npj | Digital Medicine

ARTICLE OPEN

Similar image search for histopathology: SMILY

Narayan Hegde[1], Jason D. Hipp[1], Yun Liu[1], Michael Emmert-Buck[2], Emily Reif[1], Daniel Smilkov[1], Michael Terry[1], Carrie J. Cai[1], Mahul B. Amin[3], Craig H. Mermel[1], Phil Q. Nelson[1], Lily H. Peng[1], Greg S. Corrado[1] and Martin C. Stumpe[1,4]

The increasing availability of large institutional and public histopathology image datasets is enabling the searching of these datasets for diagnosis, research, and education. Although these datasets typically have associated metadata such as diagnosis or clinical notes, even carefully curated datasets rarely contain annotations of the location of regions of interest on each image. As pathology images are extremely large (up to 100,000 pixels in each dimension), further laborious visual search of each image may be needed to find the feature of interest. In this paper, we introduce a deep-learning-based reverse image search tool for histopathology images: Similar Medical Images Like Yours (SMILY). We assessed SMILY's ability to retrieve search results in two ways: using pathologist-provided annotations, and via prospective studies where pathologists evaluated the quality of SMILY search results. As a negative control in the second evaluation, pathologists were blinded to whether search results were retrieved by SMILY or randomly. In both types of assessments, SMILY was able to retrieve search results with similar histologic features, organ site, and prostate cancer Gleason grade compared with the original query. SMILY may be a useful general-purpose tool in the pathologist's arsenal, to improve the efficiency of searching large archives of histopathology images, without the need to develop and implement specific tools for each application.

npj Digital Medicine (2019)2:56 ; https://doi.org/10.1038/s41746-019-0131-z

INTRODUCTION

The growing adoption of digital pathology[1] provides opportunities to archive and search large databases of pathology images for diagnosis, research, and education. Histopathology is the examination of biological tissue specimens for diagnostic purposes and is traditionally performed using microscopes. After digitization, images "tagged" (annotated) with clinical data such as diagnoses and patient demographics can be searched based on the text-based tags. For example, searching for "breast" and "carcinoma" in the clinical notes could yield a list of images that were diagnosed or suspected to contain breast cancer.

A relatively unique aspect of histopathology images is that they are typically much larger than those found in other imaging specialties: a typical pathology slide might be 100,000 × 100,000 pixels when digitized at high magnification. Since clinical annotations such as text reports apply to the entire image or sets of images rather than specific locations within the image, matching a search "query" with the location in the image that the search is relevant to can be challenging. For instance, a tumor in a pathology image may be only 100 pixels across, comprising *one-millionth* of the image area. A clinician, researcher, or trainee who has found this image or set of images via searching based on text would still need to visually search the image to locate the lesion before any subsequent analysis. This problem is further compounded because like many disciplines, real-world pathology cases contain multiple (e.g., 5–100) images, and the available text labels might not be specific enough in terms of a particular disease subtype of interest.

In non-medical domains, a potential solution is *reverse image search*, also termed content-based image retrieval (CBIR),[2] to find visually "similar" images. In the diagnostic workflow for example, a clinician may want to search a database for similar lesions to determine if a feature of interest is malignant or a benign histologic mimic, for example in basal cell carcinoma.[3] Relevant tools in non-medical domains include "search by image" for general images,[4] visual search[5] for retail products, and other tools for faces[6] and art.[7] In medical imaging, related works include CBIR for radiology[8–10] and pathology.[11–20] Prior machine learning-based CBIR systems have employed application-specific models, which require collecting labeled data for each application, creating a significant burden to their implementation. Furthermore, "similarity" in these works were defined along specific axes, whereas the intended meaning could vary based on the use case. For example, two images could be similar in that they originate from the same organ, same cancer, similar staining, or similar histologic features.

In this paper, we developed a histopathology similar image search tool (Similar Medical Images Like Yours, SMILY) without using labeled histopathology images. We then evaluated histopathology image search quality in several organs: breast, prostate, and colon, representing three of the four most common non-cutaneous cancer sites. Our evaluation had two components. First, we quantitatively evaluated how often a query image would be matched to an appropriate result from a dataset that was pre-annotated by pathologists. Second, in a blinded prospective study, we had pathologists evaluate how a query image compared to search results that were selected either using SMILY or randomly. In both evaluations, we assessed SMILY's ability to retrieve similar tissue types, histologic features and even disease states such as prostate cancer Gleason grading.

[1]Google AI Healthcare, Mountain View, CA 94043, USA; [2]Avoneaux Medical Institute, Baltimore, MD 21215, USA; [3]Department of Pathology and Laboratory Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA; [4]Present address: AI and Data Science, Tempus Labs Inc, Chicago, IL, USA
Correspondence: Craig H. Mermel (cmermel@google.com) or Martin C. Stumpe (stumpem@gmail.com)
These authors contributed equally: Narayan Hegde, Jason D. Hipp

