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*

October 7, 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, LGP, 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, AM, SJS. Supervision: MA, TJF. Project administration: AJS, WJ, SJC, MA, TJF. Funding acquisition: AJS, TJF.

activities, e.g. discriminating classes of vehicles. In computational pathology, such machine learning algorithms are applied 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 2,750 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 (i) accurately identify different stains, (ii) discriminate between five tissue types, and (iii) differentiate nontumoral, benign/low grade malignant potential [low grade], and malignant diseases. Using a Random Forest, we report (i) 0.960 ± 0.012 Area Under Receiver Operating Characteristic [AUROC] when differentiating between human hematoxylin and eosin [H&E] stained microscopy images from all other types of images e.g. natural scenes, and (ii) 0.996 ± 0.003 AUROC when distinguishing H&E from immunohistochemistry [IHC] stained microscopy images. Though a Support Vector Machine found color features to be important, a Random Forest surprisingly found texture features to be important, for these stain tasks. Additionally, we distinguish all pairs of breast, dermatological, gastrointestinal, genitourinary, and gynecological pathology tissue types, with mean AUROC for any pairwise comparison ranging from 0.783 to 0.873. We report 0.803 ± 0.059 AUROC when all five tissue types are considered in a single multiclass classification task. Finally, for our most difficult and clinically relevant task of distinguishing low grade from malignant tumors, we report 0.703 ± 0.058 AUROC, which marginally drops to 0.683 ± 0.056 for the 3-class classification task of distinguishing nontumoral diseases, low grade tumors, and malignant tumors. We hope this inspires other groups to use our dataset, to (i) improve performance, (ii) build upon our definition of nontumoral, low grade, and malignant in terms of diagnosis text keywords, or (iii) use our data as an independent test set for nontumoral, low grade, and malignant disease classification tasks. We provide a tool, called the Interactive Pathology Annotator, for pathologists and data scientists to browse, search, and validate the dataset. 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 [6] has millions of human-labeled images; CIFAR-10 [13] [Canadian Institute for Advanced Research] has thousands. Machine learning methods for natural vision tasks routinely use datasets like these to benchmark performance, and transfer learned representations to other tasks, such as pathology [2, 16, 25]. However, computational pathology [7] 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 studies are larger, such as the TUPAC16 [TUmor Proliferation Assessment Challenge] dataset of 821 cases [28] – all 821 cases being whole slide images from The Cancer Genome Atlas [TCGA][2]. TCGA has tens of thousands of whole slide images available in total, but these images are only hematoxylin and eosin [H&E] stained slides.

Figure 1: Thirteen pathologists over 8 countries donated cases for our study. They also answered questions that arose during manual case annotation procedures (Table 1).

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 [5, 8, 18]. 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

[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 2: *At left*: Pathologist (author S.R.A.) discusses a case. Without mentioning the diagnosis himself, he confirms the diagnosis suggested by a second pathologist, i.e. cystadenofibroma, which we explicitly annotate. *At right*: Our Interactive Pathology Annotation [IPA] tool displays an image from this case, in the context of the Tweet overall. IPA is a portal for pathologists to (i) browse Tweets and images in the dataset; (ii) validate our data annotations; (iii) check our tissue type categorization algorithm results, (iv) check our nontumor, low grade, and malignant categorization algorithm results; (v) search Tweets for specific keywords or diagnoses; (vi) filter out all cases except those from a specific pathologist; and (vii) click the link to the original Tweet on Twitter for context.

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's social media post, per Sec 2.1.1.1.
4.1.	Online question	We ask pathologists to clarify social media post (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.

dataset, which includes a variety of sections and techniques, ranging from immunohistochemistry [IHC] to fluorescence *in situ* hybridization [FISH], and a range of tissues, with linked annotations by pathologists. This initial dataset includes images donated by 13 pathologists, from 13 institutions in 8 countries (Fig 1). We annotated 2,750 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.

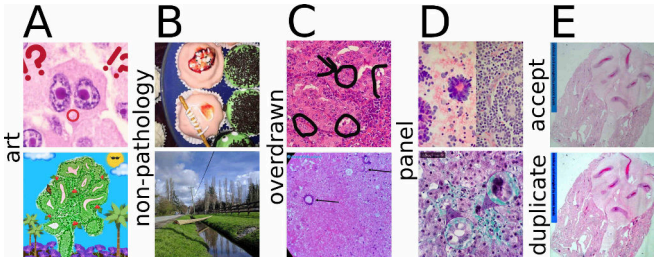

Figure 3: Examples of images that are rejected, because they are not pathology images that a pathologist would see in clinical practice. *Panel A (top M.P.P., bottom B.D.S)*: "art" rejects. *Panel B (top B.S.P., bottom S.Y.)*: "non-pathology" rejects. *Panel C (top B.X., bottom A.M.)*: "overdrawn" rejects. *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 (see Sec 2.1.1 for definition), bottom is "dup" [duplicate] rejection.

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 an acceptable H&E-stained human microscopy image or not, (ii) predicting if a microscopy image is H&E-stained or IHC-stained, (iii) predicting the histopathology tissue

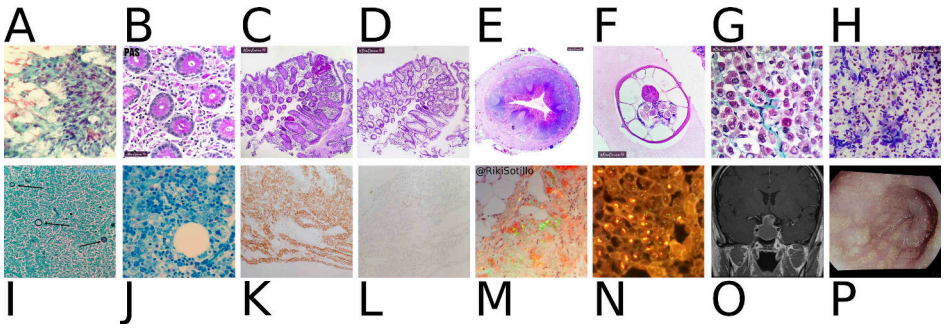

Figure 4: Our dataset includes diverse techniques. Initials indicate image ownership. *Panel A (R.S.S.)*: Papanicolaou stain, i.e. pap smear. *Panel B (L.G.P.)*: Periodic acid-Schiff [PAS] stain, glycogen in pink. *Panel C (L.G.P.)*: PAS stain at lower magnification. *Panel D (L.G.P.)*: Hematoxylin and 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 S1.2.1), 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.

type of an image, and (iv) predicting if an image shows nontumoral, benign/low grade malignant potential [low grade], or malignant disease.

We also provide a tool, called IPA, for pathologists and data scientists to browse, search, and validate the dataset (Figs 2 and S13). 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.

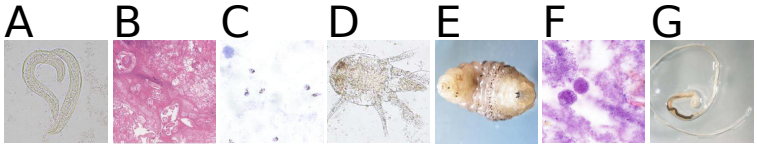

Figure 5: Our dataset includes diverse parasitology samples. *Panel A (B.S.P.)*: *Strongyloides stercoralis*, light microscopy. *Panel B (B.S.P.)*: *Dirofilaria immitis*, in human, H&E stain. *Panel C (B.S.P.)*: *Plasmodium falciparum*, in human, Giemsa stain. *Panel D (B.S.P.)*: Incidental finding of 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 Materials and Methods

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 infrequently shares tissues, e.g. neuropathology. In step 2, we 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 posted cases. In step 4, we manually annotate these posted cases for acceptability (if overdrawn, corrupt, duplicate, multi-panel, art, or non-pathology rejecting per Fig 3 and Sec 2.1.1), technique (Fig 4), species (Fig 5 and Fig 6A,B,E), and private status (e.g. 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 2.1 and 2.1.1.1. Moreover, in step 4, if the

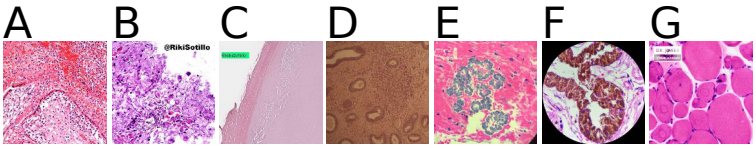

Figure 6: Our dataset includes diverse H&E-stained slide microscopy images. *Panel A (S.R.A)*: Acute villitis due to septic *Escherichia coli*. *Panel B (R.S.S.)*: Garlic. *Panel C (R.S.S.)*: "Accellular" leiomyoma after ulipristal acetate treatment. *Panel D (R.S.S.)*: Brownish appearance from dark 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.

nontumor/low-grade/malignant status in a Tweet is not clear, we read the Twitter discussion thread for this case and manually annotate the case appropriately if possible. 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. author S.J.S.) for an opinion (step 4.2).

IPA (Figs 2 and S13) is an important part of step 4, where pathologists validate tissue and disease categorization. In step 5, we aggregate all data from all pathologists and apply machine learning to make predictions. These steps were repeated as more pathologist collaborators were identified (Fig 1). We aimed to have thousands of images for a large-scale machine learning task, and with 13 pathologists we have over 2,000 images.

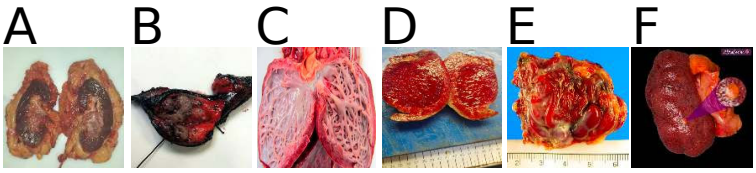

Figure 7: Gross sections are represented in our dataset, putting the slide images in context. *Panel A (M.P.P)*: Urothelial carcinoma. *Panel B (M.P.P.)*: Lung adenocarcinoma. *Panel C (S.R.A.)*: Barth syndrome. *Panel D (N.Z.)*: Enlarged spleen. *Panel E (S.R.A.)*: Arteriovenous malformation. *Panel F (L.G.P.)*: Kidney adrenal heterotopia.

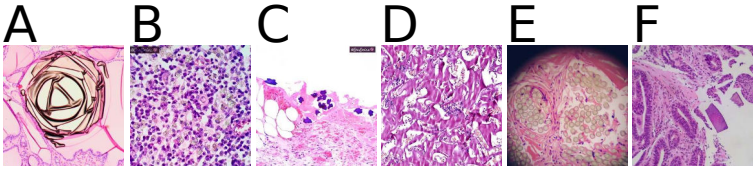

Figure 8: Our dataset includes artifacts and foreign bodies which machine learning should not consider prognostic. All panels human H&E. *Panel A (B.X.)*: Colloid. *Panel B (L.G.P.)*: Barium. *Panel C (L.G.P.)*: Oxidized regenerated cellulose, a.k.a. gauze, granuloma may mimic mass lesion [27]. *Panel 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.

2.1 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 (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 4) – even variety for a single stain, e.g. H&E stains[3] (Fig 6). Section S1.2.1 discusses intra-stain variability. Our dataset includes gross sections (Fig 7) that pathologists share alongside images of stained slides. In addition to variation in the signal of interest (i.e., stain, tissue, or disease), we find variability in the noise (i.e. pathology artifacts, Fig 8). 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 and other [micro]organisms (Fig 5, and Fig 6A,E).

2.1.1 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].

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

For our manual data curation process, we defined several rejection criteria (Fig 3), detailed in Section S2.1. Figure 3A shows examples of images rejected as "art", because they are artistically manipulated H&E pathology microscopy images. Figure 3B shows examples of images rejected as "non-pathology", e.g. parasitology-

[3]H&E stain composition may vary by country – e.g. in France, H&E typically includes saffron, which stains collagen fibers. This helps differentiate between connective tissue and muscle, or to see cell cytoplasm better against a fibrous background. This stain may be referred to as "HES", and we consider it H&E.

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 who primarily used their accounts for sharing and discussing pathology. Figure 3C 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 3D 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 3 Panel E *top* is an acceptable H&E-stained pathology image. Figure 3 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.

2.1.2 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 9) are available in Apache LiRE [17]. These features, and their dimension counts, are as follows: CEDD (144) [3], Color Correlogram (256) [11], Color Histogram (64) [17], FCTH (192) [4], Gabor (60) [17], Local Binary Patterns (256) [19], Local Binary Patterns Pyramid (756) [20], PHOG (630) [1], Rotation Invariant Local Binary Patterns (36) [20], and Tamura (18) [26].

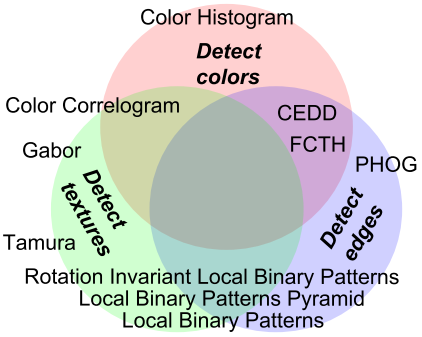

Figure 9: We use a variety of color, texture, and edge features for baseline machine learning analyses. Some features, such as color histograms, detect only color. Other features, such as Color Correlograms, detect both colors and textures. Pyramid features are scale-invariant.

2.2 Text data overview

For supervised learning, we analyze a Tweet's text with keyword-based matching, to determine the proper labels for the Tweet's images. The text and included hashtags may indicate (i) tissue type, or (ii) nontumor, low grade, or malignant disease.

2.2.1 Tissue type categories from text

Prior work has discussed pathology-related hashtags as a way to make pathology more accessible on social media[4] [21]. Pathologists use hashtags to indicate histopathology tissue types, such as "#gynpath" to indicate gynecological pathology (Fig 10). 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. As detailed in Section S2.5, we used hashtags and keywords for all Tweets in a message thread to identify the five most common tissue types, finding 57 breast Tweets, 78 dermatological Tweets, 172 gastrointestinal Tweets, 58 genitourinary Tweets, 108 gynecological Tweets.

