ARTICLES
https://doi.org/10.1038/s42256-019-0052-1

nature machine intelligence

Corrected: Publisher Correction; Publisher Correction

Pathologist-level interpretable whole-slide cancer diagnosis with deep learning

Zizhao Zhang[1], Pingjun Chen[2], Mason McGough[2], Fuyong Xing[3], Chunbao Wang[4], Marilyn Bui[5], Yuanpu Xie[2], Manish Sapkota[6], Lei Cui[2], Jasreman Dhillon[5], Nazeel Ahmad[7], Farah K. Khalil[5], Shohreh I. Dickinson[5], Xiaoshuang Shi[2], Fujun Liu[6], Hai Su[2], Jinzheng Cai[2] and Lin Yang[2]*

Diagnostic pathology is the foundation and gold standard for identifying carcinomas. However, high inter-observer variability substantially affects productivity in routine pathology and is especially ubiquitous in diagnostician-deficient medical centres. Despite rapid growth in computer-aided diagnosis (CAD), the application of whole-slide pathology diagnosis remains impractical. Here, we present a novel pathology whole-slide diagnosis method, powered by artificial intelligence, to address the lack of interpretable diagnosis. The proposed method masters the ability to automate the human-like diagnostic reasoning process and translate gigapixels directly to a series of interpretable predictions, providing second opinions and thereby encouraging consensus in clinics. Moreover, using 913 collected examples of whole-slide data representing patients with bladder cancer, we show that our method matches the performance of 17 pathologists in the diagnosis of urothelial carcinoma. We believe that our method provides an innovative and reliable means for making diagnostic suggestions and can be deployed at low cost as next-generation, artificial intelligence-enhanced CAD technology for use in diagnostic pathology.

Most carcinoma identification requires microscopy-level image assessment for early tumour discovery and for developing therapies based on diagnostic pathology. Diagnosing pathology slides is a complicated task that requires years of pathologist training. Unlike other medical image types (for example, radiology images), digital pathology slides are obtained at very high resolution (more than 10 gigapixels). However, region- and cellular-level features, which have significant implications for diagnosis and treatment, can be subtle and confusing. Facing the enormous amounts of information that exists in huge slides and the heavy workload in clinics, even experienced pathologists are prone to misdetect features and make errors[1,2]. Thus, qualified cancer diagnosis requires peer review and consensus[3], a standard that can be expensive to satisfy in hospitals and small cancer centres where experienced pathologists are scarce. We show that artificial intelligence (AI) has the potential to assure trust in computer-aided diagnosis (CAD) in diagnostic pathology, with the goal to offer reliable diagnosis, objective second opinions, strong generalizability and cost-effective deployment, all of which have the potential to greatly improve the routine pathology experience.

With the development of AI, deep learning with deep convolutional neural networks (CNNs)[4] has been shown to be a powerful algorithm for advancing biomedical image analysis[5,6]. CNNs can be applied to pathology image analysis tasks such as the detection of tumours and quantifying cellular features[7,8]. Previous work in pathology diagnosis only classified small tissue images[7,9,10], used poorly generalizable image-processing techniques[11] or lacked large-scale validation[12,13]. Unfortunately, these CNN-based methods are typically trained to process image pixels and thus predict disease labels. Despite its recognized human-level classification accuracy in relation to several diseases[5,6], using deep learning as a diagnostic prediction mechanism is discouraged due to its lack of interpretability[14,15]. Doctors prefer explicit declarative representations from machines that they can perceive and comprehend in order to determine their decision boundaries. Moreover, whole-slide diagnosis requires sophisticated algorithm design and coordination to conduct the multi-level, multi-structure pixel-level analysis of slides. Thus practical CAD for diagnostic pathology remains a substantial challenge.

We have developed a novel method to effectively automate whole-slide reading and the diagnosis processes of pathologists. Specifically, our method diagnoses a slide via region-level tumour detection, pixel-level morphological analysis of nuclear anaplasia and architectural abnormalities, and establishes slide-level diagnosis. Each process is powered by neural networks; their cascading progressively encodes enormous pixels into meaningful and compact representations. Instead of only predicting diagnosis labels, our method includes interpretability mechanisms to decode learned representations into rich interpretable predictions, which are understandable to pathologists. Specifically, the system generates natural language descriptions of microscopic findings (diagnostic tissue cell and nucleus characteristics), whose structures conform to the clinical pathology report standard. The model simultaneously visualizes feature-aware attention along with the description generation process. The description generation module is trained using tissue images and associated diagnostic reports provided by pathologists, while the visual attention module learns spontaneously only by observing the visual-semantic correspondences from image pixels to annotated description words of these images in data. For example, the system builds direct correspondence between

[1]Department of Computer Information Science Engineering, University of Florida, Gainesville, FL, USA. [2]J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA. [3]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [4]Department of Pathology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. [5]H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. [6]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. [7]James A. Haley Veterans' Hospital, Tampa, FL, USA. *e-mail: lin.yang@bme.ufl.edu

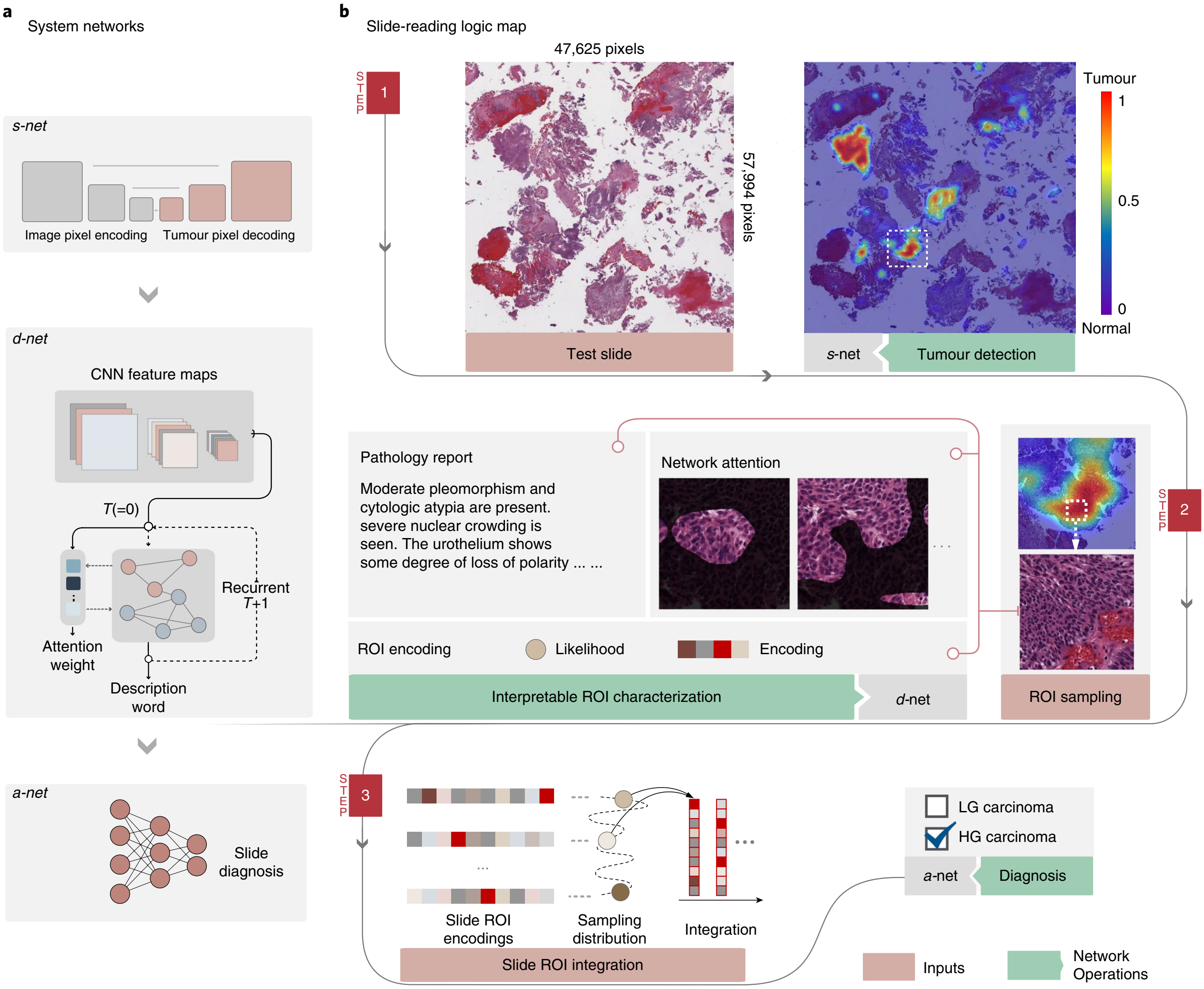
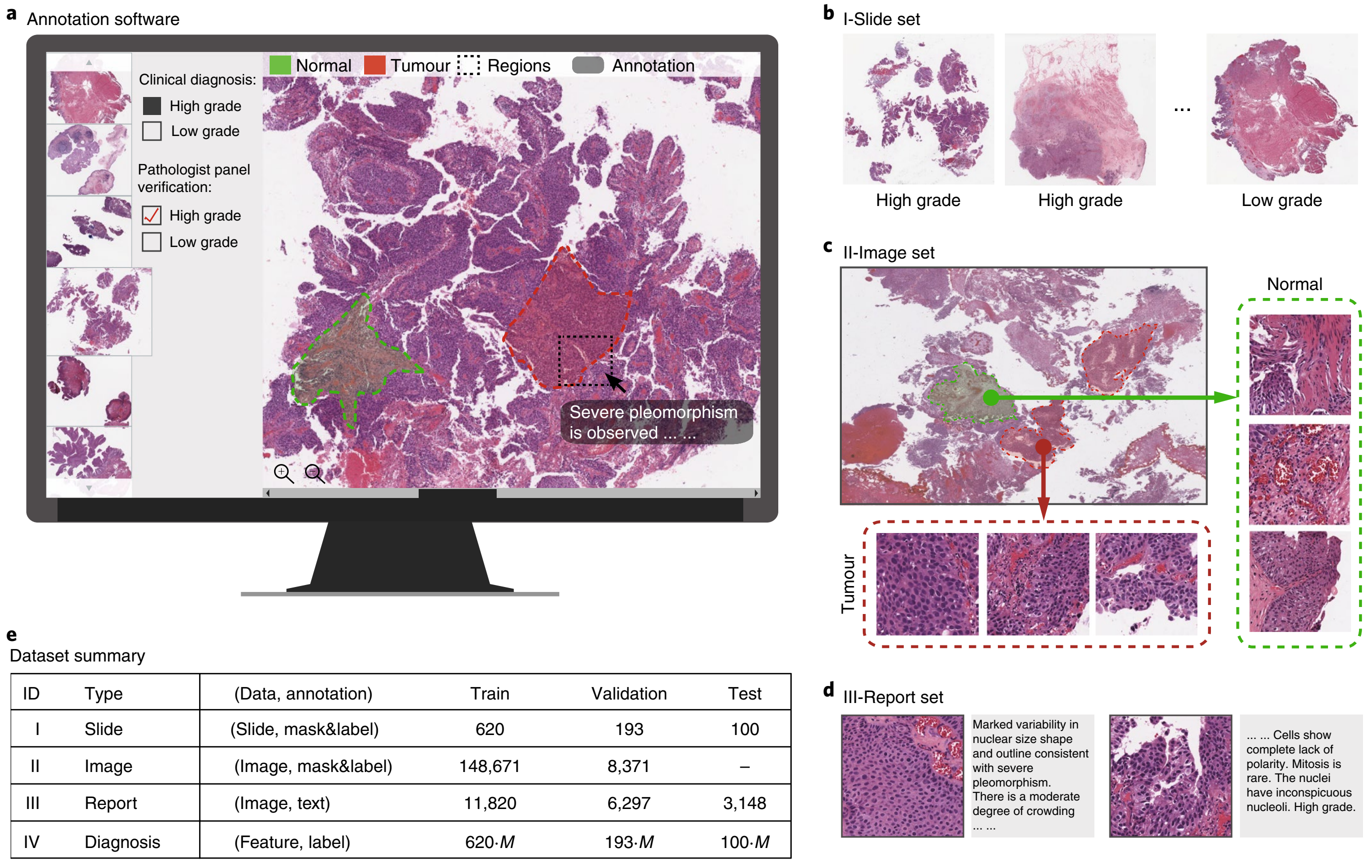


