathology.

https://doi.org/10.1016/j.ajpath.2021.01.015

considering an alternative diagnosis (false negative), and may be related to distracting factors that are increasing in modern clinical practice.[3,8,11,12] Recently, several potential diagnostic misinterpretations in expert-annotated cohorts have been highlighted, suggesting that interobserver variability may be an under-reported issue that requires addressing modern informatic solutions.[13] As new tools for addressing interobserver variability are designed, an understanding of why and where most variability exists can help in prioritizing the development of context-specific informatics tools to improve quality assurance in pathology.

Interobserver variability also remains problematic in a major field of pathology, hematopathology, specifically in myelodysplastic syndromes (MDSs). Diagnosing MDSs is one of the most challenging areas in hematopathology and involves multiple clinical and pathologic components.[14] The central and most important aspect in making a diagnosis of MDSs is cytomorphology (visual features of bone marrow cells).[14,15] In suspected cases of MDS, individual bone marrow cells are assessed by a hematopathologist for subtle and nuanced variations of normal cytomorphologic features known as *myelodysplasias*.[16] While the World Health Organization has attempted to standardize the assessment of myelodysplasias in MDS[15], significant interobserver variation remains.[16–20] Up to 12% of MDS cases might be misdiagnosed in smaller, less experienced centers due to a failure to recognize morphologic dysplasias.[21] Consequently, there is a recognized need for standardization in the assessment of cytomorphologic features in MDS.[16,22,23] Given the success of machine learning in the analysis of histomorphologic features in many nonhematopoietic pathology domains,[24–26] there is a clear opportunity to develop machine learning–based approaches to support standardized cytomorphology in MDS.

Looking at the immensity of observer variability as a pervasive source of error and ineffectiveness, the question becomes pressing: Which artificial intelligence (AI) paradigm, and to what extent, can bring about a real solution?

Supervised versus Unsupervised Solutions

In AI, different learning algorithms can be utilized, depending on the available data and the research problem at hand. The level of "supervision" used for solving the problem is certainly a key factor in categorizing AI methods. Supervised, semi-supervised, and unsupervised learning can generally be distinguished. When there is a labeled data set in which each sample is tagged/delineated either quantitatively or qualitatively, supervised learning is often a reasonable candidate solution. The goal of supervised learning is to use inputs for predicting the matching outputs. According to the output type, the prediction task can be called *regression* or *classification* for qualitative and quantitative outputs.

Supervised-learning applications in computational pathology can fall into several categories: classification,[27] segmentation,[28] detection,[29] immunostaining scoring,[30] cancer staging,[31] and survival prediction.[32,33] Although supervised-learning approaches sometimes achieve and even surpass human-level precision, the required labor-intensive and costly manual annotation and delineation procedures hinder the utilization of such methods in clinical practice. In addition, any annotated and delineated data set used for supervised learning is subject to observer variability, limiting its practical generalization beyond the experimental setting.

On the other hand, in unsupervised learning, the goal is to identify and infer salient patterns and underlying structure from the input, without a supervisor (ie, the experienced pathologist) to provide a correct answer. In other words, unsupervised learning attempts to transform the complexity of data into low-dimensional spaces that capture prominent information about the histopathologic findings. Given that there is no straightforward measure of success in unsupervised learning, it is not easy to validate the unsupervised predictions. Clustering and anomaly detection,[34] data compression, and dimensionality reduction techniques are examples of the utilization of unsupervised learning in pathology.[35]

Semi-supervised learning falls between supervised and unsupervised learning, as it combines a small amount of labeled data with a large amount of unlabeled data. Similarly, semi-supervised learning benefits from the advantages of both supervised and unsupervised learning. The labeled data are used for identifying specific groups present in the data. In contrast, the unlabeled data are used for boundary definition between the latter groups and to help identify unspecified data. Similar to supervised and unsupervised learning, semi-supervised learning is also used as a tool for representation learning in pathology.[36] However, in case of any labeled data use, observer variability biases and limits learning scope, regardless of the degree of supervision.