2.2.2 Nontumor, low grade, and malignant categories from text

We define three broad disease state categories (Fig S28) to use as labels for supervised learning. Our "nontumor" category includes normal tissue, injuries, and nontumoral diseases, e.g. Crohn's disease, herpes simplex infection, and myocardial infarction. Our "malignant" category includes all malignant disease, including carcinoma, blastoma, sarcoma, lymphoma, and metastases. Our definition of malignancy in epithelial cancers is

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

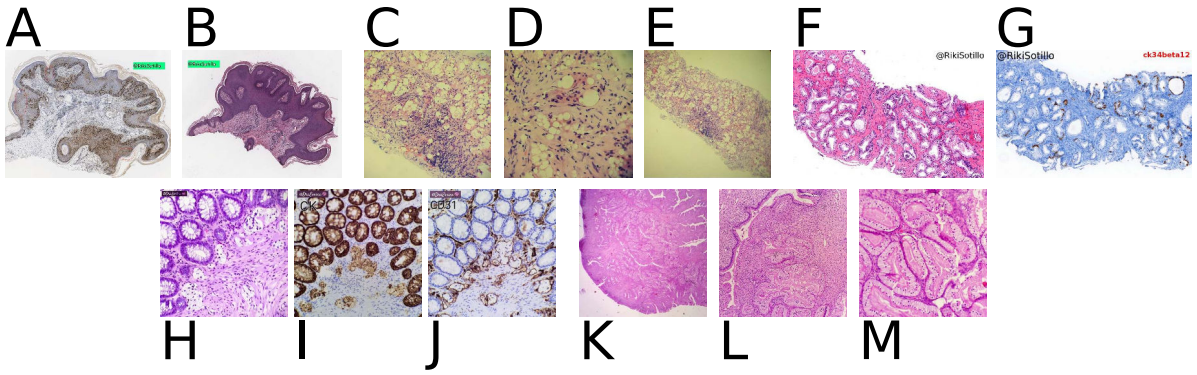

Figure 10: We use machine learning to distinguish five histopathological tissue types: dermatological, breast, genitourinary, gastrointestinal, and gynecological. *Panels A,B (R.S.S.)*: IHC- and H&E-stained dermatological pathology at low magnification, showing hallmark layering of epidermis, dermis, 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,L,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.

the ability to breach the basement membrane, i.e. a malignant tumor escapes containment and is therefore no longer treatable with surgical resection. Our "benign/low grade malignant potential" [low grade] category is then all tumors or pre-cancer/neoplastic lesions that are not yet invasive/malignant, e.g. hamartomas, carcinoid tumors, adenomas, and carcinoma *in situ*. These three categories naturally split the data into three portions of similar size, i.e 309-385 images per portion (Table 2). Details in Sec S4.

For the nontumoral vs low grade vs malignant task (Sec 3.4), text processing was more complicated because (i) of a heavy reliance on diagnosis keyword matching (flowchart in Fig S29), and (ii) additional per-Tweet and per-image annotations to clarify nontumor/low-grade/malignant state, which may involve feedback from a pathologist. Details in Sec S4.

2.3 Machine learning methods

We used a variety of baseline machine learning methods (Fig S14), to test whether more complex machine learning methods perform significantly better than simpler machine learning methods. These methods are discussed in Section S2.7. 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.

We use Weka version 3.8.1 [9] on an ASUS Intel 4-CPU laptop with 16 GB RAM. Section S2.8 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 or different histochemistry stains. 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

Task	n^{total}	n^-	n^+	ZR acc. %	RF accuracy %	ZR AUROC	RF AUROC
Acceptable H&E	2325	1153	1172	50.409 ± 0.151	91.380 ± 1.687	0.5	0.960 ± 0.012
Accept H&E (bal)	1506	753	753	49.801 ± 0.163	89.502 ± 2.293	0.5	0.954 ± 0.014
H&E vs IHC	1351	1174	177	86.899 ± 0.336	97.173 ± 1.209	0.5	0.996 ± 0.003
Breast vs Gyn	381	135	246	64.946 ± 0.848	71.871 ± 6.292	0.5	0.783 ± 0.082
Derm vs Breast	303	168	135	55.452 ± 1.329	74.088 ± 6.994	0.5	0.832 ± 0.069
Derm vs Gyn	414	168	246	59.814 ± 0.584	75.986 ± 5.158	0.5	0.847 ± 0.062
GI vs Breast	483	348	135	72.054 ± 0.892	77.978 ± 3.974	0.5	0.873 ± 0.050
GI vs Derm	516	348	168	67.443 ± 0.645	76.198 ± 5.639	0.5	0.854 ± 0.059
GI vs Gyn	594	348	246	58.192 ± 0.284	73.338 ± 5.495	0.5	0.815 ± 0.053
Breast vs GU	252	135	117	53.569 ± 1.746	74.791 ± 7.968	0.5	0.822 ± 0.081
Derm vs GU	285	168	117	58.953 ± 1.288	77.273 ± 7.203	0.5	0.871 ± 0.070
GI vs GU	465	348	117	74.843 ± 0.845	78.930 ± 2.670	0.5	0.830 ± 0.071
Gyn vs GU	363	246	117	68.131 ± 0.864	76.462 ± 4.066	0.5	0.795 ± 0.078
LowGrade vs Malignant	732	347	385	52.595 ± 0.599	65.055 ± 5.159	0.5	0.703 ± 0.058
Nontumor vs Malignant	694	309	385	55.474 ± 0.464	65.744 ± 5.131	0.5	0.700 ± 0.066
Nontumor vs LowGrade	656	309	347	52.895 ± 0.455	64.493 ± 5.536	0.5	0.704 ± 0.059
Nontumor vs Low+Mal	1041	309	732	70.317 ± 0.293	73.046 ± 2.188	0.5	0.683 ± 0.062
Nontum+Low vs Malig	1041	656	385	63.016 ± 0.459	66.551 ± 3.255	0.5	0.687 ± 0.052

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 Sections S3 and S4.

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 S1.2.1). 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 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 Sec 3.2 and 3.3, we show statistically significant performance, with room for improvement. We consider five tissue types: breast, dermatological [derm], gastrointestinal [GI], genitourinary [GU], and gynecological (Fig 10). In Sec 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 (Sec S1.2.2). In Sec 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 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[5].

Fourth, can machine learning distinguish nontumoral, low grade, and malignant disease – in acceptable H&E human tissue microscopy images (Sec 3.4)? Pathologists routinely answer this clinically important question. This is our most difficult question, and our Random Forest baseline's low though statistically

[5]A case of this from K.H. is `https://twitter.com/Ho_Khanh_MD/status/999989201734197250` (Fig S13)

significant performance (Table 2) is an open invitation to the field for improved methods on this task, such as deep learning. Section S4 discusses the keywords we use on the Tweet text to determine if an image is labeled nontumoral, low grade, or malignant.

3.1 Stain-related tasks

Because stain-related tasks had strong performance, i.e. ˜0.9 AUROC or more, we additionally interpreted what the machine learning models learned (Sec S3.1.2). Surprisingly, texture features were most important to a Random Forest. Intuitively, color features were most important to a Support Vector Machine.

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 2325 images: 1153 negative images that were not acceptable and 1172 positive images that were acceptable. Classes were essentially balanced. Accuracy is 91.380±1.687% (chance 50.409±0.151%). AUROC is 0.960 ± 0.012 (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.

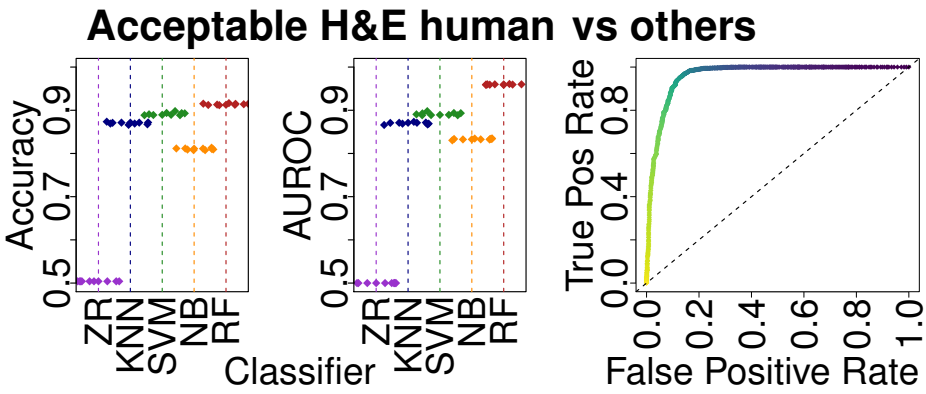

Figure 11: Predicting if an image is acceptable H&E human tissue or not. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig S14). The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.9600 here for n=2325 per Table 2.

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 S16). There were 1351 images: 1174 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 $97.173 \pm 1.209\%$ (chance $86.899 \pm 0.336\%$). AUROC is 0.996 ± 0.003 (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[6]. 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), and accept multi-panel images (Fig 3D) here because the panels describe the same tissue. For dermatological [Derm], breast, gastrointestinal [GI], genitourinary [GU], and gynecological [Gyn] types, we consider all pairwise comparisons;

[6]Section S1.2.1 has more discussion on IHC color variability.

Task	n^{total}	$n^{type} + ... + n^{type}$	ZR acc. %	RF acc. %	ZR AUROC	RF AUROC
		$n^{brst} + n^{derm} + n^{gi} + n^{gu} + n^{gyn}$				
5 tissues	1014	$135 + 168 + 348 + 117 + 246$	34.320 ± 0.438	50.827 ± 4.075	0.5	0.803 ± 0.059
5 tiss (bal)	585	$117 + 117 + 117 + 117 + 117$	18.805 ± 0.162	48.935 ± 6.089	0.5	0.786 ± 0.058
		$n^{nontumor} + n^{low\text{-}grade} + n^{malig}$				
3 diseases	1041	309 + 347 + 385	36.984 ± 0.459	51.239 ± 4.781	0.5	0.683 ± 0.056

Table 3: Random Forest [RF] machine learning analysis results for (i) 5-class tissue classification tasks, to predict if an image shows Breast, Derm, GI, GU, or Gyn tissue; and (ii) 3-class disease classification task, to predict if an image shows nontumoral, low grade, or malignant disease. 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 a particular class, e.g. the dermatological [derm] class, when majority of RF trees vote derm, 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.

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 Breast, #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 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 distinguish all five histopathology tissue types (Fig 10) simultaneously in one learning task (Fig 12 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, breast, gastrointestinal, genitourinary, or gynecological 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 and realistic because there are five possible tissue types to predict rather than two.

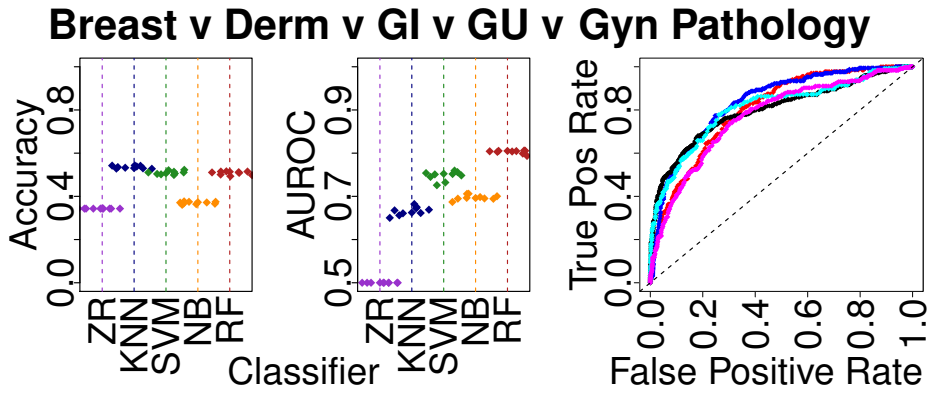

Figure 12: Predicting if an image is Breast, Derm, GI, GU, or Gyn. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig S14). The ROC curve for the highest AUROC classifier [RF] is shown at right. In the ROC plot (right), **Breast** is red (AUROC=0.8036, n=135), **Derm** is blue (AUROC=0.8352, n=168), **GI** is black (AUROC=0.7996, n=348), **GU** is cyan (AUROC=0.8078, n=117), and **Gyn** is magenta (AUROC=0.7701, n=246), with performance details in Table 3. For this trial, the weighted mean of AUROCs is 0.800, which is below the mean of 0.803 (Table 3), but within a standard deviation (0.059). 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.

3.3.1 5-class tissue classification

Our Random Forest predicts if an image is one of five possible tissue types (Figs 10 and 12). There were 1014 images: 135 breast, 168 dermatological, 348 gastrointestinal [GI], 117 genitourinary, and 246 gynecological. Classes were imbalanced (roughly one third GI images). Accuracy is $50.827 \pm 4.075\%$ (chance $34.320 \pm$

0.438%). AUROC is 0.803 ± 0.059 (chance 0.5). However, the confusion matrix (Table S6) 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.

3.4 Nontumoral, Low grade, and Malignant tasks

Our Random Forest predicts if a microscopy slide image is nontumor, low grade, or malignant per Sec 2.2.2. There were 1041 images: 309 images that were normal/nontumoral, 347 images that were low grade, and 385 images that were malignant. Classes were essentially balanced. We were interested in the low grade vs malignant comparison for its clinical relevance, so for this we ignored nontumoral images. For this binary classification task, accuracy is $65.055 \pm 5.159\%$ (chance $52.595 \pm 0.599\%$) and AUROC is 0.703 ± 0.058 (chance 0.5) (Fig S30, Table 2). Results for all pairwise comparisons of nontumor, low grade, and malignant detailed in Sections S4.2 and S4.3, including nontumor vs low-grade/malignant. For the 3-class classification task where nontumor, low grade, and malignant images must be distinguished simultaneously, accuracy is $51.239 \pm 4.781\%$ (chance $36.984 \pm 0.459\%$) and AUROC is 0.683 ± 0.056 (chance 0.5) (Fig S35, Tables 3 and S8). Though the Random Forest demonstrated statistically above chance accuracy and AUROC, performance is far too weak for clinical consideration. We provide this task as an open challenge to the field of computational pathology, and as an independent test set for machine learning researchers. Properly answering these questions, particularly low grade vs malignant, has applications for clinical decision support.

