Large-Scale Annotation of Histopathology Images from Social Media

Andrew J. Schaumberg[a,b,c,1,*], Wendy Juarez[c,d,α], Sarah J. Choudhury[c,d,α], Laura G. Pastrián MD[e,β], Bobbi S. Pritt MD DTM&H[f,β], Mario Prieto Pozuelo MD PhD[g,β], Ricardo Sotillo Sánchez MD[h,β], Khanh Ho MD[i,β], Nusrat Zahra MD[j,β], Betul Duygu Sener MD[k,β], Stephen Yip MD PhD[l,β], Bin Xu MD PhD[m,β], Srinivas Rao Annavarapu MD[n,β], Aurélien Morini MD[o,β], Karra A. Jones MD PhD[p,β], Kathia Rosado-Orozco MD[q,β], S. Joseph Sirintrapun MD[r], Mariam Aly PhD[s,2,δ,*], and Thomas J. Fuchs Dr.Sc[b,r,3,δ,*]

[α] *Equal contribution*
[β] *Generously donated pathology cases*
[δ] *Principal Investigator*
[*] *Correspondence*
[a] *Memorial Sloan Kettering Cancer Center and the Tri-Institutional Training Program in Computational Biology and Medicine, NY, USA*
[b] *Weill Cornell Graduate School of Medical Sciences, NY, USA*
[c] *Weill Cornell High School Science Immersion Program*
[d] *Manhattan/Hunter Science High School, NY, USA*
[e] *Hospital Universitario La Paz, Madrid, Spain*
[f] *Mayo Clinic, Department of Laboratory Medicine and Pathology, MN, USA*
[g] *Hospital Universitario HM Sanchinarro, Laboratorio de Dianas Terapéuticas, Madrid, Spain*
[h] *Virgen de Altagracia Hospital, Manzanares, Spain*
[i] *Centre Hospitalier de Mouscron, Belgium*
[j] *Allama Iqbal Medical College, Lahore, Pakistan*
[k] *Konya Training and Research Hospital, Konya, Turkey*
[l] *BC Cancer, British Columbia, Canada*
[m] *Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada*
[n] *Royal Victoria Infirmary, Department of Cellular Pathology, England, UK*
[o] *Université Paris Est Créteil, Faculté de médecine de Créteil, France*
[p] *University of Iowa, Department of Pathology, IA, USA*
[q] *HRP Labs, San Juan, Puerto Rico, USA*
[r] *Memorial Sloan Kettering Cancer Center, Department of Pathology, NY, USA*
[s] *Columbia University, Department of Psychology, NY, USA*
[1] `ajs625@cornell.edu orcid:0000-0001-7556-9208`
[2] `ma3631@columbia.edu orcid:0000-0003-4033-6134`
[3] `fuchst@mskcc.org orcid:0000-0001-7603-8687`

August 21, 2018

Abstract

Large-scale annotated image datasets like ImageNet and CIFAR-10 have been essential in developing and testing sophisticated new machine learning algorithms for natural vision tasks. Such datasets allow the development of neural networks to make visual discriminations that are done by humans in everyday

*Respective contributions. Conceptualization: AJS, MA. Methodology, Software, Validation, Formal analysis, Investigation, Writing original draft: AJS. Resources (pathology): LGP, BSP, MPP, RSS, KH, NZ, BDS, SY, BX, SRA, AM, KAJ, KRO. Resources (computational): AJS. Data curation: AJS, WJ, SJC, LGP, BSP, MPP, NZ, BDS, SY. Writing (reviewing): AJS, BSP, MPP, SY, AM, SJS, MA, TJF. Writing (editing): AJS, MA. Visualization, wrote annotation files: AJS, WJ, SJC. Answered annotator questions: LGP, BSP, MPP, NZ, BDS, SY, SJS. Supervision: MA, TJF. Project administration: AJS, WJ, SJC, MA, TJF. Funding acquisition: AJS, TJF.

activities, e.g. discriminating classes of vehicles. An emerging field – computational pathology – applies such machine learning algorithms to the highly specialized vision task of diagnosing cancer or other diseases from pathology images. Importantly, labeling pathology images requires pathologists who have had decades of training, but due to the demands on pathologists' time (e.g. clinical service) obtaining a large annotated dataset of pathology images for supervised learning is difficult. To facilitate advances in computational pathology, on a scale similar to advances obtained in natural vision tasks using ImageNet, we leverage the power of social media. Pathologists worldwide share annotated pathology images on Twitter, which together provide thousands of diverse pathology images spanning many sub-disciplines. From Twitter, we assembled a dataset of 2,746 images from 1,576 tweets from 13 pathologists from 8 countries; each message includes both images and text commentary. To demonstrate the utility of these data for computational pathology, we apply machine learning to our new dataset to test whether we can accurately identify different stains and discriminate between different tissues. Using a Random Forest, we report (i) 0.959 ± 0.013 Area Under Receiver Operating Characteristic [AUROC] when identifying single-panel human hematoxylin and eosin [H&E] stained slides that are not overdrawn and (ii) 0.996 ± 0.004 AUROC when distinguishing H&E from immunohistochemistry [IHC] stained microscopy images. Moreover, we distinguish all pairs of breast, dermatological, gastrointestinal, genitourinary, and gynecological [gyn] pathology tissue types, with mean AUROC for any pairwise comparison ranging from 0.771 to 0.879. This range is 0.815 to 0.879 if gyn is excluded. We report 0.815 ± 0.054 AUROC when all five tissue types are considered in a single multiclass classification task. Our goal is to make this large-scale annotated dataset publicly available for researchers worldwide to develop, test, and compare their machine learning methods, an important step to advancing the field of computational pathology.

1 Introduction

Supervised learning requires annotated data. ImageNet [7] has millions of human-labeled images; CIFAR-10 [12] [Canadian Institute for Advanced Research] has thousands. Machine learning methods for natural vision tasks routinely use datasets like these to benchmark performance, as well as transfer learned representations to other tasks, such as pathology [3, 14, 23]. However, computational pathology [8] datasets that are annotated for supervised learning are often much smaller, because obtaining annotations from a pathologist is difficult. For example, there are only 32 cases in the training data for a MICCAI challenge for distinguishing brain cancer subtypes, and this includes both pathology and radiology images[1]. Other datasets, such as the one for TUPAC16 [TUmor Proliferation Assessment Challenge], involve hundreds of cases [26]. The Cancer Genome Atlas [TCGA][2] has tens of thousands of cases, with molecular and whole slide images available, but these images are only hematoxylin and eosin [H&E] stained slides.

To overcome the main limitation of developing a pathology dataset on the scale of ImageNet or CIFAR-10 – the availability of pathologists to annotate images – we leverage the power of social media. Pathologists worldwide voluntarily use social medial platforms (e.g., Twitter) to share annotated cases [6, 16]. These cases constitute a diverse, large-scale pathology dataset, which, if curated, can be used by computational pathologists all over the world to develop their machine learning techniques. We have developed such a dataset, which includes a variety of sections and techniques, ranging from immunohistochemistry [IHC] to fluorescence *in situ* hybridization [FISH], and a range of tissues, along with linked annotations by pathologists.

Here, we describe our steps in assembling and analyzing cases shared by pathologists on social media.

[1]This MICCAI [Medical Image Computing and Computer Assisted Intervention] challenge is `http://miccai.cloudapp.net/competitions/82`

[2]TCGA available at http://cancergenome.nih.gov/

Figure 1: Thirteen pathologists over eight countries generously donated cases for our study. They also answered questions that arose during manual case annotation procedures (Table 1).

Step	Purpose	Description
1.	Find pathologist	We find pathologists who share many or under-represented pathology cases.
2.	Obtain consent	Pathologist consents to have their images included in a public database.
3.	Download data	We use custom bots and scripts to obtain the pathologist's cases.
4.	Annotate data	We write a text file to describe each case, based on its social media post, per Sec 2.3.1.
4.1.	Online question	We ask pathologists for clarification about posted cases (if needed), e.g. stain used.
4.2.	Local question	If the pathologist does not respond, we ask a local pathologist for help.
5.	Analyze all data	We aggregate data, perform machine learning, and measure performance.

Table 1: Details of each step of our pipeline.

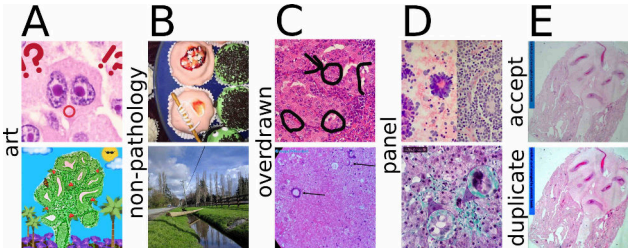

Figure 2: Examples of images that are rejected, because they are not pathology images that a pathologist would see in clinical practice. *Panel A (top M.P.P., bottom B.D.S)*: "art" rejects. *Panel B (top B.S.P., bottom S.Y.)*: "non-pathology" rejects. *Panel C (top B.X., bottom A.M.)*: "overdrawn" rejects. *Panel D (top S.R.A., bottom L.G.P.)*: "panel" rejects. *Panel G (top and bottom S.R.A.)*: top is acceptable H&E, bottom is "dup" [duplicate] rejection.

This initial dataset includes images donated by 13 pathologists, from 13 institutions in 8 countries (Fig 1). We annotated 2,746 images from 1,576 tweets from 13 pathologists, with consent and help from pathologists. The message text and hashtags posted along with the images were treated as image annotations.

Our current work makes two novel contributions to the field of computational pathology: (1) we present the first study of pathology images and annotations shared on social media by pathologists, and (2) we demonstrate the utility of these data with a variety of machine learning analyses. These analyses include (i) predicting if an image is a human H&E-stained microscopy image or not, (ii) predicting if a microscopy image is H&E-stained or IHC-stained, and (iii) predicting the histopathology tissue type of an image. Our goal is to make this large-scale annotated dataset publicly available for researchers worldwide to develop, test, and compare their machine learning methods, an important step for advancing computational pathology.

2 Materials and Methods

2.1 General workflow

We follow the procedure outlined in Table 1 to obtain and analyze pathology data. In step 1, we find pathologists on social media (Twitter) who share many pathology cases, or share tissues that are infrequently shared, such as neuropathology. In step 2, we next contact the pathologist via social media and ask for permission to use their cases. In step 3, we use a social media bot and our custom scripts to download the pathologist's social media posts, including text and images. In step 4, we manually annotate these posted cases for acceptability (if overdrawn, corrupt, duplicate, multi-panel, art, or non-pathology rejecting per Fig 2), technique (gross section, H&E stain, IHC stain, Pap smear, FISH, or CT scan per Fig 3), species (human, plant, worm, fly, tick, *Loa loa*, or *Enterobius vermicularis* per Fig 4), and private status (personally identifiable pictures of adults or pictures of children). For more information, e.g. our definition of "overdrawn" or what is [not] pathology, see Sections 2.2 and 2.3.1. Step 4 involves clarifying cases that we have trouble annotating, e.g. if it is not clear what stain was used for the image. We first ask the pathologist who posted this case to social media (step 4.1). If we do not obtain an answer from that pathologist, we ask a pathologist at our local institution (i.e. S.J.S.) for an opinion (step 4.2). In step 5, we aggregate all data from all pathologists and apply machine learning to make predictions. This process was repeated as we identified more pathologists as potential collaborators (Fig 1). We aimed to have thousands of images available for a large-scale machine learning task, and with 13 pathologists we have over 2,000 images.