Received: 31 January 2019 Accepted: 29 May 2019
Published online: 21 June 2019

npj nature partner journals

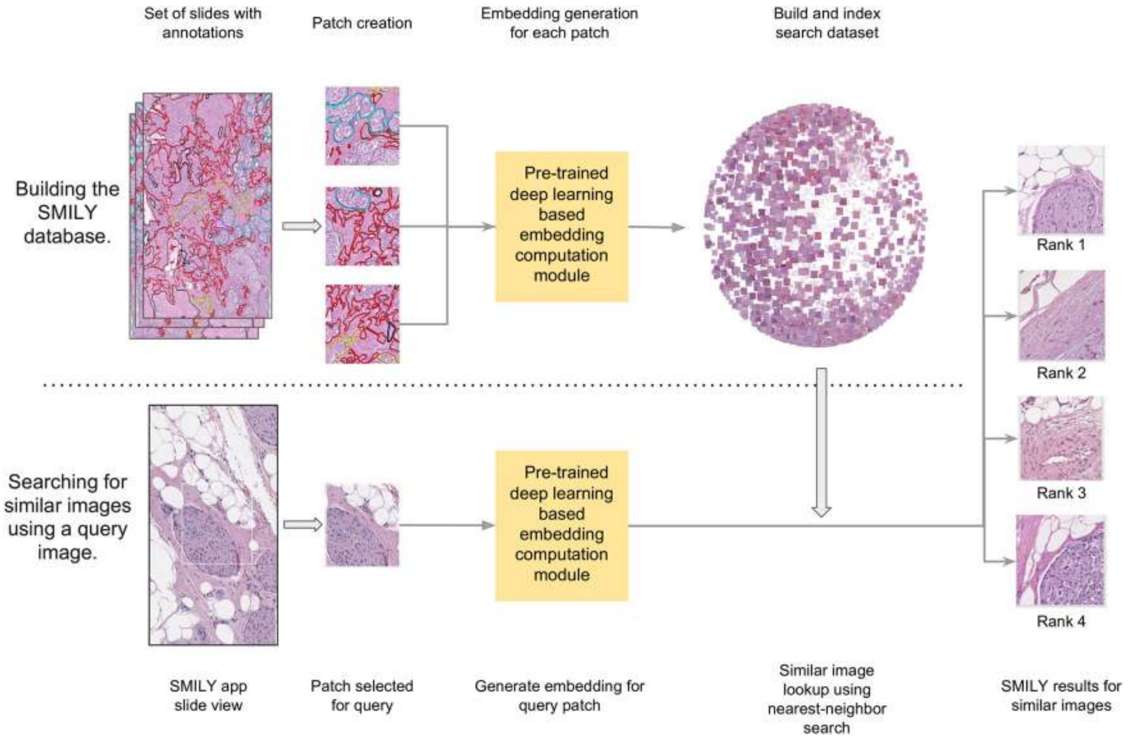

Set of slides with annotations
Patch creation
Embedding generation for each patch
Build and index search dataset
Building the SMILY database.
Pre-trained deep learning based embedding computation module
Rank 1
Rank 2
Rank 3
Rank 4
Searching for similar images using a query image.
Pre-trained deep learning based embedding computation module
SMILY app slide view
Patch selected for query
Generate embedding for query patch
Similar image lookup using nearest-neighbor search
SMILY results for similar images

Fig. 1 Overview of Similar Medical Images Like Yours (SMILY). First, a database of image patches and a numerical characterization of each patch's image contents (termed the embedding) is created. SMILY uses a convolutional neural network to compute this embedding (schematic used for illustration purposes only, see Methods for architecture descriptions). Next, when a query image is selected, SMILY computes the embedding of that query image and compares the embedding with those in the database in a computationally efficient manner. Finally, SMILY returns the k most similar patches, where k is customizable

RESULTS

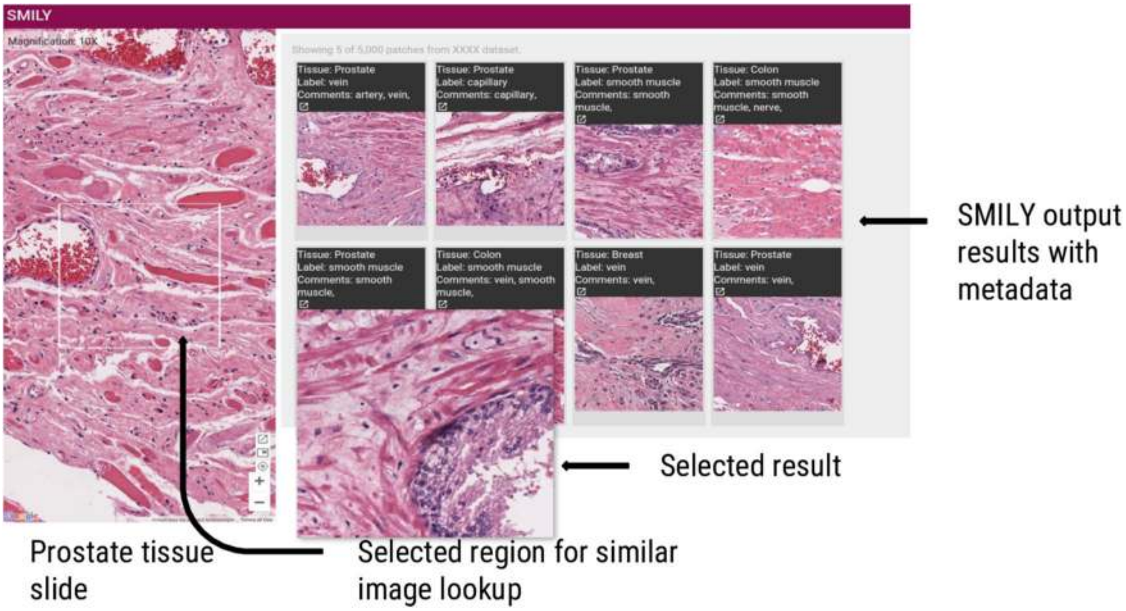
Overview of SMILY and our evaluations

Figure 1 provides an overview of the development and usage of our proposed tool, SMILY. The first step is to create the SMILY database for searching. The core algorithm of this step is a convolutional neural network that condenses image information into a numerical feature vector, termed an embedding. When computed across image patches cropped from slides, this created a database of patches and a numerical summary of each patch's information content. When a region of interest is selected for searching (termed the query image), the embedding for the query image is computed and compared with those in the database to retrieve the most similar patches (Methods). An example of the user interface for SMILY is presented in Fig. 2. In the following evaluations, the database was constructed using images at medium (×10) magnification from the publicly available TCGA (The Cancer Genome Atlas).[21] In total, the evaluations used 127,000 image patches from 45 slides and the query set consisted of 22,500 patches from another 15 slides.

Large-scale quantitative study using annotations

Labels for the first type of evaluation were prepared by pathologists annotating various histologic features in the images, such as arteries, nerves, smooth muscle, and fat. Despite non-exhaustive annotations, this process produced a total annotated area exceeding 128,000,000 $8 \times 8\,\mu m$ regions (each roughly equivalent to a lymphocyte). After sampling to ensure a balanced dataset with respect to classes of interest in each analysis, this produced thousands of image patches per class (Methods, Table 1). Next, for each query patch, we evaluated the performance of SMILY in retrieving patches of the same histologic feature in the database. We used the top-5 score, which evaluates the ability of SMILY to correctly present at least one correct result in the top five search results. This metric was chosen to mimic the standard search process, where a user evaluates a small number of search results to find matches of interest. The subsequent evaluations used image patches extracted at ×10 "medium power" magnification, which is commonly used for reviewing images. The results on other magnifications are presented in Supplementary Fig. 1, and additional performance metrics are presented in Supplementary Fig. 2. The use of multiple magnifications at the same time for additional context also improved performance (Supplementary Fig. 3).

Figure 3a illustrates the results of this large-scale quantitative analysis. When we used query images from prostate specimens, SMILY had a 62.0% top-5 score at retrieving images of the same histologic feature. This was significantly higher than a traditional image feature extractor (scale-invariant feature transform, SIFT) used in related work[17] (44.2%, $p < 0.001$ for all histologic features except nerve, which was non-significant) and random (28.3%). When SMILY retrieved results that did not exactly match the histologic feature, it commonly returned a similar feature, such as another fluid-transporting vessel: capillaries, arteries, veins, and lymphatics (Fig. 4a). Next, we expanded to queries from multiple organs: breast, colon, and prostate. For most histologic features, errors tended to occur between the same histologic features, but across organs (Fig. 4b). For example, the histologic feature match score was at 65.3%, but the combined histologic feature and organ match was lower at 40.0%. Finally, we evaluated the ability of SMILY to retrieve images of the same prostate cancer Gleason pattern (Fig. 3b). SMILY was significantly more accurate than the SIFT baselines at retrieving images with the correct Gleason patterns (73.1% vs. 62.1%, $p < 0.001$), and frequently with both the same Gleason pattern and the same histologic feature (25.3% vs. 17.6%, $p < 0.001$ for comparison with SMILY).



