Histopathology

Histopathology 2024, **85**, 116–132. DOI: 10.1111/his.15180

Deceptive learning in histopathology

Sahar Shahamatdar,[1,2] Daryoush Saeed-Vafa,[3] Drew Linsley,[4,5] Farah Khalil,[3] Katherine Lovinger,[6] Lester Li,[7] Howard T. McLeod,[8] Sohini Ramachandran[1,9,10] & Thomas Serre[4,5]

[1]*Center for Computational Molecular Biology,* [2]*The Warren Alpert Medical School, Brown University, Providence, RI,* [3]*Department of Anatomic Pathology, H. Lee Moffitt Cancer and Research Institute, Tampa, FL,* [4]*Carney Institute for Brain Science,* [5]*Department of Cognitive Linguistic and Psychological Sciences, Brown University, Providence, RI,* [6]*Department of Molecular Biology, H. Lee Moffitt Cancer and Research Institute, Tampa, FL,* [7]*University of Rochester, Rochester, NY,* [8]*Intermountain Precision Genomics, St George, UT,* [9]*Department of Ecology, Evolution and Organismal Biology, Brown University and* [10]*The Data Science Initiative, Brown University, Providence, RI, USA*

Date of submission 1 February 2024
Accepted for publication 10 March 2024

Shahamatdar S, Saeed-Vafa D, Linsley D, Khalil F, Lovinger K, Li L, McLeod H T, Ramachandran S & Serre T (2024) *Histopathology* **85**, 116–132. https://doi.org/10.1111/his.15180

Deceptive learning in histopathology

Aims: Deep learning holds immense potential for histopathology, automating tasks that are simple for expert pathologists and revealing novel biology for tasks that were previously considered difficult or impossible to solve by eye alone. However, the extent to which the visual strategies learned by deep learning models in histopathological analysis are trustworthy or not has yet to be systematically analysed. Here, we systematically evaluate deep neural networks (DNNs) trained for histopathological analysis in order to understand if their learned strategies are trustworthy or deceptive.

Methods and results: We trained a variety of DNNs on a novel data set of 221 whole-slide images (WSIs) from lung adenocarcinoma patients, and evaluated their effectiveness at (1) molecular profiling of KRAS versus EGFR mutations, (2) determining the primary tissue of a tumour and (3) tumour detection. While DNNs achieved above-chance performance on molecular profiling, they did so by exploiting correlations between histological subtypes and mutations, and failed to generalise to a challenging test set obtained through laser capture microdissection (LCM). In contrast, DNNs learned robust and trustworthy strategies for determining the primary tissue of a tumour as well as detecting and localising tumours in tissue.

Conclusions: Our work demonstrates that DNNs hold immense promise for aiding pathologists in analysing tissue. However, they are also capable of achieving seemingly strong performance by learning deceptive strategies that leverage spurious correlations, and are ultimately unsuitable for research or clinical work. The framework we propose for model evaluation and interpretation is an important step towards developing reliable automated systems for histopathological analysis.

Keywords: computational pathology, deep learning, explainable artificial intelligence, KRAS, molecular profiling

Address for correspondence: Address for corresponcence: D Linsley, Carney Institute for Brain Science, Brown University, 190 Thayer Street, Providence, RI 02912, USA. e-mail: drew_linsley@brown.edu
Sahar Shahamatdar and Daryoush Saeed-Vafa contributed equally to this work.

Abbreviations: CUP, cancer of unknown primary; DNN, deep neural network; FFPE, Formalin-fixed paraffin-embedded; H&E, hematoxylin and eosin; LCM, laser capture microdissection; ROI, region-of-interest; SVS, scanScope virtual slide file; TCGA, the cancer genome atlas; WSI, whole slide image.

Introduction

Histopathology image analysis is essential for diagnosing cancer and determining the course of treatment.[1] There is now growing evidence that deep neural networks (DNNs) can at least partially automate this procedure; for instance, DNNs rival expert pathologists at detecting malignant skin lesions,[2,3] diagnosing diabetic retinopathy[4–6] and detecting breast cancer.[7,8] In each of these cases, DNNs learned to solve straightforward but time-consuming tasks that are already within the expert physician's repertoire. However, there is now a growing number of reports that DNNs can also learn to solve tasks posed on histopathological images that are difficult or impossible for pathologists to do by visual analysis alone. These tasks include profiling the genome of a tumour[9–12] and identifying the originating tissue for cancer of unknown primary[13,14] (CUP) from tissue morphology. Underlying these findings is the assumption that histopathology images depict patterns of morphological features that are impossible for experts to detect by eye, but which DNNs have the capacity to learn to identify. This assumption has not been systematically tested, raising the possibility that DNNs solve histopathological tasks by exploiting morphology that is indirectly or spuriously related to disease.

The trustworthiness of DNNs is an open problem in computer vision because DNNs often exploit idiosyncratic or spurious correlations between images and labels to solve visual tasks,[15,16] a strategy which can lead to high accuracy on a single benchmark data set but fails to hold up under more rigorous testing of the location of mutations. Such a reliance on spurious features has caused many notable errors in estimating the abilities of animals and machines. For example, the notable case of 'Clever Hans',[17] the horse who relied on non-verbal cues from his trainer to solve simple arithmetic, has been evoked as an analogue for the propensity of the learned strategies of DNNs to be misaligned with humans.[16] To avoid such errors when developing DNNs for clinical applications, it is standard practice to validate models by visualising the morphological features that drive their decisions.[9,11,12,18–22] If a model relies on similar morphological features as an expert pathologist for a single well-defined task, the visual strategy it has learned for histopathological analysis is aligned with humans and is trustworthy. While this approach is effective for comparing the strategies of humans and machines on established histopathological analysis tasks, such as tumour detection, it is poorly suited for tasks that are difficult or impossible for expert pathologists because there are no established morphological criteria. Indeed, for tasks such as molecular profiling from haematoxylin and eosin (H&E) slides, the features DNNs use to render their decisions are often presented without any biological ground-truth for comparison,[9,11,12,18–22] possibly because the appropriate biological data is impossible or very costly to acquire. Here, we challenge this standard for assessing DNN visual strategies in histopathological image analysis to identify tasks where they are trustworthy or deceptive. If a DNN is trustworthy, its visual strategy will closely align with expert pathologists on standard tasks or reflect novel disease-relevant features on tasks that are difficult or impossible for expert pathologists. Conversely, a deceptive DNN will rely on visual features that are unrelated to the underlying biology and ultimately useless for the clinic or research.