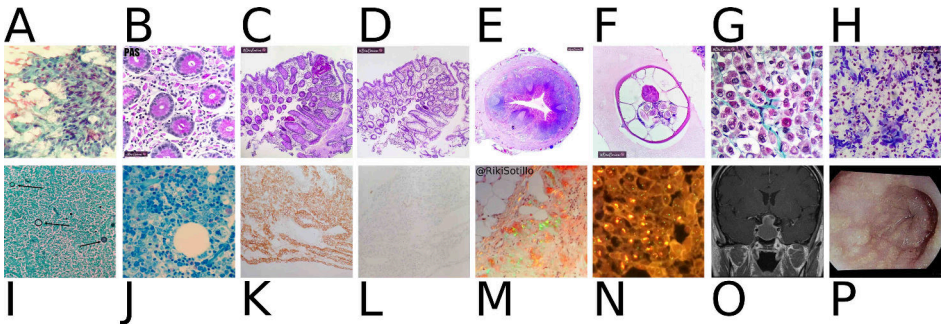

Figure 3: Our dataset includes diverse techniques. Initials indicate image ownership. *Panel A (R.S.S.)*: Papanicolaou stain, i.e. pap smear. *Panel B (L.G.P.)*: Periodic acid-Schiff [PAS] stain, glycogen in pink. *Panel C (L.G.P.)*: PAS stain at lower magnification. *Panel D (L.G.P.)*: Hematoxylin and eosin [H&E] stain, for comparison to Panel C. *Panel E (L.G.P.)*: H&E stain of human appendix, including a parasite, *Enterobius vermicularis*. *Panel F (L.G.P.)*: Higher magnification of *Enterobius vermicularis* in Panel E. *Panel G (L.G.P.)*: Gömöri trichrome, collagen in green. *Panel H (L.G.P.)*: Diff-quik stain. *Panel I (R.S.S.)*: GMS stain (see also Sec S1.1), fungi in black. *Panel J (M.P.P.)*: Giemsa stain. *Panel K (A.M.)*: Immunohistochemistry [IHC] stain, positive result. *Panel L (A.M.)*: IHC stain, negative result. *Panel M (R.S.S.)*: Congo red under polarized light, with plaques showing green birefringence. *Panel N (M.P.P.)*: Fluorescence *in situ* hybridization [FISH] indicating *Her2* heterogeneity in breast cancer. *Panel O (S.Y.)*: Head CT scan. *Panel P (L.G.P.)*: Esophageal endoscopy.

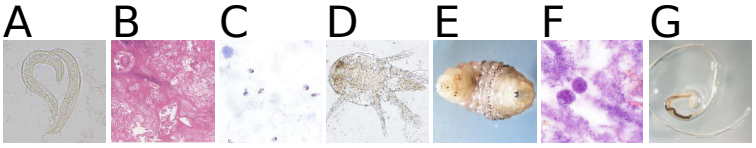

Figure 4: Our dataset includes diverse parasitology samples. *Panel A (B.S.P.)*: *Strongyloides stercoralis*, light microscopy. *Panel B (B.S.P.)*: *Dirofilaria immitis*, in human, H&E stain. *Panel C (B.S.P.)*: *Plasmodium falciparum*, in human, Giemsa stain. *Panel D (B.S.P.)*: Incidental finding of unspecified mite in human stool, light microscopy. *Panel E (B.S.P.)*: *Dermatobia hominis*, live gross specimen. *Panel F (B.S.P.)*: *Acanthamoeba*, in human, H&E of corrective contact lenses. *Panel G (B.S.P.)*: *Trichuris trichiura*, gross specimen.

2.2 Image data overview

The goal of obtaining images from practicing pathologists worldwide is to create a dataset with a diverse and realistic distribution of cases. A worldwide distribution of pathologists (Fig 1) may be appropriate to overcome potential biases inherent at any single institution, such as stain chemistries or protocols. Our dataset includes a wide variety of stains and techniques (Fig 3), and even variety within a single stain, e.g., the appearance of H&E stains (Fig 5). Section S1.1 discusses intra-stain variability further. Our dataset also includes gross sections (Fig 6) that pathologists share alongside images of stained slides (e.g., with H&E, IHC, etc). In addition to variation in the signal of interest (i.e., stain type or tissue type), we also find a great deal of variability in the noise, in terms of artifacts present in the pathology images (Fig 7). Such noise may initially seem undesirable, but is likely important for machine learning techniques to robustly predict which image motifs are relatively unimportant rather than prognostic. Finally, our dataset includes a variety of parasites (Fig 4) and other [micro]organisms, including tapeworms, pinworms, *Loa loa*, fly larvae, eggs, ticks, lice, fleas, malarial trophozoites, *Leishmania*, *Sarcina* (which is not a parasite), and septic *E. coli*.

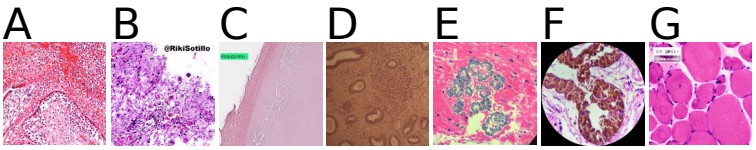

Figure 5: Our dataset includes diverse H&E-stained slide microscopy images. *Panel A (S.R.A)*: Acute villitis due to septic *Escherichia coli*. *Panel B (R.S.S.)*: Garlic. *Panel C (R.S.S.)*: "Accellular" leiomyoma after ulipristal acetate treatment. *Panel D (R.S.S.)*: Brownish appearance from dark lighting. *Panel E (K.R.O.)*: *Sarcina* in duodenum. *Panel F (B.D.S.)*: Mature teratoma of ovary, pigmented epithelium. *Panel G (K.A.J.)*: Central core myopathy.

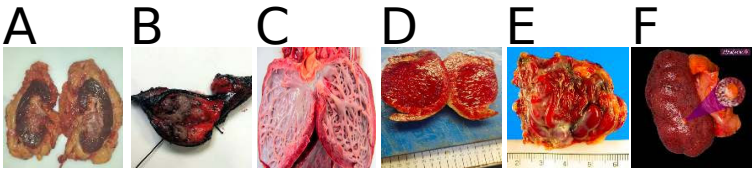

Figure 6: Gross sections are represented in our dataset, putting the slide images in context. *Panel A (M.P.P)*: Urothelial carcinoma. *Panel B (M.P.P.)*: Lung adenocarcinoma. *Panel C (S.R.A.)*: Barth syndrome. *Panel D (N.Z.)*: Enlarged spleen. *Panel E (S.R.A.)*: Arteriovenous malformation. *Panel F (L.G.P.)*: Kidney adrenal heterotopia.

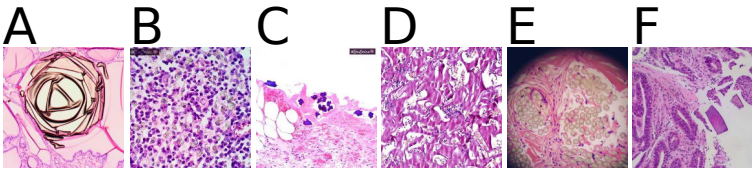

Figure 7: Our dataset includes artifacts and foreign bodies, which are generally unremarkable and should not be confused by machine learning methods as prognostic. All examples human H&E. *Panel A (B.X.)*: Colloid. *Panel B (L.G.P.)*: Barium. *Panel C (L.G.P.)*: Oxidized regenerated cellulose, a.k.a. gauze, granuloma may mimic mass lesion [25]. *Panel D (R.S.S.)*: Hemostatic gelatin sponge, a.k.a. Spongostan^{TM}, may mimic necrosis. *Panel E (S.Y.)*: Sutures, may mimic granuloma or adipocytes. *Panel F (L.G.P.)*: Crystallized kayexelate, may mimic mass lesion or parasite.

2.3 Defining an acceptable pathology image

To create our database, we first identified pathology images, and second, narrowed down the set of pathology images into those that were of sufficient quality to be used and could be shared publicly. By "pathology image", we mean images that a pathologist may see in clinical practice, e.g., gross sections, microscopy images, endoscopy images, CT scans, or X-rays. An image designated as a "pathology image" is not necessarily an image of diseased tissue. After we identified pathology images, we screened them for inclusion in our dataset. "Acceptable images" are those that do not meet rejection or discard criteria defined in the next section. If an acceptable image is personally identifiable or otherwise private (see criteria below), we retain the image for some machine learning analyses, but do not distribute the image publicly [for legal reasons].

2.3.1 Criteria for rejected, discarded, private, or acceptable images

For our manual data curation process, we defined several rejection criteria (Fig 2), detailed in Section S2.1. Figure 2A shows examples of images rejected as "art", because they are artistically manipulated H&E pathology microscopy images. Figure 2B shows examples of images rejected as "non-pathology", e.g. parasitology-inspired cupcakes (*top*) and a natural scene image (*bottom*). Non-pathology images are relatively common on pathologists' social media accounts, though we tried to minimize their frequency by searching for pathologists

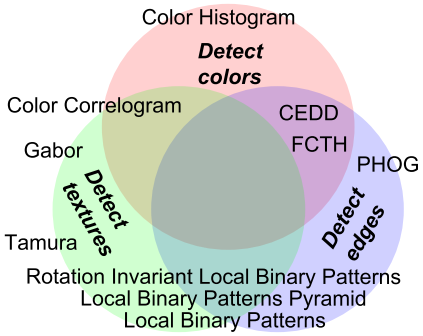

Figure 8: We use a variety of color, texture, and edge features for baseline machine learning analyses. Some features, such as color hsitograms, detect only color. Other features, such as Color Correlograms, detect both colors and textures. Pyramid features are scale-invariant.

who primarily used their accounts for sharing and discussing pathology. Figure 2C shows examples of images rejected as "overdrawn". Overdrawn images are those that have hand-drawn marks from a pathologist (which pathologists refer to as "annotations"), which prevent us from placing a sufficiently large bounding box around regions of interest while still excluding the hand-drawn marks. Section S2.2 discusses our "overdrawn" criterion in detail. Figure 2D shows examples of images rejected as "panel", because they consist of small panels (*top*) or have small insets (*bottom*); splitting multi-panel images into their constituent single-panel images would substantially increase our manual curation burden. Figure 2 Panel E *top* is an acceptable H&E-stained pathology image. Figure 2 Panel E *bottom* is rejected as a duplicate of the Panel E *top* image, though the colors have been slightly modified, and the original image is a different size.

2.3.2 Image features for machine learning

To perform baseline machine learning analyses on the images from social media, we first derive a feature representation for each image in the following manner. If a posted image is rectangular, we crop it to the center square and resize it to 512x512 pixels [px]. See Sec S2.3 for more discussion of the 512x512px image size and how it relates to the 256x256px image size for the "overdrawn" criterion. This 512x512px image is then converted to a feature vector of 2,412 dimensions. The features we use (Fig 8) are available in Apache LiRE [15]. These features, and their dimension counts, are as follows: CEDD (144) [4], Color Correlogram (256) [10], Color Histogram (64) [15], FCTH (192) [5], Gabor (60) [15], Local Binary Patterns (256) [17], Local Binary Patterns Pyramid (756) [18], PHOG (630) [2], Rotation Invariant Local Binary Patterns (36) [18], and Tamura (18) [24].

2.4 Text data overview

Prior work has discussed pathology-related hashtags as a way to make pathology more accessible on social media[3] [19]. Pathologists use hashtags to indicate histopathology tissue types, such as "#gynpath" to indicate gynecological pathology. Sometimes alternative spellings are used, such as "#ginpath". Abbreviations are also common, e.g. "#breastpath" and "#brstpath" all mean the same thing: breast pathology. Because a tweet can have more than one hashtag, we took the first tissue type hashtag to be the "primary" tissue type of the tweet, and ignored the others. Section S2.4 discusses a special case.

We found a large number of pathology-related hashtags. We opted to use the 5 most common hashtags and their alternative spellings for our analyses, to maximize the amount of data per histopathology subtype. Here, we list all the hashtags for completeness, and highlight in bold/color those that we used for histopathology tissue analyses: **146 gipath**, **77 dermpath**, **72 gynpath**, **43 breastpath**, **42 gupath**, 37 pedipath, 34 hemepath, 26 neuropath, 20 entpath, 20 endopath, 18 pulmpath, 16 bstpath, 14 grosspath, 14 cytopath, 8 surgpath, 8 ihcpath, **6 ginpath**, 5 liverpath, 4 paz_path, 4 lungpath, 3 molpath, 2 oralpath, 2 idpath, 2 eroticpath, 1 turkpath, 1 sarcomapath, 1 musclepath, 1 headneckpath, 1 fnapath, 1 eyepath, 1 cardiacpath, **1 brstpath**, 1 autopsypath, 1 artpath.