Fig. 1 | Method framework. a, Neural network system structures, including three main neural networks: the scanner network (*s*-net), the diagnoser network (*d*-net) and the aggregator network (*a*-net). Details of each network are provided in the Methods. **b**, Slide-reading workflow logic map. Given a whole slide, the *s*-net detects tumour regions. A collection of tissue images near detected tumours are automatically selected as diagnostically useful regions of interest (ROIs). The *d*-net characterizes each ROI, and the analysis interprets each ROI by describing pathological features (microscopic findings) and showing feature-aware network attention to explain what the network sees when describing observations. The information from all ROIs is encoded in a set of low-dimensional feature representations (vectors). The *a*-net integrates over all characterized features and establishes a diagnosis. See Methods for complete details.

the text 'severe crowding of nuclei' and image regions that exhibit crowded nuclei. During inference, the system is capable of interpreting its observation explicitly through its text and visual outputs. Both biologically inspired mechanisms unveil the machine's decision reasoning and deliver direct evidence (that is, second opinions) to pathologists for review and visual inspection, with the aim to help reduce variability in clinical decision making and, overall, to harmonize the pathologist's experience with machine assistance. Figure 1 demonstrates the framework in detail. The method comprises several compositional and multimodal deep neural networks: CNNs act to manage tumour detection and cellular-level characterization; fully-connected neural networks act to distil meaningful representations to diagnosis; recurrent neural networks (RNNs) act to control the visual and linguistic competencies (see Methods). These neural networks are mutually conditioned and complementary as they reach the final diagnosis.

Results

Our method is designed to be applicable to a wide variety of cancer types. Here, we have validated it on a large dataset containing 913 haematoxylin and eosin (H&E) stained whole slides (with an average slide height × width of 80,386 × 59,143 pixels) from patients with bladder cancer, obtained from multiple medical sources (this is a substantially larger data set than used in most related studies, to the best of our knowledge). In the United States, bladder cancer is the fifth most common malignancy. Approximately 90% of bladder tumours are classified as urothelial carcinomas[16,17]. Grading these carcinomas (defined by the World Health Organization system[18]) as being of low grade (LG) or high grade (HG) is essential for specified therapies. To evaluate our neural network system, 21 board-certified (genitourinary) pathologists and pathology-trained doctors (in total) participated in data annotation (4 pathologists) and diagnosis performance evaluation (17 pathologists) over approximately two



Marked variability in nuclear size shape and outline consistent with severe pleomorphism. There is a moderate degree of crowding ... ...

... ... Cells show complete lack of polarity. Mitosis is rare. The nuclei have inconspicuous nucleoli. High grade.

ID	Type	(Data, annotation)	Train	Validation	Test
I	Slide	(Slide, mask&label)	620	193	100
II	Image	(Image, mask&label)	148,671	8,371	–
III	Report	(Image, text)	11,820	6,297	3,148
IV	Diagnosis	(Feature, label)	620·*M*	193·*M*	100·*M*

Fig. 2 | Data preparation, organized in four data sets. a, Illustration of the data annotation using our developed web program. **b**, The I-Slide data set includes whole slides in bladder cancer with verified diagnosis labels from a panel of pathologists. **c**, The II-Image data set includes sampled images from annotated tumour regions (the red mask) and normal tissue regions (the green mask). **d**, The III-Report data set includes images with paired diagnostic feature descriptions (see main text for detailed explanations). **e**, Summary of each organized data set. The IV-Diagnosis data set relies on the trained *s*-net and *d*-net (as described in the Methods).

years of effort. The data set was cleaned and manually annotated by pathologists using several carefully designed procedures using our developed web-based annotation programs. For an explanation of the data preparation process and summary of the data set information see Fig. 2 and the Methods.

We conducted a variety of experiments to validate our method. The data set was split into 620 whole slides for training, 193 slides for validation and 100 slides for testing. We compared the performance of the system on the test set against the performance of board-certified pathologists (see Methods). Figure 3a presents the comparative results, which demonstrate the high reliability and accuracy of our method. For example, it achieves a 97% area under the curve (AUC) score and outperforms or matches most compared pathologists (as shown by the points falling below or on the curve). We also show the performance of the system for the larger validation and test sets shown in Fig. 3b, for which a 95% AUC score is achieved. In addition, we use the confusion matrix for comparison (Fig. 3e,f); the results show a 94.6% mean accuracy of the system, compared with an 84.3% mean accuracy for the pathologists.

Existing studies have shown that a consistent diagnosis for some types of breast cancer or prostate cancer can be reached less than 50% of the time[1,2]. Here, we also investigated pathologist variability in diagnosing bladder cancer and further compared the results to our method. We evaluated the pairwise percentage of disagreement between pathologists and the system for the 100 test slides shown in Fig. 3c; the results showed that more than 100 pathologist pairs exhibited a disagreement of more than 30%. The percentage of disagreement ranged from 8% to 46%, with an average of 23.8% disagreement. We also measured the per-slide accuracy (Fig. 3d); the results demonstrate irregular biases of the pathologists, in different classes. The results exhibit considerable variability despite the fact that some pathologists either have a similar background or work at the same institution. In contrast, the system exhibits a balanced and stable diagnosis performance.