Considering central challenges in computational pathology, such as the insufficient amount of labeled data, ground-truth errors due to observer variability, and weak data labeling due to expensive annotation and delineation processes and privacy and confidentiality concerns, semi-supervised and unsupervised learning methods hold more promise for the future. It is still not known exactly how unsupervised learning can be employed to overcome observer variability for building consensus, but it is more aligned with the philosophy of "strong AI" to be rather free from supervision. Both paradigms may be subject to biases of the AI expert/developer who selects a specific algorithm—supervised or not—to perform the desired task. However, user involvement would mostly put a cap on the performance and not introduce a new source of inconsistency, as computer algorithms are generally consistent.

Image Search

Advances in AI research have shown great promise in assisting health care professionals. However, innovative

Table 1 Commonly Used Deep Networks for Extracting Features for Image Search and Retrieval

Network	Training	Network size
DenseNet-121[39,44–46]	Pre-trained	7 million weights
ResNet-50[47,48]	Pre-trained	24 million weights
NasNet-A-Large[49]	Pre-trained	85 million weights
Fully Connected Network[45]	Fine-tuned using Motic and CAMELYON16 data sets (separate experiments)	1 million weights
Customized 16-layer CNN[50]	Trained using 5256 skeletal muscle images from MCWNL and 2904 lung cancer images from TCGA	3 million weights
Deep Ranking Model[40]	Trained on 500,000,000 natural images from 18,000 distinct classes.	Unavailable
Graph Neural Networks[44,46]	Trained using ACDC-LungHP in [B] and TCGA in [H]	1 million weights

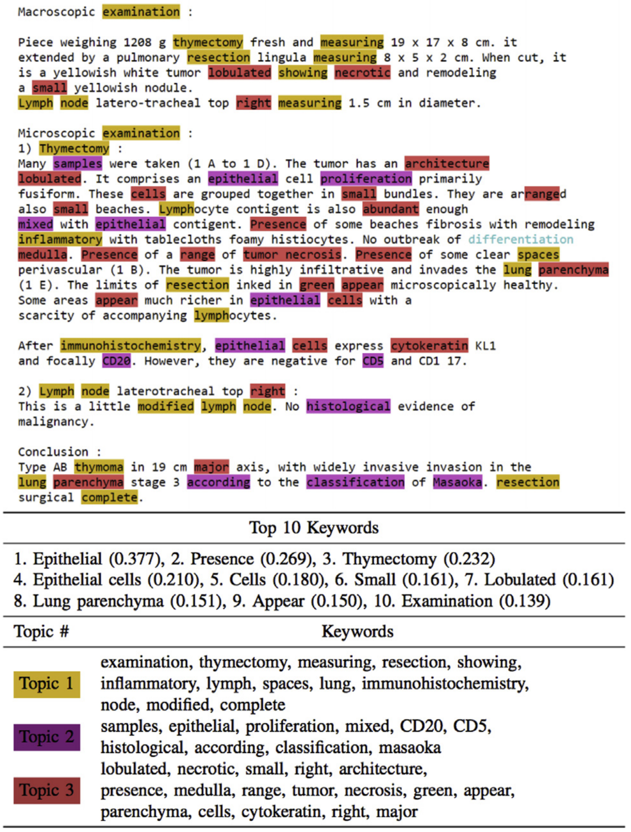
ACDC, Automatic Cancer Detection and Classification; CAMELYON, Cancer Metastases in Lymph Nodes Challenge; CNN, convolutional neural networks; MCWNL, Medical College of Wisconsin Neuromuscular Laboratory; TCGA, The Cancer Genome Atlas.

algorithms with reliable performance are necessary for gaining trust and adoption in clinical settings. Histopathology is the gold standard for the diagnosis of many diseases such as cancer, inflammations, and infections. With the widespread adoption of digital pathology, histopathology can greatly benefit from the applications of AI. When pathologists diagnose a difficult case, or when a new pathologist is in training, the identification and description of histologic features in images may become a common source of uncertainty. The conventional solutions are for pathologists to ask colleagues or to laboriously browse reference textbooks, hoping to find an image with similar visual characteristics. The general computer vision solution to similar problems is *content-based image retrieval* (CBIR), a research field with almost three decades of research.