We therefore had **146 gastrointestinal** tweets, **77 dermatological** tweets, **78 (72+6) gynecological** tweets, **44 (43+1) breast** tweets, and **42 genitourinary** tweets. To expand the per-tissue tweet counts, we moved beyond the hashtags and next searched for keywords in the tweet using Perl regular expressions, which we detail in Section S2.5. After keyword-based expansion, there were **180 gastrointestinal** tweets, **84 dermatological** tweets, **115 gynecological** tweets, **56 breast** tweets, and **58 genitourinary** tweets.

2.5 Machine learning methods

We used a variety of baseline machine learning methods (Fig 9), to test whether more complex machine learning methods perform significantly better than simpler machine learning methods. These methods are discussed in Section S2.6. Results are detailed below, but in general, Random Forest [RF] performed the best in our tasks. As expected, ZeroR [ZR] performed the worst. Also as expected, K-nearest neighbors [KNN], Naïve Bayes [NB], and Support Vector Machine [SVM] performed somewhere in between RF and ZR. It remains to be seen if neural networks will outperform RF.

[3] A pathology hashtag ontology is available here or alternatively here.

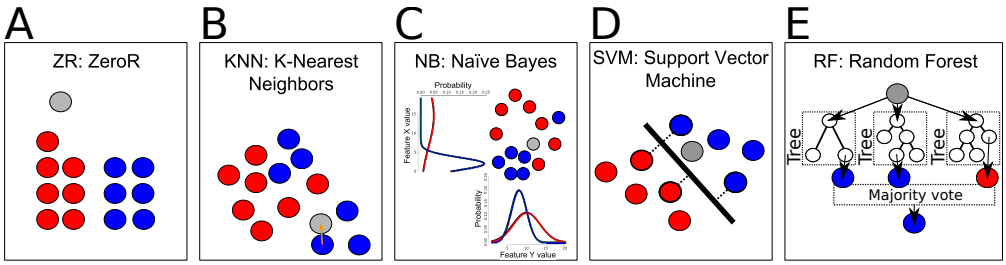

Figure 9: Machine learning methods for baseline analyses included ZeroR (Panel A), K-nearest neighbors (Panel B), Naïve Bayes (Panel C), an SMO-based support vector machine (Panel D), and Random Forest (Panel E). Visual schematics of each method shown. Weka provided all methods.

Task	n^{total}	n^{-}	n^{+}	ZR acc. %	RF accuracy %	ZR AUROC	RF AUROC
Acceptable H&E	2338	1165	1173	50	91.369 ± 1.828	0.5	0.959 ± 0.013
Accept H&E (bal)	1508	754	754	50	89.477 ± 2.691	0.5	0.953 ± 0.018
H&E vs IHC	1352	1175	177	85	97.026 ± 1.262	0.5	0.996 ± 0.004
Breast vs Gyn	391	135	256	65	72.533 ± 5.116	0.5	0.798 ± 0.075
Derm vs Breast	310	175	135	56	75.129 ± 7.932	0.5	0.840 ± 0.078
Derm vs Gyn	431	175	256	59	76.096 ± 5.762	0.5	0.833 ± 0.063
GI vs Breast	495	360	135	73	78.578 ± 3.552	0.5	0.879 ± 0.043
GI vs Derm	535	360	175	67	77.347 ± 5.736	0.5	0.853 ± 0.059
GI vs Gyn	616	360	256	58	73.029 ± 5.072	0.5	0.811 ± 0.054
Breast vs GU	251	135	116	54	72.940 ± 7.579	0.5	0.815 ± 0.078
Derm vs GU	291	175	116	60	75.843 ± 6.981	0.5	0.856 ± 0.070
GI vs GU	476	360	116	76	79.692 ± 2.609	0.5	0.816 ± 0.072
Gyn vs GU	372	256	116	69	75.230 ± 4.899	0.5	0.771 ± 0.088

Table 2: Random Forest [RF] machine learning analysis results for various binary classification tasks. Results compared to chance, i.e. ZeroR [ZR]. Accuracy [acc] and AUROC reported as mean ± stdev over 10 iterations of 10-fold cross validation. For accuracy reporting, prediction is positive class when majority of RF trees vote positive, i.e. accuracy is not calibrated/optimized. Results are detailed in Section S3.

2.6 Computational hardware and software

We use Weka version 3.8.1 [9] on an ASUS Intel 4-CPU laptop with 16 GB RAM. Section S2.7 discusses.

3 Results

To conduct preliminary tests of our dataset, we ran several baseline machine learning methods in Weka. Results are reported in Table 2. Our first question was the most basic: can machine learning distinguish pathology images from non-pathology images? In Section 3.1.1, we show acceptable H&E-stained human pathology images can be distinguished from other images – e.g., natural scenes, different histochemistry stains, or different species. Section S3.1.1 goes further with a pathologist-balanced and class-balanced analysis, sampling without replacement an equal number of acceptable images and non-acceptable images from each pathologist, to overcome possible biases from any pathologists. A classifier on this task may partially automate one of our manual data curation tasks, i.e. identifying acceptable images on social media. This task also serves as a positive control that machine learning works in our dataset. This learning task may be a "bridge" for transfer learning, when adapting a deep neural network trained on natural images to be used for pathology purposes. This task would allow the deep neural network to learn what pathology "looks like" before being re-trained on different data to learn a more specific pathology concept.

Second, can machine learning distinguish histochemistry stains, such as H&E and IHC? Section 3.1.2 shows strong performance when distinguishing these two stains of different coloration, though IHC colorations may vary (Section S1.1). H&E and IHC stain types were the most common in our dataset and are common in practice. Our classifier may be useful with large digital slide archives having a mix of H&E and IHC slides

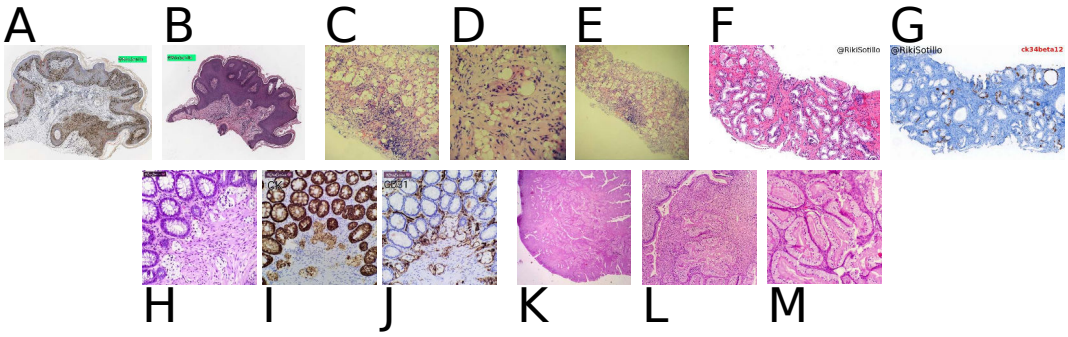

Figure 10: We use machine learning to distinguish four histopathological tissue types: dermatological, breast, gastrointestinal, and gynecological. *Panels A,B (R.S.S.)*: IHC- and H&E-stained dermatological pathology at low magnification, showing hallmark layering of epidermis, dermis, subcutaneous fat, and stroma. *Panels C,D,E (K.H.)*: H&E-stained breast needle biopsy pathology, showing hallmark adipocytes as small clear circles, because slide processing clears away the fat. *Panels F,G (R.S.S.)*: H&E- and IHC-stained prostate needle biopsy genitourinary pathology, showing atypical adenomatous hyperplasia, a departure from normal "feathery" gland structure in prostate. *Panels H,I,J (L.G.P.)*: H&E-stained gastrointestinal pathology with two different IHC stains, showing hallmark rosettes, circular regions that are cross sections of intestinal crypts. *Panels K,K,M. (M.P.P.)*: H&E-stained gynecological pathology, here an endocervical polyp, though we did not notice clear hallmark patterns across the variety of organs studied in gynecological pathology.

lacking explicit labels for staining information. Our classifier can distinguish these stains so downstream pipelines may process each stain type in a distinct way. This task serves as another positive control.

Third, can machine learning distinguish histopathology tissue types? In Sections 3.2 and 3.3, we show statistically significant discrimination performance, but with room for improvement. We consider five tissue types: breast, dermatological [derm], gastrointestinal [GI], genitourinary [GU], and gynecological (Fig 10). In Section 3.2, we consider all ten pairs of the five tissue types, using machine learning in a binary classification task for each pair. For example, the first task is to distinguish between breast and derm pathology. The differentiating histological intuition here is that breast tissue typically has many adipocytes throughout, which show as small clear circles in the image, while derm tissue is layered, from thin epidermis, to thicker dermis, to subcutaneous adipose tissue that also includes adipocytes (Fig 10). Moreover, one visual motif to distinguish GI tissue is "rosettes", circular lumen surrounded by endothelial cells (Fig 10), although we did not recognize clear identifying motifs in GU or gynecological tissues, and Section S1.2 has further discussion. In Section 3.3, we consider all five tissue types simultaneously, rather than pairwise. This is a more realistic setting because we typically cannot assume an image will be one of two possible tissue types. Moreover, ImageNet and CIFAR-10 are also multi-class classification tasks. Learning to distinguish tissue types has implications from determining tumor site of origin, e.g. whether a tumor originated in the GI or the breast. This has implications for metastasis prediction, e.g. a microscopy slide image of the GI may show morphology that appears similar to lobular breast cancer. Lobular breast cancer can metastasize to the GI.

3.1 Stain-related tasks

3.1.1 Acceptable H&E human tissue vs others task

Our Random Forest predicts if an image is an "acceptable" H&E-stained microscopy slide image or not (Fig 11). There were 2338 images: 1165 negative images that were not acceptable and 1173 positive images that were acceptable. Classes were essentially balanced. Accuracy is $91.369 \pm 1.828\%$ (chance 50%). AUROC is 0.959 ± 0.013 (chance 0.5). We believe this task is a simple positive control that the machine learning works, because H&E images are typically red and purple, while unacceptable images are typically (i) natural scenes such as outdoor photos or (ii) other histopathology techniques with different coloration. Performing well on this task is important to partially automate our otherwise manual annotation efforts on social media images. We are interested to reduce the manual data curation burden as much as possible. In Section S3.1.1 we explored pathologist-balanced and class-balanced subsampling, to potentially overcome biases in our data, but encouragingly this balanced approach did not produce a significantly different result.

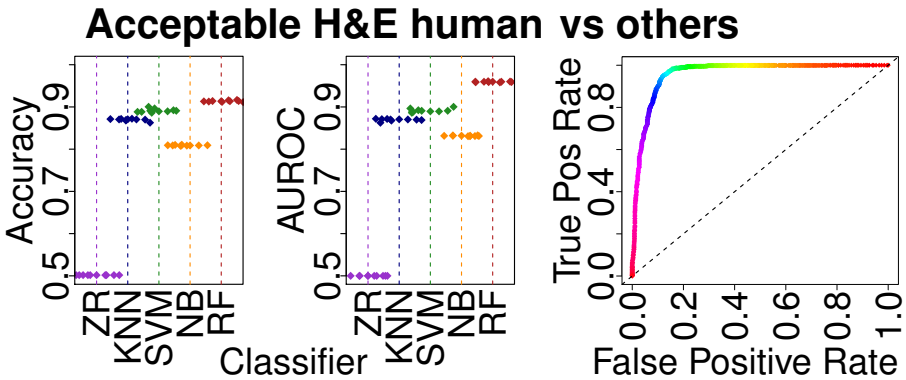

Figure 11: Predicting if an image is acceptable H&E human tissue or not. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig 9). The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.9588 here for n=2338 per Table 2.