Next, we demonstrated the slide-reading process of our method qualitatively. All types of output, including tumour detection, interpretable ROI cellular-level characterization, and diagnoses are shown in Fig. 4. Tumour detection indicates a diagnostically useful area, allowing the rapid localization of tumour regions at higher objectives (Figs. 4a,b and 5). ROI characterization is performed by analysing the tumour appearance, cell morphological patterns and so on, and these features are interpreted using natural language descriptions. The system describes a certain number of observed cellular features together with feature-aware attention maps to indicate what the network sees when describing each of these features. A strong interpretation is given regarding the type of visual information observed by the network (Fig. 4c–e). An attention map contains real-valued per-pixel weights (see Methods) to decide which pixels are more important for a given feature observation. The attention maps are visualized in a binary manner that is alpha-blended with the input image. We also compared the generated descriptions to pathologists' descriptions (Fig. 4c–e). Our method accurately describes multiple types of cell features that resemble the pathologists' interpretations.

We quantitatively validated the performance of our method component by component. First, we evaluated the tumour

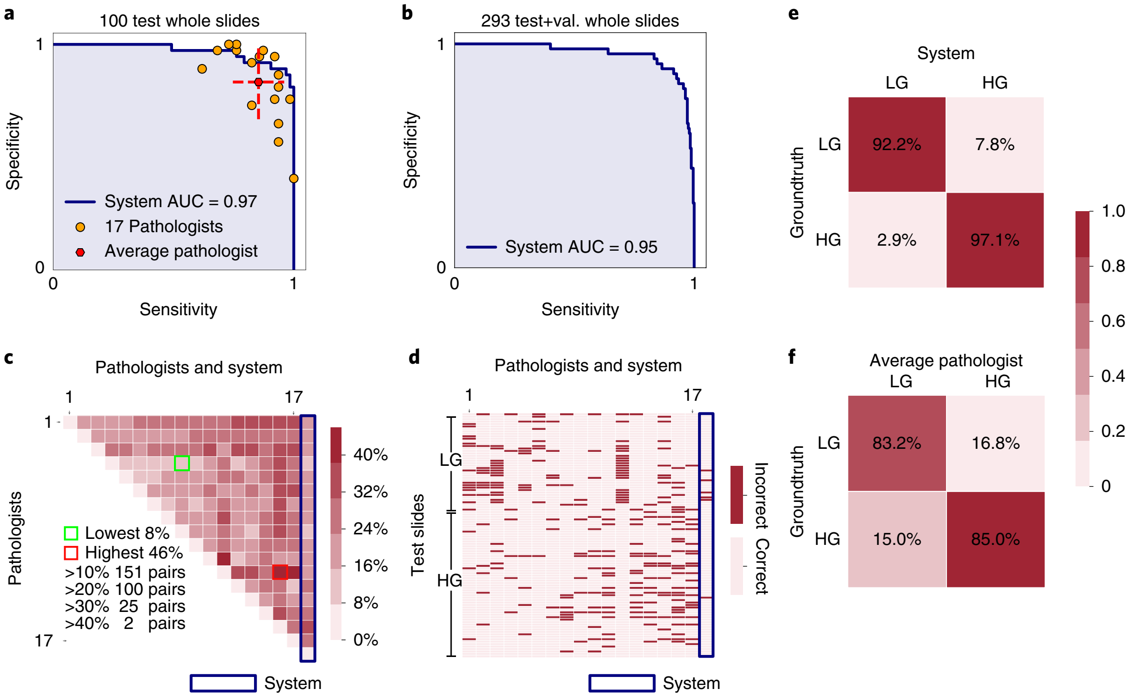


Fig. 3 | Results for the whole-slide diagnosis. a, Our method outperforms the average pathologist (17 total). The blue sensitivity–specificity curve represents the system's diagnosis. Each point on the curve represents a sensitivity (the true positive rate) and specificity (the true negative rate) score of our method, which are computed using predictions of *a*-net with a threshold applied to its probability output. Here, HG is treated as the positive recall versus LG. Sweeping a threshold over the interval 0–1 results in the points composing the curve. The 100 test slides include 37 slides obtained from the LG class and 63 obtained from HG. Each orange point represents a pathologist with a single sensitivity–specificity score. The red point is the average score for all pathologists, and the red-dashed error bars represent standard deviations. Our method outperforms the points falling below the curve and matches the points on the curve. **b**, The sensitivity–specificity curve of the system for the larger group of 293 slides in our test and validation sets. The system also achieves an excellent AUC score. **c**, Percentage of pairwise disagreement for the pathologists and the system, showing an average 23.8% disagreement over all pathologist pairs. **d**, Per-slide diagnosis for each pathologist and the system. The biases of the pathologists and the system can be observed. **e**,**f**, Confusion matrices for the system and the average pathologist (1,700 instances of diagnosis). We set a 0.79 threshold to obtain the *a*-net predictions. Each entry (r, c) on a confusion matrix represents the percentage of predictions for label c that matches the groundtruth label r.

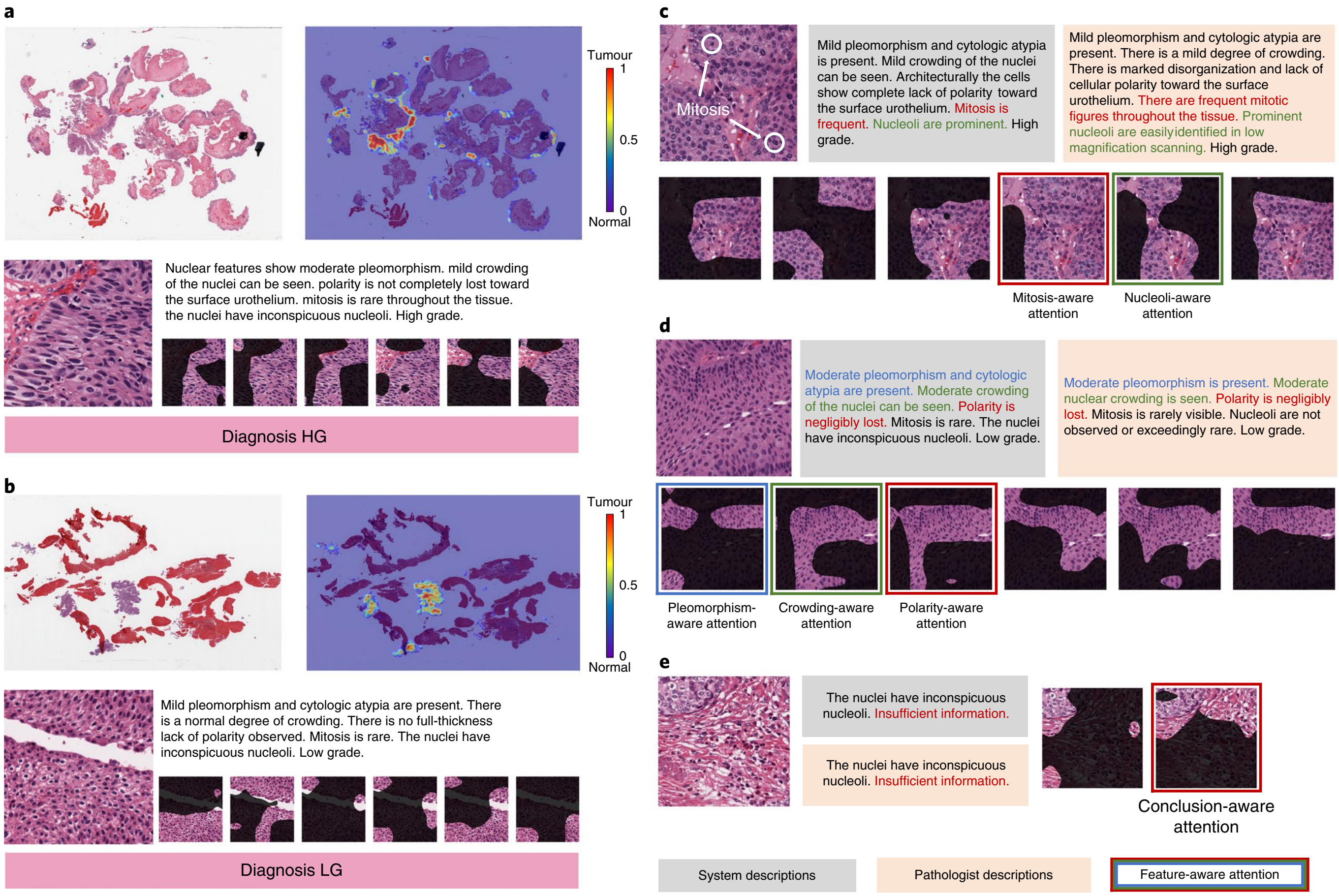
detection recall rate of the *s*-net for both tumour and non-tumour images (a non-tumour image represents a cropped slide tissue region that has no prominent tumour inside), as shown in Fig. 6a. A high positive rate setting and a high true negative rate setting of the network configuration are highlighted. For example, *s*-net achieves a 94% true positive (tumour) recall rate (no. of detected tumour pixels/total annotated tumour pixels) and simultaneously maintains a 95.3% negative (non-tumour) recall rate. Second, we validated the quality of the generated diagnostic descriptions using two evaluation metrics: Bilingual Evaluation Understudy (BLEU)[19] and Consensus-based Image Description Evaluation (CIDEr)[20]. Finally, to demonstrate the superiority of the results, we also compared the results obtained using *d*-net to those obtained using a well-known (in the field of computer vision) image-to-text translation method[21] as the baseline (see Methods). Our method is elaborately designed to enhance the effective combination of modules for learning from multimodal diagnosis and report data (see Methods) in network training and it outperforms the baseline on both metrics (Fig. 6b,c).