CBIR systems use a search engine to retrieve similar images when the pathologist provides a query image, instead of using text for search in an archive. CBIR systems of medical images have been well researched.[37] Only with the emergence of digital pathology and deep learning, research has begun to focus on image search and analysis for histopathological images.[38–41] There are two major hurdles in the extensive usage of CBIR systems for digital pathology. First, pathology images exhibit an intractable level of variability in visual features which makes their identification, compared with natural images, much more challenging. The computational representation of histopathologic images for search purposes requires capturing the sematic high-level patterns of whole slide images (WSIs). Second, WSIs are gigapixel images of huge dimensionality (ie, larger than 50,000 × 50,000 pixels). Most research is focused on resolving these issues to make CBIR systems feasible for use in digital pathology.

Recent digital pathology studies have reported the success of supervised AI algorithms for use in classification and segmentation.[42] Compared to other AI algorithms, this success is related to the ease of design and in-laboratory validation in generating highly accurate results. However, compared to other methods of computer-vision algorithms, CBIR offers a new approach to computational pathology. To facilitate image search, CBIR algorithms essentially describe the image content with nontextual attributes, generally with a vector of real numbers known as a *feature vector*. A set of AI algorithms can be trained to transform an image into a feature vector to serve as its representation. If a feature vector encompasses the descriptive visual properties of an image, then searching for similar images becomes a "nearest-neighbor" problem, that is, a matching task to find similar histologic features. Images with similar content can be retrieved based on a comparison of their feature vectors and not on the associated textual metadata or manual delineations subject to observer variability, and often come from a small number of pathologists. Image comparison is generally possible if a feature vector encodes the semantic structures of an image invariant to scale, rotation, translation, and deformation to some degree. Such rich and descriptive features, identified in an unsupervised manner, represent images for comparison and matching, which is the core task of any CBIR system.

Recently, the search engine Yottixel was proposed, enabling image search in large archives of histopathology images at both the patch/tile and WSI levels.[39] The underlying technology behind Yottixel consists of a series of unsupervised AI algorithms, including clustering techniques, deep networks, and gradient barcoding. Yottixel indexes a WSI by converting tissue patterns into a set of barcodes, a process that is both storage friendly and computationally efficient. Every WSI is then represented as a *bunch of barcodes* (BoB), and thereby the image search can be performed by simpler and fast binary matching. After a large-scale validation of the search engine, the Yottixel results demonstrated that the image archives size and diversity play a role in image search.[38] The largest public repository of WSIs, The Cancer Genome Atlas[43] (provided by the National Cancer Institute/NIH) has been used for a



```
Macroscopic examination :

Piece weighing 1208 g thymectomy fresh and measuring 19 x 17 x 8 cm. it
extended by a pulmonary resection lingula measuring 8 x 5 x 2 cm. When cut, it
is a yellowish white tumor lobulated showing necrotic and remodeling
a small yellowish nodule.
Lymph node latero-tracheal top right measuring 1.5 cm in diameter.

Microscopic examination :
1) Thymectomy :
Many samples were taken (1 A to 1 D). The tumor has an architecture
lobulated. It comprises an epithelial cell proliferation primarily
fusiform. These cells are grouped together in small bundles. They are arranged
also small beaches. Lymphocyte contigent is also abundant enough
mixed with epithelial contigent. Presence of some beaches fibrosis with remodeling
inflammatory with tablecloths foamy histiocytes. No outbreak of differentiation
medulla. Presence of a range of tumor necrosis. Presence of some clear spaces
perivascular (1 B). The tumor is highly infiltrative and invades the lung parenchyma
(1 E). The limits of resection inked in green appear microscopically healthy.
Some areas appear much richer in epithelial cells with a
scarcity of accompanying lymphocytes.

After immunohistochemistry, epithelial cells express cytokeratin KL1
and focally CD20. However, they are negative for CD5 and CD1 17.

2) Lymph node laterotracheal top right :
This is a little modified lymph node. No histological evidence of
malignancy.

Conclusion :
Type AB thymoma in 19 cm major axis, with widely invasive invasion in the
lung parenchyma stage 3 according to the classification of Masaoka. resection
surgical complete.
```