SMILY

Magnification: 10X

Showing 5 of 5,000 patches from XXXX dataset.

Tissue: Prostate
Label: vein
Comments: artery, vein,

Tissue: Prostate
Label: capillary
Comments: capillary,

Tissue: Prostate
Label: smooth muscle
Comments: smooth muscle,

Tissue: Colon
Label: smooth muscle
Comments: smooth muscle, nerve,

Tissue: Prostate
Label: smooth muscle
Comments: smooth muscle,

Tissue: Colon
Label: smooth muscle
Comments: vein, smooth muscle,

Tissue: Breast
Label: vein
Comments: vein,

Tissue: Prostate
Label: vein
Comments: vein,

SMILY output results with metadata
Selected result
Prostate tissue slide
Selected region for similar image lookup

Fig. 2 Sample view of the SMILY user interface. Sample query from a prostate specimen and search results. One of the search results has been magnified for better visualization. Clicking a search result opens a new viewer centered on the result that can be zoomed in for detail or zoomed out for context. Additional examples of queries and search results are presented in Supplementary Fig. 4, including an additional interface for scoring the quality of each search result for the prospective studies with pathologists

Table 1. Summary of data used in large-scale quantitative study

Dataset	Organ site(s)	Categories assessed	Number of slides in the database	Number of patches in the database	Number of slides in the query set	Number of patches in the query set
Organ-specific	Prostate	9 histologic features	15	45,000 (5,000 per feature)	5	9,000 (1000 per feature)
Multi-organ	Prostate, breast, colon	10 histologic features	45	87,000 (3,000 per feature/organ[a])	15	14,500 (500 per feature/organ[a])
Gleason grading[b]	Prostate	Non-tumor and Gleason Patterns 3,4,5 (NT, GP3, GP4, GP5)	20	40,000 (10,000 in each category)	5	8,000 (2,000 in each category)

To avoid biases in the evaluation, we randomly subsampled the original annotated regions, resulting in 5,000 patches per histologic feature per organ

[a] In our study, no lymphocytes were found upon non-exhaustive review of the prostate specimens, so the number of patches exclude this

[b] Not every patch in this dataset was concurrently labeled with histologic features, so 4,000 database patches and 1,600 query patches with both types of annotations were used for assessing the simultaneous match of both Gleason pattern and histologic feature

Study with pathologists

The previous analyses used a large number of patches for evaluations: >20,000 patches in the query set and five times that number in the database. However, one limitation was that they were based on non-exhaustive annotations of histologic features and Gleason patterns. For example, if a query image contains fat only, a retrieved image search result that contains both fat and an artery (but is only annotated as "artery") will be considered an error during the prior evaluation. Thus, our second set of evaluations involved a study with pathologists to assess if at least one histologic feature or cancer grade present in the query image exists in the SMILY search results. As a control to ensure that graders were not artificially scoring SMILY results highly, some search results were from a random search instead of SMILY. The graders were blinded to the source of the search results: SMILY or random. Analogous to the large-scale quantitative stores above, we then assessed search results along multiple axes: histologic feature, organ site, and Gleason grading (Table 2).

Using queries from prostate specimens, SMILY had an average score (Methods) of 62.1% for finding similar histologic features, significantly higher than the random search results (26.8%, $p < 0.001$) (Fig. 5a). "Random" performance exceeds the inverse of the number of categories ($1/9 = 11\%$) because each search result can contain multiple histologic features. When we queried from multiple organs, SMILY's score for histologic feature match was similarly significantly higher than random (57.8% vs. 18.3%, $p < 0.001$, Fig. 5b). In this study, pathologists reported the organ site as unambiguous for only 32.0% of the individual search results. Among these search results, 68.3% were from the same organ site as the query image (Fig. 5c). Finally, graders provided a 0–100 "match quality score" for prostate cancer patches, based on tumor presence, Gleason pattern and histologic features (Methods and Table 3). In this analysis, SMILY scored an average of 61.0%, compared with 30.0% for random results (Fig. 5d, $p < 0.001$).

Visualization of learned embeddings

To better understand SMILY, we visualized the embeddings using t-SNE, a common tool for understanding where data lie in a high-dimensional embedding space.[22] Figure 6a shows that image patches from the same organ site can lie in very different areas in the embedding space. When colored based on histologic features, the clusters tend to have more distinct colors (Fig. 6b). For