Furthermore, the inner working of *d*-net is a direct translation from image pixels to report words. Beyond the image-to-text generation shown already, a trained *d*-net also supports text-to-image retrieval. This capability provides a solution for doctors to query reference tissue images from databases by simply giving wanted feature descriptions (Fig. 7). The text-to-image retrieval evaluation is also an exact measure of the translation quality of *d*-net, because it indicates a failure to retrieve a normal tissue image given a diseased tissue query description and vice versa. Our method exhibits more stable and accurate retrieval performance than the compared baseline (Fig. 6d).

The *d*-net encodes the visual-semantic information of ROIs in a low-dimensional feature vector (that is, the combination of multiple network layer outputs). Finally, the goal of *a*-net is to aggregate such diagnostic information together in all slide ROIs and establish a final diagnosis (Fig. 1b). *a*-net is implemented as a three-layer fully connected neural network that receives slide ROI encodings and predicts the probability of cancer classes. Our method presents a novel stochastic sampling algorithm to effectively join ROI encodings and support robust network training and inference (Fig. 1b and Methods). We demonstrate the effectiveness of *a*-net by examining the behaviours of its hidden layers using the *t*-distributed Stochastic Neighbour Embedding (t-SNE) visualization algorithm[22]. Figure 6e visualizes the two-dimensional distributions of ROI encodings of LG and HG test slide samples. The data dimensions in the three layers are reduced progressively, from the 6,144-dimensional data at the input layer to the class probability at the output layer, with increasingly decentralized distributions between different classes and clustered distributions within each class.

Conclusions

In summary, we present a novel interpretable diagnosis method for diagnostic pathology, which shows unprecedented advantages over previous work. The proposed method interprets predictions through natural language descriptions and visual attention, which are understandable to pathologists when conducting a second review



Nuclear features show moderate pleomorphism. mild crowding of the nuclei can be seen. polarity is not completely lost toward the surface urothelium. mitosis is rare throughout the tissue. the nuclei have inconspicuous nucleoli. High grade.

Mild pleomorphism and cytologic atypia are present. There is a normal degree of crowding. There is no full-thickness lack of polarity observed. Mitosis is rare. The nuclei have inconspicuous nucleoli. Low grade.

Mild pleomorphism and cytologic atypia is present. Mild crowding of the nuclei can be seen. Architecturally the cells show complete lack of polarity toward the surface urothelium. Mitosis is frequent. Nucleoli are prominent. High grade.

Mild pleomorphism and cytologic atypia are present. There is a mild degree of crowding. There is marked disorganization and lack of cellular polarity toward the surface urothelium. There are frequent mitotic figures throughout the tissue. Prominent nucleoli are easilyidentified in low magnification scanning. High grade.

Moderate pleomorphism and cytologic atypia are present. Moderate crowding of the nuclei can be seen. Polarity is negligibly lost. Mitosis is rare. The nuclei have inconspicuous nucleoli. Low grade.

Moderate pleomorphism is present. Moderate nuclear crowding is seen. Polarity is negligibly lost. Mitosis is rarely visible. Nucleoli are not observed or exceedingly rare. Low grade.

The nuclei have inconspicuous nucleoli. Insufficient information.

The nuclei have inconspicuous nucleoli. Insufficient information.

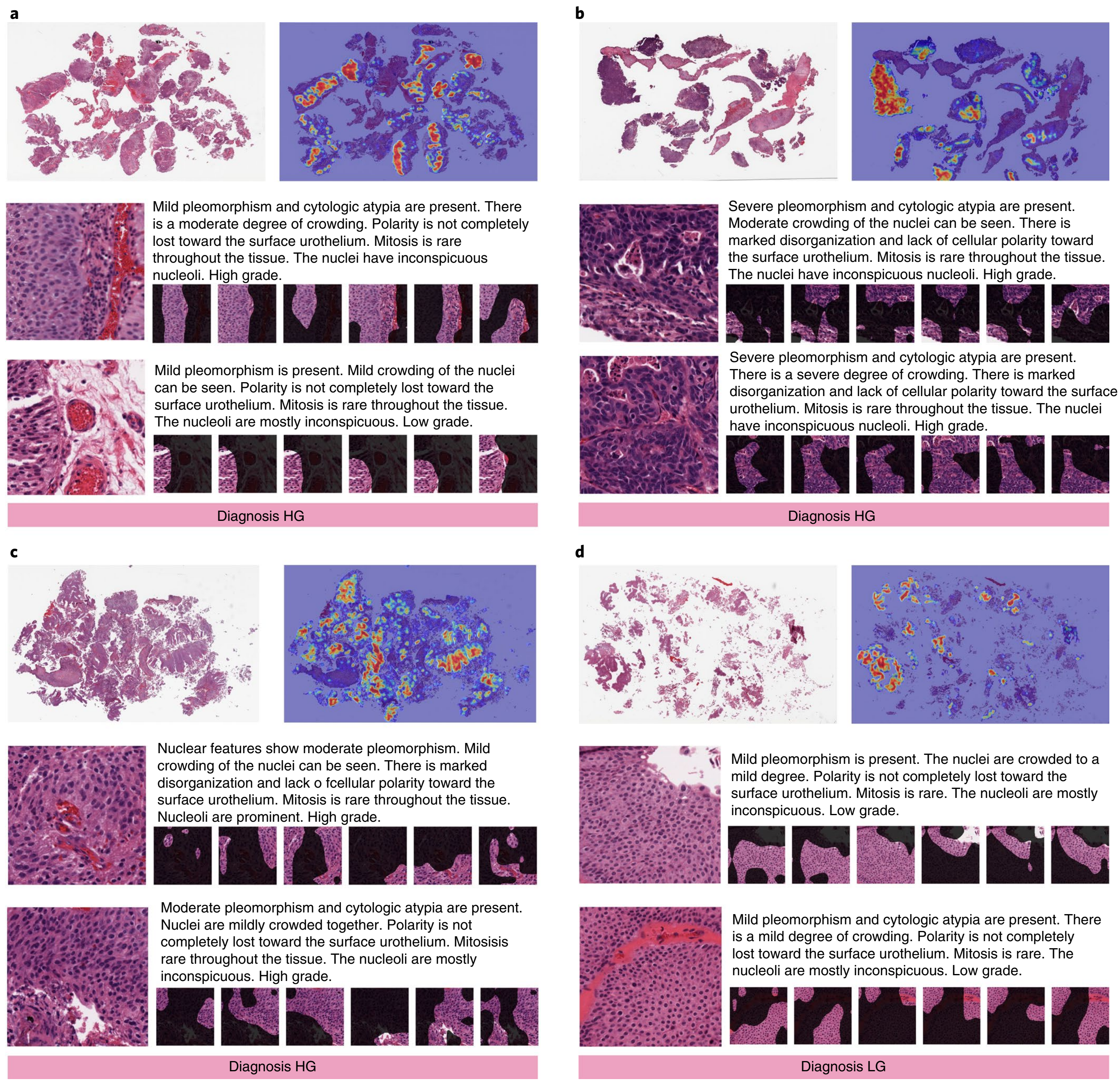
Fig. 4 | Visualization of interpretable predictions of the method. a,**b**, Whole-slide tumour detection (heat maps and close-up tumour region segmentation) and diagnosis results. A representative tissue image and the generated descriptions of the images and feature-aware attention maps are shown underneath (see Fig. 5 for more results). **c**–**e**, Description and feature-aware attention results. For each sample tissue image, the description generated by our method and the groundtruth description written by a pathologist are compared. The system generates an arbitrary number of sentences to describe the observed features. Each sentence has a corresponding attention map to visually highlight what the network sees. Each sentence and attention map pair is highlighted with the same color. For instance, in **e**, attention maps show that the network is observing nuclei at the upper-left corner. Then, 'insufficient information' is generated as a conclusion to indicate that insufficient diagnostic features are observed to make a diagnosis. To visualize a real-valued attention map, we convert the information to a binary map (using a threshold of 0.5) and alpha-blend this map with the corresponding input image (see Methods).

and visual inspection. This mechanism will encourage harmony in the routine experience of clinical pathology. Comprehensive validation on a large-scale bladder cancer data set demonstrates that the performance is similar to that of a wide range of pathologists. Our method allows diagnostic consistency and cost-effective system deployment to meet clinically demanding needs, for example in a cloud-based host for providing objective second opinions and consensus in small cancer centres. Our method is data-agnostic. With the success of deep learning, we believe that our method has strong generalizability for learning complex tissue structures and cell patterns in different cancer types. Our method has great potential to profoundly alleviate the above-demonstrated problems in diagnosis and will inspire the emergence of new AI-based CAD systems for various types of cancer. Further research will be useful to demonstrate its performance for other types of cancer (for example, lung, stomach and so on). In addition, in the context of precision medicine, we acknowledge the diagnostic value of contextual clinical information of patients beyond pixel knowledge or one type of stain in isolation. For example, the immunohistochemistry (IHC) procedure is also commonly used in clinics to confirm specific tumours or disease processes along with H&E-stained morphological analysis, although H&E-stained whole-slide morphological analysis remains the foundation for diagnostic medicine[3]. Future exploration in utilizing multiple types of information for diagnosis is clinically desirable.