Top 10 Keywords
1. Epithelial (0.377), 2. Presence (0.269), 3. Thymectomy (0.232) 4. Epithelial cells (0.210), 5. Cells (0.180), 6. Small (0.161), 7. Lobulated (0.161) 8. Lung parenchyma (0.151), 9. Appear (0.150), 10. Examination (0.139)

Topic #	Keywords
Topic 1	examination, thymectomy, measuring, resection, showing, inflammatory, lymph, spaces, lung, immunohistochemistry, node, modified, complete
Topic 2	samples, epithelial, proliferation, mixed, CD20, CD5, histological, according, classification, masaoka
Topic 3	lobulated, necrotic, small, right, architecture, presence, medulla, range, tumor, necrosis, green, appear, parenchyma, cells, cytokeratin, right, major

Figure 1 The top 50 key words in a report were identified using simple natural language processing methods. The key words are color-coded as per the abstract "topics." Each topic is given a separate color scheme.[71]

first validation. Almost 30,000 WSI files of 25 primary anatomic sites and 32 cancer subtypes were processed by dismantling of the large slides into almost 20 million image patches/tiles that were then individually indexed by employing approximately 3 million barcodes. The validation results show that image search can indeed provide a base for computational consensus. The primary diagnoses of evidently diagnosed cases can be used after they are identified by the search engine and listed as the best matches for the query WSI to decide based on the majority vote. Finding similar images opens the way for the development of new approaches to resolving observer variability through what might be called *virtual peer review*: Image search provides access to the reports of evidently diagnosed cases with similar histopathologic features.

The Google team recently published another search tool for use in digital pathology, *similar image search for histopathology* (SMILY).[40] SMILY uses a deep-learning model, trained using 500 million natural, nonpathology images (dogs, trees, human-made objects, etc.), to compress images into a feature vector. During the training process, the network learned to distinguish similar images from dissimilar ones by computing and to compare their embeddings. This model is then used for generating a database of image patches and their associated features using a subset of WSIs from The Cancer Genome Atlas data set. SMILY uses a dense sampling, which may be advantageous in some situations, such as mitotic counting or finding some specific lesion within the same WSI, but is quite computationally expensive.

Any search framework depends on robust and expressive image representation, which are called *deep features*. Hence, extracting these deep features is the main step toward capable image search engines. Table 1 provides an overview of the most commonly used deep networks for this purpose.

One should bear in mind that the purpose of the image search as an unsupervised framework is not to get pathologists to agree on individual case diagnoses through interactions with each other (although this may very well be a possible and promising venue to investigate). Rather, the goal of any computer-driven solution is to maximize diagnostic accuracy. Whereas the high accuracy of supervised AI is subject to limited generalization due to the scarcity and variability-prone nature of labels, it is expected that unsupervised AI offers a more reliable path toward high accuracy by acting as a mediator for building computational consensus using evidently diagnosed cases from large-scale archives.

NLP and Pathology

Using a sophisticated CBIR system to search for similar histopathologic features in a large-scale archive may be an impressive task demonstrating the human-like ability of AI to identify images. However, it is of little value if the software shows the pathologist only similar images retrieved. The histologic matching results contribute to consensus and higher accuracy only if pathology reports and treatment outcomes accompany retrieved images of evidently diagnosed cases. Such information may not be stored where images are stored but elsewhere (eg, in laboratory information systems). These metadata are generally available in the form of text documents. Another set of AI algorithms, *natural language processing* (NLP), is required for processing the textual metadata.