4 Discussion

We mined social media to obtain pathology images shared by pathologists worldwide, and organized them into a diverse dataset that can be used to rigorously test computational pathology methods. See Lepe [15] for social media pathology collaboration. We report 0.954 ± 0.014 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 ± 0.003 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 distinguish all pairs of breast, dermatological, gastrointestinal, genitourinary, and gynecological pathologies, with AUROC ranging from 0.783 to 0.873. Breast vs gynecological 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 (0.832) across binary classifications. We report 0.803 ± 0.059 AUROC when all five tissue types are considered in a five-class classification task.

We also distinguish all pairs of nontumoral, low grade, and malignant disease states, with mean AUROC ranging from 0.700 to 0.704. When we grouped low grade with malignant, or low grade with nontumoral, mean AUROC was marginally lower at 0.683 and 0.687, respectively. Due to its clinical implications, we were most interested in distinguishing low grade from malignant, where nontumoral images are ignored, and we report 0.703 ± 0.058 AUROC for this task. In a three-class classification task to simultaneously distinguish nontumoral, low grade, and malignant, we report 0.683 ± 0.056 AUROC. Though our findings are statistically significant, they are far too weak for clinical application, so we leave these tasks as defined, open, and clinically relevant computational pathology questions for this dataset.

Section S5 discusses future directions of this study. Section S6 discusses caveats of this study.

5 Conclusion

We mined social media to obtain, curate, and perform baseline 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. thanks Dr. Sanjay Mukhopadhyay of Cleveland Clinic for pathology discussion and suggesting a "benign vs malignant" task. A.J.S. thanks Dr. Joanna Cyrta of Institut Curie for H&E-saffron [HES] discussion.

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.

S.Y. is a consultant and advisory board member for Bayer, receiving an honorarium and travel allowance.

T.J.F. is a founder, equity owner, and Chief Scientific Officer of Paige.AI.

References

[1] 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, Nov. 2009. ISBN 1522-4880. doi: 10.1109/ICIP.2009.5413938. URL `http://dx.doi.org/10.1109/ICIP.2009.5413938`.

[2] N. Bayramoglu and J. Heikkila. Transfer Learning for Cell Nuclei Classification in Histopathology Images. pages 532–539. Springer International Publishing, 2016. ISBN 978-3-319-49409-8. doi: 10.1007/978-3-319-49409-8_46. URL `http://dx.doi.org/10.1007/978-3-319-49409-8_46`.

[3] 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`.

[4] 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`.

[5] 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`.

[6] 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`.

[7] 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`.

[8] J. Gardner and T. Allen. Keep Calm and Tweet On: Legal and Ethical Considerations for Pathologists Using Social Media. *Archives of pathology & laboratory medicine*, Aug. 2018. ISSN 1543-2165. URL `http://view.ncbi.nlm.nih.gov/pubmed/30132683`.

[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] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. Dec. 2015. URL `http://arxiv.org/abs/1512.03385`.

[11] 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`.

[12] 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`.

[13] 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`.

[14] 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`.

[15] M. Lepe, R. Wu, P. Oltulu, D. Alex, M. Canepa, A. Deeken, E. Doxtader, V. Fitzhugh, J.-B. Gibier, D. Jain, N. Janaki, A. Jelinek, S. Kumar, T. Labiano, V. L'Imperio, C. Michael, S. Mukhopadhyay, F. Pagni, A. Panizo, L. Pijuan, L. Quintana, S. Roy-Chowdhuri, A. Sanchez-Font, I. Valero, J. Sauter, D. Skipper, L. Spruill, V. Torous, J. Gardner, and X. Jiang. #EBUSTwitter: Novel Use of Social Media for Conception, Coordination and Completion of an International, Multi-Center Pathology Study. *Journal of the American Society of Cytopathology*, 7(5):S88–S89, Sept. 2018. ISSN 2213-2945. doi: 10.1016/j.jasc.2018.06.015. URL `http://dx.doi.org/10.1016/j.jasc.2018.06.015`.

[16] 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`.

[17] 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`.

[18] 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`.

[19] 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, Oct. 1994. doi: 10.1109/ICPR.1994.576366. URL `http://dx.doi.org/10.1109/ICPR.1994.576366`.

[20] 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, July 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1017623. URL `http://dx.doi.org/10.1109/TPAMI.2002.1017623`.

[21] 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`.

[22] 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`.

[23] 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`.

[24] 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.

[25] 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.

[26] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, June 1978. ISSN 0018-9472. doi: 10.1109/TSMC.1978.4309999. URL http://dx.doi.org/10.1109/TSMC.1978.4309999.

[27] 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.

[28] 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.

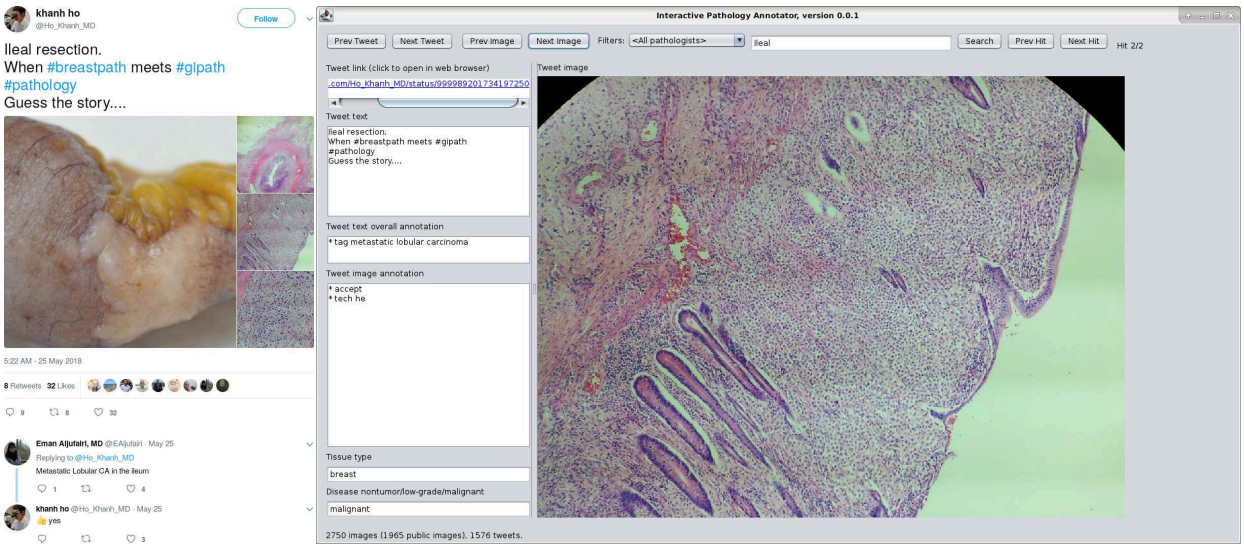

Figure S13: *At left*: Pathologist (author K.H.) discusses case. Without mentioning the diagnosis himself, he confirms diagnoses suggested by other pathologists, i.e. lobular breast carcinoma metastasized to ileum, which we explicitly annotate. *At right*: IPA shows that our tissue type categorization algorithm categorizes this Tweet as breast pathology rather than gastrointestinal. The primary tumor is in breast. We define the tissue classification task this way to have applications for tumor site of origin prediction.

Supporting Information

S1 Supplementary discussion

S1.1 Interactive Pathology Annotator discussion

For completeness, we show another example of the use of our Interactive Pathology Annotator [IPA] tool (Fig S13). This is a case of metastatic disease, from breast to gastrointestinal tissue, showing a diffuse pattern of lobular carcinoma that is more common in breast.

S1.2 Data diversity discussion

S1.2.1 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 instance 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.

S1.2.2 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 S2.6. 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 2.1.1.1 – 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[7] 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.

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

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[8]. 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 3 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 [16] typically require 227x227px (i.e. AlexNet [14] or CaffeNet [12]) or 224x224px (i.e. ResNet [10]) images, and we have used these sizes in the past [22, 23]. We note the Inception [24] family of deep convolutional neural networks takes a 299x299px image input, which is larger than 256x256px and is also frequently used in computational pathology [16]. Ideally, each image would have ROIs and hand-drawn arrows/circles annotated at the pixel level, so each image could be 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 3 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 accommodate 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.

[8]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.

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 Tissue hashtags and keywords

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 S2.6. Further, if a Tweet's tissue type could not be determined by hashtags and keywords, we assigned the tissue type of any other Tweet in the message thread of Tweets. For example, if a Tweet of unknown tissue type were a reply to a Tweet of known genitourinary type, then we considered both Tweets to be genitourinary. After keyword-based and message-thread-based expansion, there were **172 gastrointestinal** Tweets, **78 dermatological** Tweets, **108 gynecological** Tweets, **57 breast** Tweets, and **58 genitourinary** Tweets. We do not count Tweets that have zero images usable for the tissue type discrimination task.

S2.6 Regular-expression-based tissue type keywords

Expanding our text processing discussion in Section 2.2.1, 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/ior` `/intestin(?:e|al)/i` or `/\banal\b/i` or `/perianal/ior` `/perine(?:um|al)/i` or `/esophag(?:us|eal|itis)/i` or `/ileum/i` or `/gall\s?(?:bladder|stone)/i` or `/liver/i` or `/ascaris/i` or `/pancrea(?:s|tic)/` or `/colitis/i` or `/hepat[ieo]c/i` or `/cholecystitis/i` or `/crohn/i` or `/ca?ec(?:um|al)` or `/jejunum/i`

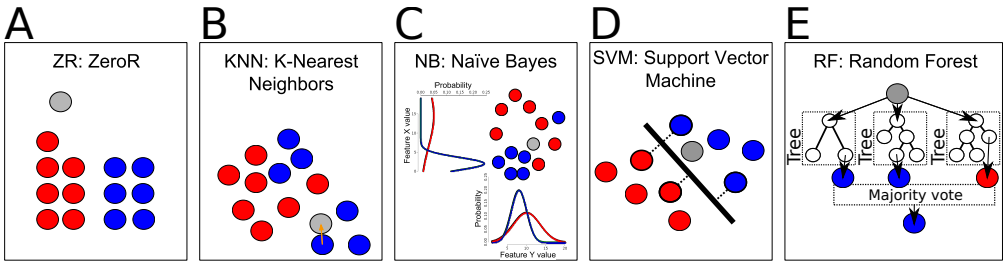

Figure S14: Machine learning methods for baseline analyses included ZeroR (Panel A), K-nearest neighbors (Panel B), Naïve Bayes (Panel C), an SMO-based support vector machine (Panel D), and Random Forest (Panel E). Visual schematics of each method shown. Weka provided all methods.

  - This matches colon, duodenum, duodenal, anal, rectal, cecum, 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, 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 uterine, ovarian, fallopian, endometrial, and other gynecological-related keywords.

Section S4 discusses text matching for nontumoral, low grade, and malignant discrimination.

S2.7 Machine learning methods discussion

Expanding on our discussion of machine learning methods in Section 2.3, ZeroR is the simplest method (Fig S14), 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. Both SVM and RF performed well on stain tasks, and in Sec S3.1.2 we interpreted the concepts they learned.

S2.8 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 S14) 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

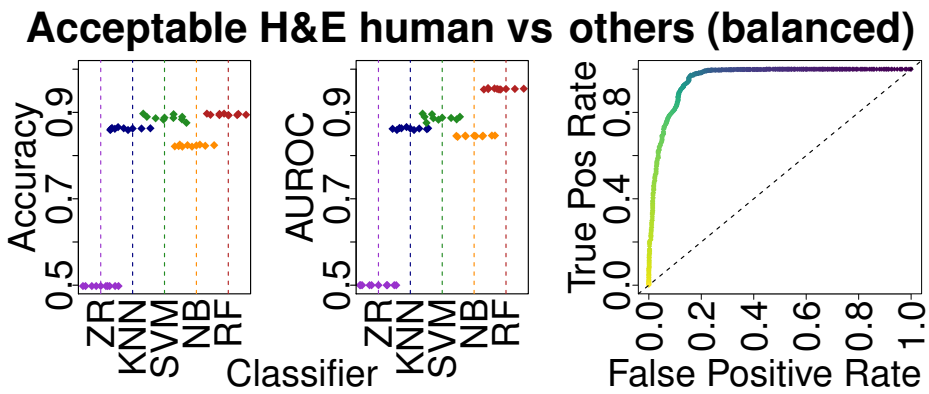

Figure S15: Predicting if an image is acceptable H&E human tissue or not, in a pathologist-balanced and class-balanced manner, for comparison to Fig 11. Here, AUROC=0.9545 for n=1506 per Table 2.

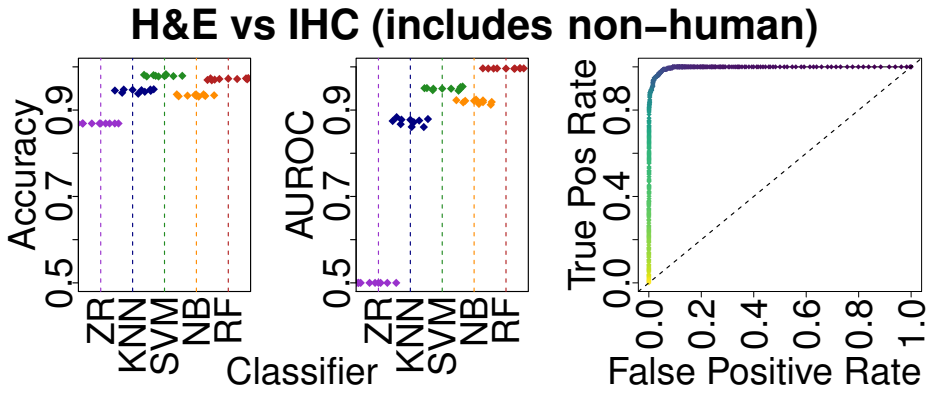

Figure S16: Predicting if an image is H&E or IHC. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in (Fig S14). The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.9960 here for n=1351 per Table 2.