Methods

Dataset. We collected our bladder whole-slide data set in the main from two medical resources: The Cancer Genome Atlas (open-sourced; the cancer genomic data commons (tgdc) data portal) and UF Health Shands Hospital in the United States. The slides were H&E stained and scanned with a ×40 objective. The data set contains 913 patient-exclusive slides for non-invasive low-grade papillary urothelial carcinoma (102) and non-invasive or invasive high-grade papillary urothelial carcinoma (811), which are the most common types of urothelial carcinoma in clinics[23]. The average slide dimension (height × width) was $80{,}386 \pm 36{,}812 \times 59{,}143 \pm 25{,}060$ (mean ± standard deviation). The data were prepared in several carefully designed steps with the participation of pathologists and pathology-trained doctors in a period of about two years using our developed web programs (Fig. 2a), resulting in four specialized data sets for network training (Fig. 2e).

To construct the I-Slide data set (Fig. 2b), we conducted a strict diagnosis label verification process. The original clinical diagnosis of a slide (acquired along with the slide acquisition) was checked by a panel of two experienced pathologists: the auxiliary pathologist was involved in the discussion if the major pathologist disagreed with the original label. We removed slides for which there was controversy with the diagnosis after panel discussion. Low-quality slides (for example, large blurry regions and tissue fold artefacts) were removed. One board-certified pathologist and one pathology-trained doctor annotated



a

Mild pleomorphism and cytologic atypia are present. There is a moderate degree of crowding. Polarity is not completely lost toward the surface urothelium. Mitosis is rare throughout the tissue. The nuclei have inconspicuous nucleoli. High grade.

Mild pleomorphism is present. Mild crowding of the nuclei can be seen. Polarity is not completely lost toward the surface urothelium. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. Low grade.

b

Severe pleomorphism and cytologic atypia are present. Moderate crowding of the nuclei can be seen. There is marked disorganization and lack of cellular polarity toward the surface urothelium. Mitosis is rare throughout the tissue. The nuclei have inconspicuous nucleoli. High grade.

Severe pleomorphism and cytologic atypia are present. There is a severe degree of crowding. There is marked disorganization and lack of cellular polarity toward the surface urothelium. Mitosis is rare throughout the tissue. The nuclei have inconspicuous nucleoli. High grade.

c

Nuclear features show moderate pleomorphism. Mild crowding of the nuclei can be seen. There is marked disorganization and lack o fcellular polarity toward the surface urothelium. Mitosis is rare throughout the tissue. Nucleoli are prominent. High grade.

Moderate pleomorphism and cytologic atypia are present. Nuclei are mildly crowded together. Polarity is not completely lost toward the surface urothelium. Mitosisis rare throughout the tissue. The nucleoli are mostly inconspicuous. High grade.

d

Mild pleomorphism is present. The nuclei are crowded to a mild degree. Polarity is not completely lost toward the surface urothelium. Mitosis is rare. The nucleoli are mostly inconspicuous. Low grade.

Mild pleomorphism and cytologic atypia are present. There is a mild degree of crowding. Polarity is not completely lost toward the surface urothelium. Mitosis is rare. The nucleoli are mostly inconspicuous. Low grade.

Fig. 5 | Visualization of more interpretable predictions of the method. a–d, Four whole-slide tumour detection (heat maps) and diagnosis results. Below each slide, two representative tissue images with generated descriptions and feature-aware attention maps are also shown.

slide regions (because of the great workload, the slides were partially annotated). For each slide they were asked to annotate at most eight tumour regions, which they believed to contain diagnostically useful information, and eight non-tumour regions. To construct the II-Image data set (Fig. 2c), we randomly sampled a set of images with 1,024 × 1,024 resolution (this resolution is suggested by pathologists for cellular feature analysis) around the annotated tumour and non-tumour regions. Each tissue image had a tumour region binary mask and was assigned with a label equivalent to the diagnosis label of its source slide. For instance, the 'tumour' label was given if the image had been sampled from tumour regions. Conversely, the 'non-tumour' label was given if sampled from non-tumour regions. Note that this label assignment was 'coarse', because a tissue image from a slide of a class could exhibit pathological features of a different class. However, because this II-Image data set had many more images than the III-Report data set, we found it was still beneficial to use this data set to pre-train the image model of *d*-net and then fine-tune using the III-Report data set, which has fine tissue labels. To construct the III-Report data set, we selected 221 non-invasive HG and LG papillary urothelial carcinoma slides from train and test sets in total, and sampled 4,253 1,024 × 1,024 images. Two experienced pathologists provided a paragraph of pathology report descriptions for each image, known as microscopic findings, which mainly describes five types of cellular feature—state of nuclear pleomorphism, cell crowding, cell polarity, mitosis and prominence of nucleoli—which are the key morphological visual features that need to be observed to classify urothelial carcinoma[17]. Each full paragraph was ended with one of four suspected conclusions: normal, LG papillary urothelial carcinoma, HG papillary urothelial carcinoma or insufficient information. Following the same procedure, two doctors and two trained students rephrased four additional reports for each image with their own interpretation. Thus, there were five reports per image and 21,265 image–report pairs in total. The vocabulary size was 112. The IV-Diagnosis data set (Fig. 2e) was collected with the assistance of our trained *s*-net and *d*-net. Each slide is represented as a bag of ROI embedded features; collection details are described in the section 'IV-Diagnosis dataset'.

The core of our method comprises several deep neural networks (Fig. 1). All the neural networks work cooperatively to achieve the overall diagnosis. We describe the detail of each network as follows.