Nowadays, NLP is no longer about translating or interpreting text or speech based on some key words, but about understanding its meaning. The first steps in this path were taken by using convolutional neural networks,[51–53] recurrent neural networks,[54,55] graph-based neural networks,[56–58] and attention models.[59,60] Using these deep neural models helps resolve the handcrafted-features problem by making this step automated, which is necessary for eliminating another aspect of observer variability in AI research. These features are learned in specific NLP tasks. Despite all of these improvements, the data used for training models were lacking, which forced networks to use shallower architectures (ie, smaller number of weights to adjust) to avoid the necessity of large data sets.

Recent work has shown that pre-trained models on a large amount of text data can be beneficial for many types of downstream NLP tasks. Deep models (ie, transformers[60]) alongside improvement in training skills have helped to generate a better pre-trained model. In this context, *pre-trained* means that an NLP method is trained with nonmedical text (abundantly available) and then used for medical cases, perhaps with a reduced need for a large-scale archive of clinical text, such as pathology reports.

The first generation of these pre-trained models, such as Skip-Gram and Glove, tried to capture the semantic meanings of words. Despite some advantages, these models failed to capture higher-level concepts in a text, such as disambiguation and syntactic structures. In the second-generation pre-trained models, such as embeddings from language models (ELMo),[61] generative pre-trained transformer (OpenAI GPT),[62] bidirectional encoder representations from transformers (BERT),[63] OpenAI GPT-2,[64] language models (Megatron-LM)[65] and OpenAI GPT-3,[66] researchers tried to overcome this problem by using learning of contextual word embeddings. Although these models have mostly been used for general domain text, of late they have been increasingly used for the biomedical domain. For instance, ClinicalBert[67] trained the BERT model with clinical records to predict the probability of patient readmission. Reports and clinical notes contain valuable information that can help unsupervised search to offer a baseline for consensus.

BioBert[68] attempted to fine-tune the BERT model with varied medical corpora types such as PubMed abstracts, PubMed Central full-text articles, and other general corpora to be used in downstream NLP tasks such as entity recognition, relation extraction, and image search. Si et al [69] compared different NLP methods for clinical concept extractors. Text and image are two significant resources of data. Therefore, many methods use these two types of data together. Image captioning and visual question answering are two examples. This combination can have far-reaching effects in digital pathology. MDNet[70] is one of these AI models that can generate a pathology report for image patches that have been retrieved by symptoms descriptions and visual attention to provide justification for a diagnosis.

Using NLP models in a specific domain can be challenging. The medical domain, especially pathology, is one of them. Pathologists can use different words to describe the same observation or use rare words, making it difficult to train an AI model to represent the images correctly. Keyword and topic selection are readily available for pathology reports (Figure 1). Moreover, a lack of a clean, large-scale, and universal data set for this domain is another challenge in using NLP methods for digital pathology.

The availability of NLP systems like BioBERT and MDNet is quite promising. Assuming that a pathologist is looking at the retrieved images through a capable CBIR system, the NLP can be applied to retrieved metadata of evidently diagnosed patients, that is, reports, to build an amalgamated caption or summarized description of the input image for the undiagnosed patient. This would be a computational consensus that can reduce observer variability if trusted by the pathologist.

Summary and Conclusion

Observer variability is not only difficult to quantify but also a major challenge in establishing diagnostic consensus. The progress and the astonishing success of AI in recent years provides a new paradigm for addressing this challenge. Unsupervised AI methods may be proven to be the solution if combined with the existing evidence in both images and reports in hospitals and laboratories. However, the unsupervised linkage of images and reports is impeded by a lack of access by the research community to large-scale clinical archives for exploiting the potentials of both image search and natural language processing.