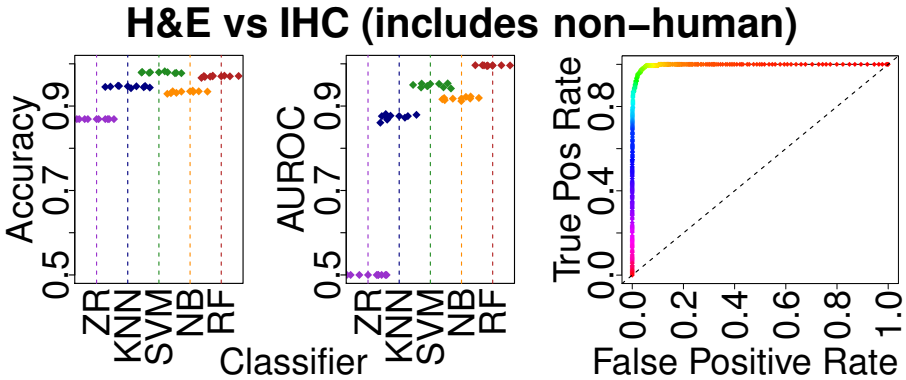

Figure 12: Predicting if an image is H&E or IHC. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig 9). The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.9963 here for n=1508 per Table 2.

3.1.2 H&E vs IHC task

Our Random Forest predicts if a microscopy slide image shows staining of H&E or IHC (Fig 12). There were 1352 images: 1175 negative images that were H&E and 177 positive images that were IHC (the choice of which stain is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state). Accuracy is $97.026 \pm 1.262\%$ (chance 87%). AUROC is 0.996 ± 0.004 (chance 0.5). Despite the marked class imbalance of ˜6.6:1, the Random Forest demonstrated statistically above chance accuracy and AUROC, with strong effect sizes. This task is a very simple positive control, because H&E images are typically red and purple, while IHC images are typically brown and blue[4]. This classifier may be useful when processing a digital archive of microscopy images having a mix of H&E and IHC slides, so that these images may be subsequently analyzed in a stain-specific manner.

3.2 Histopathological tissue type binary classification tasks

Next, we make a variety of histopathology tissue type discriminations (Fig 10). For dermatological [Derm], breast [Brst], gastrointestinal [GI], genitourinary [GU], and gynecological [Gyn] types, we consider all pairwise comparisons; these pairwise comparisons are detailed in Section S3. To determine the type of tissue, we used hashtags in the accompanying tweet, e.g. #dermpath indicates Derm, #breastpath indicates Brst, #gipath indicates GI, #gupath indicates GU, and #gynpath indicates Gyn. We also included common variants of these hashtags, such as #brstpath and #ginpath. If no hashtags were present, we used regular expressions

[4]Section S1.1 has more discussion on IHC color variability.

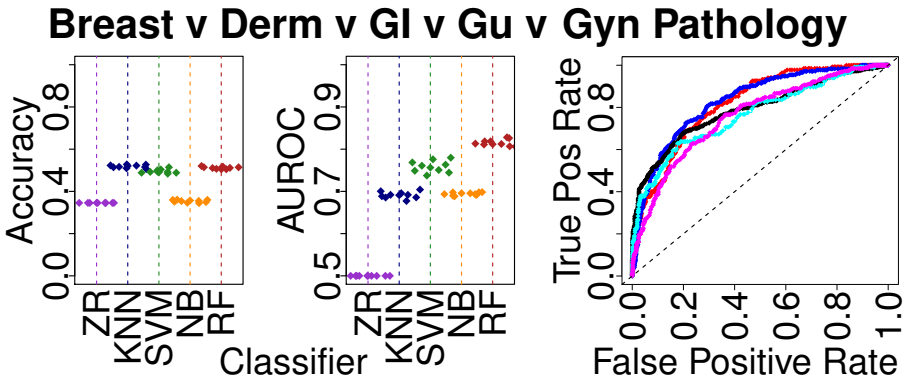

Figure 13: Predicting if an image is Breast, Derm, GI, GU, or Gyn. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig 9). The ROC curve for the highest AUROC classifier [RF] is shown at right. In the ROC plot (right), **Breast** is red (AUROC=0.8183, n=135), **Derm** is blue (AUROC=0.8306, n=175), **GI** is black (AUROC=0.7890, n=360), **GU** is cyan (AUROC=0.7655, n=116), and **Gyn** is magenta (AUROC=0.7585, n=256), with performance details in Table 3. For this trial, the weighted mean of AUROCs is 0.790, which is below the mean of 0.815 (Table 3), but within a standard deviation (0.054). ROC is calculated as the tissue versus all other tissues, e.g. in red is Breast vs all other tissues, and in blue is Derm vs all other tissues.

Task	n^{total}	n^{brst}	n^{derm}	n^{gi}	n^{gu}	n^{gyn}	ZR acc. %	RF acc. %	ZR AUROC	RF AUROC
5 tissues	1042	135	175	360	116	256	35	51.162 ± 4.153	0.5	0.815 ± 0.054
5 tiss (bal)	580	116	116	116	116	116	20	47.724 ± 5.057	0.5	0.784 ± 0.053

Table 3: Random Forest [RF] machine learning analysis results for 5-way tissue classification tasks, to predict if an image shows Breast, Derm, GI, GU, or Gyn tissue. Results compared to chance, i.e. ZeroR [ZR]. Accuracy [acc] and AUROC reported as mean ± stdev over 10 iterations of 10-fold cross validation. For accuracy reporting, prediction is positive class when majority of RF trees vote positive, i.e. accuracy is not calibrated/optimized. AUROC is calculated for each class independently – e.g. Breast vs others, or Derm vs others – and then a weighted average of all five independent AUROCs is calculated, based on how many examples were really of that tissue type. This weighted average AUROC is the default method in Weka to calculate AUROC for these multiclass classification tasks.

to perform a keyword search on the tweet's text, e.g. "duodenal" indicates GI and "ovarian" indicates Gyn. Accurately determining histopathology tissue type has implications for detecting tumor site of origin.

3.3 Histopathological tissue type multiclass classification tasks

Next, we attempt to distinguish all five histopathology tissue types (Fig 10) simultaneously in one learning task (Fig 13 and Table 3). This poses an important test for our dataset, because ImageNet and CIFAR-10 are multi-way classifications tasks. Success on this discrimination would be a critical step towards building an ImageNet-type database for computational pathology. For this task, an image could be any one of dermatological [Derm], breast [Brst], gastrointestinal [GI], genitourinary [GU], or gynecological [Gyn] tissue types, and the learning task was to predict which one of these five types the image is. While this task benefits from having more data than the other comparisons, it is more difficult because there are five possible tissue types to predict rather than two; this also makes the task much more realistic.

3.3.1 5-class tissue classification

Our Random Forest predicts if an image is one of five possible tissue types (Figs 10 and 13). There were 1042 images: 135 breast, 175 dermatological, 360 gastrointestinal [GI], 116 genitourinary, and 256 gynecological. Classes were imbalanced (roughly one third GI images). Accuracy is $51.162 \pm 4.153\%$ (chance 35%). AUROC is 0.815 ± 0.054 (chance 0.5). However, the confusion matrix (Table S1) suggests that because so much of the data is GI, many of the predictions are GI. A class-balanced sampling approach, or class weighting approach, may remedy the GI false positives. We explored this in Section S3.3.1

4 Discussion

We mined social media to obtain pathology images shared by pathologists across the world, and organized them into a diverse dataset that can be used to rigorously test computational pathology methods. We report 0.959 ± 0.013 AUROC when using this dataset to train a Random Forest to identify single-panel human H&E-stained slides that are not overdrawn. We also report 0.996 ± 0.004 AUROC when distinguishing H&E from IHC slides – almost perfect performance on this simple task. We consider both these tasks to be positive controls for machine learning methods on these data.

We also distinguish all pairs of breast, dermatological, gastrointestinal, genitourinary, and gynecological pathologies, with AUROC ranging from 0.771 to 0.879. Gynecological vs genitourinary is the most difficult discrimination; gastrointestinal vs breast is the easiest. Dermatology is the easiest to discriminate from any other pathology, with the highest minimum mean AUROC (0.833) across binary classifications. Finally, we report 0.815 ± 0.054 AUROC when all five tissue types are considered in a 5-way multi-class classification task rather than in 2-way binary classification.

4.1 Future Directions

To increase the granularity and accuracy of tissue type predictions, we first plan to expand the size of this dataset by recruiting more pathologists via social media, aiming to have representative images for each organ. Second, we will advocate for data sharing of normal tissue. Third, we will advocate for an expanded, more precise ontology of tweet hashtags to more fully describe images in a standard way, which will reduce our manual annotation burden, and can allow us to complement histology with molecular hashtags. Finally, we will use advanced techniques, e.g. deep learning, to improve performance. Section S4 discusses further.

4.2 Caveats

A number of caveats exist in our dataset, most of which can be remedied. First, a particular patient may be represented with more than one image, and more than one tweet. To control for this, we can consider at most one image per patient (although this increases our manual data curation burden). Second, there is a risk of error in our data because many different pathologists share cases, and they may disagree on the most appropriate hashtags or diagnosis. Third, there may be sampling bias if we typically have unusual cases that pathologists consider worth sharing, and our cases by necessity only come from pathologists on social media. Fourth, our pipeline crops images, potentially losing important information. Finally, our quality control pipeline does not filter out pathologist markings on these images. Section S4.1 discusses further.

5 Conclusion

We mined social media to obtain, curate, and perform preliminary machine learning analyses on pathology images shared by pathologists across the world. Our dataset includes a diverse, realistic, and comprehensive snapshot of pathology, spanning multiple image modalities, stain types, and pathology sub-specialties, along with text annotations from practicing pathologists. To our knowledge, this is the first study of pathology text and images shared on social media. Our goal in sharing this dataset is to advance the next generation of computational pathology machine learning methods.

Acknowledgments

A.J.S. thanks Dr. Marcus Lambert and Pedro Cito Silberman for organizing the Weill Cornell High School Science Immersion Program. A.J.S. thanks Terrie Wheeler and the Weill Cornell Medicine Samuel J. Wood Library for providing vital space for A.J.S., W.C., and S.J.C. to work early in this project.

A.J.S. was supported by NIH/NCI grant F31CA214029 and the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937). This research was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748.

References

[1] Deep Residual Learning for Image Recognition. Dec. 2015. URL http://arxiv.org/abs/1512.03385.

[2] Y. Bai, L. Guo, L. Jin, and Q. Huang. A novel feature extraction method using Pyramid Histogram of Orientation Gradients for smile recognition. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 3305–3308. ISBN 1522-4880. doi: 10.1109/ICIP.2009.5413938. URL http://dx.doi.org/10.1109/ICIP.2009.5413938.

[3] N. Bayramoglu and J. HeikkilÃ? Transfer Learning for Cell Nuclei Classification in Histopathology Images. pages 532–539. Springer International Publishing, 2016. ISBN 978-3-319-49409-8.

[4] S. Chatzichristofis and Y. Boutalis. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In A. Gasteratos, M. Vincze, and J. Tsotsos, editors, *Computer Vision Systems*, volume 5008, pages 312–322. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-79547-6_30. URL http://dx.doi.org/10.1007/978-3-540-79547-6_30.

[5] S. Chatzichristofis and Y. Boutalis. FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on*, pages 191–196. IEEE, May 2008. ISBN 978-0-7695-3344-5. doi: 10.1109/wiamis.2008.24. URL http://dx.doi.org/10.1109/wiamis.2008.24.

[6] G. Crane and J. Gardner. Pathology Image-Sharing on Social Media: Recommendations for Protecting Privacy While Motivating Education. *AMA journal of ethics*, 18(8):817–825, Aug. 2016. ISSN 2376-6980. URL http://view.ncbi.nlm.nih.gov/pubmed/27550566.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. pages 248–255. IEEE, June 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/cvpr.2009.5206848. URL http://dx.doi.org/10.1109/cvpr.2009.5206848.

[8] T. Fuchs and J. Buhmann. Computational pathology: challenges and promises for tissue analysis. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35(7-8):515–530, Oct. 2011. ISSN 1879-0771. doi: 10.1016/j.compmedimag.2011.02.006. URL http://dx.doi.org/10.1016/j.compmedimag.2011.02.006.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009. doi: 10.1145/1656274.1656278. URL http://dx.doi.org/10.1145/1656274.1656278.