We systematically test the trustworthiness of DNNs by measuring the visual strategies they learn after being trained to solve multiple histopathology analyses ranging in difficulty, from straightforward tumour detection to tasks that are potentially impossible for expert pathologists: determining the molecular profile and primary tissue of a tumour (CUP). Next, we develop a standardised evaluation framework to distinguish between trustworthy and deceptive visual strategies of DNNs trained to automate these analyses. Our general approach is to infer the tissue morphology driving DNN decisions, then compare this morphology to gold-standard diagnostic criteria for the same task. For molecular profiling, where there is no such gold standard—common wisdom is that expert pathologists cannot do this from tissue morphology alone—we turn to laser capture microdisection (LCM[23]) to collect regions of interest (ROI)-based genomic labels on adjacent WSIs to the ones that have full-slide genomic labels. For determining the primary tissue of a tumour, we restrict our analysis to tumour subtypes that contain no tissue-specific features, which controls for known shortcuts. Our approach fills a void in the rapidly growing field of deep-learning-based histopathological image analysis: revealing tasks that are vulnerable to deceptive learning and as of now unsuitable for automation with state-of-the-art deep learning models and training routines.

Subjects and Methods

PATIENT COHORTS

Moffitt cohort

The Moffitt cohort includes patients at Moffitt Cancer Center (MCC) diagnosed with adenocarcinoma of

pulmonary origin that also have associated molecular profile data from 01/01/2011 to 06/09/2020. Participants were included if they satisfied the following criteria. (i) The patient must have a pathological diagnosis of adenocarcinoma with a pulmonary origin. The adenocarcinoma may involve any organ, as long as pathology reports it as primary lung adenocarcinoma (e.g. a brain tumour that is pathologically confirmed as metastatic lung adenocarcinoma). (ii) The aforementioned adenocarcinoma must have had molecular profiling demonstrating either the presence or absence of KRAS mutations or EGFR mutations. (iii) A tissue glass slide associated with the formalin-fixed paraffin-embedded (FFPE) tissue block sent for molecular testing must be readily available through the MCC Pathology Department. Participants were excluded from the cohort if they satisfied any of the following criteria: (i) they had a concurrent pathological diagnosis of another cancer, (ii) no associated molecular data (iii) or mutations in HRAS and NRAS genes.

Slides that passed the inclusion/exclusion criteria were reviewed for quality and scanned at ×20 via the Aperio AT2 high-volume digital whole slide scanner. This digitisation process produced a digital ScanScope Virtual Slide (SVS) file per slide scanned. These SVS files are referred to as whole-slide images (WSIs). The WSIs were annotated for tumour regions of interest areas using a virtual pen via the Aperio ImageScope pathology slide viewing software. The final annotations were stored in XML files.

The Cancer Genome Atlas cohort

Diagnostic slides for adenocarcinoma patients from The Cancer Genome Atlas (TCGA) lung were downloaded from the GDC (https://gdc.cancer.gov) in SVS format. Seven patients were excluded secondary to an unacceptable prior treatment, an item not meeting the study protocol. Nine SVS files without magnification data were excluded. Ten SVS files with extensive pen marks were excluded.

ETHICS STATEMENT

Our Moffitt cohort consisted of existing tissue specimens and medical records information at the H. Lee Moffitt Cancer Center in Tampa Bay, Florida. All patients had previously provided written informed consent and a signed waiver of Health Insurance Portability and Accountability Act (HIPAA) authorisation, and no new samples were gathered for the sole purpose of this study. The study was approved by the Moffitt Institutional Review Board (IRB no. 00000971).

MOLECULAR LABELS

Moffitt cohort

The associated molecular profiles of each patient's WSIs were obtained through requests to Moffitt's Collaborative Data Services (CDS). The data came from four different sequencing strategies: FoundationOne CDx assay (Foundation Medicine, Cambridge, MA, USA), Moffitt STAR, TruSight Tumour 15 (Illumina, San Diego, CA, USA) and in-house pyrosequencing.[24] Only mutations that were labelled as clinically significant by each respective kit were included in the final data set.

TCGA cohort

TCGA molecular alteration labels were derived from the public mutation annotation file prepared by Ellrott *et al*.[25] All intronic and silent mutations were excluded. Patients with mutations in EGFR as well as in a RAS family gene (KRAS, NRAS, HRAS) were excluded. To match our Moffitt cohort, we excluded patients without clinically significant mutations in either EGFR or KRAS [as determined by Catalogue of Somatic Mutations in Cancer (COSMIC) annotations].

Laser capture microdissection

Twenty-one slides with minimal histological artefacts and abundant tissue were chosen (11 with KRAS mutations and 10 EGFR mutations). For each slide, 10–20 1-mm × 1-mm regions of interest (ROI) in the tumour-annotated area were further annotated. The ROIs were spatially distributed throughout each slide at randomly selected locations within the tumour.

Unstained FFPE tissue sections, prepared on polyethylene napthalate membrane slides, were deparaffinised and dehydrated by dipping in 100% xylene for 2 min. Slides were allowed to air-dry for 5 min. Selected ROIs were microdissected from the tissue using an Acturus XT LCM system (Life Technologies Corp., Carlsbad, CA, USA) with UV laser cut only. The UV laser precisely cut the PEN membrane around the ROI and the microdissected tissue was transferred to a 0.5 ml microfuge tube. All ROIs microdissected in the study were 1 mm^2 in size and matched to ROIs annotated by the study pathologist on WSIs representing a sequential section of the tissue. Each ROI was sequenced for KRAS and EGFR mutations using pyrosequencing.