References

1. Weydert JA, De Young BR, Cohen MB: A preliminary diagnosis service provides prospective blinded dual-review of all general surgical pathology cases in an academic practice. Am J Surg Pathol 2005, 29:801−805
2. Weir MM, Jan E, Colgan TJ: Interinstitutional pathology consultations: a reassessment. Am J Clin Pathol 2003, 120:405−412
3. Renshaw AA, Gould EW: Comparison of disagreement and amendment rates by tissue type and diagnosis: identifying cases for directed blinded review. Am J Clin Pathol 2006, 126:736−739
4. Tomaszewski JE, LiVolsi VA: Mandatory second opinion of pathologic slides: is it necessary? Cancer 1999, 86:2198−2200
5. Raub SS, Nakhleh RE, Ruby SG: Patient safety in anatomic pathology: measuring discrepancy frequencies and causes. Arch Pathol Lab Med 2005, 129:459−466
6. Renshaw AA, Gould EW: Correlation of workload with disagreement and amendment rates in surgical pathology and nongynecologic cytology. Am J Clin Pathol 2005, 125:820−822
7. Renshaw AA, Pinnar NE, Jiroutek MR, Young ML: Quantifying the value of in-house consultation in surgical pathology. Am J Clin Pathol 2002, 117:751−754
8. Renshaw AA, Gould EW: Measuring errors in surgical pathology in real-life practice: defining what does and does not matter. Am J Clin Pathol 2007, 127:144−152
9. Valenstein PN, Raub SS, Walsh MK; College of American Pathologists: Identification errors involving clinical laboratories: a College of American Pathologists Q-Probes study of patient and specimen identification errors at 120 institutions. Arch Pathol Lab Med 2006, 130
10. Djuric U, Zadeh G, Aldape K, Diamandis P: Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis Oncol 2017, 1:22
11. Schulte MA: Distraction index, part II: inflammation in nongynecologic cytology. Diagn Cytopathology 2000, 23:149−150
12. Neal MH, Kline TS: Distraction index, part I: the elusive trich. Diagn Cytopathol 1999, 21:367−369
13. Faust K, Roohi A, Leon AJ, Leroux E, Dent A, Evans AJ, Pugh TJ, Kalimuthu SN, Djuric U, Diamandis P: Unsupervised resolution of histomorphologic heterogeneity in renal cell carcinoma using a brain