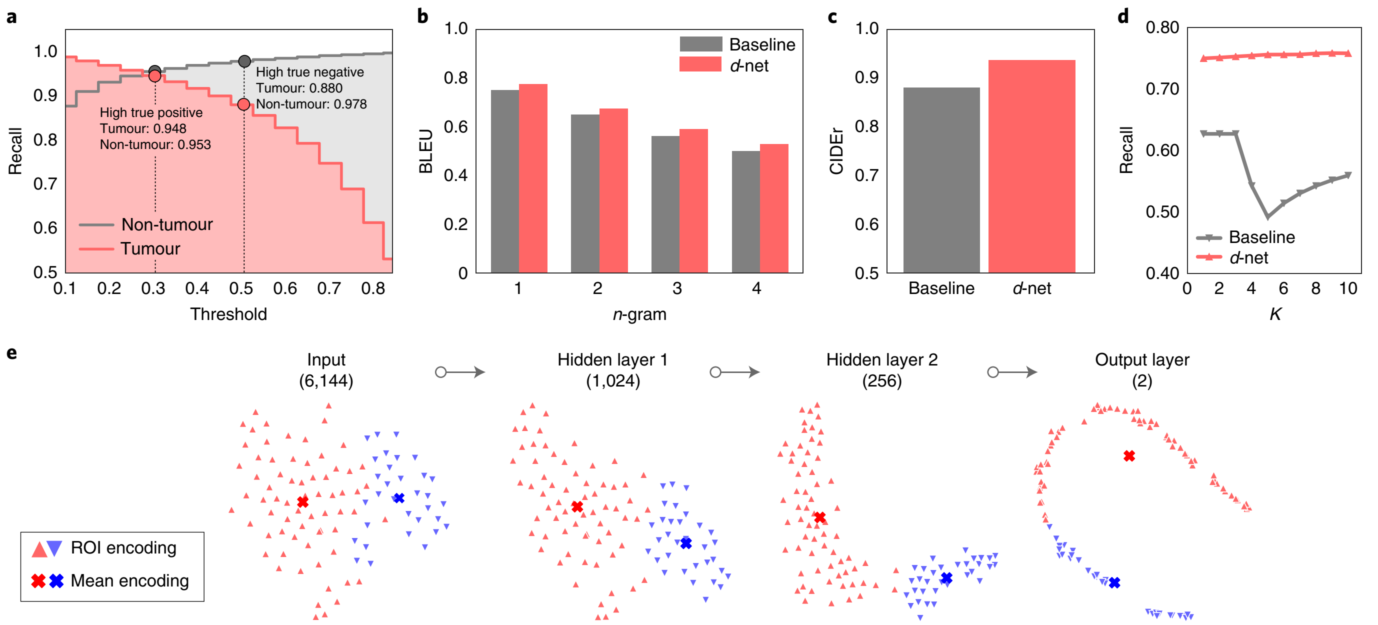


Fig. 6 | Evaluation of the network components. a, Recall rates of tumour and non-tumour regions. The curve is computed by varying the thresholds of the *s*-net probability output. The figure also highlights two settings to maintain a high true positive rate and a high true negative rate. **b**, BLEU indicates the evaluation (*n*-gram) of the quality of the description generation. **c**, CIDEr evaluation of the quality of description generation. **d**, The Recall@*K* metric for text-to-image retrieval. *d*-net is compared with a baseline method (see Methods). Our *d*-net outperforms the baseline for all evaluation metrics and shows more stable results on Recall@*K*. **e**, t-SNE visualization of *a*-net layer representations. Two-dimensional embedded distributions of vectors in the input layer (6,144-dimensional), the two hidden layers and the output layer (two-dimensional) are shown. Blue and red points indicate LG and HG classes, respectively.

Tumour detection. The *s*-net conducts tumour detection by classifying each pixel as tumour or non-tumour, represented as a probability value. The architecture of *s*-net resembles U-net[24], which is a type of end-to-end fully CNN for pixel-wise classification. The II-Image data set is used to train *s*-net. Because the tumour region is partially annotated, unannotated region pixels have unknown classes. To bypass this problem, during network training we compute the losses of *s*-net only for annotated region pixels, while ignoring unannotated pixels. At the inference time, processing a whole slide is decomposed into two steps: (1) divide the slide into computationally memory-affordable tiles and (2) detect tumours within each slide tile using *s*-net. As a result, *s*-net generates a probability map as the tumour region detection. The performance is evaluated in Fig. 6a.

Based on the tumour detection result, the system conducts the following steps to automatically select a set of diagnostically useful tissue images (termed ROIs in the main text) around detected tumours from the whole slide, denoted as $\{R_1, \ldots, R_D\}$. These ROIs are the inputs of *d*-net. First, to decide the sampling location, the system computes the average pixel probability of each tile; this step estimates which tile contains more tumours so that more ROIs can be sampled from this tile. The average probabilities of all tiles are then normalized so that they all sum to one; the normalized value × D is the number of ROIs needed to sample from a tile. We set $D=200$ empirically. A pixel is treated as a tumour pixel if its probability of being tumour is greater than 0.5 while the others are not considered as candidate pixels. Finally, using probabilities of tumour pixels as weights, the system simply applies weighted sampling to select central candidate pixels and crop ROIs.

Cellular-level ROI characterization. The *d*-net plays the core role in the system in characterizing ROIs, generating an interpretable diagnosis and encoding observed visual information. It is a composite neural network that can combine multimodal information. Specifically, it includes an image model to represent visual knowledge by encoding image pixels into feature maps. It also includes a language model to generate diagnostic descriptions and network visual attention. The overall framework is inspired by advances in image captioning in computer vision[4,21,25]. To build our image model, we utilize the Inception-v3 CNN[26] with initial weights from a pretrained model on the ImageNet data set[27]. Input images are resized to 256 × 256 (from the original 1,024 × 1,024); so a slide is seen at the ×10 objective by the network). This downsampling rate does not cause a loss of critical information, as confirmed by pathologists. The image model produces feature maps

$$V = [\mathbf{v}_1, \ldots, \mathbf{v}_S]^T \in \mathbb{R}^{2048\times(6\times6)}$$

with its last convolution layer, that is 2,048 feature maps where each has 6 × 6 resolution.

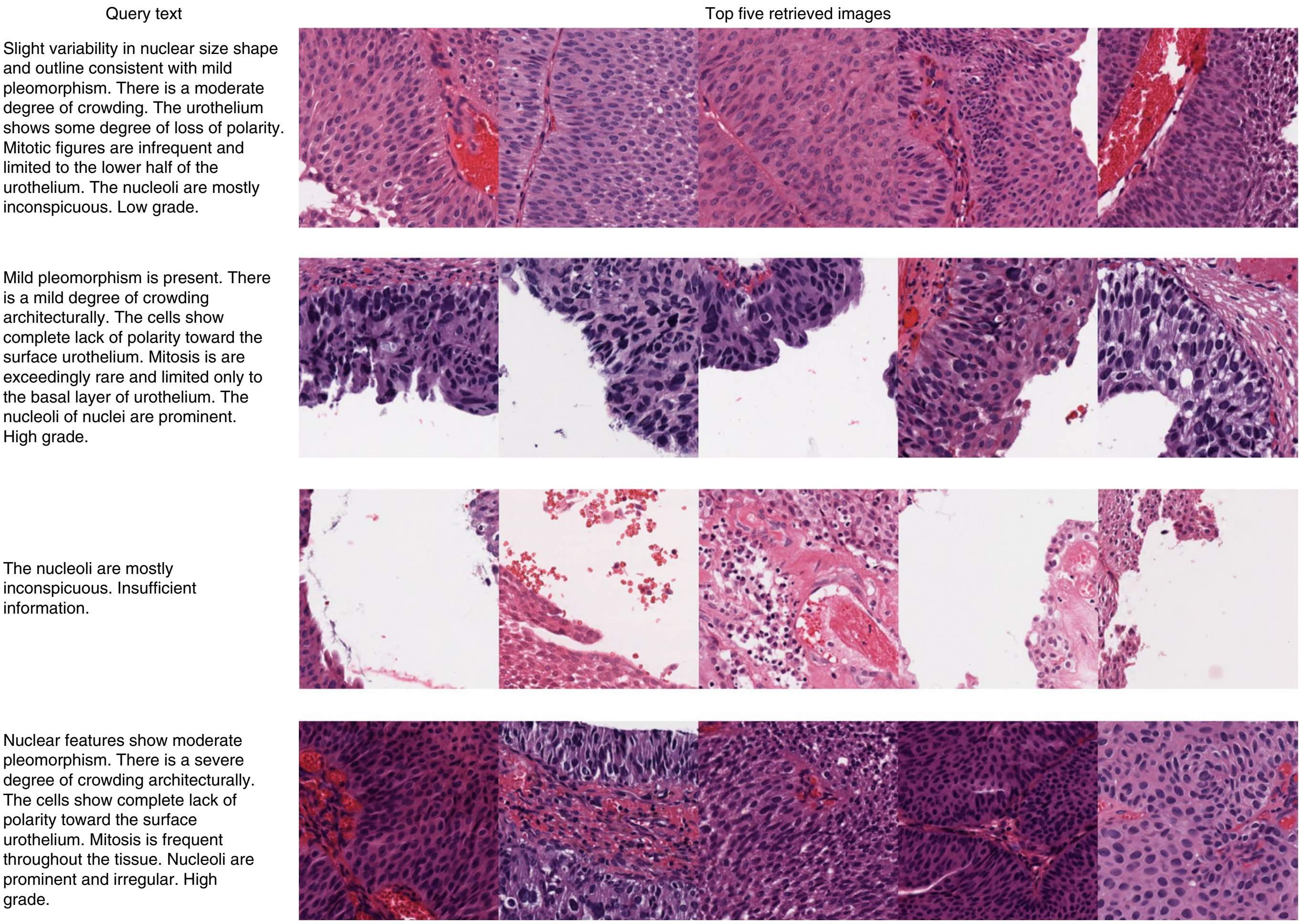
As we introduced in the section 'Data set', the pathology reports contain observations of five morphological features and a conclusion (six diagnostic concepts in total). Each concept is a condition to determine the final diagnosis[17]. To effectively model such report structure priors, we designed a language model with two long short-term memory (LSTM) modules to model the concepts and then descriptions. This design is inspired by the design of hierarchical LSTM[28].

The language model is conditioned on these visual features to generate descriptions. Our method is based on the advanced RNN LSTM[29], which is widely used for sequence-to-sequence modelling (for example, machine translation[30]). Basically, LSTM is a computation unit that holds a hidden state vector $\mathbf{h}_t$ and a memory state vector $\mathbf{m}_t$ at each time step to integrate spatiotemporal information. The unit is defined as

$$\mathbf{h}_{t+1}, \mathbf{m}_{t+1} = \mathrm{LSTM}(\mathbf{x}_t, \mathbf{h}_t, \mathbf{m}_t) \qquad (1)$$

To represent a word embedding vector $\mathbf{x}_t$, a word is first represented as a one-hot vector that has number of elements equal to the word vocabulary size of the training data and has only one element, which corresponds to the representing word, equal to 1. Then, to represent the vector as a real valued encoding, the vector is multiplied by a word-embedding matrix to the hidden state dimension. The matrix is learned together with the full model. We use LSTM with 256 hidden state units. The initial states h_0 and m_0 are zero-filled vectors.