Image preprocessing

For the Moffitt cohort, H&E FFPE tissue slides were scanned using the Aperio AT2 high-volume digital whole slide scanner at ×20 magnification, corresponding to a resolution of 0.5 μm $pixel^{-1}$. For the TCGA cohort, the

H&E FFPE tissue slides were downloaded in SVS format from the GDC. Each WSI was divided into 512-pixel × 512-pixel tiles with no overlap between adjacent tiles. Tiles with more than 50% background pixels were removed ('background pixel' is defined as a pixel with a grey-scale pixel value greater than 220, which was chosen based on visual inspection). Macenko stain normalisation[26] was applied to each tile image using a reference WSI to correct for differences in the staining process. Tile images with low contrast or empty tissue masks were removed. For each WSI, the tumour area was annotated by our pathologist (D.S.V.) at a 2× to 4× magnification using the Aperio ImageScope pathology slide viewing software. The tile images were assigned a 'tumour' or 'benign' label if the entire image was inside or outside the tumour-annotated area, respectively. The 'tumour' tile images were assigned a 'primary' or 'metastatic' label if the WSI tissue was derived from the lung or another tissue, respectively.

Model training and hyperparameter tuning

We trained and tested five different deep neural networks (DNNs) on tile images for the three histopathological analysis tasks described in Results using PyTorch. The five different architectures—ResNet18, ResNet34, ResNet50,[27] Shufflenet-version 2[28] and Inception version 3—were chosen because of their widespread use in computer vision and their recent applications to histopathology data. All models were pretrained on the ImageNet data set.[29] The last classification layer in each model was replaced with a fully connected linear layer with one output node, and the weights for this linear layer were initialised randomly.

The training data was augmented with versions of each tile image that were flipped either horizontally or vertically, or both at once. Training tile images were sampled inversely proportional to the frequency of their labels to control for imbalance labels in individual tasks. Models were trained with batches of 64 tile images (ResNet50 runs had a mini-batch size of 32 tile images due to memory constraints). Model weights were optimised using binary cross-entropy, the Adam optimiser,[30] and a learning rate of $1 \times 10e\text{-}4$, which was the best-performing learning rate on pilot experiments on the molecular profiling task.

For the tumour detection and primary tissue identification tasks, each neural network was trained for three epochs and tested on validation tile images every 200 mini-batches (400 for ResNet50). The final weights for each model were chosen based on the training step that yielded the lowest validation loss. For the molecular classification task, the final weights for each model were chosen based on the training step that yielded the highest slide-level area under the receiver operating curve (AUROC), as was carried out in prior work on this task.[9]

Cross-validation

To ensure the robustness of model results, cross-validation data folds were created for each classification task using WSIs from the Moffitt cohort. In brief, cross-validation involves splitting up a larger image data set into smaller groups for training, validating and testing models. We split the Moffitt cohort by patients, such that WSIs from a single patient could were only used in either training, validation or testing within a single iteration of cross-validation. Models were trained using the training set and weights which achieved the minimum loss on the validation set were used for testing, then model performance on the test set was recorded. The cross-validation procedure was iterated until all patients were included in a test set, which allows us to report test performance across all images. Details on the composition of Moffitt data used in each experiment varied, and are reported below.

Molecular profiling in WSIs

The Moffitt data set consisted of 221 total FFPE WSIs, spanning 182 patients. For analysis, we filtered the original data set by excluding patients and slides that also had LCMs, which we set aside as challenging test data sets. We also excluded three WSIs with annotated tumours that were too small to extract tile images from. This resulted in 182 remaining WSIs. Of those, we made the decision to keep patients with multiple WSIs (46) or metastatic WSIs (72) in the training set for all folds of cross-validation, to control contamination of tissue from the same patient appearing in a train and test set and so that any negative results would not be to out-of-domain generalisation issues from non-lung tissue. The remaining 85 WSIs consisted of 33 EGFR WSIs and 52 KRAS WSIs. To balance the labels of WSIs entered into our cross-validation procedure, 19 KRAS mutated WSIs were chosen randomly and added permanently to the training. The rest of the WSIs were divided into six different splits of 138 training WSIs, 22 validation WSIs and 22 test WSIs. The validation and test sets contained an equal number of KRAS mutated and KRAS wild-type WSIs.

Molecular profiling in LCMs

To evaluate the localisation accuracy of models trained for molecular profiling, we tested their ability to generalise to LCMs after training on WSIs. We did this by augmenting the cross-validation folds used to train models for molecular profiling, mixing in LCMs

with WSI image tiles in the training split and replacing the validation and test splits with LCMs. This strategy gave our DNNs the best chance to learn to generalise to the LCMs.

Because we had 21 WSIs with LCM data (11 with KRAS mutations and 10 without), we had to take steps to ensure that an equal number of these slides were included in training for each fold (note that we controlled for imbalances in the test set during evaluation through resampling, which we elaborate on below). Specifically, for all but one cross-validation split we selected one additional WSI with LCM data and a KRAS mutation at random to include in the training set.

Primary tissue identification

For the task in which models were trained to identify the primary tissue of a tumour, we once again took additional steps to ensure that WSI labels were balanced in our cross-validation procedure. After adding 60 WSIs depicting metastatic lung adenocarcinomas in brain, liver, lymph node and bone to the 221 WSIs used for molecular profiling, 72 WSIs from patients with multiple WSIs were added permanently to the training set. One metastatic WSI and 45 primary lung WSIs were chosen randomly and added permanently to the training set. The remaining 100 WSIs (50 primary and 50 metastatic) were divided into five different splits of 80 training WSIs, 10 validation WSIs and 10 test WSIs. The validation and test partitions each contain five WSIs from primary lung tissue and five WSIs from metastatic sites to ensure balanced data sets for evaluation. Like the tumour-normal folds, each of the 100 WSIs was in a validation split and a test split exactly once.

Tumour detection

For this task, we considered 148 WSIs from primary lung tissue with high-quality tumour annotations. Any patient with multiple WSIs or at least one WSI with tile images only inside or outside the tumour annotated area was added permanently to the training set. Of the remaining 93 WSIs, three were randomly chosen to be permanently in the training set. The 90 remaining WSIs were divided into five different splits of 54 training WSIs, 18 validation WSIs and 18 test WSIs. Each of the 90 WSIs was used in a validation split and a test split exactly once.