tumor−educated neural network. JCO Clin Cancer Inform 2020, 4: 811−821

14. Steensma DP: Myelodysplastic syndromes current treatment algorithm 2018. Blood Cancer J 2018, 8:47
15. Arber DA, Hasserjian RP: Reclassifying myelodysplastic syndromes: what's where in the new WHO and why. Hematology 2015, 2015: 294−298
16. Della Porta MG, Travaglino E, Boveri E, Ponzoni M, Malcovati L, Papaemmanuil E, M Rigolin G, Pascutto C, Croci G, Gianelli U, Milani R, Ambaglio I, Elena C, Ubezio M, C Da Via' M, Bono E, Pietra D, Quaglia F, Bastia R, Ferretti V, Cuneo A, Morra E, J Campbell P, Orazi A, Invernizzi R, Cazzola M: Minimal morphological criteria for defining bone marrow dysplasia: a basis for clinical implementation of WHO classification of myelodysplastic syndromes. Leukemia 2015, 29:66
17. Sasada K, Yamamoto N, Masuda H, Tanaka Y, Ishihara A, Takamatsu Y, Yatomi Y, Katsuda W, Sato I, Matsui H: Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome. Leuk Res 2018, 69:54−59
18. Font P, Loscertales J, Benavente C, Bermejo A, Callejas M, Garcia-Alonso L, Garcia-Marcilla A, Gil S, Lopez-Rubio M, Martin E, Muñoz C: Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification. Ann Hematol 2013, 92:19−24
19. Font P, Loscertales J, Soto C, Ricard P, Muñoz-Novas C, Martín-Clavero E, López-Rubio M, Garcia-Alonso L, Callejas M, Bermejo A, Benavente C: Interobserver variance in myelodysplastic syndromes with less than 5 % bone marrow blasts: unilineage vs. multilineage dysplasia and reproducibility of the threshold of 2% blasts. Ann Hematol 2015, 94:565−573
20. Goasguen JE, Bennett JM, Bain BJ, Brunning R, Vallespi M-T, Tomonaga M, Zini G, Renault A: Proposal for refining the definition of dysgranulopoiesis in acute myeloid leukemia and myelodysplastic syndromes. Leuk Res 2014, 38:447−453
21. Naqvi K, Jabbour E, Bueso-Ramos C, Pierce S, Borthakur G, Estrov Z, Ravandi F, Faderl S, Kantarjian H, Garcia-Manero G: Implications of discrepancy in morphologic diagnosis of myelodysplastic syndrome between referral and tertiary care centers. Blood 2011, 118:4690−4693
22. Senent L, Arenillas L, Luño E, Ruiz JC, Sanz G, Florensa L: Reproducibility of the World Health Organization 2008 criteria for myelodysplastic syndromes. Haematologica 2013, 98:568−575
23. Parmentier S, Schetelig J, Lorenz K, Kramer M, Ireland R, Schuler U, Ordemann R: Assessment of dysplastic hematopoiesis: lessons from healthy bone marrow donors. Haematologica 2012, 97:723−730
24. Chang HY, Kwon Jung C, Isaac Woo J, Lee S, Cho J, Kim SW, Kwak T-Y: Artificial intelligence in pathology. J Pathol Transl Med 2018, 53:1−12
25. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A: Artificial intelligence in digital pathology–new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019, 16:703−715
26. Tizhoosh HR, Pantanowitz L: Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 2018, 9: 38
27. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018, 24:1559−1567
28. Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghighi M, Heng C: Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nat Methods 2019, 16: 1247−1253
29. Ludovic R, Racoceanu D, Loménie N, Kulikova M, Irshad H, Klossa J, Capron F, Genestie C, Le Naour G, Gurcan MN: Mitosis detection in breast cancer histological images: an ICPR 2012 contest. J Pathol Inform 2013, 4:8
30. Qaiser T, Mukherjee A, Reddy Pb C, Munugoti SD, Tallam V, Pitkäaho T, Lehtimäki T: HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 2018, 72:227−238
31. Shaban M, Awan R, Fraz MM, Azam A, Tsang Y-W, Snead D, Rajpoot NM: Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans Med Imaging 2020, 39:2395−2405
32. Tabibu S, Vinod PK, Jawahar CV: Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. Sci Rep 2019, 9:10509
33. Yu K-H, Zhang CE, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016, 7:1−10
34. Faust K, Xie Q, Han D, Goyle K, Volynskaya Z, Djuric U, Diamandis P: Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC Bioinformatics 2018, 19:173
35. Hu B, Tang Y, Chang EIC, Fan Y, Lai M, Xu Y: Unsupervised Learning for cell-level visual representation in histopathology images with generative adversarial networks. IEEE journal of biomedical and health informatics., 2019, 1316−1328.
36. Sparks R, Madabhushi A: Out-of-sample extrapolation utilizing semi-supervised manifold learning (OSE-SSL): content based image retrieval for histopathology images. Sci Rep 2016, 6:1−15
37. Muller H, Michoux N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. Int J Med Inform 2004, 73: 1−23
38. Kalra S, Tizhoosh HR, Shah S, Choi C, Damaskinos S, Safarpoor A, Shafiei S, Babaie M, Diamandis P, Campbell CJ, Pantanowitz L: Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ digital medicine 2020, 3:1−15
39. Kalra S, Tizhoosh H, Choi C, Shah S, Diamandis P, Campbell CJ, Pantanowitz L: Yottixel–an image search engine for large archives of histopathology whole slide images. Med Image Analysis 2020, 65: 101757
40. Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D, Terry M, Cai CJ, Amin MB, Mermel CH, Nelson PQ: Similar image search for histopathology: SMILY. NPJ digital medicine 2019, 2:1−9
41. Shi X, Xing F, Xu K, Xie Y, Su H, Yang L: Supervised graph hashing for histopathology image retrieval and classification. Med Image Analysis 2017, 42:117−128
42. Komura D, Ishikawa S: Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 2018, 16:34−42
43. Tomczak K, Czerwi nska P, Wiznerowicz M: The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015, 19:A68
44. Zheng Y, Jiang B, Shi J, Zhang H, Xie F: Encoding histopathological WSIs using GNN for scalable diagnostically relevant regions retrieval. International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 (MICCAI 2019). Cham, Switzerland: Springer; 2019
45. Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, Zhao Y: Histopathological whole slide image analysis using context-based CBIR. IEEE Trans Med Imaging 2018, 37:1641−1652
46. Adnan M, Kalra S, Tizhoosh HR: Representation learning of histopathology images using graph neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020 June 14-19, Seattle, WA IEEE/CVF; 2020. pp. 988−989