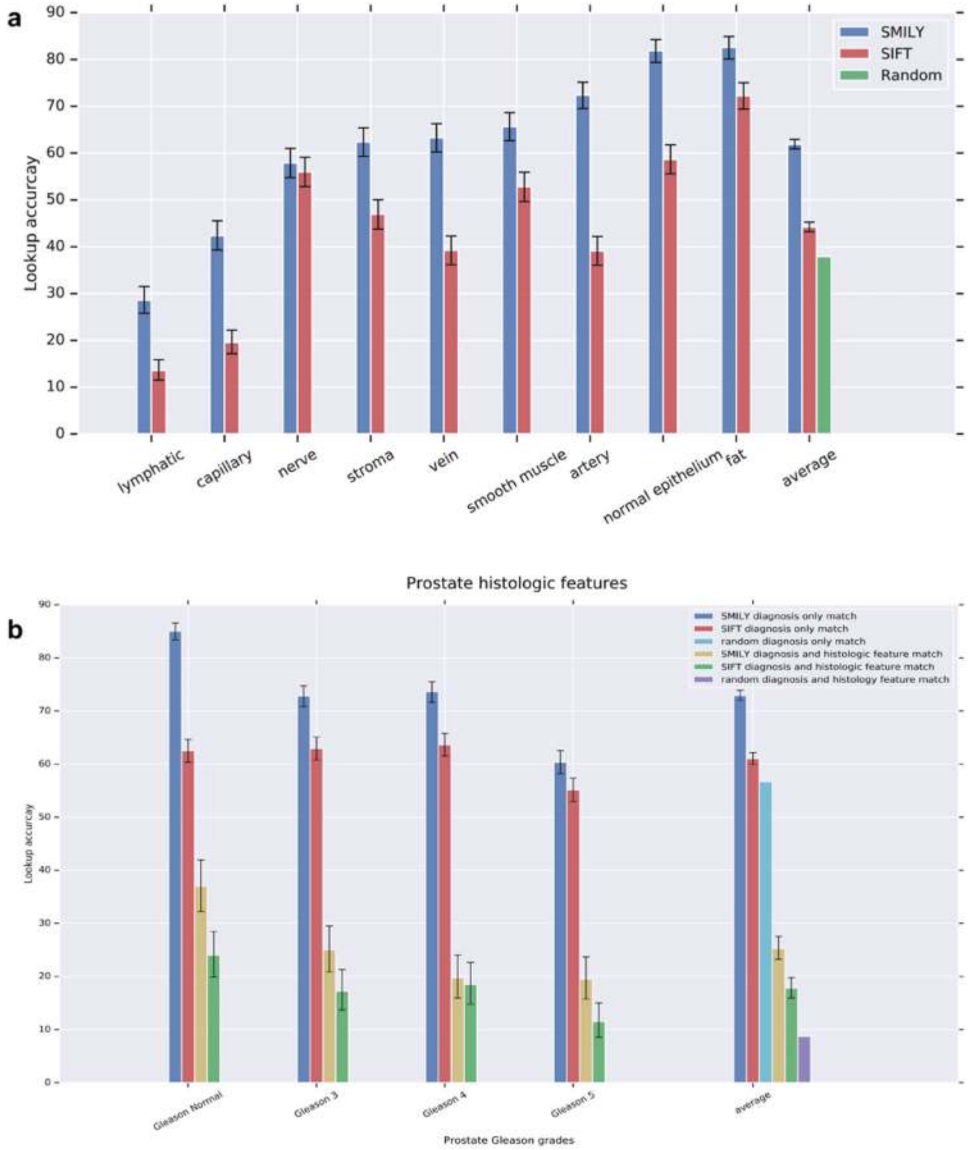


Fig. 3 SMILY search accuracy from large-scale quantitative evaluation using pathologist-provided annotations. **a** Results for histologic feature match in prostate specimens, in comparison with a traditional image feature extractor (scale-invariant feature transform, SIFT) and random search. **b** Results for prostate cancer Gleason grade and histologic feature match, in comparison with the same baselines. Error bars indicate 95% confidence intervals (Methods)

example, the bottom left prostate cluster in Fig. 6a is composed of a mixture of histologic features such as arteries, lymphatic vessels, and capillaries in Fig. 6b.

Computational efficiency of searching

Finally, to investigate the computational efficiency of SMILY for datasets of realistic sizes, we created a database for all prostate, breast, lung, and colon specimens from TCGA, at four magnifications: ×40, ×20, ×10, and ×5. This generated about 10^9 image patches. Using 400 computers with ten compute threads each, and some optimizations to use a hash table instead of the kd-tree depending on the local embedding density,[23] queries had a median query time of 1.3 s. This can in principle be further accelerated using text-based search to filter images, and real-time updating of search results to present preliminary results before the search completes. By contrast, a naive implementation on a single machine with 10^7 image patches (100 times fewer than above) required a significantly slower 25 s per query.

DISCUSSION

This study presents SMILY, a tool to search for similar histopathology images using an image as the query. To our knowledge, we have performed the most comprehensive evaluation of a reverse image search tool for histopathology. SMILY retrieves image search results with similar histologic features, organ site, and cancer grades, based on both large-scale quantitative analysis using annotated tissue regions and prospective studies with pathologists blinded to the source of the search results. In the rest of this discussion, we will discuss some nuanced issues regarding similar image search: what 'similarity" means; what a tool like SMILY can be used for; comparison with "traditional" application-specific approaches; how SMILY was developed and what that means for future applications not covered in our evaluations; comparison with prior work; and finally technical implementation considerations.

First, the meaning of "accuracy" in the setting of a similar image search tool deserves some thought. From first principles, the ideal

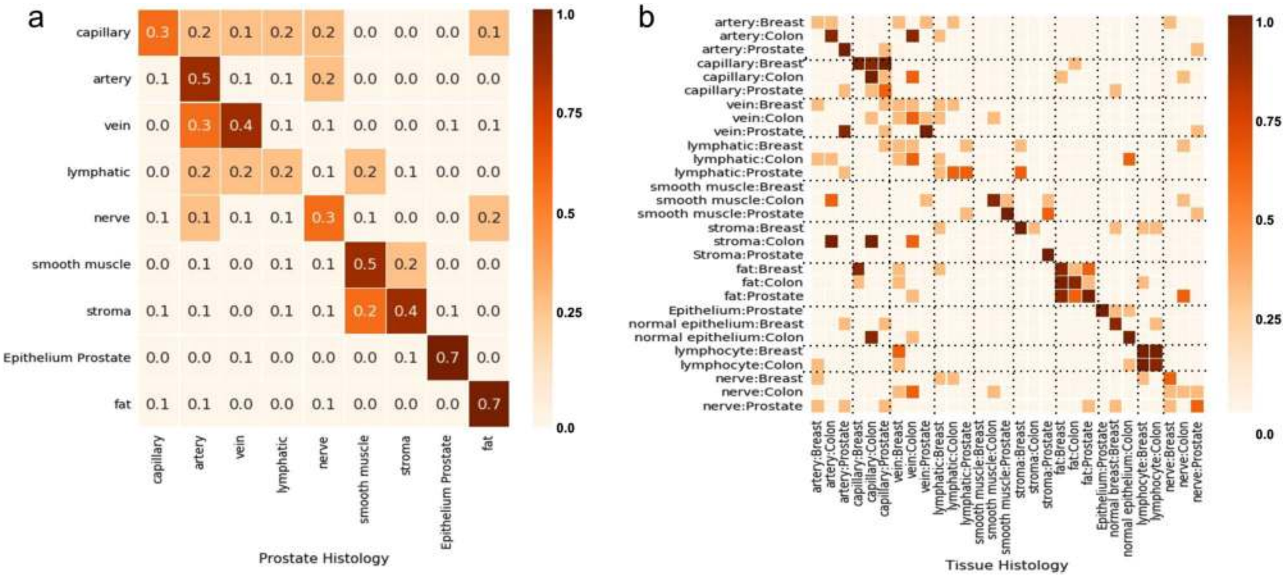


Fig. 4 Confusion matrix from SMILY search. An element in row *i*, column *j* indicates the fraction of search results for query *i* that result in a "hit" based on the top-5 score for the category *j*. **a** Confusion matrix for the results from Fig. 3a: histologic feature match in prostate specimens. **b** Confusion matrix for histologic feature match across prostate, breast, and colon specimens. To improve visual contrast and highlight trends better, only discrete colors and rounded-off values are used

Table 2. Summary of data used in the studies with pathologists

Dataset	Organ site	Categories assessed	No. of patches[a] queried by SMILY	No. of patches[a] queries by random search (negative control)	Scoring system (see Methods)
Organ-specific	Prostate	9 histologic features	270 by 2 pathologists	90 by 2 pathologists	0 or 100 for histologic feature match
Multi-organ	Prostate, breast, colon	10 histologic features	410 by 2 pathologists	145 by 2 pathologists	2 scores: 0 or 100 for histologic feature match, and 0 or 100 or "unclear" for organ match
Gleason grading	Prostate	Non-tumor and Gleason Patterns 3,4,5 (NT, GP3, GP4, GP5)	250 by 2 pathologists	90 by 2 pathologists	0 to 100 for tumor grade and histologic feature match

The database used for this study is identical to Table 1, while the query set was subsampled to retain a tractable number for manual evaluations

[a] Whether search results were returned by SMILY versus random were randomized based on the ratios specified in this table and so the final numbers of patches are approximate

search tool displays what you are searching for. However, this goal is ambiguous because the intent of the search depends on the use case: searching for other images with the same stain, similar stain intensity, same histologic feature, or similar lesion in the most general sense. As such, in the absence of information about search intent, the ideal tool should surface a breadth of search results instead of focusing on any single axis of similarity. To address the lack of algorithmic awareness for the search intent, advances in human-computer interaction may enable interactive refinement of search results based on certain desired axes of similarity.[24]