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.

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 S15), 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 1506 images (64.8% of the 2325 images total for the unbalanced case, per Table 2): 753 negative images that were not acceptable and 753 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. Classes were exactly balanced. However, we find accuracy is $89.502 \pm 2.293\%$ (chance $49.801 \pm 0.163\%$) and is not significantly worse than the prior non-balanced analysis accuracy of $91.380 \pm 1.687\%$. Moreover, AUROC here is 0.954 ± 0.014 (chance 0.5), which again is not significantly worse that the prior non-balanced analysis AUROC of 0.960 ± 0.012. 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.

S3.1.2 Feature importance for stain comparisons

One way to interpret what a Random Forest has learned is to compute the Mean Decrease in Gini Impurity [MDI] of each feature used by the Random Forest[9] [RF]. Highly important features have higher MDI. MDI for various learning tasks shown in Table S4. Because the RF demonstrates such strong performance for these stain tasks (˜0.9 AUROC or more), these MDI feature importances may be especially meaningful – more meaningful that feature importance interpretations for tasks with weaker performance. To measure these importances, Weka trains a RF over the entire dataset, then for each feature computes the mean Gini Impurity decrease from decision tree splits on that feature, averaged over all decision trees in the RF.

We find texture features are important to the Random Forest [RF], namely Local Binary Patterns and Local Binary Patterns Pyramid (Table S4). The latter is scale-invariant. We interpret this to mean that for stain-related tasks, the RF will incorrectly classify an example if texture features change. This surprised us, because intuitively color distinguishes H&E from IHC stains. However, we note Fig 6D shows a case of an H&E image being poorly lit such that the image appears brown, not completely unlike the brown from DAB in IHC (Figs 4K and 10I), though the texture of the image suggests H&E staining rather than IHC, because the poorly lit H&E is brown throughout the image, whereas IHC has brown foci where the DAB-conjugated antibody binds to a molecular target. In this way the texture of dark/brown pixels suggests IHC staining or not, and a change in texture may lead the RF to classify an image incorrectly. However, in the future we may investigate RF performance when using only Local Binary Pattern Pyramid [LBPP] features, versus performance using only Color Correlograms, to test whether or not important features predict strong RF performance empirically. We note that, over all 1000 decision trees in the RF where each tree may be grown to an arbitrary depth of decision tree nodes, Weka reports only 1-2 decision tree nodes using these most important LBPP features – which may by chance give splits for label-pure leaves in a decision tree, artificially inflating MDI measurements. The intuition here is that an average over only 1-2 measurements gives a poor estimate of a true value. To further investigate whether or not color may be important for stain prediction tasks, we additionally interpreted feature importance for a Support Vector Machine.

A Support Vector Machine [SVM] learns a weight vector from the data for classification. The weight on each feature may be interpreted. Features are more important as the absolute value of the weight on the feature is greater. The top ten weights are shown for the three stain tasks in Table S5. The SVM performed significantly worse than the Random Forest, but much better than random chance (Table 2, Fig 11, Fig S15, and Fig S16). For feature importance interpretation, we trained and tested an SVM with one iteration of ten-fold cross validation. On the Acceptable H&E task, the SVM had accuracy of 88.8602% and AUROC of 0.889 (also Fig 11). On the Acceptable H&E class-balanced task, the SVM had accuracy 88.9774% and AUROC of 0.890 (also Fig S15). On the H&E vs IHC task, the SVM had accuracy of 98.1495 and AUROC of 0.951 (also Fig S16).

For the H&E vs IHC task, we find color features are important to the SVM, namely Color Correlogram and Color Histogram, and both of these are rotation invariant features (Table S5). We interpret this to mean

[9]The other way, mean decrease in accuracy, permutes feature values and measures the impact on accuracy of permutation, which is more computationally intense than MDI. When the value of an increasingly important feature is permuted, the Random Forest's accuracy is more greatly reduced.

Task	Top feature	2nd	3rd	4th	5th
	6th	7th	8th	9th	10th
Acceptable H&E	0.9183 LBPP631	0.9183 CEDD11	0.9183 LBPP642	0.8372 LBP43	0.8113 FCTH107
	0.8097 LBPP419	0.7307 LBPP487	0.6733 LBPP393	0.6448 FCTH83	0.6359 LBPP743
Accept H&E (bal)	1 LBP69	0.9183 CEDD41	0.9183 FCTH74	0.9183 LBPP91	0.9183 LBPP455
	0.8976 LBPP161	0.8648 LBPP119	0.8201 LBPP449	0.7542 LBPP661	0.7334 LBP41
H&E vs IHC	1 LBP86	1 LBPP89	1 LBP21	1 LBP174	1 LBPP452
	1 LBPP138	0.971 LBPP209	0.971 LBPP125	0.9183 LBPP30	0.9183 LBP101

Table S4: Random Forest feature importance interpretations as measured by Mean Decrease in Gini Impurity [MDI], for stain tasks in Table 2. Important features have higher MDI, with max MDI of 1.0 an minimum MDI of 0.0. The top ten features with highest MDI are shown, for each of the three stain tasks. An entry such as "0.9183 LBPP631" means Local Binary Patterns Pyramid feature 631 has Gini Impurity of 0.9183, which is close to 1.0, so this feature is important. Other feature abbreviations include CEDD for Color and Edge Directivity Descriptor, LBP for Local Binary Patterns, and FCTH for Fuzzy Color and Texture Histogram. Features outlined in Sec 2.1.2. Surprisingly, rather than color features being most important, texture features Local Binary Patterns and Local Binary Patterns Pyramid features are most important to a Random Forest to distinguish acceptable H&E images from all other images, and to distinguish H&E from IHC stains.

Task	Top feature	2nd	3rd	4th	5th
	6th	7th	8th	9th	10th
Acceptable H&E	1.0887 FCTH162	1.0559 FCTH27	-0.9778 CEDD72	-0.9658 FCTH22	0.9498 FCTH12
	0.8989 FCTH97	0.8904 PHOG159	0.8748 PHOG338	0.8691 CEDD24	-0.8385 PHOG506
Accept H&E (bal)	-0.6667 LBP250	0.6322 FCTH12	0.6313 FCTH27	-0.6271 CC164	-0.6253 CH38
	0.6208 LBPP296	0.618 PHOG269	0.6004 FCTH162	0.5977 CEDD24	-0.5828 FCTH101
H&E vs IHC	-0.4263 CC236	-0.3785 CC204	-0.3534 CC238	-0.3478 CC237	0.3211 CC140
	-0.2816 CC205	0.2815 CH63	-0.2815 CC239	0.2622 FCTH0	0.2575 FCTH1

Table S5: Support Vector Machine [SVM] feature importance interpretations as measured by the absolute value of the feature weight, for stain tasks in Table 2. Important features have higher absolute weight, with max weight of infinity, and minimum weight of negative infinity, where a weight of zero means the feature has no influence on the SVM's predictions. The top ten features with highest absolute weight are shown, for each of the three stain tasks. An entry such as "-0.4263 CC236" means Color Correlogram feature 236 has a weight of -0.4236 (so this feature predicts the negative class), which is far from zero, so this feature is important. Other feature abbreviations include CEDD for Color and Edge Directivity Descriptor, CH for Color Histogram, FCTH for Fuzzy Color and Texture Histogram, LBP for Local Binary Patterns, LBPP for Local Binary Patterns Pyramid, and PHOG for Pyramid Histogram of Oriented Gradients. Features outlined in Sec 2.1.2. Intuitively, color features are most important for an SVM to distinguish H&E from IHC images, in contrast to texture features being important to a Random Forest.

that for the H&E vs IHC task, the Support Vector machine's predictions are influenced the most by color feature changes. We found this to be an intuitive result, because we notice H&E images are typically red (from eosin) and purple (from hematoxylin), while IHC images are typically brown (from DAB) and blue (from hematoxylin).

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 S17). There were 381 images: 135 negative images (from 57 Tweets) that were breast pathology and 246 positive images (from 108 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 ˜1.8:1. Accuracy is $71.871 \pm 6.292\%$ (chance $64.946 \pm 0.848\%$). AUROC is 0.783 ± 0.082 (chance 0.5). Of all tissue type binary comparisons, this was the most challenging pair to compare, in terms of mean AUROC. 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

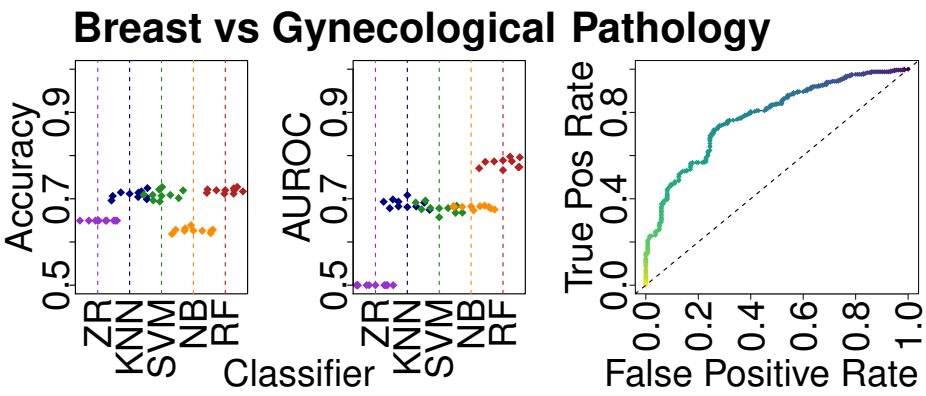

Figure S17: Predicting if an image is Breast or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

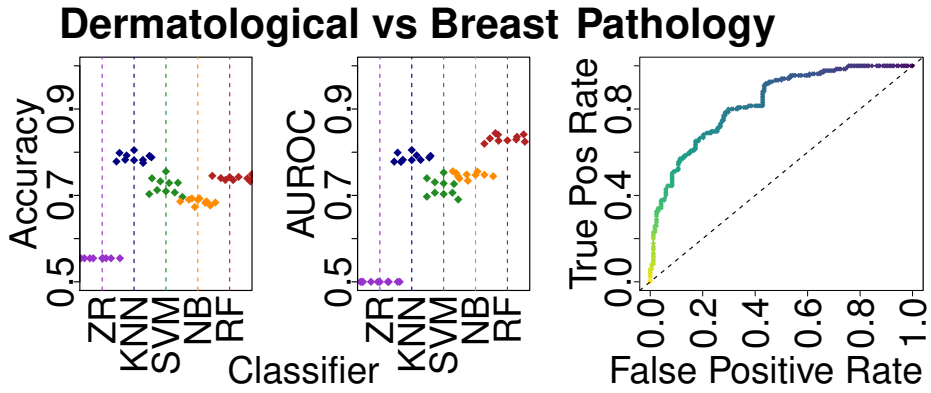

Figure S18: Predicting if an image is Derm or Breast. RF classifier had greatest AUROC. RF ROC curve at right.

performance may improve with more data. In contrast, adipocytes may be a hallmark of breast tissue.

S3.2.2 Derm vs Breast task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, breast pathology (Fig S18). There were 303 images: 168 negative images (from 78 Tweets) that were dermatological pathology and 135 positive images (from 57 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 $74.088 \pm 6.994\%$ (chance $55.452 \pm 1.329\%$). AUROC is 0.832 ± 0.069 (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.

S3.2.3 Derm vs Gyn task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, gynecological pathology (Fig S19). There were 414 images: 168 negative images (from 78 Tweets) that were dermatological pathology and 246 positive images (from 108 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 $75.986 \pm 5.158\%$ (chance $59.814 \pm 0.584\%$). AUROC is 0.847 ± 0.062 (chance 0.5). There is room to improve performance in this task. Of all the pairwise comparisons involving gynecological pathology, this comparison to dermatological

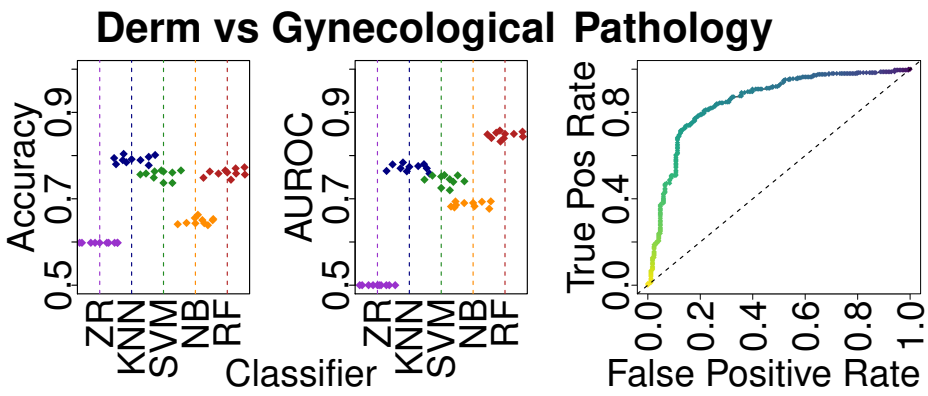

Figure S19: Predicting if an image is Derm or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

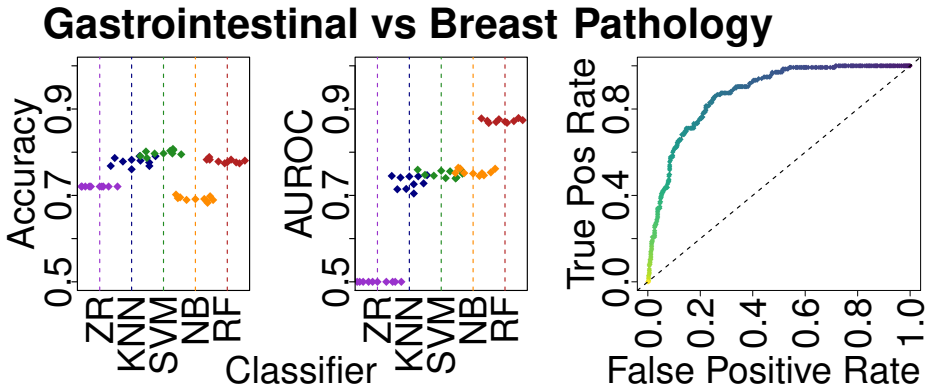

Figure S20: Predicting if an image is GI or Breast. RF classifier had greatest AUROC. RF ROC curve at right.