Performance metrics

For each task we calculated balanced accuracy and AUROC, following the procedure laid out in recent work in automating histopathological analysis.[9] Balanced accuracy is defined as the average of the true positive rate and true negative rate; a model score of 0.5 is used as the boundary between negative and positive. For the primary-metastatic task and KRAS-mutation task, we also calculated the performance metrics at the slide level. Each WSI is assigned a score based on the median of all the tile-level model scores. All reported performance metrics were computed using the WSIs in the held-out test set. The model scores were concatenated across all cross-validation test folds. If a tile image had multiple model scores because it appeared in more than one split of cross-validation, the median model score was assigned as the final model score.

Confidence intervals (CIs) for each metric were calculated with bootstrapping (1000 iterations). For an experiment with n exemplars in the test set, we sampled n exemplars with replacement. We performed this sampling scheme for 1000 iterations and calculated balanced accuracy and AUROC for each iteration. The distribution of bootstrapped balanced accuracy and AUROC metrics was used to generate 95% CIs. In experiments where there was potential for imbalanced labels at test time, we additionally ran a version of the bootstrapping procedure where every iteration consisted of class-balanced samples. Both balanced and standard versions of bootstrapping yielded the same results on every experiment, so we report results using the standard method from Coudray *et al.*[9] for consistency with the result of the field.

Grad-CAM

We visualised the learned analysis strategies of each trained model using generated guided Grad-CAM.[31] This method generates a feature importance map of the same size as an input image, which indicates the pixels that contributed to the model's decision for that image. The final convolutional layer in each model ('conv5' for ShuffleNet and 'layer4' for the ResNet models) was used to generate the Grad-CAM mask, which ignores noisy locations in the feature importance map. We generated these maps for all tile images in test sets using code from: https://github.com/kazuto1011/grad-cam-pytorch. Given an input tile image of size $512 \times 512 \times 3$ (three colour channels: red, green, blue), the output Grad-CAM map has the same dimensions of $512 \times 512 \times 3$. We processed each Grad-CAM map using the following sequence: (i) each value in the map was set to its absolute value, (ii) the mean value across feature channels at every location was stored, yielding a $512 \times 512 \times 1$ map, (iii) outlier values that were more than three standard deviations away from the mean pixel value

in the map were clamped to three standard deviations and finally (iv) the map was normalised to [0, 1] for visualisation.

Nuclei segmentation

We used the PyTorch implementation of HoVer-Net to segment and type nuclei in WSIs. The model was trained on the PanNuke data set.[32] All tile images in the tumour annotated regions of WSIs and all LCM patches were processed using HoVer-Net. The model labels each segmented nuclei as belonging to one of the following categories: unknown, neoplastic, inflammatory, connective, dead or non-neoplastic epithelial. We combined the 'dead' category with the 'inflammatory' category based on observations by expert pathologists that the model was misclassifying lymphocytes as 'dead'.

Nuclei-gradient score calculation

The nuclei-gradient score is based on the intersection over union (IoU) metric that is used to assess object segmentation accuracy in computer vision.[33] For a tile image, we used its Grad-CAM map and the nuclei segmentation results to calculate a nuclei-gradient score for each type of nuclei. Given a threshold value between zero and one, all pixels with Grad-CAM map values greater than or equal to the threshold value were identified. The threshold value was set at 0.5 for results reported in the main text. To compute the IoU, we first counted the suprathreshold Grad-CAM map values. This count was used as the denominator in the IoU. Next, we counted the suprathreshold Grad-CAM map values that also fell within a segmented nucleus, which was used as the numerator in the IoU.

Results

OVERVIEW

We investigated the trustworthiness of DNNs trained for histopathological analysis by developing a standardised evaluation framework, which consists of two steps. For step 1, models are tested within and outside of the training distribution to identify trivial shortcut strategies by measuring generalisation performance; for step 2, the morphological features that each model relies on for solving a task are inferred, validated for their importance to a model's decisions, then compared to the task's gold-standard diagnostic criteria to detect deceptive visual strategies. We scrutinised four popular DNNs trained to solve multiple image analysis tasks posed on a novel data set of 221 histology WSIs. These WSIs depicted H&E-stained formalin-fixed paraffin-embedded (FFPE) tissue from 182 patients diagnosed with lung adenocarcinoma at the Moffitt Cancer Center (see Subjects and methods). We refer to this data set as the Moffitt data set. Our DNNs were 18, 34 and 50-layer versions of the ResNet version 2[27] and the ShuffleNet version 2.[9,11,27]

Each of the DNNs was trained to solve three analysis tasks posed on lung adenocarcinoma slides that range in difficulty for expert pathologists, from potentially impossible to a standard histopathological analysis: (i) molecular profiling of KRAS versus EGFR-mutated tumour (KRAS-mutation), (ii) determining the primary tissue of a tumour (CUP) and (iii) classifying lung adenocarcinoma versus benign tissue (tumour detection, task denoted as tumour-normal). As detailed in the results, we utilised a combination of molecular and expert pathologist annotations in order to pose these three distinct tasks on the same slides. For each task, we analysed the performance and interpretability of the four DNN architectures on WSIs from our novel Moffitt data set, which was held out from training or model selection (Figure 1). To measure out-of-distribution generalisation, we tested models on the public TCGA slide image data set (173 total slides; https://www.cancer.gov/tcga), which includes clinical, genomic and histological data for lung adenocarcinoma.

Each WSI in the Moffitt data set and TCGA yielded, on average, approximately 2000 non-overlapping 512-pixel × 512-pixel image patches. We partitioned WSIs from the Moffitt data set into separate cross-validation folds for training, model selection (validation) and testing (each fold contained unique WSIs, see Subjects and Methods). Models were then trained and evaluated on image patches extracted from these WSIs, as is standard practice in the field due to the computational complexity of training directly on WSIs.[9,11,12,18–22]

TASK 1: DNNS TRAINED FOR MOLECULAR PROFILING ARE VULNERABLE TO DECEPTIVE LEARNING

We began by analysing the performance of DNNs trained to do molecular profiling on H&E WSIs. Molecular profiling is an important step for determining the course of treatment for the disease, which requires expensive and time-consuming genomic panels. Recent work has shown that genomic mutations can be reliably decoded from morphology by DNNs.[9,11,12] Are these DNNs and their putative successes trustworthy or deceptive?