The potential use cases for a tool like SMILY can be categorized into diagnosis, research, and education. In diagnosis, SMILY could be a helpful tool to search for similar lesions within the same slide or in other patients. For example, the pathologist might want to know the frequency of that lesion or histologic feature's occurrence in the specimen, such as searching for the region of highest mitotic activity and then counting mitoses in ten high-powered fields-of-view for breast cancer.[25] Searching for rare features in other slides may be helpful in rare diagnoses, to better understand the prognosis of other patients with similar features, including potentially rare pathologies from historic cases with known intervention and treatment response. For research, a clinician might have a hypothesis: occurrence of a certain histopathological feature in the slide is correlated with clinical endpoints. However, an adequately powered study may require a large number of patients, rendering the manual search for these features highly labor intensive. SMILY could enable significant speedups in this search via computer-assisted search. Finally, trainees are frequently confronted with unknown lesions. Manual searching of pathology textbooks, atlases, and other resources for similar lesions can be time-consuming; SMILY could reduce this process to an image-based query and manual assessment for the most relevant result. Importantly, these searches could also leverage large publicly available databases such as the TCGA, as we have done here.

Indeed, with respect to specific applications such as mitotic counting,[26] approaches that have been developed specifically for that application may result in higher accuracy for that purpose. However, developing and implementing specific but separate approaches for every possible task of interest is impractical. Some challenges are: expensive data collection and labeling, difficulty of workflow integration and potential legal or commercial issues, and

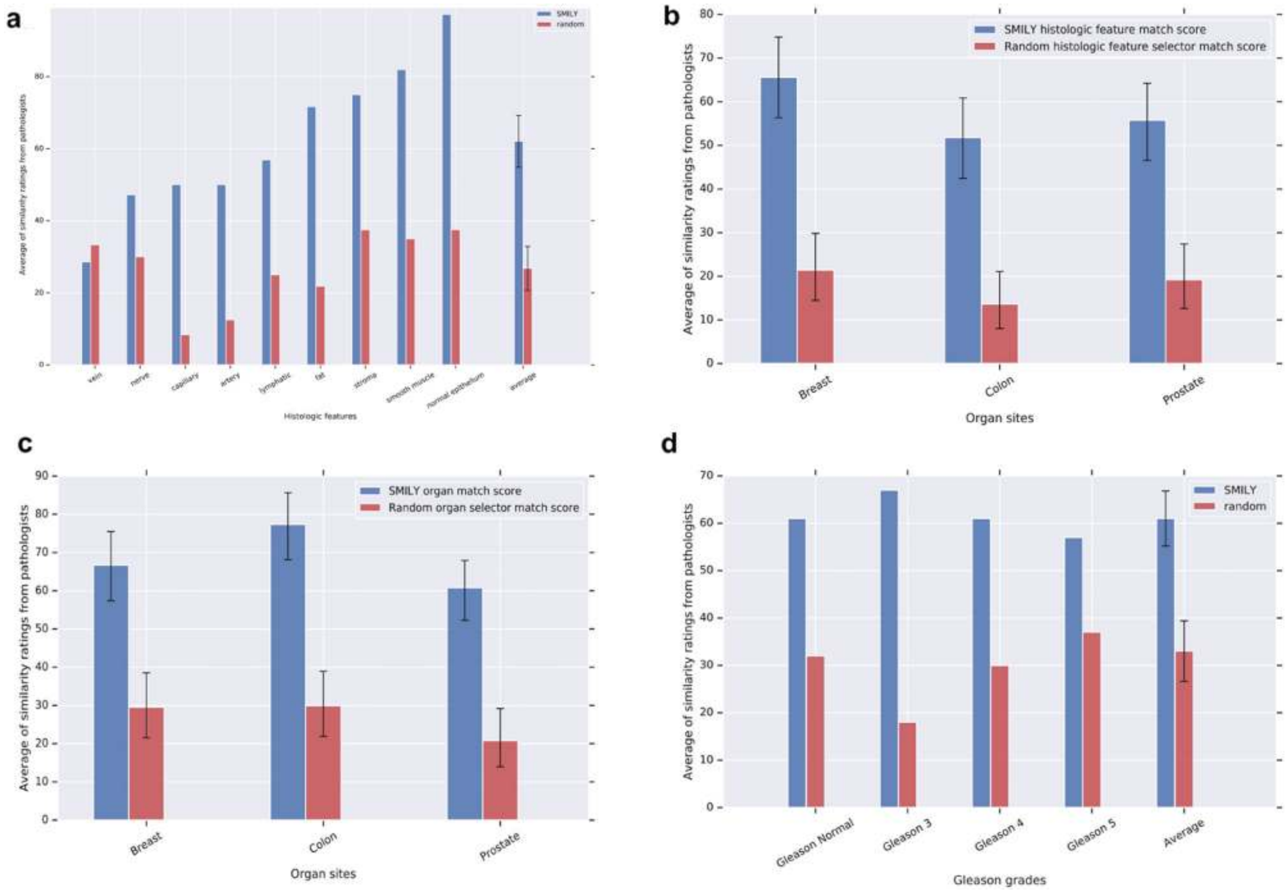


Fig. 5 Evaluation of SMILY from studies with pathologists. The pathologists evaluated search results, blinded to whether the results were retrieved by SMILY versus a negative control, random selection. The similarity scoring rubrics are detailed in the "Prospective studies with pathologists" subsection in the Methods. **a** Histologic feature match in prostate specimens. **b** Histologic feature match in prostate, breast, and colon specimens. **c** Organ site match in prostate, breast, and colon specimens. **d** Overall match score (Table 3) in prostate specimens for similarity in histology and prostate cancer Gleason grade. Error bars indicate 95% confidence intervals (Methods)

Table 3. Overall match quality score for multi-aspect similarity evaluation

Score	Criteria
0	If the presence/absence of tumor in both patches do not match and they look visually different
25	If the presence/absence of tumor in both patches do not match and but histologic features match
50	If the presence of tumor in both patches match but not the tumor grade
75	If the diagnostic grades match or both patches are normal (e.g. Gleason grade for prostate)
100	If the diagnostic grades match or both patches are normal (e.g. Gleason grade for prostate) in addition to at least one histologic feature match

lack of machine learning, software or hardware expertize for development or implementation. As such, the availability of a general-purpose tool like SMILY that can be used in multiple applications, can be helpful despite having lower accuracy than an application-specific tool.