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 S20). There were 483 images: 348 negative images (from 172 Tweets) that were gastrointestinal pathology and 135 positive images (from 57 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 ˜2.6:1. Accuracy is $77.978 \pm 3.974\%$ (chance $72.054 \pm 0.892\%$). AUROC is 0.873 ± 0.050 (chance 0.5). Of all six tissue pairs (Table 2), our Random Forest performed best on this pair, GI vs Breast, 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 differences between these tissue types, beyond intuitive hallmarks and the Random Forest.

S3.2.5 GI vs Derm task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, dermatological pathology (Fig S21). There were 516 images: 348 negative images (from 172 Tweets) that were gastrointestinal pathology and 168 positive images (from 57 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 $76.198 \pm 5.639\%$ (chance $67.443 \pm 0.645\%$). AUROC is 0.854 ± 0.059 (chance 0.5).

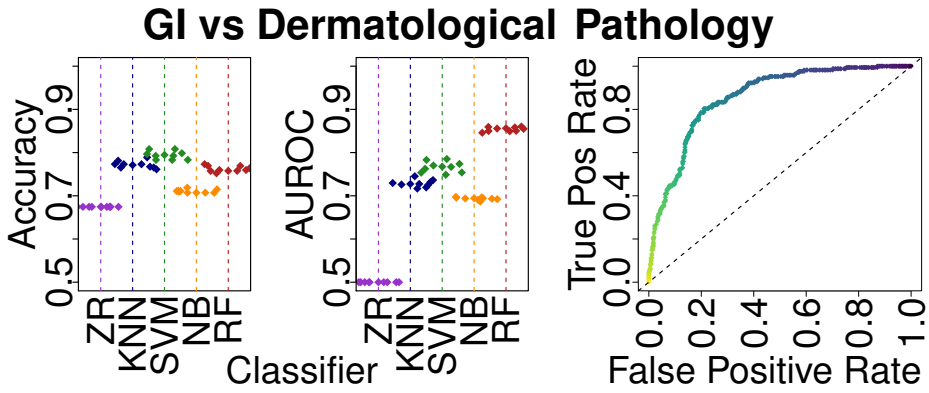

Figure S21: Predicting if an image is GI or Derm. RF classifier had greatest AUROC. RF ROC curve at right.

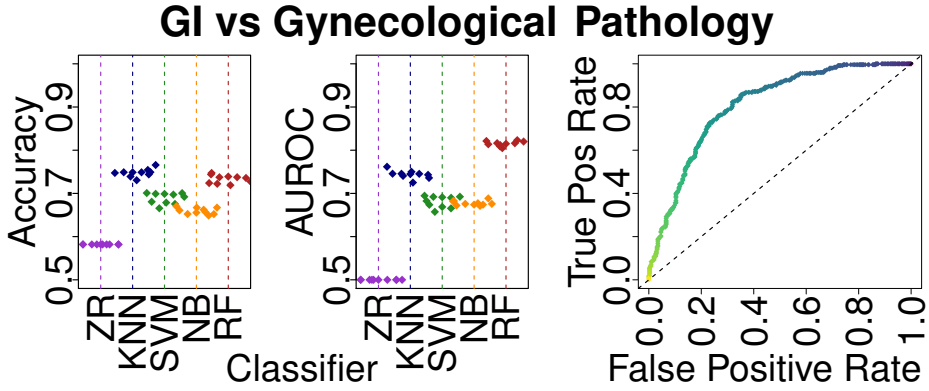

Figure S22: Predicting if an image is GI or Gyn. RF classifier had greatest AUROC. RF ROC curve at right.

S3.2.6 GI vs Gyn task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, gynecological pathology (Fig S22). There were 594 images: 348 negative images (from 180 Tweets) that were gastrointestinal pathology and 246 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 $73.338 \pm 5.495\%$ (chance $58.192 \pm 0.284\%$). AUROC is 0.815 ± 0.053 (chance 0.5).

S3.2.7 Breast vs GU task

Our Random Forest predicts if an image is breast pathology, or alternatively, genitourinary pathology (Fig S23). There were 252 images: 135 negative images (from 56 Tweets) that were breast pathology and 117 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 $74.791 \pm 7.968\%$ (chance $53.569 \pm 1.746\%$). AUROC is 0.822 ± 0.081 (chance 0.5).

S3.2.8 Derm vs GU task

Our Random Forest predicts if an image is dermatological pathology, or alternatively, genitourinary pathology (Fig S24). There were 285images: 168 negative images (from 57 Tweets) that were breast pathology and 117 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 $77.273 \pm 7.203\%$ (chance

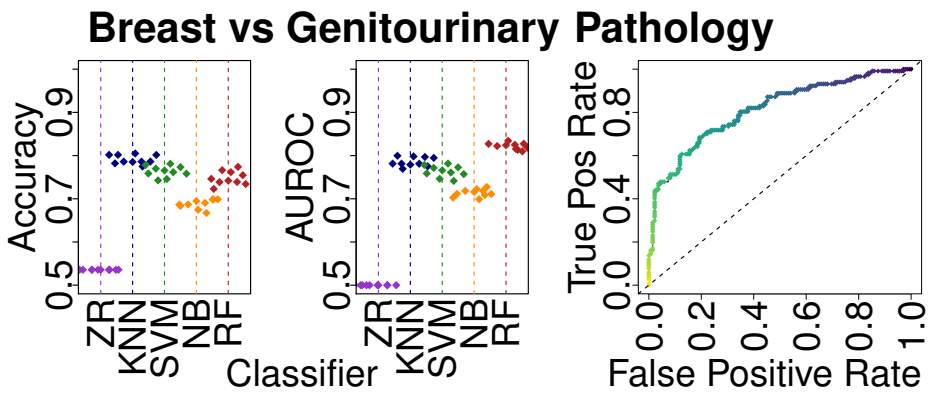

Figure S23: Predicting if an image is Breast or GU. RF classifier had greatest AUROC. RF ROC curve at right.

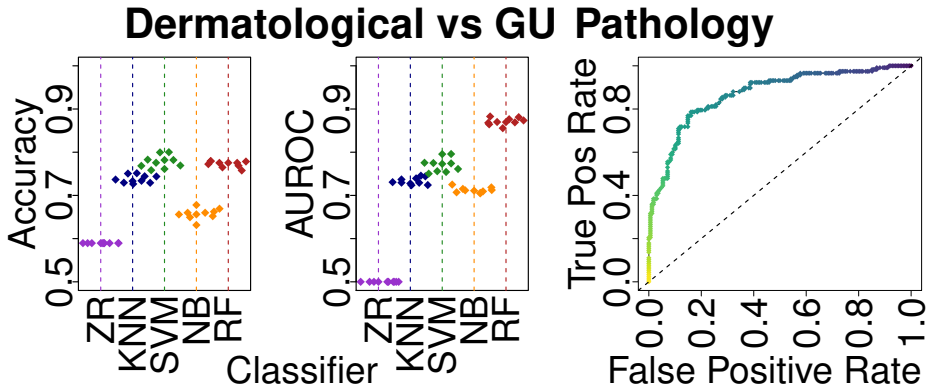

Figure S24: Predicting if an image is Derm or GU. RF classifier had greatest AUROC. RF ROC curve at right.

$58.953 \pm 1.288\%$). AUROC is 0.871 ± 0.070 (chance 0.5).

S3.2.9 GI vs GU task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, genitourinary pathology (Fig S25). There were 465 images: 348 negative images (from 57 Tweets) that were breast pathology and 117 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 ˜3.0:1. Accuracy is $78.930 \pm 2.670\%$ (chance $74.843 \pm 0.845\%$). AUROC is 0.830 ± 0.071 (chance 0.5).

S3.2.10 Gyn vs GU task

Our Random Forest predicts if an image is gastrointestinal pathology, or alternatively, genitourinary pathology (Fig S26). There were 363images: 246negative images (from 57 Tweets) that were breast pathology and 117positive 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 ˜2.1:1. Accuracy is $76.462 \pm 4.066\%$ (chance $68.131 \pm 0.864\%$). AUROC is 0.795 ± 0.078 (chance 0.5).

S3.3 5-way comparison details

The confusion matrix for Breast vs Derm vs GI vs GU vs Gyn is in Table S6, and Table S7 shows a similar confusion matrix but for class-balanced sampling.

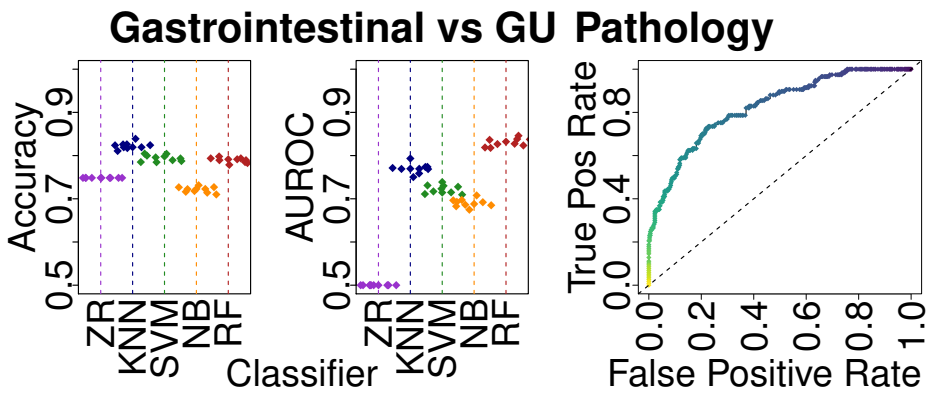

Figure S25: Predicting if an image is GI or GU. RF classifier had greatest AUROC. RF ROC curve at right.

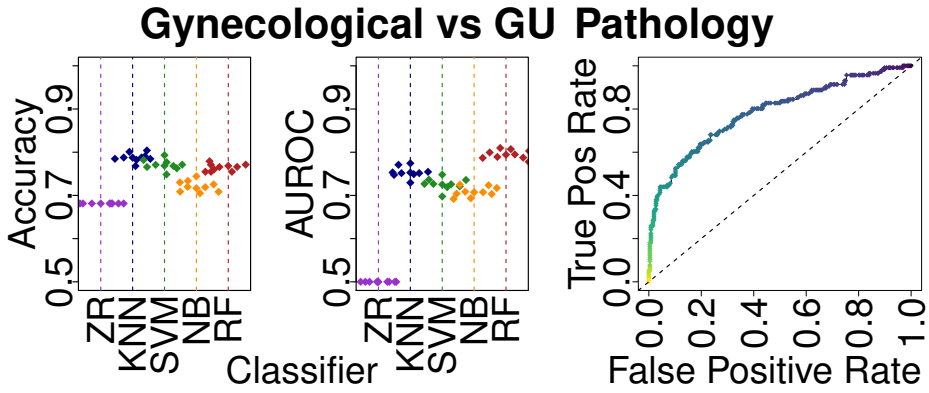

Figure S26: Predicting if an image is Gyn or GU. RF classifier had greatest AUROC. RF ROC curve at right.

S3.3.1 5-way tissue classification (balanced)

Our Random Forest predicts if an image is one of five possible tissue types (Figs 10 and S27) after sampling tissues in a balanced manner. There were 585 images, with 117 images from each of the five tissue types: breast, dermatological, gastrointestinal, genitourinary, and gynecological. Classes were exactly balanced. At 585 images, this was 57.7% of the amount of data we used for unbalanced sampling in the prior task, namely 1014 images. Accuracy is $48.935 \pm 6.089\%$ (chance $18.805 \pm 0.162\%$). AUROC is 0.786 ± 0.058 (chance 0.5).

Moreover, the confusion matrix (Table S7) 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 (Table 3).

a	b	c	d	e	⟵	classified as
29	12	56	1	37	—	a = breast
4	72	66	0	26	—	b = derm
3	22	278	0	45	—	c = gi
3	6	64	16	28	—	d = gu
4	15	98	1	128	—	e = gyn

Table S6: RF confusion matrix for Breast 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.

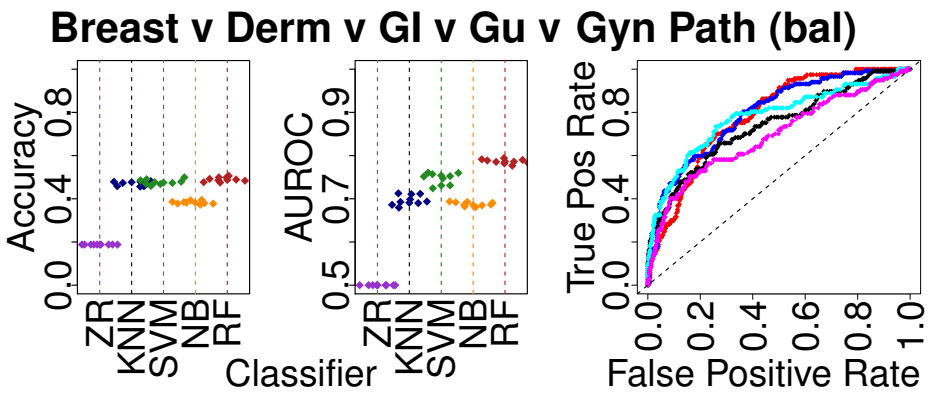

Figure S27: Predicting if an image is Breast, Derm GI, GU, or Gyn. RF classifier had greatest AUROC. RF ROC curve at right. In the ROC plots, **Breast** is red, **Derm** is blue, **GI** is black, **GU** is cyan, and **Gyn** is magenta. ROC is calculated as the tissue versus all other tissues, e.g. in red is Breast vs all other tissues, and in blue is Derm vs all other tissues.

a	b	c	d	e	⟵	classified as
57	25	5	14	16	—	a = breast
18	63	17	8	11	—	b = derm
21	20	52	12	12	—	c = gi
16	12	14	58	17	—	d = gu
25	17	10	16	49	—	e = gyn

Table S7: RF confusion matrix for Breast vs Derm vs GI vs GU vs Gyn comparison under class-balanced subsampling, showing the enrichment of GI false positives has been reduced.