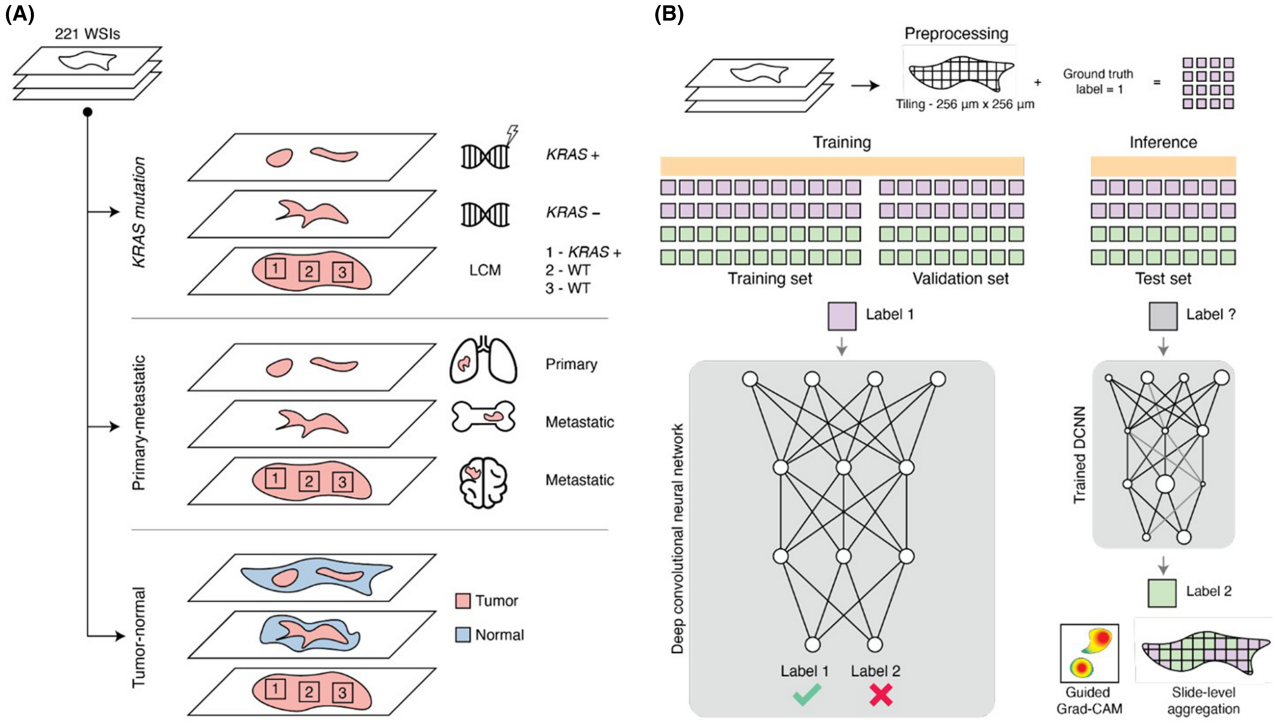

Figure 1. Overview of classification tasks and deep learning framework. **A,** Schematic showing the three classification tasks. For the KRAS-mutation task data, we divided our data set into whole-slide images (WSIs) from tissue that had mutations in the KRAS gene (KRAS+) and WSIs from tissue that did not have mutations in the KRAS gene (KRAS−). We collected additional sequencing information using laser capture microdissection (LCM) to dissect and sequence arbitrarily selected regions of interest (depicted by boxes 1–3) for mutations in KRAS. For the primary-metastatic task data, we identified WSIs that were biopsied from the lung (primary) or from other sites but with a known pulmonary origin (metastatic). Only tissue inside the tumour annotated area was used in this task. Lastly, for the tumour-normal task data, we identified tumour and normal tissue in all WSIs using pathologist tumour annotations. Tumour tissue is coloured pink and normal tissue is coloured blue. **B,** All WSIs were tiled into non-overlapping 256 µm × 256 µm tile images, corresponding to 512-pixel × 512-pixel image patches. Each image patch was assigned a different ground-truth label for each task. The image patches were divided into training, validation and test folds; each patient's image patches belonged to a single fold. The image patches in the training and validation sets were used to train and select hyperparameters, whereas image patches from the test fold were used for model evaluation.

We modelled clinically significant genes to remain consistent with prior work on the task,[9,11,12] focusing specifically on whether or not DNNs could discriminate mutations of genes that do not co-occur in the Moffitt data set or TCGA: KRAS and EGFR. We trained models for binary classification, where KRAS mutations were the positive class and EGFR mutations were the negative class. All image patches for training and evaluation were taken from inside tumour-annotated regions of WSIs. Models were evaluated by recording area under the receiver operator characteristic curve–area under the curve (ROC–AUC) and class-balanced accuracy.

To evaluate DNNs on molecular profiling, we first measured their performance within and outside the training distribution. All DNNs were significantly above chance at detecting KRAS mutations at the slide level in the Moffitt data set ($P < 0.05$ over 1000 bootstrap replicates, where chance = 0.5 for both balanced accuracy and ROC–AUC; Table 1). There were no significant differences between the performance of the models and they all rendered similar decisions on the task, indicating that model architecture differences did not translate into qualitatively different task strategies (all model-to-model Pearson correlations exceeded $\rho = 0.68$, $P < 0.001$; Supporting information, Table S1). The models also performed significantly above chance when tested out-of-distribution on the TCGA (Supporting information, Table S1). In other words, all the DNNs we tested learned similarly generalisable strategies for detecting KRAS at the WSI level. These results are consistent with multiple other applications of DNNs for molecular profiling,[9–12] which indicate that DNNs are significantly above chance in profiling WSI genomes while still far below the error rate of standard molecular tests (most