[10] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 0, pages 762–768. IEEE, June 1997. ISBN 0-8186-7822-4. doi: 10.1109/cvpr.1997.609412. URL http://dx.doi.org/10.1109/cvpr.1997.609412.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. June 2014. URL http://arxiv.org/abs/1408.5093v1.pdf.

[12] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Apr. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[14] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. SÃ¡nchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005. URL http://dx.doi.org/10.1016/j.media.2017.07.005.

[15] M. Lux and S. Chatzichristofis. Lire: Lucene Image Retrieval: An Extensible Java CBIR Library. pages 1085–1088. ACM, 2008. ISBN 978-1-60558-303-7. doi: 10.1145/1459359.1459577. URL http://dx.doi.org/10.1145/1459359.1459577.

[16] J. Nix, J. Gardner, F. Costa, A. Soares, F. Rodriguez, B. Moore, M. Martinez-Lage, S. Ahlawat, M. Gokden, and D. Anthony. Neuropathology Education Using Social Media. *Journal of neuropathology and experimental neurology*, 77(6):454–460, June 2018. ISSN 1554-6578. URL http://view.ncbi.nlm.nih.gov/pubmed/29788115.

[17] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585 vol.1, . doi: 10.1109/ICPR.1994.576366. URL http://dx.doi.org/10.1109/ICPR.1994.576366.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, . ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1017623. URL http://dx.doi.org/10.1109/TPAMI.2002.1017623.

[19] P. Oltulu, A. A. S. R. Mannan, and J. Gardner. Effective use of Twitter and Facebook in pathology practice. *Human pathology*, 73:128–143, Mar. 2018. ISSN 1532-8392. URL http://view.ncbi.nlm.nih.gov/pubmed/29307629.

[20] A. Schaumberg, M. Rubin, and T. Fuchs. H&E-stained Whole Slide Deep Learning Predicts SPOP Mutation State in Prostate Cancer. *bioRxiv*, page 064279, July 2016. doi: 10.1101/064279. URL http://dx.doi.org/10.1101/064279.

[21] A. Schaumberg, S. Sirintrapun, H. Al-Ahmadie, P. Schueffler, and T. Fuchs. DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope. *bioRxiv*, page 097246, Dec. 2016. doi: 10.1101/097246. URL http://dx.doi.org/10.1101/097246.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. Sept. 2014. URL http://arxiv.org/abs/1409.4842v1.pdf.

[23] N. Tajbakhsh, J. Shin, S. Gurudu, T. Hurst, C. Kendall, M. Gotway, and Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, May 2016. ISSN 1558-254X. URL http://view.ncbi.nlm.nih.gov/pubmed/26978662.

[24] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473. ISSN 0018-9472. doi: 10.1109/TSMC.1978.4309999. URL http://dx.doi.org/10.1109/TSMC.1978.4309999.

[25] T. Tefik, O. Sanli, T. Oktar, O. B. Yucel, Y. Ozluk, and I. Kilicaslan. Oxidized regenerated cellulose granuloma mimicking recurrent mass lesion after laparoscopic nephron sparing surgery. *International journal of surgery case reports*, 3(6):227–230, 2012. ISSN 2210-2612. URL http://view.ncbi.nlm.nih.gov/pubmed/22472162.

[26] M. Veta, Y. Heng, N. Stathonikos, B. Bejnordi, F. Beca, T. Wollmann, K. Rohr, M. Shah, D. Wang, M. Rousson, M. Hedlund, D. Tellez, F. Ciompi, E. Zerhouni, D. Lanyi, M. Viana, V. Kovalev, V. Liauchuk, H. Phoulady, T. Qaiser, S. Graham, N. Rajpoot, E. Sjöblom, J. Molin, K. Paeng, S. Hwang, S. Park, Z. Jia, E. Chang, Y. Xu, A. Beck, P. van Diest, and J. Pluim. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. July 2018. URL http://arxiv.org/abs/1807.08284.

Supporting Information

S1 Data diversity discussion

S1.1 Intra-stain diversity

There is an art and variability in histochemical stains that we have not discussed in the main text, but for completeness mention here. We note that in clinical practice we have observed high variability stains, for isntance H&E stains that appear almost neon pink, to GMS stains (discussed below) that had silver (black) deposition throughout the slide. One reason for this is that there are a number of reagents that may be used for staining, each with different qualities that can make the stain darker, brighter, pinker, bluer, etc.

IHC stains typically use an antibody conjugated to a brown stain, namely 3,3'-Diaminobenzidine [DAB]. The blue counterstain is typically hematoxylin. However, some laboratories conjugate the antibody to a red stain instead. As we acquire more data, we expect to have both types of IHC stains. Currently we only see DAB.

There is counterstain variability in Grocott's modification of the Gömöri methenamine silver stain [GMS stain]. Typically the counterstain is green, but a pink counterstain is also available. We may see the pink variant as we acquire more data. Currently we see only green.

S1.2 Intra-tissue-type diversity

The tissue type hashtags we use are very broad, e.g. #gipath encompasses several organs, such as stomach, small intestine, large intestine, liver, gallbladder, and pancreas. This is also noted in Section S2.5. We note, for instance, liver morphology looks nothing like the stomach. Moreover, gynecological pathology, i.e. #gynpath, includes vulva (which looks just like skin, i.e. dermatological pathology, #dermpath), vagina, cervix, uterus, fallopian tubes and ovaries. Again, vulva looks nothing like uterus. A number of tissue features also overlap, such as adipocytes in breast tissue and adipocytes in the subcutaneous fat layer in skin. The amount and distribution of adipocytes typically differs between these tissues however. However, a lipoma in any tissue has a great deal of adipocytes and should not strictly be confused with breast tissue. For all these motivating reasons, we have a future direction to sample every organ within a tissue type hashtag category, for all hashtag categories we study.

S2 Supplementary materials and methods

S2.1 Criteria details for rejected, discarded, private, or acceptable images

Though criteria are outlined in Section 2.3.1 – more formally, we reject the following image types, during our manual data curation process:

1. Non-pathology images, such as pictures of vacations or food.
2. Multi-panel images, such as a set of 4 images in a 2x2 grid. Images with insets are also rejected. We only accept single-panel images, and leave for future work the complexities of splitting multi-panel images into sets of single-panel images. Multi-panel images may have black dividers, white dividers, no dividers, square insets in a corner, or floating circular insets somewhere in the image. There may be two or more panels/insets. Per-pixel labels for each panel may be the best solution here, and would support a machine learning approach to split multi-panel images to reduce this additional manual data curation burden.
3. Overdrawn images, where a 256x256px region could not bound all regions of interest in an image. This occurs most frequently if a pathologist draws by hand a tight circle around a region of interest, preventing image analysis on the region of interest in a way that completely avoids the hand-drawn marks.
4. Images that manipulate pathology slides into artistic motifs, such as smiley faces or trees. In contrast, a picture of a painting would be a non-pathology image.

Moreover, we completely discard from analysis certain types of images:

1. Duplicate images, according to identical SHA1 checksums or by a preponderance of similar pixels.
2. Corrupt images, which either could not be completely downloaded or employed unusual JPEG compression schemes that Java's ImageIO[5] library could not open for reading.
3. Pathology images that are owned by pathologists who have not given us explicit written permission to use their images. Consider the following example. When a pathlogist gives us permission to download data, our software bot downloads thousands of that pathologists's social media posts regardless if some of the images in those posts are actually owned by a different pathologist who did not give us permission. We detect these cases when we manually curate the pathologist's data, and discard these images belonging to pathologists who have not given us permission. To elaborate, pathology images that are taken by pathologists and shared on social media are treated the same way as pathology images taken from case reports or copyrighted manuscripts, i.e. if the pathologist or publisher has not provided us explicit written permission to use the image, we discard the pathology image and do not use it.

Images that are not rejected or discarded are deemed "acceptable" pathology images. However, for legal reasons, we cannot distribute all of the images we have from social media, namely:

1. Pathology images obtained from children (including fetuses), which may be identifiable. The data shared on social media are anonymized; thus, we do not have contact information for the child's parent and therefore cannot obtain consent to distribute a picture of e.g., a child's X-rays or gross specimens. Although unlikely to be identified by the parent if these images were made public, we prefer to err on the side of caution. However, microscopy slide images are not personally identifiable, so we may distribute these.
2. Personally identifiable pictures involving adults, because they have the right to consent or not to their likeness being distributed. We consider faces, body profiles, automobile license plates, etc to all be personally identifiable pictures involving adults, especially because these data may be cross-referenced against timestamp, location, clinician, institution, medical condition, other people in the picture, etc.
3. Copyrighted content, which includes images of copyrighted manuscripts, pictures of slideshow presentations, and pictures of any brand or logo. A lab picture that includes boxes bearing logos would be a non-pathology image that we cannot distribute, because we do not have permission to distribute any images with the protected logos. A picture of a powerpoint slide at a conference that shows some text outlining a new way to make a clinical decision would also be a non-pathology image that we hold privately and do not distribute. We similarly hold privately an image of text taken from a copyrighted manuscript because it may not be possible to identify the original source to provide a proper citation, and even if we could, this poses an additional data curation burden that we would rather avoid. Moreover, we prefer to err on the side of caution and not distribute these images, rather than rely on "fair use" or similar law that may expose us to legal challenges and costs[6]. By retaining these images privately, we can train a machine learning classifier to detect these types of images and potentially reduce our manual data curation burden.

S2.2 Overdrawn rejection criterion

Here we discuss the details of rejecting images as "overdrawn". Figure 2 Panel C *top* is rejected as "overdrawn", because the regions of interest [ROIs] in the H&E image that the pathologist refers to in the social media post's text have hand-drawn circles and arrows such that it is not possible to place a 256x256px square over all ROIs without including these circle and arrow marks. We chose 256x256px because deep convolutional neural networks in computational pathology [14] typically require 227x227px (i.e. AlexNet [13] or CaffeNet [11]) or 224x224px (i.e. ResNet [1]) images, and we have used these sizes in the past [20, 21]. We note the Inception [22] family of deep convolutional neural networks takes a 299x299px image input, which is larger than 256x256px and is also frequently used in computational pathology [14]. Ideally, each image would have ROIs and hand-drawn arrows/circles annotated at the pixel level, so each image could be

[5] ImageIO documentation available here: `https://docs.oracle.com/javase/7/docs/api/javax/imageio/ImageIO.html`

[6] Courts in the United States have ruled that images posted to social media are still owned by their authors and are not public domain. Indeed, in *Morel v. AFP*, AFP was ordered to pay Morel $1.2 million for copyright infringement because AFP used images that Morel posted to social media.

annotated as "overdrawn" to arbitrary bounding box sizes, whether 256x256px or 299x299px, and we leave this to future work. Smaller "overdrawn" bounding boxes may allow more images to pass as acceptable, rather than be rejected. A 256x256px image size allows minor rotations and crops for deep learning data augmentation using 224x224px input image sizes. Minor upsampling and/or image reflection at the image's outer boundaries may allow a 256x256px image to work for 299x299px input image sizes. Figure 2 Panel C *bottom* is rejected as "overdrawn", because this image was originally 783x720px and the arrow marks prevent us from capturing each of the two indicated regions of interest in their own 256x256px square.

S2.3 Uniform cropping and scaling of original images

Images shared on social media may be any rectangular shape. However, machine learning methods typically require all images be the same size. To accomplish this, we use the following procedure:

1. Take the minimum of two numbers: the original image's height and width.
2. Crop from the center of the original image a square with a side whose length is the minimum from the prior step.
3. Scale this square to 512x512px.

This square is intended to be large enough to represent small details, such as arrows and circles drawn one pixel wide by the pathologist. Such arrows and circles may then be used to predict if an image is "overdrawn" or not. Ideally, the tweet's text would be available alongside the image to give the machine learning the fullest information possible about potential ROIs in the image, for "overdrawn" prediction, but for simplicity here we perform only image-based machine learning.