S4 Nontumoral, Low grade, and Malignant task details

Tasks involving distinguishing nontumoral disease, low grade tumors, and malignant tumors (Fig S28) are our most difficult tasks. The acknowledged definition of "malignant" in epithelial cancers is the ability to breach the basement membrane, i.e. a malignant tumor escapes containment and is therefore no longer "treatable with surgical resection". A malignant tumor can invade into the adjacent tissue, lymphatics, and blood vessels. For machine learning, we define a three categories of disease: (a) normal tissue and nontumoral disease; (b) benign, low grade, and oncovirus-driven tumoral disease; and (c) malignant tumors – but there are number of caveats with this, because:

1. there is a spectrum of pathology rather than an oversimplified 3-class nontumoral/low-grade/malignant system.
2. a two-class dichotomy is simpler, e.g. "malignant vs everything else", but in practice we find a 3-class nontumoral/low-grade/malignant classification problem does not perform significantly differently, so we believe this finer-grained 3-class model is preferable. We are most interested in "low grade vs malignant", and nontumoral is discarded for this.
3. the benign/malignant dichotomy may be more vague in certain tissues e.g. central nervous system [CNS] primary tumors such as chordomas.
4. vague terms like adenoma are typically benign but may be malignant, and likewise vague terms like anaplasia are more often associated with malignancy but not always.
5. vague terms like anaplasia and neoplasia make no real reference to the malignancy of lesions i.e. there are benign anaplastic lesions, while neoplasia is almost synonymous with tumor.
6. terms like tumor do not provide information about benign or malignant state, though normal/nontumoral can be ruled out.
7. there may be some disagreement if some terms, e.g. "carcinoma *in situ*", are more appropriate to include as low grade, or if instead should be considered malignant due to their malignant potential or

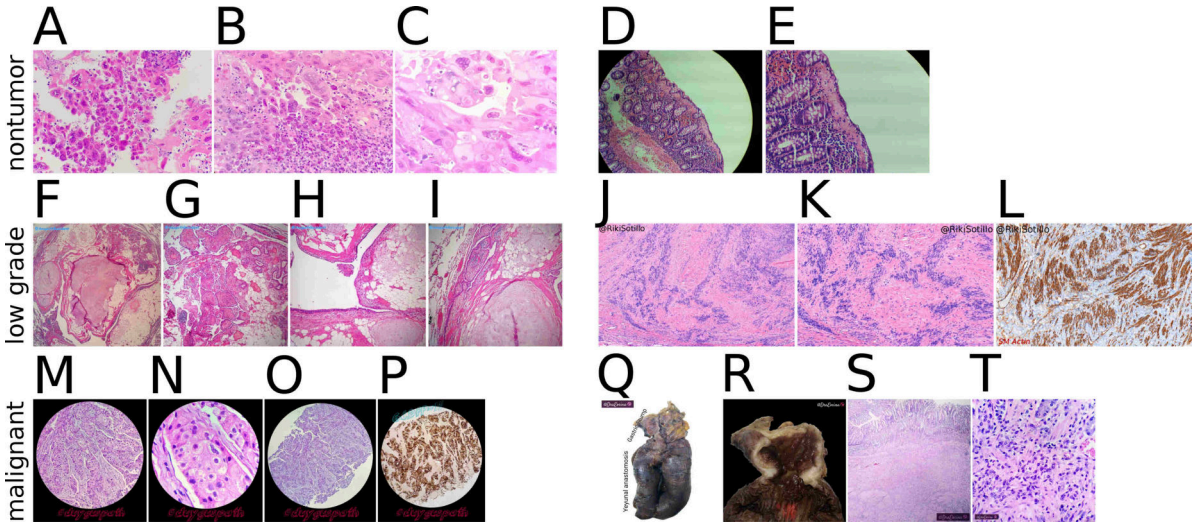

Figure S28: We use machine learning to distinguish nontumoral disease, benign/low grade malignant potential [low grade] tumor, and malignant tumor. *Panels A,B,C (M.P.P.)*: Nontumoral disease, i.e. herpes esophagitis with diagnostic Cowdry A inclusions. *Panels D,E (K.H.)*: Nontumoral disease, i.e. collagenous colitis showing thickened irregular subepithelial collagen table with entrapped fibroblasts, vessels and inflammatory cells. *Panels F,G,H,I (A.M.)*: Low grade tumor, i.e. pulmonary hamartoma showing entrapped clefts lined by respiratory epithelium. *Panels J,K,L (R.S.S.)*: Low grade tumor, i.e. leiomyoma showing nuclear palisading. Panel L is IHC rather than H&E, so Panel L is shown here for completeness but not included in this machine learning analysis. *Panels M,N,O,P (B.D.S.)*: Malignant tumor, i.e. breast cancer with apocrine differentiation. Panel P is IHC and not included in analysis here, but is shown for completeness. *Panels Q,R,S,T (L.G.P.)*: Malignant tumor, i.e. relapsed gastric adenocarcinoma with diffuse growth throughout the anastomosis and colon. Panels Q,R are gross not H&E so they are not included in the analysis here, but shown for completeness.

treatment implications. For instance, ductal carcinoma *in situ* [DCIS] typically needs to be removed with surgery or radiotherapy, whereas lobular carcinoma *in situ* [LCIS] typically does not. DCIS's lower grade counterpart, atypical ductal hyperplasia, may get surgery or not. We believe treatment implications are a separate task. Typically, Tweets do not include a decision to perform surgery or not, so additional annotations may be needed for the surgery task. We assign all pre-cancer and tumoral disease with malignant potential to the "low grade"' category, in light of these benign/malignant ambiguities and data limitations.

8. the diagnosis should be known before deciding benign/malignant, but it is very difficult to know the full diagnosis from the brief, generic, descriptive terms in the Tweet.

S4.1 Text processing for Nontumoral, Low grade, and Malignant tasks

To determine if an acceptable H&E human microscopy image is nontumoral, low grade, or malignant – like in Sec S2.6 we relied on regular-expression-based keyword matching in Tweet text. However, keywords differed and we considered all Tweets in a message thread per Sec 2.2.1. To infer these message threads of Tweets, we downloaded from Twitter each Tweet's metadata (in JavaScript Object Notation [JSON] format), which describes the parent Tweet for each Tweet. If Tweet A is a reply to Tweet B, then Tweet A is the parent of Tweet B, and both Tweets are in the same message thread. In a message thread from a consenting pathologist, we only considered that pathologist's Tweets, not other users.

Our heirarchical algorithm for nontumor/low-grade/malignant keyword-matching shown in Fig S29, and details for each step follow. First, to determine if a single Tweet indicated nontumoral, low grade, or malignant, we looked for specific hashtags in a Tweet's text that indicated malignancy, tumoral status, or nontumoral status.

1. Malignant: `/#[a-z]*cancer/i` or `/#metastas[ei]s/i`
    - The first regular expression in this set matches #ANYcancer, where ANY can be any non-

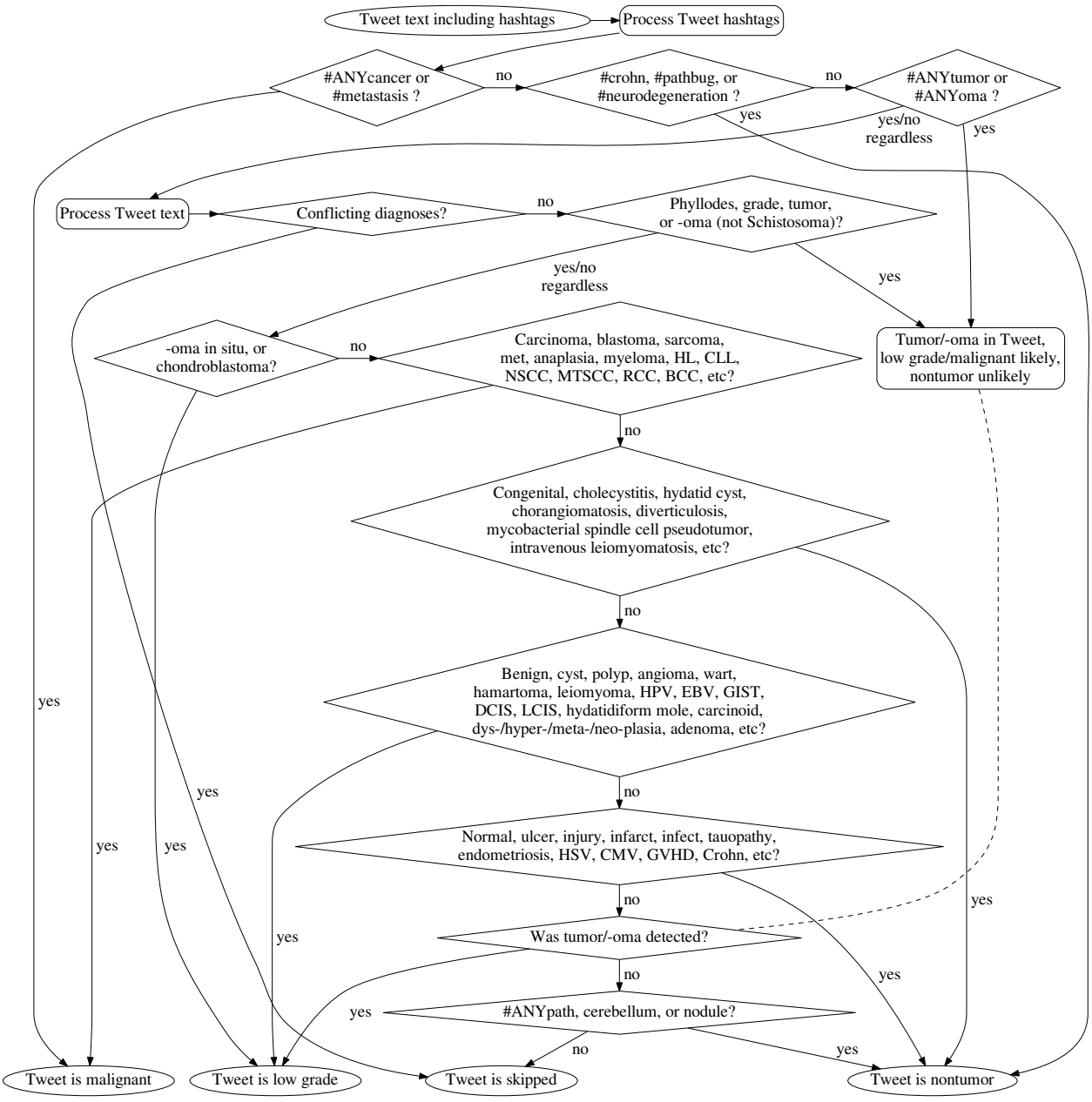

Figure S29: Flowchart of algorithm that processes a single Tweet's text to categorize it as nontumor (309 images), benign/low grade malignant potential [low grade] (347 images), or malignant (385 images). A Tweet may be skipped (132 images, i.e. 11.3% of images) when the pathologist discusses multiple possible diagnoses for this case or when no pathology keywords are found. Dashed line indicates early steps where tumor/-oma detected, and a later step where detected tumor/-oma considered for possible low grade categorization. Nontumor, low grade, and malignant are defined in Sec S4. Flowchart steps are detailed in Sec S4.1. The algorithm has many steps in order to parse overlapping words that have different diagnoses. For instance, if "Lobular carcinoma in situ of the breast" (which is a low grade disease) was the Tweet text, the algorithm has an early step to categorize "carcinoma in situ" as low grade (which is correct here) because a later step categorizes "carcinoma" as malignant (which is not correct here). Indeed, Tweet text "Carcinoma of the breast" describes a malignant disease and the algorithm categorizes it malignant because "in situ" is absent. Besides "carcinoma in situ" (low grade) and "carcinoma" (malignant), the algorithm distinguishes "chorangiomatosis" (nontumor) from "angioma" (low grade), "hydatidiform mole" (low grade) from "hydatid cyst" (nontumor) from "ovarian cyst" (low grade) from "cholecystitis" (nontumor), and "intravenous leiomyomatosis" (nontumor) from "leiomyoma" (low grade).

whitespace characters, e.g. "#bladdercancer" and "#breastcancer" both match, as well as "#cancer".
- Metastasis is a sign of malignant cancer, so Tweets with #metastasis or #metastases hashtags are malignant.
- If any matching keyword is detected, no further keyword processing is performed. The Tweet is malignant.

2. Nontumoral: `/#crohn/i` or `/#neurodegeneration/i` or `/#pathbug/i`

   - Crohn's disease and neurodegeneration are not tumoral diseases, so this Tweet is in the nontumoral/normal category. This `/#crohn/i` regular expression is case-insensitive, so it matches "#crohn", "#Crohn", and "#CROHN". The #pathbug hashtag indicates a parasite or other microorganism is in the image, which is also nontumoral.
   - If any matching keyword is detected, no further keyword processing is performed. The Tweet is nontumoral.

3. Tumoral status (ambiguously low grade or malignant): `/#[a-z]*tumou?r/i` or `/#[a-z]*oma/i`

   - The first regular expression in this set matches #ANYtumor or #ANYtumour, where ANY can be any non-whitespace characters, e.g. "#BrainTumor" and "#phyllodestumour" both match, as well as "#tumor".
   - The second regular expression matches #ANYoma, e.g. #Lymphoma and #leiomyoma both match.
   - Because "tumor" and "-oma" do not necessarily mean a tumor is low grade or malignant, further keyword matching is performed. It is unlikely that the Tweet is nontumoral. If no other specific information is found after all further keyword matching is performed, the tumor is presumed to be low grade.

Second, if no hashtags matched, we then analyzed keywords in the Tweet text.