Table 1. Deep neural networks (DNNs) successfully detect KRAS mutations at the whole-slide level, but fail to localise KRAS mutations within a slide. Slide predictions are calculated from the median of logit scores across all image patches in a whole-slide image. Laser capture microscopy-captured image patches are evaluated independently. For each metric, we report the 95% confidence interval using 1000 bootstrap replicates

Analysis	Score	ResNet-18	ResNet-34	ResNet-50	ShuffleNet
Whole-slide level (WSI)	Weighted accuracy	0.59 [0.52–0.68]*	0.60 [0.53–0.68]*	0.60 [0.53–0.69]*	0.63 [0.54–0.71]*
	ROC–AUC	0.69 [0.59–0.78]**	0.69 [0.59–0.78]**	0.70 [0.60–0.78]**	0.67 [0.57–0.77]**
LCM-patch localisation	Weighted accuracy	0.51 [0.45–0.57]	0.38 [0.33–0.44]	0.45 [0.40–0.50]	0.47 [0.41–0.53]
	ROC–AUC	0.50 [0.43–0.57]	0.40 [0.33–0.47]	0.44 [0.37–0.50]	0.46 [0.39–0.53]

Statistical testing against chance accuracy (0.5) is denoted by asterisks: $^{**}P < 0.01$, $^{*}P < 0.05$.
ROC–AUC, receiver operating curve–area under the curve.

assays can detect somatic mutations at greater than 95% sensitivity and specificity[34]).

The above-chance generalisation performance of the DNNs means that it is unlikely that they learned to detect KRAS through a trivial shortcut,[15] such as experimental batch effects or systematic variations in how the slides of different patients were handled and prepared. However, it is still possible that the DNNs learned to rely on a deceptive visual strategy, exploiting visual features that are correlated to KRAS mutations but not related to the underlying biology. Because there are no gold-standard morphological phenotypes for different mutations, it is difficult to detect such deceptive learning on this task.

To investigate whether our DNNs learned deceptive visual strategies for detecting KRAS mutations, we gathered additional sequencing data on 21 patient WSIs. We collected 10–20 1×1 mm regions-of-interest (ROIs; $n = 216$) at positions distributed within the tumour of each of these WSIs using LCM. Next, we sequenced each ROI for KRAS and EGFR mutations (see Subjects and Methods), providing us with an estimate of the spatial distribution of these mutations in the WSIs. There were 90 patches with KRAS mutations, 109 patches with EGFR mutations and 17 wild-type patches without mutations in either gene. KRAS mutations were spatially heterogeneous, and four of the slides that were labelled as KRAS mutation based on their whole-slide panel contained regions of wild type (17 total patches).

No DNN exceeded chance-level on the LCM patches, and all performed significantly worse than on WSIs (Figure 2). These results raised the question: what morphology had the DNNs learned to rely on to perform above chance at the level of WSIs?

A popular approach to interpret DNN decisions is to compute the gradient of a model's output (z) with respect to the input image (x): $\frac{\partial z}{\partial x}$. This vector of gradients captures the importance, or 'saliency', of every pixel in the input image to the model's decision for that image, where larger (absolute) gradient values denote more importance.[35–38]

Model saliency maps are noisy estimates of feature importance, and subsequent advancements in explainable artificial intelligence have involved modifications of the gradient to reduce noise in these maps. A particularly effective approach for biomedical imaging[39] is guided gradient-weighted class activation mapping (GradCAM[31]), which we used here to identify morphological features that drove DNN molecular profiling decisions. We specifically focused on interpreting Shufflenet's decision-making, because it performed the best of the models that we tested at detecting KRAS mutations on WSIs.

We applied GradCAM to the predictions of Shufflenet on WSIs in the Moffitt test set. In order to identify the types of features the model relied on for its decisions, we also passed these slides through HoVer-Net,[40] a DNN trained to segment nuclei in H&E images. Nuclear features in H&E slides are commonly used to assess a host of different diseases.[41] By measuring the overlap between GradCAM maps for KRAS detection and nuclear features, we reasoned that it would be possible to measure the extent to which model decisions relied on each type of feature. On average, only 14% of model GradCAM pixels for KRAS detection overlapped with tumour nuclei predictions (this metric ranged from 11 to 15% for the three other models; Figure 3). We also explored the correlation between model decisions and histological subtype, a morphological feature that is known to be weakly associated with KRAS mutations.[21] Histological subtyping, unlike molecular profiling, is straightforward for expert pathologists and is detectable by

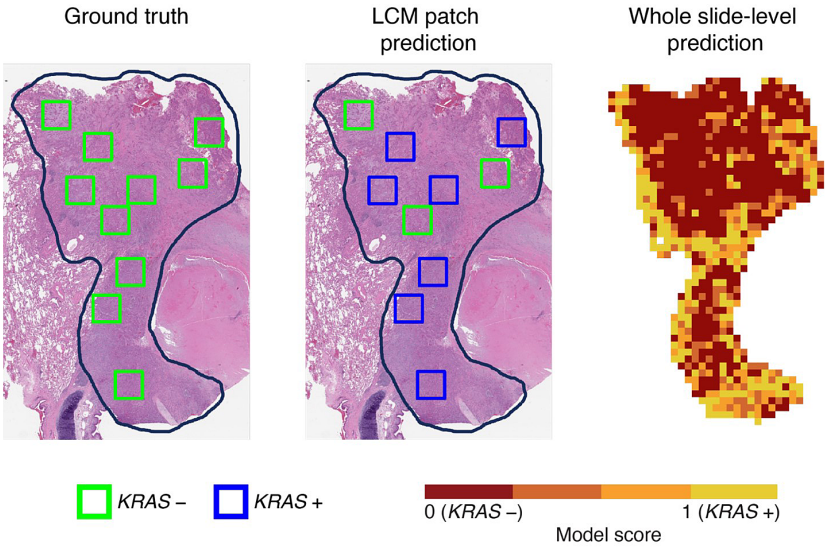

Figure 2. Deep neural network (DNN) KRAS mutation localisation is inconsistent with laser capture microscopy (LCM) sequencing. Left and middle: haematoxylin and eodin (H&E) stained whole-slide images with tumour outlined in black, and boxes representing locations of LCM, which are coloured according to their molecular labels. The left panel depicts the ground truth labels, and the middle the predictions of the ShuffleNet model, which accurately classifies three LCM patches as KRAS negative (green) and incorrectly classifies the rest of the seven patches as KRAS positive (blue). On the right, the heatmap is coloured according to the ShuffleNet model score; a model score of 0 (maroon squares) corresponds to no KRAS mutation detected and 1 (yellow squares) corresponds to KRAS mutation detected. Areas outside of the annotated tumour or where there were no cells are white. Despite the heterogeneity of the model's predictions at the LCM level, the model's whole slide decision correctly identifies this as a KRAS wild-type tumour.