In our language model, the two LSTM modules are a conceptional LSTM (c-LSTM) and a descriptive LSTM (d-LSTM). c-LSTM maintains a concept state to represent each concept in the latent space conditioned on V. d-LSTM maintains a word state to decode each concept state to sentences. To be specific, given an encoded image representation in a 2,048-dimensional vector (by averaging V along the 6 × 6 spatial dimension), c-LSTM takes the vector as the input $\mathbf{x}_0^c$ at the first time step and begins to generate concept states, $\{\mathbf{h}_1^c, \ldots \mathbf{h}_t^c\}$, from the second time step. The concept state carries two types of information: (1) indicating whether all salient feature concepts are necessarily described and thereby the generation stop criterion is satisfied; (2) encoding global visual knowledge $\mathbf{h}_t^c$ to help generate visual context knowledge $\boldsymbol{z}_t$ (see definition in the following) and as inputs of d-LSTM. c-LSTM has a stop controller to estimate $\Pr(\mathrm{stop}|\mathbf{h}_t^c)$ (c-LSTM terminates when $\Pr(\mathrm{stop}|\mathbf{h}_t^c) > 0.5$). This controller is a fully connected layer followed by a sigmoid activation function, which takes $\mathbf{h}_t^c$ as input and produces a scalar. Thus the training supervision of the controller is from the number of feature sentences described in groundtruth reports. d-LSTM receives both types of information and then decodes the concept state to sentences. Given a concept state $\mathbf{h}_t^c$, the inputs of d-LSTM at the first two time steps are $\mathbf{x}_0^d = \mathbf{z}_t$ and $\mathbf{x}_1^d = \mathbf{h}_t^c$. d-LSTM generates a set of word states, $\{\mathbf{h}_0^d, \ldots, \mathbf{h}_t^d\}$. From the second time step, a fully connected layer



Slight variability in nuclear size shape and outline consistent with mild pleomorphism. There is a moderate degree of crowding. The urothelium shows some degree of loss of polarity. Mitotic figures are infrequent and limited to the lower half of the urothelium. The nucleoli are mostly inconspicuous. Low grade.

Mild pleomorphism is present. There is a mild degree of crowding architecturally. The cells show complete lack of polarity toward the surface urothelium. Mitosis is are exceedingly rare and limited only to the basal layer of urothelium. The nucleoli of nuclei are prominent. High grade.

The nucleoli are mostly inconspicuous. Insufficient information.

Nuclear features show moderate pleomorphism. There is a severe degree of crowding architecturally. The cells show complete lack of polarity toward the surface urothelium. Mitosis is frequent throughout the tissue. Nucleoli are prominent and irregular. High grade.

Fig. 7 | Text-to-image retrieval results. In each row, the top five retrieved images (right) of each query text (left) are shown.

is applied to transform the hidden state to a distribution over the vocabulary, Pr(word|$\mathbf{h}_t^{\mathrm{d}}$). The element of a word with the highest probability is chosen as the prediction. At the training stage, the input of d-LSTM is a groundtruth word from the previous time step. In the inference stage, it is the predicted word from the previous time step. All concept states share one d-LSTM network.

We designed the method to generate attention interpretation for the prediction of every concept, which allows visual inspection by end-users. Figure 4 demonstrates the attention results. To this end, we proposed a concept-aware attention mechanism to build the correlation between semantic concepts and visual pixels. The overall attention mechanism includes two stages. The first stage is to generate a first-level attention guidance by taking advantage of the implicit class-aware localization property of CNNs[31]. Specifically, this attention map is defined as $\boldsymbol{\alpha}_{\mathrm{g}} = \mathrm{softmax}(w_1^I \mathbf{v}_1 + \ldots + w_S^I \mathbf{v}_S) \in \mathbb{R}^S$, which provides class-aware weights on each spatial location. Note that the weights only condition on the class of images without text intervention. To compute $\mathbf{w}_I = [w_1^I, \ldots, w_S^I]$, we stack a fully connected layer on the top of the image model, aiming to classify the cancer labels (in three classes: LG, HG and merged insufficient information/normal) of input images. Thus the weight matrix of this stacked layer has dimension 2,048 × 3. $\mathbf{w}_I$ is the column that corresponds to the highest class probability. In addition, this class probability will be used during the training of *a*-net to integrate slide ROIs (see the section 'Slide diagnosis'). With this attention guidance to help distil class-aware visual information via $V\boldsymbol{\alpha}_{\mathrm{g}}$, we can assist attention generation in the second stage. We found that this technique assists in generating focal feature-aware attention on informative tissue regions.

In the second stage, for the attention of the language model, at time t the attention module takes a concept state $\mathbf{h}_t^{\mathrm{c}}$, α_{g}, and the feature maps V to compute visual context knowledge $\mathbf{z}_t$. The attention module is specifically defined as

$$\begin{aligned} \mathbf{e}_t &= W_{\mathrm{e}} \tanh([W_{\mathrm{v}} V;\ W_{\mathrm{v}'} V \boldsymbol{\alpha}_{\mathrm{g}}] + (W_{\mathrm{h}} \mathbf{h}_t) 1^T) \\ \boldsymbol{\alpha}_t &= \mathrm{softmax}(\mathbf{e}_t) \end{aligned} \quad (2)$$

where W_{e}, W_{v}, $W_{\mathrm{v}}{}'$ and W_{h} are learnable parameters. $1 \in \mathbb{R}^{1\times(S+1)}$ is a vector with all elements equal to one ($S = 6 \times 6$ here). $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is an activation function that scales the input value to the range of −1 to 1. softmax($\mathbf{x}$) normalizes each element x_i of the input vector $\mathbf{x}$ to be $e^{x_i} / \sum_i (e^{x_i})$, so the summation of elements is 1. $[\cdot;\cdot]$ is the concatenation operation. $\boldsymbol{\alpha}_t$ is an $(S+1)$-dimension vector, where each of the first S element $\boldsymbol{\alpha}_t^S \in \mathbb{R}^{36}$ is an attention weight corresponding to a spatial location. The context vector is computed as

$$\mathbf{z}_t = V \boldsymbol{\alpha}_t^S \quad (3)$$

where $\boldsymbol{z}_t \in \mathbb{R}^{2048}$. $\boldsymbol{\alpha}_t^S$ is a real-valued 6 × 6 map with all element weights are summed to be 1. Thus each attention map location represents pixel information in a $\left\lceil \frac{256}{6} \times \frac{256}{6} \right\rceil$ image grid area. To visualize an attention map on the input image, we directly upscale to the size of the image plane. A larger weight indicates higher attention focus on grid pixels.

In the training stage, *d*-net minimizes three cross-entropy loss terms simultaneously: the loss for classification at the image model, the loss for word prediction at the language model (d-LSTM outputs) and the loss for the stop criterion at the c-LSTM outputs. The overall training is end-to-end with stochastic gradient descent and backpropagation[4]. We pre-trained the image model of *d*-net using the II-Image data set first. It achieved a 86.6% classification accuracy on the validation set, and then the overall *d*-net was further trained using the III-Report data set.

IV-Diagnosis dataset. To conduct slide-level diagnosis, *a*-net aggregates encoded information from all ROIs $\{R_1, \ldots, R_D\}$ in a slide and establishes slide diagnosis. For each of the slide ROIs, there are two types of information extracted from *d*-net: encoded features $\mathbf{f}^R$ and raw class probability $\mathbf{p}^R$. We extract feature maps at several layers of Inception-v3[26] (namely mixed10, mixed9, mixed8 and mixed7; we use the Keras implementation https://github.com/tensorflow/tensorflow/blob/master/tensorflow/python/keras/applications/inception_v3.py of Inception-v3 CNN, where the naming of each used layer is defined.). Convolutional feature maps are averaged along the spatial dimension to obtain feature vectors, which are then concatenated together as the encoded feature of an ROI, that is, $\mathbf{f}_R \in \mathbb{R}^{6144}$. All D ROIs in a whole slide are organized as $R = \{(\mathbf{f}_1^R, \mathbf{p}_1^R), \ldots, (\mathbf{f}_D^R, \mathbf{p}_D^R)\}$. We follow this step to process slides and organize the IV-Diagnosis data set $\{(R_i, l_{R_i})\}_{\mathrm{train}}$ and $\{(R_i, l_{R_i})\}_{\mathrm{val}}$, for the training of *a*-net (see Fig. 2e).