An interesting aspect of SMILY is that the core neural network algorithm was not trained using histopathology images. Instead, the network was trained using a dataset of images, including people, animals, and man-made and natural objects (see Methods). Thus, our approach does not require the use of large, pixel-annotated datasets such as those used for breast cancer mitotic figure detection,[27] breast cancer metastasis,[28] or "image-level" labels such as those extracted from pathology reports.[29] Although the network was not exposed during training to imaging characteristics from different laboratories or slide scanners, the study images we tested SMILY on were digitized on several scanners (Aperio, 3DHISTECH, and Hamamatsu), different magnifications (× 40 and × 20), different color spaces (RGB and YUV), and different compressions schemes (JPEG and JPEG2000). In principle, the development of a similar histopathology "similarity" dataset could further improve the embeddings learned by the model, and is the subject of future work. The results of using several other "pre-trained" neural networks are presented in Supplementary Fig. 5.

CBIR has been studied extensively in medical imaging,[9,10,30] and in histopathology for both slides[14,18] and image patches.[11–13,15,19,20,31] However, the models underlying these CBIR systems require pathologist-annotated labels for development,

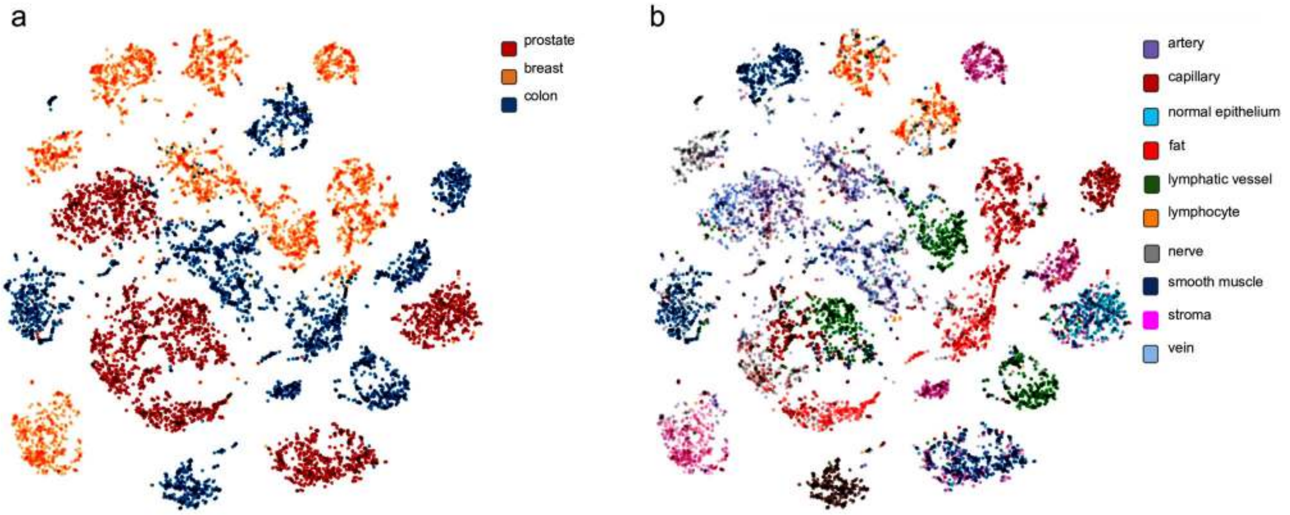


Fig. 6 Visualizations of the embeddings of image patches in the SMILY database. Each dot represents an image patch. **a** Colored by organ site, indicating that patches from the same organ were distributed among different clusters. **b** Colored by histologic feature, indicating a more distinct separation between histologic features

which is both costly and non-scalable. These annotations also in turn restrict the concept of similarity to be along a few predetermined axes, such as cancer grade and histologic features. In prior work, lookup accuracies ranged from 60 to 80% for breast, prostate, necrotic and leukocyte organ sites and histologic features.[13,15] By contrast, SMILY achieved comparable performance without the use of application-specific labeled data for development, and thus can potentially be applied to applications without labeled training data. Specifically, although the neural network underlying SMILY was trained using supervised learning based on non-histopathological images (Methods), the histopathology annotations we collected were used exclusively for evaluating SMILY and not training. Alternative histopathology CBIR approaches in the absence of labeled training data include SIFT, kernel and Fourier functions.[17,32] We have performed a quantitative comparison with SIFT to evaluate the added value of using a neural network-based system like SMILY for automatic feature extraction, showing a significant improvement in lookup accuracy across multiple image retrieval experiments.

The large size of each histopathology image and the scale of typical histopathology databases (10^3–10^6 images) raise important technical considerations for real-world use. First, the embedding of each patch needs to be calculated as a one-time computation cost. This incurs a delay to compute the embeddings for the 10^3–10^6 patches in each newly digitized slide before the slide can be searched across. Second, these embeddings need to be stored to avoid repeated embedding computation. Although this overhead was only 0.4% per gigabyte-sized image in our studies, this storage requirement increases with the number of magnifications of interest, and density of sampling each slide. Finally, the search phase requires comparing the query image embedding with millions or billions of other embeddings. For example, a naive implementation of this process on the entirety of the publicly available The Cancer Genome Atlas (TCGA, contains over 33,000 slides) dataset[33] will incur an impractical, half-minute latency on a modern desktop computer. To support real-world usage, we have optimized this process to require only seconds on a web interface (Methods).

This study contains limitations, such as those discussed in-depth above regarding accuracy of a similar image search tool and limitation of a general-purpose search tool versus application-specific tools. In addition, the number of slides could be increased to better capture the breadth of tissue processing conditions and resulting images. We have evaluated "similarity" in terms of histologic features, organ site, and prostate cancer grade, but there are other axes of similarities that SMILY will need to be further validated on. Additional studies will be needed to fully assess and improve the intraobserver and interobserver variability for pathologists' scoring of "similarity". Lastly, future work will also need to tackle the 'refinement" of search results along specific similarity axes of interest, to enable more targeted image-based search for histopathology.

METHODS

Neural network architecture

SMILY is based on a convolutional neural network architecture called a deep ranking network.[34] We chose to use neural networks for this task because of their success in recent years in image-related tasks such as classification and object detection, by learning the discriminative features instead of needing to be specifically designed.[35] Briefly, the deep ranking network is based on an embedding-computing module that compresses input image patches (of dimensions width x height x channels) into a fixed-length vector. This module contains layers of convolutional, pooling, and concatenation operations. During training, the network was fed labeled sets of three images: a reference image I of a certain class, a second image I_+ of the same class, and a third image I_- of a different class. The network then uses the modules to compute the embeddings of each of the three images. The network is then trained to assign a lower distance between the embeddings of (I, I_+) than the embedding distances of (I, I_-). Our network was trained on about 500,000,000 "natural images" (e.g., dogs, cats, trees, man-made objects etc) from 18,000 distinct classes. In this way, the network learned to distinguish similar images from dissimilar ones by computing and comparing the embeddings of input images.[34,36] This network was successfully leveraged to generate embeddings that discriminated between cellular phenotypes in high-content screening.[37]

Building the SMILY embedding database

For the experiments described in this paper, we used slides from The Cancer Genome Atlas (TCGA).[21] TCGA was used because it is publicly available and widely used for histopathology studies. TCGA tissue samples were collected with approval of local Institutional Review Boards (IRBs), with the informed consent of patients. Ethics review and IRB exemption for the use of de-identified images in this study was obtained from Quorum Review IRB (Seattle, WA).