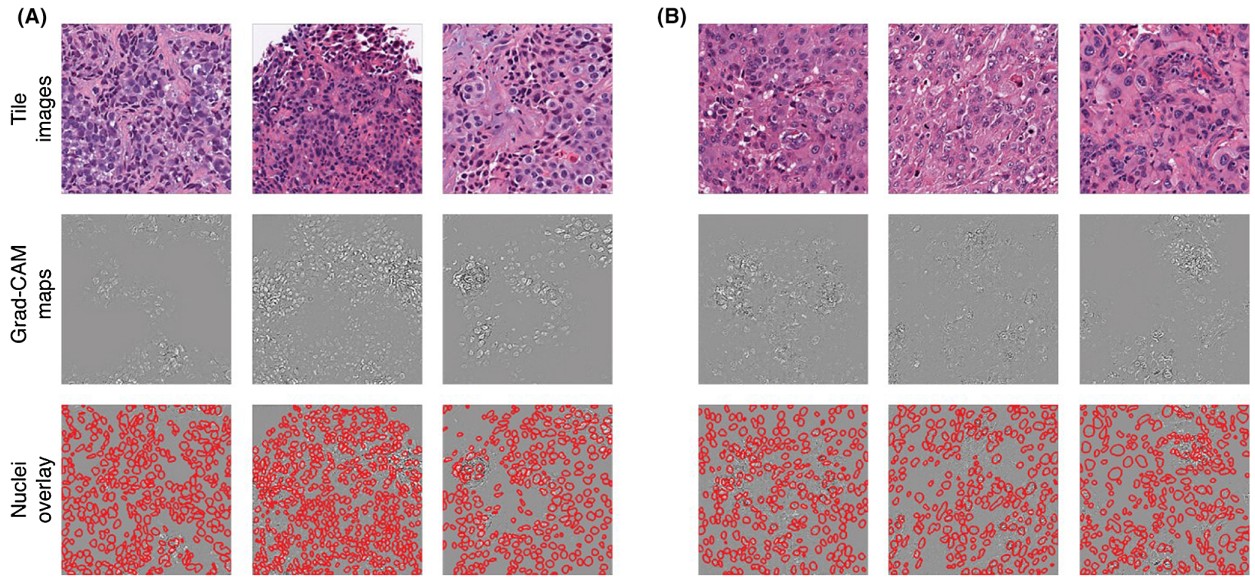

Figure 3. Deep learning models use nuclear and non-nuclear features to identify the primary tissue of a tumour. The first row shows three representative tile images from (**A**) metastatic and (**B**) primary tile images (WSIs). Guided gradient classification activation maximisation (GradCAM) maps from ResNet34 for each tile image are shown in the second row. The last row depicts GradCAM maps with segmented nuclei outlined in red. [Colour figure can be viewed at wileyonlinelibrary.com]

DNNs.[42] It is possible that DNNs trained for KRAS detection could not find strong morphological signals for KRAS, and instead learned to rely on weakly correlated subtypes. If this is the case, the drop in DNN performance when tested on WSIs versus LCMs could be explained by the correlation strength—or

lack thereof—between subtypes and mutations on WSIs versus LCMs. To test for this possibility, we annotated the histological subtypes present in each LCM patch image and the WSIs they came from. We found six total subtypes: acinar, lepidic, solid, papillary, micropapillary and mucinous. We next computed the correlations between these subtypes and the presence of KRAS mutated versus KRAS wild-type in the LCMs and WSIs.

We used logistic models to regress the presence or absence of a KRAS mutation in each WSI onto its annotated histological subtypes. Consistent with prior work,[21] we found a significant (one-tailed) association between the micropapillary subtype and KRAS wild-type ($z = -1.675$, $P = 0.044$), and a significant correlation between the solid subtype and KRAS mutations ($z = 2.026$, $P = 0.026$). However, after repeating this analysis on the LCM patch images, we found no significant correlations between histological subtype and KRAS mutated or KRAS wild-type. This means that a model which has learned to detect KRAS mutations by focusing on subtype at the whole slide level should fail to generalise to LCM patches, where these correlations are absent. We verified that the Shufflenet adopted this strategy by fitting a linear model to regress the WSI histological subtypes onto its KRAS mutation predictions (in logits). The model's predictions were significantly correlated with the solid subtype ($z = 1.814$, $P = 0.035$) despite being trained to detect KRAS mutations, validating our hypothesis that these models tend to rely on histological subtype to detect mutations.

DNNs are significantly above chance at detecting mutations from histology in WSIs because they learn to classify histological subtypes that are correlated with those mutations. Subtypes are straightforward to classify by eye for expert pathologists and are only weakly associated with genetic mutations. As we demonstrate, these correlations exist at the whole-slide level but disappear at granular levels of analysis. That DNNs learn to rely on subtypes as a shortcut for molecular profiling means that their predictions of morphology related to genetic mutation are deceptive. This is especially an issue when models are asked to render decisions at a different level of analysis than they were trained, such as predicting mutations for particular cells after being trained on whole slides.