1. Skip: `/mistake/i` or `/misinterpret/i` or `/confuse/i` or `/suspect/i` or `/worry/i` or `/surprise/i` or `/mimic/i` or `/simulate/i` or `/lesson/i` or `/\bhelp\b/i` or `/usually/i` or `/difficult/i` or `/pathart/i` or `/pathchallenge/i` or `/pathquiz/i` or `/pathgame/i` or `/^http/`

   - We skip Tweets where (i) the pathologist discusses points of the case which may be easily mistaken – instead of providing a single diagnosis, (ii) the pathologist provides a diagnosis but may suspect an alternative diagnosis, or (iii) the Tweet is simply a link to another Tweet. No further keyword matching is performed for this Tweet.

2. Tumoral status (ambiguously low grade or malignant): `/phyllod/i` or `/\bgrade\b/i` or `/tumou?r/i` or (`/[a-z]{3,}oma\b/i` and not `/schistost?oma/i`)

   - Phyllodes tumors, mentions of "tumor" or "tumour", mentions of tumor "grade", and mentions of words that end in "oma" but are not "Schistosoma" – are all detected here.
   - Loosely speaking, phyllodes tumors are only slightly more likely to be low grade than malignant. Because "tumor", "-oma", and "grade" do not necessarily mean a tumor is low grade or malignant, further keyword matching is performed. It is unlikely that the Tweet is nontumoral. If no other specific information is found after all further keyword matching is performed, the tumor is presumed to be low grade.
   - *Schistosoma* (and its misspelling "Schistostoma") refers to a genus of parasitic worm, rather than a tumor, though *Schistosoma* ends in "oma" like many tumor types.

3. Low grade: `/oma in situ/i` or `/chondroblastoma/i`

   - If we did not skip this Tweet, but the Tweet does mention "oma *in situ*", e.g. "carcinoma *in situ*" or "melanoma *in situ*", then we consider this Tweet and images to represent low grade disease. Carcinoma *in situ* is pre-cancer, and we consider it more low grade than malignant. If a Tweet contains only "carcinoma" but not "*in situ*", subsequent steps will consider the Tweet as malignant.

- If the Tweet includes "chondroblastoma", this Tweet is low grade. This is not to be confused with other blastomas, such as glioblastoma or lymphoblastoma, which are malignant and matched in subsequent steps.
- No further keyword matching is performed if these patterns match. The Tweet is low grade.

4. Malignant: `/malignant/i` or `/malignancy/i` or `/cancer/i` or `/\bCA\b/i` or `/carc?inoma/i` or `/sarcoma/i` or `/blastoma/i` or `/\bWilms/i` or `/GBM/i` or `/anaplas(?:ia|tic)/i` or `/metastas[ie]s/i` or `/metastatic/i` or `/\bmets?\b/i` or `/adenoca/i` or `/melanoma/i` or `/seminoma/i` or `/lymphoma/i` or `/leuka?emia/i` or `/mesothelioma/i` or `/myeloma/i` or `/hodgkin/i` or `/\bHL\b/i` or `/burkitt/i` or `/plasmoc[yi]toma/i` or (`/paget/i` and `/breast/i`) or `/\bCLL\b/i` or `/PCNSL/i` or `/NSCHL/i` or `/\bCHL\b/i` or `/NSCC/i` or `/\bI[LD]C\b/i` or `/\bASPS\b/i` or `/mtscc/i` or `/sq?cc/i` or `/rcc/i` or `/bcc/i`

   - Many diagnoses and abbreviations may indicate cancerous malignancy, e.g. carcinoma, sarcoma, Wilms' tumor, leukemia, RCC [renal cell carcinoma], NSCC [non-small cell lung carcinoma], or the stand-alone abbreviation "CA" [cancer].
   - We consider "anaplastic/anaplasia" to be more malignant than low grade disease.
   - No further keyword matching is performed if these patterns match. The Tweet is malignant.

5. Nontumoral: `/congenital/i` or `/cholecystitis/i` or `/chorangiomatosis/i` or `/mycobacteri(?:um|al)\s*spindle\s*cell\s*pse?udotumor/i` or `/intravenous\s*leiomyomatosis/i` or `/helicobacter/i` or `/dirofilaria/i` or `/tuberculo/i` or `/enterobius/i` or `/echinococcus/i` or `/hydatid\s*cyst/i` or `/giardia/i` or `/cryptosporidium/i` or `/ascaris/i` or `/sarcina/i` or `/worm/i` or `/spiroquet(?:osis|es)/i` or `/diverticulosis/i` or `/villitis/i` or `/colitis/i` or `/gastritis/i` or `/esophagitis/i` or `/appendicitis/i` or or `/xanthoma/i`

   - Many diagnoses and abbreviations may indicate nontumoral disease, e.g. congenital conditions, *Helicobacter* infection, and villitis. Nontumoral disease keywords that contain "cyst", e.g. "cholecystitis" and "hydatid cyst", are detected here, because subsequent keyword matching steps will detect "cyst" as a sign of low grade tumoral disease.
   - If one of these nontumoral keywords matches, no further keyword matches are attempted, and the Tweet is considered nontumoral, even if prior steps detected "tumor" or "-oma". For instance, a "xanthoma" is a lipid aggregate, not a tumoral disease, even though xanthoma ends in -oma.

6. Low grade: `/benign/i` or `/cyst/i` or `/polyp/i` or `/hamartoma/i` or `/chorangioma/i` or `/ha?ematoma/i` or `/cylindroma/i` or `/fibroma/i` or `/luteoma/i` or `/c[yi]toma/i` or `/cond[yi]loma/i` or `/neoplas(?:ia|tic|m)/i` or `/LCIS/i` or `/DCIS/i` or `/\b[LD]IN\b/i` or `/lipoma/i` or `/carcinoid/i` or `/neuroma/i` or `/meningioma/i` or `/perineurioma/i` or `/cavernoma/i` or `/\bLGG\b/i` or `/\bODG\b/i` or `/oligodendroglioma/i` or `/craniopharyngioma/i` or `/le[yi]om[iy]oma/i` or `/schwannoma/i` or `/osteochondroma/i` or `/ependymoma/i` or `/angioma/i` or `/syringoma/i` or `/acanthoma/i` or `/collagenoma/i` or `/hidradenoma/i` or `/papilloma/i` or `/pilomatrixoma/i` or `/hydatidiform\s*mole/i` or `/wart/i` or `/molluscum/i` or `/\bHPV\b/i` or `/\bEBV\b/i` or `/kerat?osis/i` or `/fibrokeratoma/i` or `/melanoc[iy]tosis/i` or `/brenner/i` or `/granular\s+cell\s+tumou?r/i` or `/metaplas(?:ia|tic)/i` or `/dysplas(?:ia|tic)/i` or `/dysembryoplas(?:ia|tic)/i` or `/hyperplas(?:ia|tic)/i` or `/\bLFH\b/i` or `/\bDNE?T\b/i` or `/\bNET\b/i` or `/\bPTC\b/i` or `/\bGIST\b/i` or `/\bSTIC\b/i` or `/\b[LD]ISN\b/i` or `/adenoma/i` or `/adenosis/i`

   - Many diagnoses may indicate benign tumor, e.g. hamartoma, fibroma, condyloma, papilloma, lipoma, adenoma, adenosis, or cyst.
   - We consider "neoplastic/neoplasia", "metaplastic/metaplasia", "hyperplastic/hyperplasia", and "dysplastic/dysplasia" to be more indicative of benign/low-grade/non-invasive/pre-malignant disease than malignant disease, but these terms are vague.
   - We broadly consider oncovirus-driven tumors and wart-like growths to be in this low grade category also, e.g. HPV [human papilloma virus] warts and *Molluscum contagiosum* "water warts".
   - We similarly consider abbreviations "LCIS" [lobular carcinoma *in situ*], "DCIS" [ductal carcinoma *in situ*], "LISN" [lobular *in situ* neoplasia], and "DISN" [ductal *in situ* neoplasia] to be more benign than malignant disease, so we categorize them as low grade. Though DCIS may require

surgical or radiological intervention to be removed while LCIS may not, we consider our "low grade" and "malignant" categories to be defined by the apparent histopathology rather than the appropriate medical intervention. Predicting appropriate medical intervention would be a different machine learning task.

- If one of these keywords match, the Tweet is considered low grade and no further keyword matching is performed.

7. Nontumoral: `/normal/i` or `/ulcer/i` or `/embolism/i` or `/thromb/i` or `/rupture/i` or `/infarct/i` or `/aneurysm/i` or `/ha?emorrhag/i` or `/injur(?:y|ed)/i` or `/inflam/i` or `/swell/i` or `/balloon\s*cell\s*na?ev(?:us|i)/i` or `/decidua/i` or `/foreign/i` or `/lymphadenopath?y/i` or `/vasculopathy/i` or `/vasculitis/i` or `/synovitis/i` or `/pulmonary\s*interstitial\s*glycogenosis/i` or `/essential\s*thrombocythemia/i` or `/endometriosis/i` or `/mastoc[iy]tosis/i` or `/castleman/i` or `/herpe(?:s|tic)/i` or `/\bHSV\b/i` or `/\bCMV\b/i` or `/cytomegalovir/i` or `/viral/i` or `/bacteri(?:a|um)/i` or `/fung(?:al|us)/i` or `/mycetoma/i` or `/myco(?:sis|tic)/i` or `/infect(?:ion|ed)/i` or `/tauopathy/i` or `/amyloidosis/i` or `/neurodegen/i` or `/\brabies\b/i` or `/hemosiderosis/i` or `/polymicrogyria/i` or `/status\s*verrucosus/i` or `/\bIUGR\b/i` or `/storage\s*dis(?:ease|order)/i` or `/athero(?:sis|ma)/i` or `/atherosclero(?:sis|tic)/i` or `/gauzoma/i` or `/colchicine/i` or `/\bIBD\b/i` or `/GVHD/i` or `/crohn/i`

   - Many diagnoses may indicate normal tissue of nontumoral disease, e.g. normal, embolism, decidua, tauopathy, foreign body, mycetoma, CMV [cytomegaolovirus] infection, GVHD [graft versus host disease], and Crohn's disease.
   - If one of these nontumoral keywords matches, no further keyword matches are attempted, and the Tweet is considered nontumoral, even if prior steps detected "tumor" or "-oma". For instance, a mycetoma is not a tumor, even though mycetoma ends with -oma.

8. Nontumoral: (not tumor/oma) and (`/#[a-z]*path/i` or `/cerebell(?:um|ar)/i` or `/nodul(?:e|arity)/i`). Low grade if tumor/oma.

   - If the Tweet does not have tumor or "-oma" keywords detected from prior steps, and if the Tweet has a #ANYpath hashtag (e.g. "#pulmpath" or "#pathology"), mention of "nodule"/"nodularity", or mention of the cerebellum, then we consider the Tweet to be nontumoral. If instead the Tweet has tumor or -oma keywords, then we consider the Tweet to be low grade. The Tweet is skipped if no steps identified the Tweet as nontumoral, low grade, or malignant.
   - Cerebellum is mentioned in several Tweets, e.g. to depict normal cerebellar tissue[10]. Currently, we group normal tissue with tissue having nontumoral disease. We expect more tissue-based keywords may be used here in the future, as we expand our study to include more pathologists, tissues, and normal cases.
   - In practice, we manually inspect all Tweet message text to minimize the number of cases that are classified as nontumoral here. We typically write regular expressions to match specific keywords that indicate if a Tweet represents nontumoral, low grade, or malignant disease.
   - Recall that if any Tweets in a message thread are malignant, then all images for all Tweets in the message thread are considered malignant[11]. Moreover, if there are no malignant Tweets in a message thread, the message thread is considered low grade if any Tweets in the message thread are low grade. Additionally, if any vague tumoral keywords, e.g. tumor and -oma, are detected in any Tweet in the message thread, then the message thread is either low grade or malignant – not nontumoral.
   - As part of our manual data curation, if on Twitter there was discussion among pathologists, and a different pathologist mentioned a correct diagnosis, and our consenting contributing pathologist concurred, then we write an auxiliary annotation file for the Tweet with a summarized diagnosis[12].

[10] Normal cerebellum case by S.Y. at `https://twitter.com/Sty_md/status/821840894634565632`

[11] A case of this is from author B.X., the initial Tweet asked a question `https://twitter.com/BinXu16/status/992958513193238528` while providing an acceptable image, then a subsequent Tweet reply provided the NUT midline carcinoma diagnosis `https://twitter.com/BinXu16/status/993075257308205056` which we consider malignant by keyword matching.

[12] A case of this is from author K.H., where a different pathologist gave the diagnosis, and he agreed. We summarized this as "metastatic lobular carcinoma" in the auxiliary annotation file for the Tweet `https://twitter.com/Ho_Khanh_MD/status/999989201734197250`.

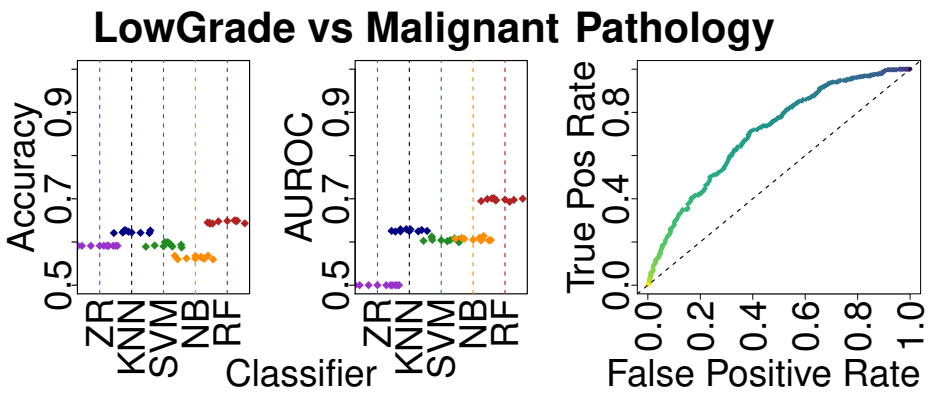

Figure S30: Predicting if an image is low grade or malignant, with nontumoral images skipped. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.7051 here for n=732 per Table 2, which is not significantly different from the mean AUROC=0.703.

This summary is also used for pattern matching. This is an additional way that we minimize how many cases are handled at this late step.