Additional details about each experiment's dataset are provided in the respective study sections. SMILY uses the embedding-computing module from the deep ranking network (Fig. 1) to compress input image patches (300 × 300 pixels 3-channel RGB (red-green-blue) images) into embeddings vectors of size 128. As histopathology images are orientation-independent, we additionally generate the four 90-degree rotations of each input image, and the mirrored and rotated versions for a total of eight orientations, and correspondingly eight 128-sized embeddings per image patch, a 260-fold dimensionality reduction. Even in the absence of any additional compression, storing these embeddings required a reasonable additional 0.4% storage overhead compared to storing the original images alone.

To create image patches at a given image magnification (e.g., high magnifications like × 40 and × 20, or medium magnification like × 10), we extracted thousands of non-overlapping patches per category (Table 1). For a real use case, overlaps can be used to ensure each histologic feature is contained entirely in a patch of the appropriate magnification, instead of being potentially bisected into two patches.

Querying the database

To retrieve matches from the database, SMILY first computes the embedding for a selected query image patch, and then compares that embedding with the embeddings stored in the database. For this work, our comparison function was the L_2 distance across pairs of 128-sized embedding vectors. To handle the eight orientations (see above), we filtered the search results such that the most similar orientation was returned, and only one orientation for each distinct image patch was presented in each set of search results. In addition, to enhance diversity of the search results, we filtered the results to ensure that no results were within 1,000 pixels of each other.

Our experiments (described below) required large numbers of embedding comparisons, ranging from 40,000 to 90,000. To enable efficient lookups, we used k-dimensional (k-d) trees[38] with a leaf size of 40 and depth of 6; this is customizable to fit computation resources and speed requirements. To further optimize lookup speed, we parallelized the comparisons across multiple machines (Results). These steps provided a lookup time sublinear in the number of comparisons.

SMILY's user interface

SMILY was implemented as a web-based whole-slide viewer (Fig. 2). To conduct a search, a user selects a rectangular image patch between 200 and 400 pixels in height and width. For a query patch that is not 300 × 300 pixels, SMILY resizes to 224 × 224 pixels using bilinear interpolation before computing the embedding. The embedding is then used to search the database based on the current magnification in the selected region, and the results are displayed as a customizable number of image patches. Optionally, any existing pixel-level annotations or slide-level metadata such as the original diagnosis can be displayed as well.

Evaluations

To evaluate the utility of SMILY, we conducted several experiments by building a SMILY database using the TCGA dataset, and then conducting studies to examine the quality of SMILY image search results.

Large-scale quantitative studies

Our large-scale quantitative experiments were based on regions annotated by pathologists with various labels (Table 1). In each case, pathologists annotated slides with various histologic features (up to ten categories: artery, capillary, fat, lymphatic vessel, lymphocyte, nerve, normal epithelium, smooth muscle, stroma, vein) or Gleason patterns (four categories: non-tumor and three Gleason patterns). Annotations were performed by three pathologists using the Hamamatsu NDPview2 whole-slide image viewer,[39] using the free-hand outlining and labeling tool. The pathologists were requested to annotate about 80% of each slide, and to look for and annotate rarer histologic features as data collection proceeded and significant skew emerged between categories. Because of the large size of each slide and the complexity in the appearance of each feature of interest, annotations were not 100% precise. Instead, annotations were requested to be precise enough to have ~70% "purity" with respect to the label of interest. A sample of these annotations are shown in the leftmost image in Fig. 1. These annotations were used only for evaluating SMILY, and not for developing the SMILY embedding neural network.

Patches of size 300 × 300 pixels for each histologic feature or Gleason pattern category were then extracted based on the annotations and stored in the SMILY database along with their embeddings computed at the appropriate magnification. To avoid class imbalance for these experiments, we subsampled hundreds to thousands of patches without replacement for each category (Table 1). One exception was for the experiments assessing both Gleason pattern and histologic feature match, where the patches were filtered for only those that contained both types of annotations.

Prospective studies with pathologists

In addition to the large-scale studies, we conducted similar studies by asking pathologists to rate the quality of search results. As our annotations (for the large-scale quantitative studies) were non-exhaustive, these studies with pathologists allowed regions containing multiple labels (but annotated only as one label) to be assessed correctly. The same database from the large-scale quantitative studies were used, but the query data were subsampled to hundreds instead of thousands, to retain a tractable number of search results for manual evaluation by pathologists. To mimic the use case of a user assessing multiple search results for a single query, each query generated four search results, and the pathologist rated each of the four results (the scoring system for each experiment is described below). The final average score for each study is the average score across all search results for all queries.

We used several scoring rubrics for similarity. Similarity along histologic feature, organ, and Gleason pattern were assessed analogously to the large-scale studies using binary scores (100 for "match" and 0 for "not-match"). For an overall "match quality score" combining multiple axes, we devised a 100-point score in 25-point increments (Table 3). A few samples of the user interface for the histologic feature and organ similarity assessment are presented in Supplementary Fig. 3.

In total, three anatomic pathologists from diverse backgrounds participated in this study: 1 U.S. board-certified, 1 non-U.S. board-certified, and 1 U.S. residency-trained. As a negative control to ensure that our pathologists were not artificially rating SMILY search results highly, 25% of the queries returned search results from random selection (i.e., all four images were from SMILY or all four images were randomly selected). The pathologists were blinded to the source of the search results: SMILY versus random.

Statistical analysis

To assess the statistical significance of our results, we used the McNemar test for differences in "binary" accuracy metrics, and the Mann–Whitney *U*-test for differences in non-binary accuracy metrics (avoiding assumptions of normality). Because of the large size of each study, most differences were statistically significant except where noted in one instance (nerve in histologic feature match). 95% confidence intervals were computed using the Clopper Pearson interval for binary metrics (which use the top-5 accuracy to summarize the results of each query) and ± 1.96 standard error for the non-binary accuracy metrics (which are averaged for the search results for each query).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Our study used images accessible from the Genomic Data Commons portal,[40] which is based upon data generated by the TCGA Research Network.[33]

CODE AVAILABILITY

Code for the neural networks evaluated in this work are based on publicly available architectures from various authors on Github as "Image Similarity Deep Ranking", and Inception (V3).

ACKNOWLEDGEMENTS

Funding for this study was provided by Google LLC. For technical advice and discussion, we thank the following, who are all employees of Google LLC: Alvin Rajkomar, MD, Daniel Tse, MD. For software infrastructure, logistical support, and slide digitization services, we thank members of the Google AI Healthcare Pathology team. Lastly, we are deeply grateful to the pathologists Isabelle Flament, Trissia Brown, and Stanislaw Krajewski who provided annotations for this study, or participated in the "prospective study with pathologists".