47. Peng T, Boxberg M, Weichert W, Navab N, Marr C: Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval. International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 (MICCAI 2019) 2019 October 13-17, Shenzhen, China. Springer, Cham, Switzerland; 2019
48. Cheng S, Wang L, Du A: Histopathological image retrieval based on asymmetric residual hash and DNA coding. IEEE Access 2019, 7: 101388−101400
49. Komura D, Kawabe A, Fukuta K, Sano K, Umezaki T, Koda H, Suzuki R: Deep texture representations as a universal encoder for pan-cancer histology. bioRxiv 2020. [Preprint] doi:10.1101/2020.07.28.224253
50. Shi X, Sapkota M, Xing F, Liu F, Cui L, Yang L: Pairwise based deep ranking hashing for histopathology image classification and retrieval. Pattern Recognition 2018, 81:14−22
51. Kalchbrenner N, Grefenstette E, Blunsom P: A convolutional neural network for modelling sentences. arXiv 2014. [Preprint] doi:1404.2188
52. Kim Y: Convolutional neural networks for sentence classification. arXiv 2014. [Preprint] doi:1408.5882
53. Garg S, Vu T, Moschitti A: Tanda: transfer and adapt pre-trained transformer models for answer sentence selection. arXiv 2019. [Preprint] doi:1911.04118
54. Liu P, Qiu X, Huang X: Recurrent neural network for text classification with multi task learning. arXiv 2016. [Preprint] doi:1605.05101
55. Sutskever I, Vinyals O, Le QV: Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014 December 8-13, Montreal, QC, Canada. Curran Associates. 2014. pp. 3104−3112
56. Marcheggiani D, Bastings J, Titov I: Exploiting semantics in neural machine translation with graph convolutional networks. arXiv 2018. [Preprint] doi:1804.08313
57. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C: Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013 October. Seattle, WA, Association for Computational Linguistics, 2013. pp. 1631−1642
58. Tai KS, Socher R, Manning CD: Improved semantic representations from tree structured long short-term memory networks. arXiv 2015. [Preprint] doi:1503.00075
59. Bahdanau D, Cho K, Bengio Y: Neural machine translation by jointly learning to align and translate. arXiv 2014. [Preprint] doi:1409.0473
60. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: Attention is all you need. Advances in Neural Information Processing Systems (NIPS 2017). 2017 December 4-9. Long Beach, CA. Curran Associates, 2017. pp. 5998−6008
61. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L: Deep contextualized word representations. arXiv 2018. [Preprint] doi:1802.05365
62. Radford A, Narasimhan K, Salimans T, Sutskever I: Improving language understanding by generative pre-training. OpenAI, 2018:12
63. Devlin J, Chang M-W, Lee K, Toutanova K: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018. [Preprint] doi:1810.04805
64. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I: Language models are unsupervised multitask learners. OpenAI, 2019:8
65. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B: Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv 2019. [Preprint] doi:1909.08053
66. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Grueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D: Language models are few-shot learners. arXiv 2020. [Preprint] doi:2005.14165
67. Huang K, Altosaar J, Ranganath R: Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv 2019. [Preprint] doi:1904.05342
68. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020, 36: 1234−1240
69. Si Y, Wang J, Xu H, Roberts K: Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019, 26: 1297−1304
70. Zhang Z, Xie Y, Xing F, McGough M, Yang L: Mdnet: a semantically and visually interpretable medical image diagnosis network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017 July 21-26. Honolulu, HI IEEE Computer Society, 2017. pp. 6428−6436
71. Kalra S, Li L, Tizhoosh HR: Automatic classification of pathology reports using TF-IDF features. arXiv 2019. [Preprint] doi:1903.07406