- Moreover, if the contributing pathologist wrote diagnostic text directly in the image, we will write this text in the auxiliary annotation file for text matching also.[13]
- The way this "default nontumoral or low grade" rule is intended to be used is as a catch-all for unusual but non-malignant conditions[14]. Our motivation for this rule is to minimize our manual data curation burden. We do not wish to write an auxiliary annotation file or make a new regular expression for each unusual type of case, and we observe many of these cases are not malignant. It remains important to inspect the cases manually for correctness.

Tweets that do not match any nontumoral, low grade, or malignant rules are skipped in the same manner that Tweets matching skip rules are skipped. An additional caveat is this keyword matching may need refinement as we accumulate data, because we expect to encounter terms that are low grade or malignant in a context-dependent manner, e.g. tissue type or genetic sequencing.

S4.2 Pairwise Nontumoral, Low grade, and Malignant task details

We first considered all pairwise comparisons: low grade vs malignant, nontumoral vs malignant, nontumoral vs low grade, nontumoral+low grade vs malignant, and nontumoral vs low grade+malignant.

S4.2.1 Low grade vs Malignant task details

Our Random Forest predicts if an image is low grade or malignant (Fig S30). There were 732 images: 347 negative images that were low grade and 385 positive images that were malignant. Classes were essentially balanced. Accuracy is $65.055 \pm 5.159\%$ (chance $52.595 \pm 0.599\%$). AUROC is 0.703 ± 0.058 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical consideration. We are most interested in this task for clinical decision support, because malignancy impacts clinical decisions, and a clinician may already know that the patient has no nontumoral disease.

S4.2.2 Nontumor vs Malignant task details

Our Random Forest predicts if an image is nontumor or malignant (Fig S31). There were 694 images: 309 negative images that were nontumor and 385 positive images that were malignant. Classes were essentially

[13] A case of this is from author M.P.P., where M.P.P. wrote "IDC DIN LISN" directly on a shared histology image in the Tweet `https://twitter.com/dr_MPrieto/status/890118713155997696` so we wrote this text in the auxiliary annotation file for the Tweet.

[14] A case of this is from K.H., observing iron fill lesions in stomach biopsy `https://twitter.com/Ho_Khanh_MD/status/963800933716123648`.

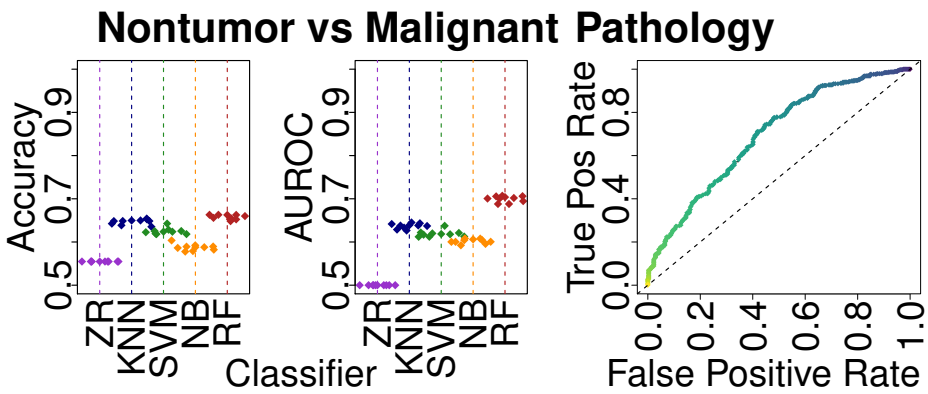

Figure S31: Predicting if an image is nontumor or malignant. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.6918 here for n=694 per Table 2, which is not significantly different from the mean AUROC=0.700.

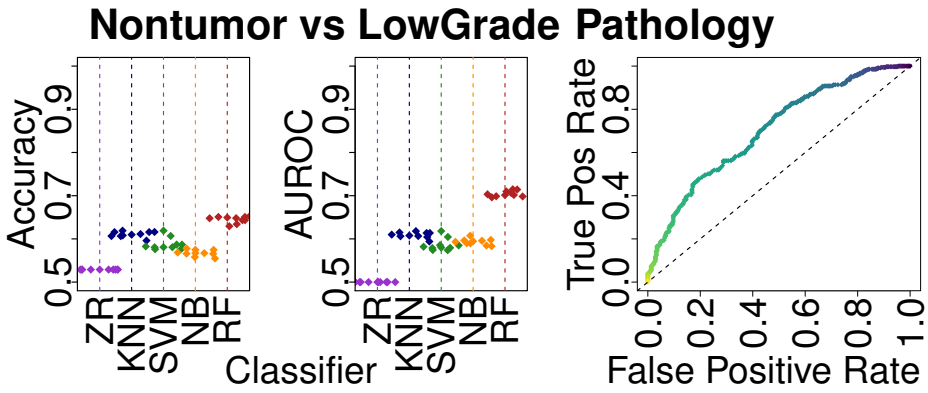

Figure S32: Predicting if an image is nontumor or low grade. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.7012 here for n=656 per Table 2, which is not significantly different from the mean AUROC=0.704.

balanced. Accuracy is $65.744 \pm 5.131\%$ (chance $55.474 \pm 0.464\%$). AUROC is 0.700 ± 0.066 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical consideration.

S4.2.3 Nontumor vs Low grade task details

Our Random Forest predicts if an image is nontumor or low grade (Fig S32). There were 656images: 309 negative images that were nontumor and 347 positive images that were low grade. Classes were essentially balanced. Accuracy is $64.493 \pm 5.536\%$ (chance $64.493 \pm 5.536\%$). AUROC is 0.704 ± 0.059 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical consideration.

S4.2.4 Nontumor vs LowGrade+Malignant task details

Our Random Forest predicts if an image is nontumor or low-grade/malignant (Fig S33). There were 1041 images: 309 negative images that were nontumor, and 732positive images that were low grade or malignant. There was mild class imbalance of ˜2.4:1. Accuracy is $73.046 \pm 2.188\%$ (chance $70.317 \pm 0.293\%$). AUROC is 0.683 ± 0.062 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical

a	b	⟵	classified as
187	160	—	a = 1
99	286	—	b = 2

Table S8: RF confusion matrix for low grade vs malignant.

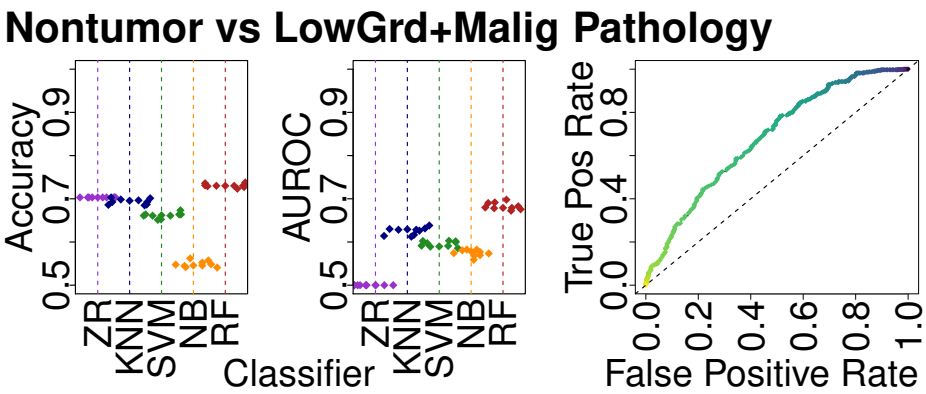

Figure S33: Predicting if an image is (i) nontumor, or (ii) low grade or malignant. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.6804 here for n=1041 per Table 2, which is not significantly different from the mean AUROC=0.683.

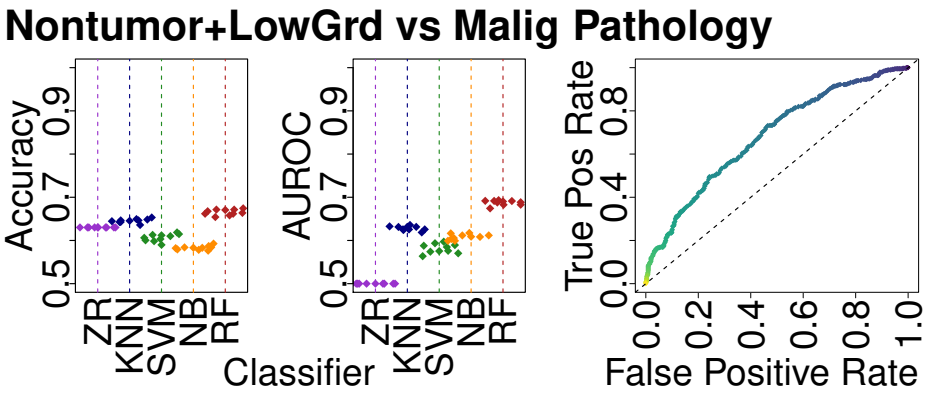

Figure S34: Predicting if an image is (i) nontumor or low grade, or (ii) malignant. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.6761 here for n=1041 per Table 2, which is not significantly different from the mean AUROC=0.687.

consideration. This task, to distinguish tumoral from nontumoral disease, did not perform significantly better than the other nontumor/low-grade/malignant tasks.

S4.2.5 Nontumor+LowGrade vs Malignant task details

Our Random Forest predicts if an image is nontumor/low-grade or malignant (Fig S33). There were 1041images: 385 negative images that were nontumor or low-grade, and 385 positive images that were malignant. There was mild class imbalance of ˜1.7:1. Accuracy is $66.551 \pm 3.255\%$ (chance $63.016 \pm 0.459\%$). AUROC is 0.687 ± 0.052 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical consideration. This task, to distinguish malignancy from everything else, did not perform significantly better than the other nontumor/low-grade/malignant tasks.

S4.3 3-class Nontumoral, Low grade, and Malignant task details

Our Random Forest predicts if an image is nontumoral, low grade, or malignant (Fig S35). There were 1041 images: 309 images that were nontumor, 347 images that were low grade, and 385 images that were malignant. Classes were essentially balanced. Accuracy is $51.239 \pm 4.781\%$ (chance $36.984 \pm 0.459\%$). AUROC is 0.683 ± 0.056 (chance 0.5). Performance in AUROC is statistically significant but too weak for clinical consideration. This may be an interesting task if there is not strong prior information about whether or not the patient has nontumoral disease, unlike our use case for low grade vs malignant.

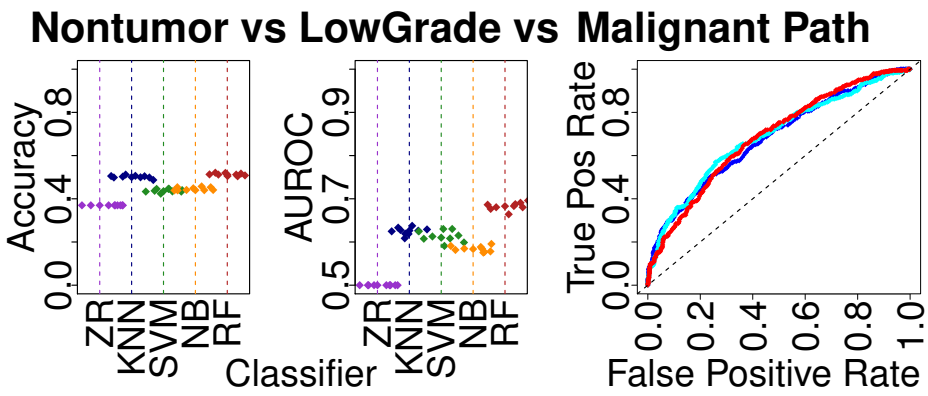

Figure S35: Predicting if an image is nontumor, low grade, or malignant. Plots show accuracy (left) and AUROC (middle) for the classifiers shown in Fig S14. The ROC curve for the highest AUROC classifier [RF] is shown at right, showing AUROC=0.6807,0.6892,0.6877 (for nontumor, low grade, and malignant, respectively – with weighted average AUROC of 0.686) here for n=1041 per Table 3, which is not significantly different from the mean AUROC=0.683.

S5 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 S5.1 discusses further.

S5.1 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.

S6 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 (a) consider at most one image per patient, or (b) allocate an entire message thread to a cross validation fold – but these either reduce statistical power or complicate the baseline analysis by departing from Weka's default cross validation scheme. 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, our nontumor/low-grade/malignant keyword rules may be incorrect, and explicit nontumor/low-grade/malignant annotations for cases from a pathologist may help. Fourth, 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. Fifth, our pipeline crops images, potentially losing important information. Sixth, this is a

retrospective study, so for improved validation, we could make (a) a prospective study using additional data from contributing pathologists, or (b) an independent test set from additional pathologists. Finally, our quality control pipeline does not filter out pathologist markings on these images. Section S6.1 discusses further.

S6.1 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.

There is room for improvement in automated duplicate detection methods. A pathologist may first Tweet an image that has no hand-drawn marks, but later reply with an image that includes hand-drawn marks such as circles and arrows to indicate a region of interest. In future work, these near-duplicates should be automatically detected. Duplicates may artificially inflate performance metrics.

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.

We crop images to convert rectangular images to be uniformly square for the machine learning. However, pathologists may include diagnostic information only at the extreme edges of an image that are cropped out. A case of this from B.X. involves a hydatid cyst in the extreme right of an image, which would be cropped out[15]. This hydatid cyst indicates *Echinococcus* infection, so the case is nontumoral. Learning over random crops of the image, as is commonly done in deep learning data augmentation, may help attend to the image's extreme edges too, rather than systematically ignore them. Another caveat is that our baseline Random Forest method does not consider all images for a patient together as a "bag" for multiple instance learning or similar methods. Multiple instance learning may be especially important for this case, because the "answer" is in only one of the three provided pictures. The other two images provoked discussion about the "intense necrotizing granulomas" in this case.

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.

[15] Case at `https://twitter.com/BinXu16/status/980404471833313280` "Kudo to @drkennethtang @luishcruzc and @DrGeeONE The answer of this case can be seen in the right corner of the 3rd picture. Dx: Echinococcus (hydatid cyst) with necrotizing pneumonia, abscess, and granulomatous inflammation. Additional high power pictures attached."