AUTHOR CONTRIBUTIONS

L.H.P. and M.C.S. conceived the idea; N.H., J.D.H., Y.L. and M.T. designed the experiments; N.H., E.R., D.S. and C.J.C. wrote code for various parts of the work; L.H.P., M.C.S., P.Q.N. and G.S.C. acquired the tissue samples for use in the study and provided strategic support; M.B.A., M.E.B. and J.D.H. created the guidelines for the studies with pathologists and designed the evaluation scoring systems; C.H.M. and M.C.S. supervised the project; N.H., Y.L., C.H.M. and J.D.H. wrote the manuscript with the assistance and feedback of all other co-authors. N.H. and J.D.H. made complementary contributions based on their respective backgrounds in software engineering and pathology and are co-first authors.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Digital Medicine* website (https://doi.org/10.1038/s41746-019-0131-z).

Competing interests: N.H., J.D.H., Y.L., E.R., D.S., M.T., C.J.C., C.H.M., P.Q.N., L.H.P., G.S.C. and M.C.S. are employees of Google LLC and own Alphabet stock. M.E.B. and M. B.A. were compensated for their expertize and time as pathologists.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Mukhopadhyay, S. et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am. J. Surg. Pathol.* **42**, 39–52 (2018).
2. Lew, M. S., Sebe, N., Djeraba, C. & Jain, R. Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans. Multimed. Comput. Commun. Appl.* **2**, 1–19 (2006).
3. Stanoszek, L. M., Wang, G. Y. & Harms, P. W. Histologic mimics of basal cell carcinoma. *Arch. Pathol. Lab. Med.* **141**, 1490–1502 (2017).
4. Google Images. https://images.google.com/. (Accessed 17 Jan 2019).
5. Amazon Flow. *A9*. https://a9.com/what-we-do/visual-search.html. (Accessed 17 Jan 2019).
6. Borovikov, E. & Vajda, S. Facematch: Real-world Face Image Retrieval. In *2016 International Conference on Recent Trends in Image Processing and Pattern Recognition* (RTIP2R). (Springer, Singapore, 2016).
7. Ivanova, K. Content-Based Image Retrieval in Digital Libraries of Art Images Utilizing Colour Semantics. (eds Gradmann, S., Borri, F., Meghini, C., Schuldt, H.) 2011 Research and Advanced Technology for Digital Libraries (TPDL). Lecture Notes in Computer Science, Vol. 6966. (Springer, Berlin, Heidelberg, 2011).
8. Sklan, J. E. S., Plassard, A. J., Fabbri, D. & Landman, B. A. Toward content based image retrieval with deep convolutional neural networks. (eds Gimi, B., Molthen, R. C.) *Proc. SPIE—the International Society for Optical Engineering*. Vol. 9417 (SPIE, Bellingham, WA, 2015).
9. Ahmad, J., Sajjad, M., Mehmood, I. & Baik, S. W. SiNC: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. *PLoS ONE* **12**, e0181707 (2017).
10. Müller, H., Michoux, N., Bandon, D. & Geissbuhler, A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *Int. J. Med. Inform.* **73**, 1–23 (2004).
11. Wang, J. Z. Pathfinder: multiresolution region-based searching of pathology images using IRM. (ed. Overhage, J.M.) *Proc. AMIA Symp*. 883–887 (Hanley & Belfus, Inc, Philadelphia, PA, 2000).
12. Komura, D. et al. Luigi: Large-scale histopathological image retrieval system using deep texture representations. *bioRxiv* 345785 (2018). https://doi.org/10.1101/345785
13. Qi, X. et al. Content-based histopathology image retrieval using CometCloud. *BMC Bioinforma.* **15**, 287 (2014).
14. Zheng, Y. et al. Histopathological whole slide image analysis using context-based CBIR. *IEEE Trans. Med. Imaging* **37**, 1641–1652 (2018).
15. Sridhar, A., Doyle, S. & Madabhushi, A. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. *J. Pathol. Inform.* **6**, 41 (2015).
16. Mosquera-Lopez, C., Agaian, S., Velez-Hoyos, A. & Thompson, I. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE Rev. Biomed. Eng.* **8**, 98–113 (2015).
17. Mehta, N., Alomari, R. S. & Chaudhary, V. Content based sub-image retrieval system for high resolution pathology images using salient interest points. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2009**, 3719–3722 (2009).
18. Kwak, J. T., Hewitt, S. M., Kajdacsy-Balla, A. A., Sinha, S. & Bhargava, R. Automated prostate tissue referencing for cancer detection and diagnosis. *BMC Bioinforma.* **17**, 227 (2016).
19. Babaie, M. et al. Classification and retrieval of digital pathology scans: a new dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. (Honolulu, HI, 2017).
20. Otalora, S., Schaer, R., Atzori, M., del Toro, O. A. J. & Muller, H. Deep learning based retrieval system for gigapixel histopathology cases and open access literature. *bioRxiv* 408237 (2018). https://doi.org/10.1101/408237
21. The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas-National Cancer Institute* (2011). https://cancergenome.nih.gov/. (Accessed 13 Dec 2018).
22. Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
23. Jegou, H., Douze, M. & Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. 10th European Conference on Computer Vision*. Part I 304–317 (Springer, Berlin, Beidelberg, 2008).
24. Cai, C. J. et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems.* (ACM, New York, NY, 2019).
25. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403–410 (1991).
26. Veta, M. et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image. Anal.* **20**, 237–248 (2015)
27. Veta, M. et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med. Image. Anal.* **54**, 111–121 (2019).
28. Litjens, G. et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience* **7**, giy065. (2018).
29. Campanella, G., Silva, V. W. K. & Fuchs, T. J. Terabyte-scale deep multiple instance learning for classification and localization in pathology. Preprint at *arXiv [cs.CV].* https://arxiv.org/abs/1805.06983 (2018).
30. Akgül, C. B. et al. Content-based image retrieval in radiology: current status and future directions. *J. Digit. Imaging* **24**, 208–222 (2011).
31. Sparks, R. & Madabhushi, A. Out-of-sample extrapolation utilizing semi-supervised manifold learning (OSE-SSL): content based image retrieval for histopathology images. *Sci. Rep.* **6**, 27306 (2016).
32. Caicedo, J. C., González, F. A. & Romero, E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. *J. Biomed. Inform.* **44**, 519–528 (2011).
33. The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas-National Cancer Institute* (2011). https://cancergenome.nih.gov/. (Accessed 13 Dec 2018).
34. Wang, J. et al. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1386–1393 (2014).
35. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
36. Image Similarity Data. https://sites.google.com/site/imagesimilaritydata. (Accessed 17 Jan 2019).
37. Michael Ando, D., McLean, C. Y. & Berndl, M. Improving phenotypic measurements in high-content imaging screens. *bioRxiv* 161422 (2017). https://doi.org/10.1101/161422
38. Friedman, J. H., Bentley, J. L. & Finkel, R. A. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.* **3**, 209–226 (1977).
39. NDP. view2 Viewing software U12388-01. https://www.hamamatsu.com/jp/en/product/type/U12388-01/index.html. (Accessed 13 Dec 2018).
40. GDC. https://portal.gdc.cancer.gov/. (Accessed 1 Feb 2019).