The motivation for the 256x256px image for the "overdrawn" criteria In Sec S2.2 is that there may be an attention layer that scans the original image for 256x256px squares that have no marks from the pathologist. Such marks including circles or arrows for ROI indication or the pathologist's name to indicate copyright/ownership. Such mark-free 256x256px images may then be used for machine learning on only patient pathology pixels.

S2.4 Hashtag special case

A hashtag special case is "#bstpath", bone and soft tissue pathology, which we include in our breast pathology category only when the social media post's text also includes the word "breast" or other breast-related keywords. Such keywords are listed further below in this subsection. Examples of such tweets are *"Pleomorphic lobular carcinoma of the breast: Beautiful cells but nasty tumour #pathology #pathologists #BSTPath"* and *"Now at my desk, W(47y-o) breast nodule...Could be it siliconoma?? But it isn't noted giant cells #pathology #pathologists #BSTpath"*.

S2.5 Regular-expression-based tissue type keywords

Expanding our text processing discussion in Section 2.4, the regular expressions used for each tissue type were:

- Breast: `/breast/i` or `/nipple/i` or `/mastectomy/i` or `/phyllod/i`
  - These regular expressions match breast, nipple, mastectomy, and phyllodes mentions in a social media post's text. Phyllodes tumor is a type of breast cancer. Nipple pathology may have some overlap with dermatological pathology, but for our purposes we consider it breast pathology.
- Dermatological: `/skin/i` or `/epiderm(?:oid|is|al)/i` or `/derma(?:l|to)/i` or `/melanoma/i` or `/keratosis/i` or `/bcc/i`
  - This matches "skin" and other dermatological keywords in a social media post's text, in a case-insensitive manner. BCC is a type of skin cancer.

- Gastrointestinal: `/colon/i` or `/duoden(?:um|al)/i` or `/appendix/i` or `/rectal/i` or `/gastric/i` or `/stomach/i`or `/intestin(?:e|al)/i` or `/\banal\b/i` or `/perianal/i`or `/perine(?:um|al)/i` or `/cecum/i` or `/esophag(?:us|eal|itis)/i` or `/ileum/i` or `/gall\s?(?:bladder|stone)/i` or `/liver/` or `/ascaris/` or `/pancrea(?:s|tic)/` or `/colitis/` or `/hepat[ieo]c/` or `/cholecystitis/i` or `/crohn/i` or `/jejunum/i`
  - This matches colon, duodenum, duodenal, appendix, rectal, and other GI-related keywords.
- Genitourinary: `/urotheli(?:um|al)/i` or `/seminal/i` or `/prostate/i` or `/kidney/i` or `/renal/i` or `/mtscc/i` or `/rcc/i` or `/bladder/i` or `/test(?:[ei]s|icular)/i` or `/sperm/i`.
  - This matches urothelium, urothelial, seminal, prostate, renal, bladder, and other GU-related keywords.
- Gynecological: `/cervix/i` or `/uteri(?:us|ine|o)/i` or `/ovar(?:y|ian)/i` or `/fallopi/i` or `/adenomyosis/i` or `/fo?et(?:al|us)/i` or `/trophoblast/i` or `/embryo/i` or `/placenta/i` or `/villitis/i` or `/umbilical/i` or `/amniotic/i` or `/anhydramnios/i` or `/chorioamnionitis/i` or `/h[yi]sterectomy/i` or `/endocervical/i` or `/endometriosis/i` or `/(?:myo|endo)metri(?:al|um|oid)/i`.
  - This matches cervix, uterus, uterine, utero, ovarian, fallopian, and other gynecological-related keywords.

S2.6 Machine learning methods discussion

Expanding on our discussion of machine learning methods in Section 2.5, ZeroR is the simplest method, which always predicts the majority class, i.e. if there are more gynecological data than breast data, every prediction will be gynecological. ZeroR [ZR] is our model of statistical "chance", i.e. if a machine learning method does not outperform ZeroR, then the machine learning's predictions may be due to chance alone rather than a learned concept. K-nearest neighbors [KNN] is slightly more complex, which calculates the feature vector of a given example and finds the single closest neighbor in the training data, predicting the class label that this closest neighbor has. KNN is our crude test for a preponderance of duplicates, e.g. if there were many duplicates in the data, and these duplicates were spread between cross validation folds, then KNN would have strong performance, because KNN would find the duplicates and make the correct predictions. Naïve Bayes [NB] is a simple probabilistic model that assumes independence between all the features, fits a Gaussian distribution over each feature, and predicts the most likely class. Despite its simplicity, NB may show unexpectedly strong performance on some tasks. A support vector machine [SVM] is more complex than NB in that SVM allows nonlinear interactions between the features, and for this we use a polynomial kernel. SVM finds the maximum margin hyperplane that divides the data space, and its predictions depend on which side of this hyperplane an example is. Finally, Random Forest [RF] random samples both the data and features to construct an ensemble of 1000 fully-grown decision trees. These trees vote to make an overall prediction, i.e. the prediction from a RF is the majority vote of its constituent decision trees.

S2.7 Computational hardware and software discussion

We use Weka version 3.8.1 [9] on a ASUS Intel core i7-6700HQ 2.6GHz 4-CPU laptop with 16GB RAM for baseline analyses and comparison of several machine learning methods (Fig 9) on each of our prediction tasks. This laptop was also used for software development and automatically downloading Twitter data from participating pathologists. This laptop ran the Windows 10 operating system, which in turn ran the Oracle VirtualBox virtual machine manager, which in turn ran Debian Jessie 3.16.7-ckt20-1+deb8u3 and Linux kernel 3.16.0-4-amd64. Weka and our other pipeline components ran within Debian.

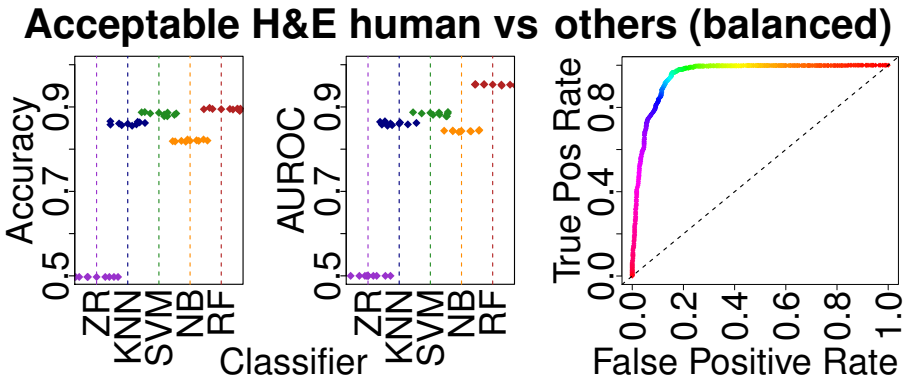

Figure S1: Predicting if an image is acceptable H&E human tissue or not, in a pathologist-balanced and class-balanced manner, for comparison to Fig 11.

S3 Supplementary results

S3.1 Pairwise stain comparisons

S3.1.1 Acceptable H&E human tissue vs others task (balanced)

Our Random Forest predicts if an image is an "acceptable" H&E-stained microscopy slide image or not (Fig S1), but in a pathologist-balanced and class-balanced manner. For example, it could be a pathologist uses a low-resolution camera and a large number of mostly natural scene pictures, in which case the machine learning may learn that low-resolution pictures predict that an image is not acceptable, rather than learning an intended concept that an image is not acceptable if it does not show H&E-stained human morphology.

The only difference from the unbalanced task is that here, from each pathologist independently we randomly sample without replacement an equal number of acceptable and not acceptable images. This addresses a potential confound in the analysis, where certain pathologists may share images with particular biases, such as low-lighting in microscopy images or low resolution images. If such confounds were prominent in the data, the machine learning would overfit to these confounds rather than learn the intended task of distinguishing acceptable H&E human tissue from other images. So, if performance in terms of accuracy or AUROC are significantly worse in this pathologist-balanced and class-balanced analysis, then there is evidence to suggest such confounds exist and such overfitting is occurring. In this pathologist-balanced and class-balanced analysis, there were 1508 images: 754 negative images that were not acceptable and 754 positive images that were acceptable. The choice of whether or not acceptable images are labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. This was 64.5% of the data available for the prior analysis of 2338 images. Classes were exactly balanced. However, we find accuracy is $89.477 \pm 2.691\%$ (chance 50%) and is not significantly worse than the prior non-balanced analysis accuracy of $91.369 \pm 1.828\%$. Moreover, AUROC here is 0.953 ± 0.018 (chance 0.5), which again is not significantly worse that the prior non-balanced analysis AUROC of 0.959 ± 0.013. So, we find that performance differences may be due to chance alone, rather than due to overfitting pathologist-specific imaging confounds. We did not perform this type of pathologist-balanced and class-balanced analysis for the other tasks, because there was an order of magnitude fewer data in the minority class on the other tasks, so performance drops may be due to insufficient data being available for machine learning. Additional balanced analyses may be more appropriate after have more data available for these tasks.

S3.2 Pairwise tissue comparisons

S3.2.1 Breast vs Gyn task

Our Random Forest predicts if an image is breast pathology, or alternatively, gynecological pathology (Fig S2). There were 391 images: 135 negative images (from 56 tweets) that were breast pathology and

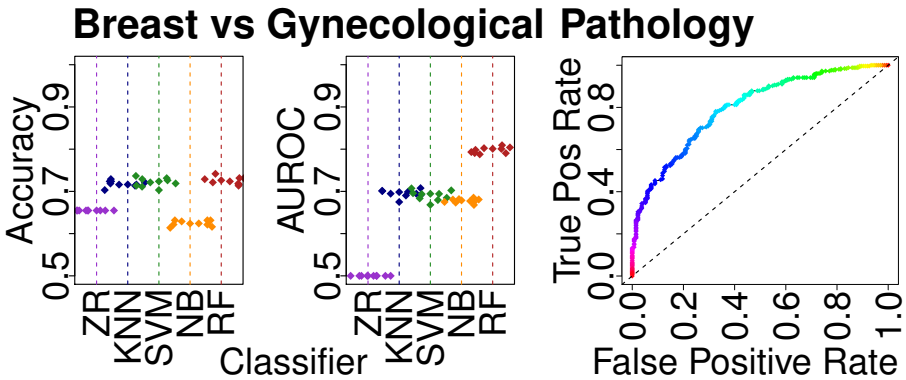

Figure S2: Predicting if an image is Breast or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

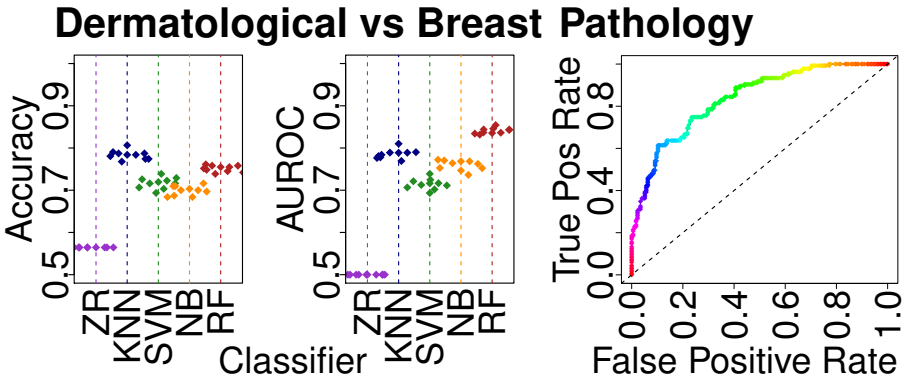

Figure S3: Predicting if an image is Derm or Breast. RF classifier had greatest AUROC. RF ROC curve at right.

256 positive images (from 115 tweets) that were gynecological pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced at a ratio of of ˜1.9:1. Accuracy is $72.533 \pm 5.116\%$ (chance 65%). AUROC is 0.798 ± 0.075 (chance 0.5). This is a difficult task. Though performance is statistically significant, performance is not strong (mean AUROC < 0.8). We do not notice clear hallmarks of gynecological pathology, which may include a variety of tissues, including ovary and cervix, so performance may improve with more data. In contrast, adipocytes may be a hallmark of breast tissue. In practice, pairwise comparisons involving gynecological pathology had lower AUROC than those without.