Slide diagnosis. The *a*-net is implemented as a three-layer fully connected neural network. It takes integrated ROI feature encodings and predicts slide cancer labels. We propose a stochastic feature sampling mechanism to effectively augment training data through random feature combination so as to improve the model generalization. Algorithm 1 describes the training details of *a*-net using the IV-Diagnosis data set. At the inference stage, *a*-net repeats this stochastic sampling-and-prediction process described in Algorithm 1. Note that the groundtruth label l_{R_i} to compute the sampling probability $\mathbf{p}_{R_i}[l_{R_i}]$ is unknown; we simply alternate l of all classes (that is, HG and LG) over multiple sampling. Ten repeats are performed in total. The final diagnosis is the maximum class probability response of the accumulated probability of the 10-time predictions.

Algorithm 1. Training *a*-net with stochastic sample generation

```
Input: The IV-Diagnosis training set $\{(R_i, l_{R_i})\}_{\mathrm{train}}$, validation set $\{(R_i, l_{R_i})\}_{\mathrm{val}}$, sampling ratio $M=0.2$, a randomly initialized a-net, batch size $B$, early stop iteration $E$.
Output: A trained a-net
for each iteration do
    Randomly select $B$ data from the training set $\{(R_1, l_{R_1}), \ldots, (R_B, l_{R_B})\}$ as a training batch
    # stochastic feature sampling
    for each slide $R_i$ in the batch do
        1. Sample $(M \cdot D)$ ROI vectors $[\mathbf{f}_1^{R_i}, \ldots, \mathbf{f}_{M \cdot D}^{R_i}]$ weighted by the probability $\mathbf{p}^{R_j}[l_{R_i}]$
        2. Compute the element-wise mean of sampled ROI vectors as a training point $(\hat{\mathbf{f}}^{R_i}, l^{R_i})$.
    end
    # actual training
    Update a-net parameters using the batch of training data $\{(\hat{\mathbf{f}}^{R_i}, l^{R_i})\}_{\mathrm{batch}}, i=1, \ldots, B$.
    # early stop
    Stop training when the validation accuracy does not increase further in $E$ iteration.
end
```

Implementation details. We implement TensorFlow library[32]. Each network of our framework was trained end-to-end using the standard stochastic gradient descent algorithm[4] with the Adam optimizer. Dropout and weight decay were used to regularize *a*-net. We use standard data augmentation techniques including rotation, horizontal and vertical flips, and random crop.

Human comparison. We acquired the participation of 17 board-certified pathologists (excluding the pathologists providing annotations) in the human comparison experiment (Fig. 3). The participating pathologists are from a total of seven medical centres (either cancer centres or hospitals) in the United States and China. They are either specialized in genitourinary or have over five years of clinical experience in diagnosing slides of the urinary system (some have been in practice for over 20 years). We organized a formal 'bladder cancer human–machine competition' to collect pathologists' results. Slide reading was completed independently and remotely using our developed browser-based slide viewer.

Compared baseline. In the main text we compare the proposed *d*-net with a well-known image-to-text generation framework[21] used in computer vision (the open-source code is available at https://github.com/karpathy/neuraltalk2). This method has a simpler network design than our specialized language model design. It also uses the same Inception-v3 CNN to encode images and is followed by an LSTM to generate descriptive sentences. We use the same model hyper-parameters and training settings as those used in *d*-net to guarantee a fair comparison.

Text-based image retrieval. In the main text we evaluated the text-based image retrieval for *d*-net. Given a query text with words $\boldsymbol{r}_{1:L}$, *d*-net retrieves the top-ranked image I from the database by calculating the maximized joint probability $\Pr(\boldsymbol{r}_{1:L}|I)$ that the model can generate the query text. A retrieved image is counted as successful recall if its associated cancer label matches the cancer type that the query text describes. We calculate top-K (from 1 to 10) image recall rates (Fig. 6d), using 1,890 groundtruth reports as query text to query 378 test images in the test set.

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

Data availability

The data that support the findings of this study are available from Figshare: https://figshare.com/projects/nmi-wsi-diagnosis/61973.

Code availability

Source code are available from the Github repository: https://github.com/zizhaozhang/nmi-wsi-diagnosis.

Received: 2 January 2019; Accepted: 5 April 2019;
Published online: 13 May 2019

References

1. Brimo, F., Schultz, L. & Epstein, J. I. The value of mandatory second opinion pathology review of prostate needle biopsy interpretation before radical prostatectomy. *J. Urol.* **184**, 126–130 (2010).
2. Elmore, J. G. et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015).
3. Djuric, U., Zadeh, G., Aldape, K. & Diamandis, P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precis. Oncol.* **1**, 22 (2017).
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
5. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
6. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
7. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
8. Litjens, G. et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).
9. Araújo, T. et al. Classification of breast cancer histology images using convolutional neural networks. *PloS ONE* **12**, e0177544 (2017).
10. Xu, Y. et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics* **18**, 281 (2017).
11. Yoshida, H. et al. Automated histological classification of whole slide images of colorectal biopsy specimens. *Oncotarget* **8**, 90719 (2017).
12. Han, Z. et al. Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci. Rep.* **7**, 4172 (2017).
13. Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2424–2433 (IEEE, 2016).
14. Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? Preprint at https://arxiv.org/abs/1712.09923 (2017).
15. Lipton, Z. C. The mythos of model interpretability. *Queue.* **16**, 30 (2018).
16. Pasin, E., Josephson, D. Y., Mitra, A. P., Cote, R. J. & Stein, J. P. Superficial bladder cancer: an update on etiology, molecular development, classification, and natural history. *Rev. Urol.* **10**, 31–43 (2008).
17. Zhou, M. & Magi-Galluzzi, C. *Genitourinary Pathology* (Foundations in Diagnostic Pathology, Saunders, 2015).
18. Humphrey, P. A., Moch, H., Cubilla, A. L., Ulbright, T. M. & Reuter, V. E. The 2016 WHO classification of tumours of the urinary system and male genital organs—Part B: Prostate and bladder tumours. *Eur. Urol.* **70**, 106–119 (2016).
19. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* 311–318 (Association for Computational Linguistics, 2002).
20. Vedantam, R., Lawrence Zitnick, C. & Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4566–4575 (IEEE, 2015).
21. Karpathy, A. & Fei-Fei, L. Deep visual–semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3128–3137 (IEEE, 2015).
22. Maaten, Lvd & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
23. Miyamoto, H. et al. Non-invasive papillary urothelial neoplasms: the 2004 WHO/ISUP classification system. *Pathol. Int.* **60**, 1–8 (2010).
24. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* 234–241 (Springer, 2015).
25. Xu, K. et al. Show, attend and tell: neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057 (JMLR, 2015).
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
27. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
28. Krause, J., Johnson, J., Krishna, R. & Fei-Fei, L. A hierarchical approach for generating descriptive image paragraphs. Preprint at https://arxiv.org/abs/1611.06607 (2016).

29. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
30. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. Preprint at https://arxiv.org/abs/1409.0473 (2016).
31. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929 (IEEE, 2016).
32. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* Vol. 16 265–283 (USENIX Association, 2016).

Acknowledgements

The authors thank the Department of Pathology, University of Florida (UF), and UF Health Shands Hospital for support with data collection. The authors also thank members of the Moffitt Cancer Center and the Department of Pathology, the First Affiliated Hospital of Xi'an Jiaotong University, for their participation in this research, and thank all participating pathologists for their valuable suggestions and active involvement. Thanks also go to Y. Cai for assistance with figure production. The research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under award no. 5R01AR065479-05. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Z.Z. led the development and evaluation. Z.Z., C.W. and L.Y. designed the research. Z.Z. implemented the algorithm. Z.Z., P.C., M.M. and M.S. collected and cleaned the data and developed the annotation software. L.Y. and M.B. recruited pathologists for annotation and machine–human comparison. L.C. and P.C. managed the machine–human competition. J.D., N.A., F.K.K. and S.I.D. participated in the competition. Z.Z. wrote the manuscript. M.M., F.X., Y.X., X.S., F.L., H.S. and J.C. provided valuable comments on the algorithm design and the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0052-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.Y.

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

nature research

Corresponding author(s): Zizhao Zhang

Last updated by author(s): YYYY-MM-DD

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection tools were implemented using open-source Javescripts and Python libraries. The implemented annotation tool for whole slide region annotation is based on the open-source toolbox MicroDraw (https://github.com/r03ert0/microdraw).
Data analysis	Training and validating the system was implemented using open-source Python language libraries including Tensorflow-v1.2 (https://www.tensorflow.org/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study will be available online upon publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size.
Data exclusions	Data with low image quality and with ambiguous diagnosis by a panel of two pathologists are excluded and not used for study
Replication	All experiments are replicated to support the conclusions of the manuscript.
Randomization	Not applicable.
Blinding	Pathologists who participated in the human-computer diagnosis comparision are blind to data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Subjects with over 18 years old.
Recruitment	This is a retrospective study of existing, archived materials that involves no new procedures or other patient interventions.
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.