S3.2.2 Derm vs Breast task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, breast pathology (Fig S3). There were 310 images: 135 negative images (from 84 tweets) that were dermatological pathology and 135 positive images (from 56 tweets) that were breast pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced. Accuracy is $72.533 \pm 5.116\%$ (chance 56%). AUROC is 0.798 ± 0.075 (chance 0.5). There is room to improve performance in this task. The layered structure of dermis, subcutaneous fat, and stromal tissue may be a hallmark of dermatological pathology. In addition, adipocytes may be a hallmark of breast tissue. More advanced methods, such as deep learning, may be better able to recognize differences between these tissue types, beyond intuitive hallmarks and the Random Forest.

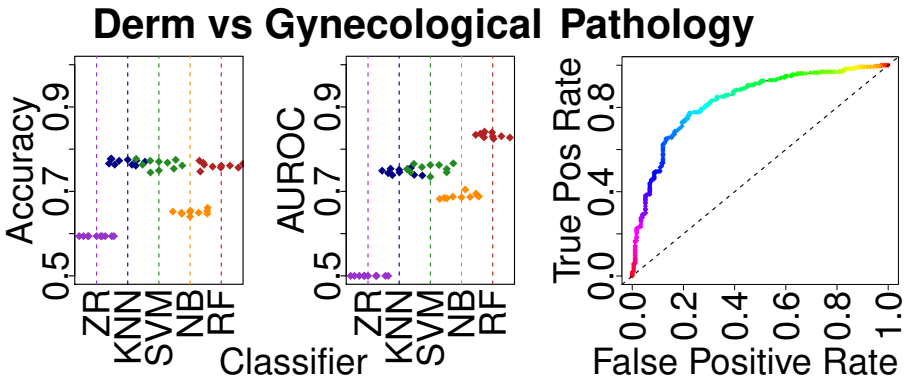

Figure S4: Predicting if an image is Derm or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

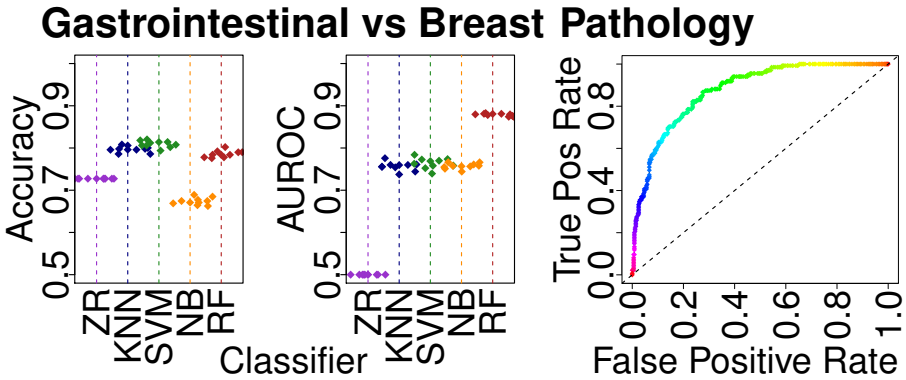

Figure S5: Predicting if an image is GI or Breast. RF classifier had greatest AUROC. RF ROC curve at right.

S3.2.3 Derm vs Gyn task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, gynecological pathology (Fig S4). There were 431 images: 175 negative images (from 84 tweets) that were dermatological pathology and 256 positive images (from 115 tweets) that were gynecological pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced. Accuracy is $76.096 \pm 5.762\%$ (chance 59%). AUROC is 0.833 ± 0.063 (chance 0.5). There is room to improve performance in this task. Of all the pairwise comparisons involving gynecological pathology, this comparison to dermatological pathology had the highest accuracy and AUROC.

S3.2.4 GI vs Breast task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, breast pathology (Fig S5). There were 495 images: 360 negative images (from 180 tweets) that were gastrointestinal pathology and 135 positive images (from 56 tweets) that were breast pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. There was mild class imbalance of ˜2.7:1. Accuracy is $78.578 \pm 3.552\%$ (chance 73%). AUROC is 0.879 ± 0.043 (chance 0.5). Of all six tissue pairs, our Random Forest performed best on this pair, GI vs Brst, though there still is room to improve on this task. Rosette structures from cross sections of intestinal crypts may be a hallmark of gastrointestinal pathology. Meanwhile adipocytes may be a hallmark of breast pathology. More advanced methods, such as deep learning, may be better able to recognize difference between these tissue types, beyond intuitive hallmarks and the Random Forest.

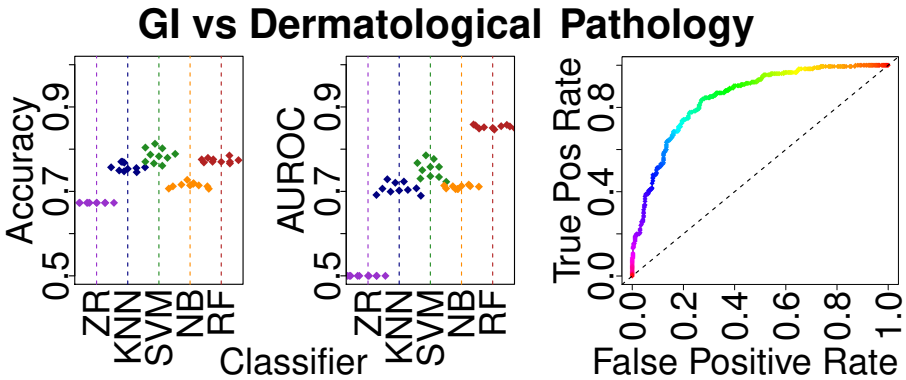

Figure S6: Predicting if an image is GI or Derm. RF classifier had greatest AUROC. RF ROC curve at right.

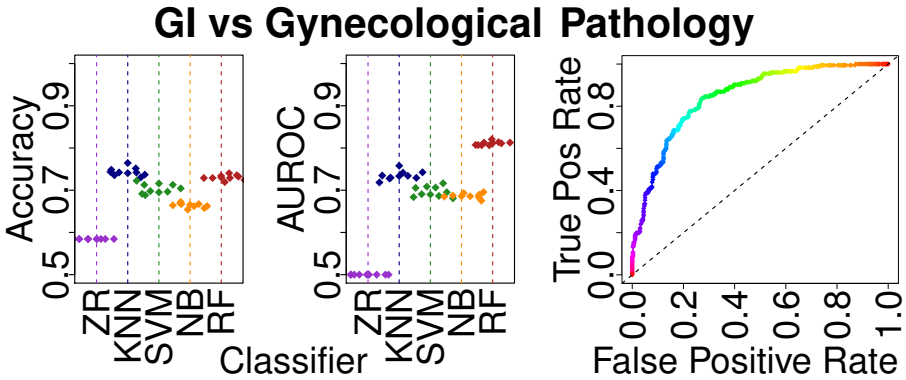

Figure S7: Predicting if an image is GI or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

S3.2.5 GI vs Derm task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, dermatological pathology (Fig S6). There were 535 images: 360 negative images (from 180 tweets) that were gastrointestinal pathology and 175 positive images (from 56 tweets) that were breast pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. There was mild class imbalance of ˜2.1:1. Accuracy is $77.347 \pm 5.736\%$ (chance 67%). AUROC is 0.853 ± 0.059 (chance 0.5). Performance was not significantly different than the (i) GI vs Brst comparison, or (ii) the Derm vs Brst comparison, suggesting these three histpathological subtypes may be readily distinguished.

S3.2.6 GI vs Gyn task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, gynecological pathology (Fig S7). There were 616 images: 360 negative images (from 180 tweets) that were gastrointestinal pathology and 256 positive images (from 115 tweets) that were gynecological pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced. Accuracy is $73.029 \pm 5.072\%$ (chance 58%). AUROC is 0.811 ± 0.054 (chance 0.5). This is a difficult task, like the Brst vs Gyn comparison, but there are more data available, and mean AUROC > 0.8. There is room to improve here.

S3.2.7 Breast vs GU task

Our Random Forest predicts if an image is breast pathology, or alternatively, genitourinary pathology (Fig S8). There were 251 images: 135 negative images (from 56 tweets) that were breast pathology and

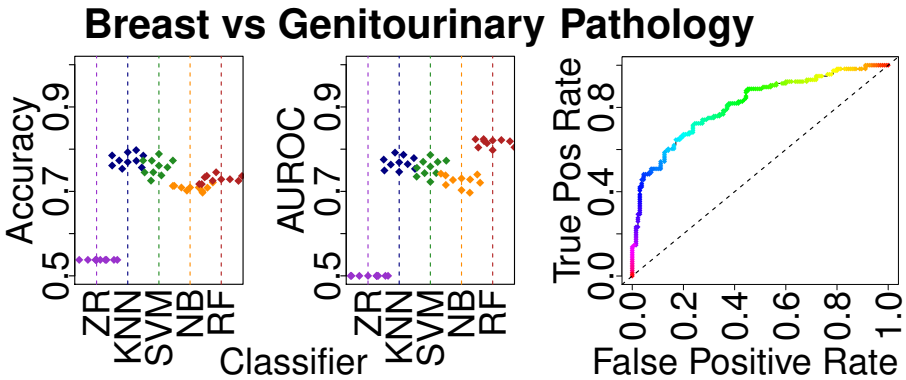

Figure S8: Predicting if an image is Breast or GU. RF classifier had greatest AUROC. RF ROC curve at right.

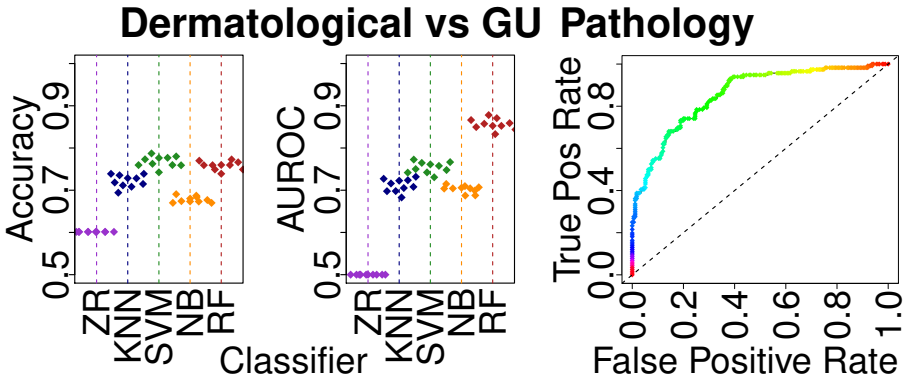

Figure S9: Predicting if an image is Derm or GU. RF classifier had greatest AUROC. RF ROC curve at right.

116 positive images (from 58 tweets) that were genitourinary pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced. Accuracy is $72.940 \pm 7.579\%$ (chance 54%). AUROC is 0.815 ± 0.078 (chance 0.5). Perhaps more data and deep learning can improve performance here.

S3.2.8 Derm vs GU task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, genitourinary pathology (Fig S9). There were 291 images: 175 negative images (from 84 tweets) that were breast pathology and 116 positive images (from 58 tweets) that were genitourinary pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. Classes were essentially balanced. Accuracy is $75.843 \pm 6.981\%$ (chance 60%). AUROC is 0.856 ± 0.070 (chance 0.5). Pairwise comparision AUROC tends to be higher if dermatology tissue in on of the tissues being compared, and this appears to hold for Derm vs GU as well.

S3.2.9 GI vs GU task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, genitourinary pathology (Fig S10). There were 476 images: 360 negative images (from 180 tweets) that were breast pathology and 116 positive images (from 58 tweets) that were genitourinary pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. There was mild class imbalance of ˜3.1:1. Accuracy is $79.692 \pm 2.609\%$ (chance 76%). AUROC is 0.816 ± 0.072 (chance 0.5).

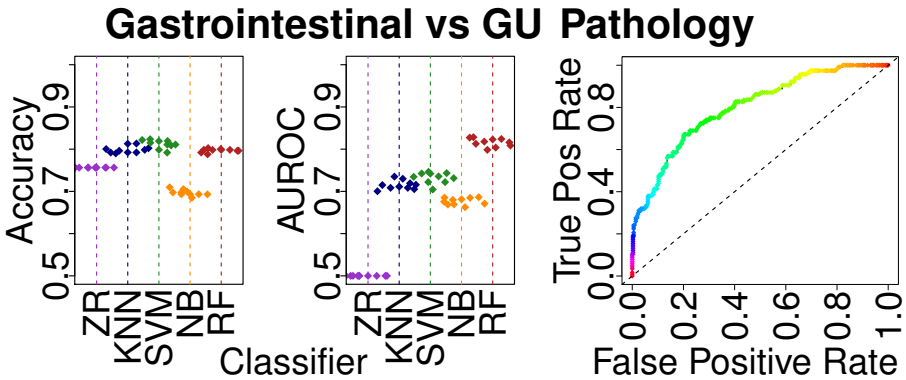

Figure S10: Predicting if an image is GI or GU. RF classifier had greatest AUROC. RF ROC curve at right.

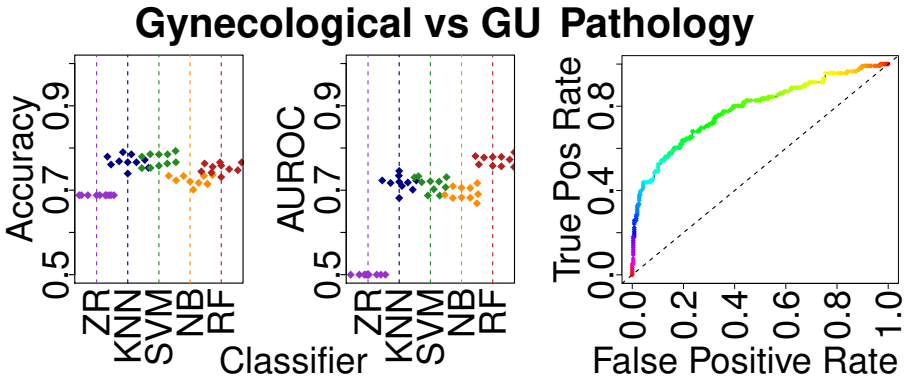

Figure S11: Predicting if an image is Gyn or GU. RF classifier had greatest AUROC. RF ROC curve at right.

S3.2.10 Gyn vs GU task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, genitourinary pathology (Fig S11). There were 372 images: 256 negative images (from 115 tweets) that were breast pathology and 116 positive images (from 58 tweets) that were genitourinary pathology. The choice of which tissue is labeled as the positive or negative class is arbitrary, does not impact performance, and does not necessarily imply any particular disease state. There was mild class imbalance of ˜2.2:1. Accuracy is $75.230 \pm 4.899\%$ (chance 69%). AUROC is 0.771 ± 0.088 (chance 0.5). This was the most challenging pair to compare, of all tissue type binary comparisons, and may suffer from a lack of samples, mildly class-imbalanced sampling, and no broad visual motifs to define each of these tissue types.

S3.3 5-way comparison details

The confusion matrix for Brst vs Derm vs GI vs GU vs Gyn is in Table S1, and Table S2 shows a similar confusion matrix but for class-balanced sampling.

a	b	c	d	e	<−−	classified as
24	11	66	0	34	—	a = breast
5	74	63	0	33	—	b = derm
2	29	290	1	38	—	c = gi
3	9	63	17	24	—	d = gu
5	17	112	1	121	—	e = gyn

Table S1: RF confusion matrix for Brst vs Derm vs GI vs GU vs Gyn comparison, showing many tissue types are predicted as GI. About one third of the data are GI.

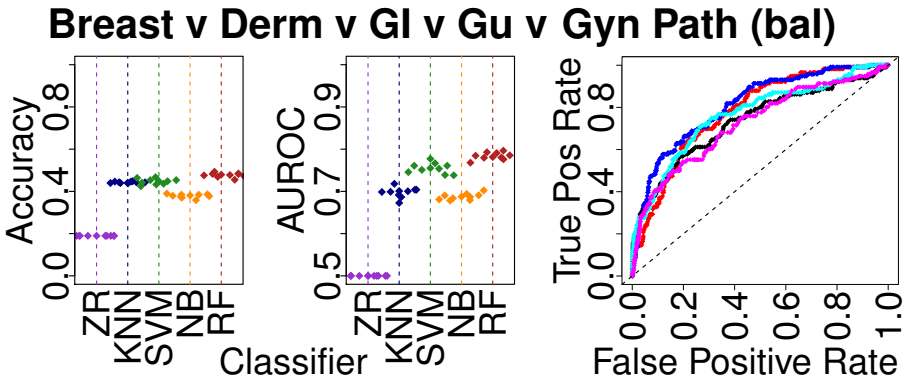

Figure S12: Predicting if an image is Breast, Derm GI, GU, or Gyn. RF classifier had greatest AUROC. RF ROC curve at right. In the ROC plots, **Breast** is red, **Derm** is blue, **GI** is black, **GU** is cyan, and **Gyn** is magenta. ROC is calculated as the tissue versus all other tissues, e.g. in red is Breast vs all other tissues, and in blue is Derm vs all other tissues.

a	b	c	d	e	<--	classified as
65	14	6	12	19	—	a = breast
15	72	11	10	8	—	b = derm
23	19	47	14	13	—	c = gi
13	19	17	54	13	—	d = gu
23	22	8	15	48	—	e = gyn

Table S2: RF confusion matrix for Brst vs Derm vs GI vs GU vs Gyn comparison under class-balanced subsampling, showing the enrichment of GI false positives has been reduced.

S3.3.1 5-way tissue classification (balanced)

s Our Random Forest predicts if an image is one of five possible tissue types (Fig s10 and S12) after sampling tissues in a balanced manner. There were 580 images, with 116 images from each of the five tissue types: breast, dermatological, gastrointestinal, genitourinary, and gynecological. Classes were exactly balanced. At 580 images, this was 56% of the amount of data we used for unbalanced sampling in the prior task, namely 1042 images. Accuracy is $47.724 \pm 5.057\%$ (chance 50%). AUROC is 0.784 ± 0.053 (chance 0.5).

Moreover, the confusion matrix (Table S2) suggests this class balanced sampling reduces the enrichment of GI false positives. However, this class-balanced subsampling comes at a cost of slightly reduced AUROC, to the the extend that AUROC is now below 0.8.

S4 Future direction details

The first step is to expand the size of this dataset by recruiting more pathologists via social media. With more data, we hope to improve performance on discriminations that were the most difficult (e.g., those involving gynecological pathology). More data may facilitate machine learning methods that discriminate between similar but less frequently used stains, such as H&E vs Diff-quik, rather than H&E vs IHC. More data might also enable us to distinguish particular organs or tissues within a histopathology tissue type, e.g. distinguish kidney tissue from bladder tissue. With increased sample size and increased tissue of origin granularity, it may be possible to predict metastatic tissue of origin. Finally, a larger dataset might also include more rare cases that can be useful for machine learning techniques that can support diagnoses.

A second step is advocacy on social media, for (i) sharing normal tissue data, and (ii) expanded pathology hashtags. Normal tissue complements our existing "relatively unimportant" artifact and foreign body data, such as colloids and gauze, which are typically not prognostic of disease. Normal tissue also complements the description of tissue morphology in our data, if we tend to have only cancerous or diseased tissue. Separately, more descriptive hashtags may reduce our manual annotation burden, and obviate the need for us to ask pathologists to clarify what stain was used or what the tissue is. Moreover, molecular hashtags may complement the histology we see. However, we understand that for pathologists sharing cases on social

media is probably a fun and voluntary activity, rather than a rigorous academic endeavor, so it may not be appropriate for us to suggest pathologists use terms from synoptic reporting in hashtag format in their tweets. Moreover, the size of tweets is limited to 280 characters, so more than 3-4 hashtags per tweet is probably infeasible. Some pathologists are already close to this limit without using additional hashtags.

We encourage the adoption of hashtags that explicitly identify what stains or techniques are used (this is not an exhaustive list):

1. #he indicates there are one or more H&E-stained images in the tweet.
2. #ihc indicates there are one or more IHC-stained images in the tweet.
3. #pas indicates there are one or more periodic acid-Schiff images in the tweet.
4. #diffq indicates there are one or more diff-quik images in the tweet. There is a common misspelling of diff-quick, so our hashtag avoids this misspelling.
5. #gross indicates one or more gross section images are in the tweet.
6. #endo indicates one or more endoscopy images are in the tweet.
7. #ct indicates one or more CT scan images are in the tweet.
8. #xray indicates one or more X-ray images are in the tweet.

We encourage the adoption of hashtags that explicitly identify any artifacts, art, or pathologist annotations/marks on the image.

1. #artifact or #artefact indicates there are artifacts or foreign bodies in one or more images, such as colloids, barium, sutures, Spongostan^{TM}, gauze, etc. We encourage the tweet message text to identify what the artifact or foreign body is.
2. #pathart is a hashtag in use today, but unfortunately it is used in two ways: (i) to identify naturally-occurring and unmodified pathology images that are "pretty" or "interesting" as natural works of art, and (ii) to identify images that have been modified by the pathologist herself/himself to be "funny" or "interesting". The trouble is (i) is "acceptable" pathology for analysis while (ii) is not. We advocate for the continued use of the #pathart hashtag, but with clarification, below:
3. #drawn or #annotated indicates the pathologist made hand-drawn marks on one or more images, such as arrows, circles, or artistic manipulations. Artistic manipulations may include drawing exclamation points, question marks, eyes, mouths, faces, skulls, cartoon bodies, etc on the image. So, "#pathart #drawn" is likely a pathology image with artistic drawn marks that prevents the image from being an "acceptable" pathology image for analysis, while "#pathart" without "#drawn" is likely a pathology image that is a naturally occurring unmodified histology image that is an "acceptable" pathology image for analysis.

We encourage the adoption of hashtags that give other information about the image.

1. #pathbug is an existing hashtag that indicates a parasite or other co-occurring non-human organism is depicted in one or more images in the tweet.
2. #panel indicates one or more multi-panel images are in the tweet.

We encourage hashtags to describe not only the histological features of a case, but also the molecular features of a case. Again, this hashtag list is far from exhaustive.

1. #braf indicates the BRAF gene is known to be mutated, perhaps through sequencing.
2. #msi indicates micro-satellite instability, which again may be evident from sequencing.
3. #desmin indicates that the IHC used targets desmin.

A third important future direction is to determine whether our machine learning performance can be improved, perhaps by use of advanced methods such as deep learning.

S4.1 Caveats details

A feature of these data is that a particular patient might be represented in more than one image. Any given patient might be discussed in multiple tweets, each of which contain one or more images. So far, our machine learning analyses have not controlled for the number of images shared for a particular patient. Future machine learning analyses can take at most one image per patient, to avoid overfitting to the peculiarities of a particular patient who is the subject of multiple images in a given class discrimination problem. However, it will be challenging to determine which images belong to which patient, because often other cases are mentioned alongside a particular patient, to provide context or comparison. Thus a Twitter thread for a particular case might include more images from that same patient as well as images from different patients. Sorting images into different patients will therefore be a challenge that will require perhaps hundreds of hours of manual curation.

Our dataset is only as good as the accuracy of the hashtags and diagnoses made by the contributing pathologists. The more pathologists that contribute to the database, the higher the risk for errors and inconsistencies. Indeed we note some uses of the #bstpath hashtag to describe breast pathology (Section S2.4). We should remember the fun and voluntary nature of sharing cases on social media.

Finally, the size of the dataset is both a blessing and a curse. A large and diverse dataset is required to provide the most benefit to computational pathology. However, quality control for such large datasets is most feasible if done automatically, and automated quality control cannot deal with all issues. For example, some pathology images include marks designating a particular pathologist as the contributor of that image. Other pathology images have been marked by pathologists with arrows and circles. Our automated quality control pipeline enables us to rapidly discriminate pathology from non-pathology images, but is not able to address these other challenges. Future steps will need to be taken for more specialized quality control.