To understand more clearly if our findings in this controlled setting are predictive of the behaviour of existing models in the literature, we next turned to the Inception version 3 DNN used in Coudray *et al.*,[9] the first work to our knowledge that reported the ability of DNNs to predict molecular profiles from H&E morphology in WSIs of lung adenocarcinoma. We worked from the publicly available GitHub repo (https://github.com/ncoudray/DeepPATH) and received model weights trained on data from TCGA provided by the corresponding author. We began by testing this model on our FFPE cohort from the TCGA, where we found it was significantly above chance in discriminating between KRAS and EGFR mutations (ROC–AUC: 0.594, CI = 0.526–0.683, $P < 0.05$). However, the model dropped to chance accuracy when tested on our WSIs (ROC–AUC: 0.415, CI = 0.300–0.501) from our Moffit cohort. Given this failure in generalisation to even WSIs from Moffit, we can conclude that the model is not trustworthy; we do not need to analyse its learned strategies any further.

To understand more clearly if the failure in generalisation of the Inception version 3 from Coudray *et al.*[9] is due to the specific training data set and routine they used, a problem with the model architecture, or a fundamental limitation of DNNs, we next trained a new Inception version 3 from an ImageNet initialisation with the same procedure we used for all models in our manuscript. This model was significantly above chance in detecting KRAS mutations on WSI TCGA (ROC–AUC: 0.663; $P < 0.01$) and Moffit (ROC–AUC: 0.688; $P < 0.01$) data sets but dropped to chance when tested on laser capture microscopy images from Moffit (ROC−AUC: 0.445; not significant). To understand more clearly the provenance of these failures, we moved to the second stage of our deceptive learning framework and analysed the model's learned strategies. The model's decisions on Moffit WSIs correlated with the presence or absence of micropapillary and acinar tumour subtypes (both $P < 0.05$), but these correlations disappeared in the LCM image data set, which indicates that this model, like the others we describe in our manuscript, had learned the deceptive shortcut of predicting mutation from subtype. In other words, DNNs trained to predict molecular data from H&E-stained histology are dangerously vulnerable to deceptive learning.

TASK 2: DNNS CAN RELIABLY IDENTIFY THE PRIMARY TISSUE OF A TUMOUR

Are there other tasks that are difficult or impossible for expert pathologists that DNNs can learn to solve reliably? To address this question, we next asked how effectively DNNs can classify the primary tissue of a tumour. Cancer of unknown primary (CUP) describes tumours for which the primary anatomical site cannot be determined. Because this information is critical

for modern therapeutics, clinicians often turn to expensive and time-consuming genetic or transcriptomics analyses. However, it has recently been suggested that DNNs can diagnose the origin of the primary tumour from morphology alone,[13] even though expert pathologists are incapable of performing this task by eye.

We tested whether the ability of DNNs to detect the primary tumour origin is trustworthy or not by posing a version of the task on the same lung adenocarcinoma slides that we used to investigate molecular profiling. We first split the slides into primary and metastatic lung adenocarcinoma. Next, we restricted our set of slides for model training to contain only those which had tumours composed of the solid histological subtype, which controls against the presence of trivial features for detecting tissue type. We extracted image patches from tumours in these slides for model training. Finally, we developed an additional test set of slides containing image patches only from sheets of tumour cells in the solid subtype tumour, which is the strictest possible criterion for controlling tissue-specific shortcuts for solving the task.

As with the molecular profiling task, we first measured the generalisation performance of DNNs trained to solve this task. All the DNNs performed significantly above chance in differentiating primary and metastatic lung adenocarcinoma in the Moffitt data set, reaching between 60 and 72% accuracy (Table 2) and rendering similar decisions to each other: the Pearson correlation coefficient between all models' decisions ranged between 0.74 and 0.77 ($P < 0.001$ for all model pairs according to randomisation tests[43]).

We next evaluated out-of-distribution generalisation of these DNNs on the TCGA data set. All subtypes were included, as we did not have histological subtype annotations for many of the WSIs in the TCGA data set. In addition, as the TCGA only contains primary lung adenocarcinoma, we measured performance by recording true positive rates. All models maintained true positive rates that were significantly greater than chance in this generalisation task ($P < 0.05$), with tile true positive rates ranging from 70 to 84%.

We focus our analyses here on the best performing ResNet34 (see Supporting information for the other models) and adopted the same evaluation strategy from the molecular profiling task to identify and analyse morphological features that the DNNs use for differentiating primary from metastatic lung adenocarcinoma. However, while it is apparent that the model learned to target tumour cells, the differences between the features selected in its GradCAM maps for primary and metastatic tissue was small, with less fibrotic tissue in the metastatic versus primary images (Figure 4).

In the absence of obvious morphological criteria, we next tested whether nuclear features were driving model decisions. We found that model performance was significantly reduced when we masked DNN activities corresponding to the locations of nuclei segmented by HoVer-Net (unmasked balanced accuracy 71.3% versus nuclei masked balanced accuracy 63.9%, $P < 0.05$) but the performance remained significantly above chance. These results suggest that the model learned a visual strategy for classifying the primary tissue of a tumour, which relied on both nuclear and non-nuclear features.

To summarise, we found that DNNs can identify a tumour's primary tissue by leveraging a strategy that is at least partially trustworthy. When tested on sheets of solid subtype tumour cells, models were surprisingly significantly above chance, utilising a combination of nuclear and non-nuclear features to achieve this performance. Our findings on this task indicate a more bullish outlook for DNNs trained to identify a tumour's primary tissue than those trained for molecular profiling (Task 1). Our models predict that there are reliable morphological features for identifying a tumour's primary tissue even though the task is exceedingly difficult or impossible for expert pathologists.

Table 2. All four deep neural networks can differentiate primary lung adenocarcinoma images from metastatic adenocarcinoma images. For each metric, we report the 95% confidence interval using 1000 bootstrap replicates

Analysis	Score	ResNet-18	ResNet-34	ResNet-50	ShuffleNet
Image tile	Weighted accuracy	0.60 [0.60–0.61]**	0.72 [0.71–0.72]**	0.69 [0.68–0.69]**	0.71 [0.71–0.72]**
	ROC–AUC	0.62 [0.61–0.62]**	0.82 [0.81–0.82]**	0.74 [0.73–0.74]**	0.73 [0.72–0.73]**

Statistical testing against chance accuracy (0.5) is denoted by asterisks: $^{**}P < 0.01$.
ROC–AUC, receiver operating curve–area under the curve.

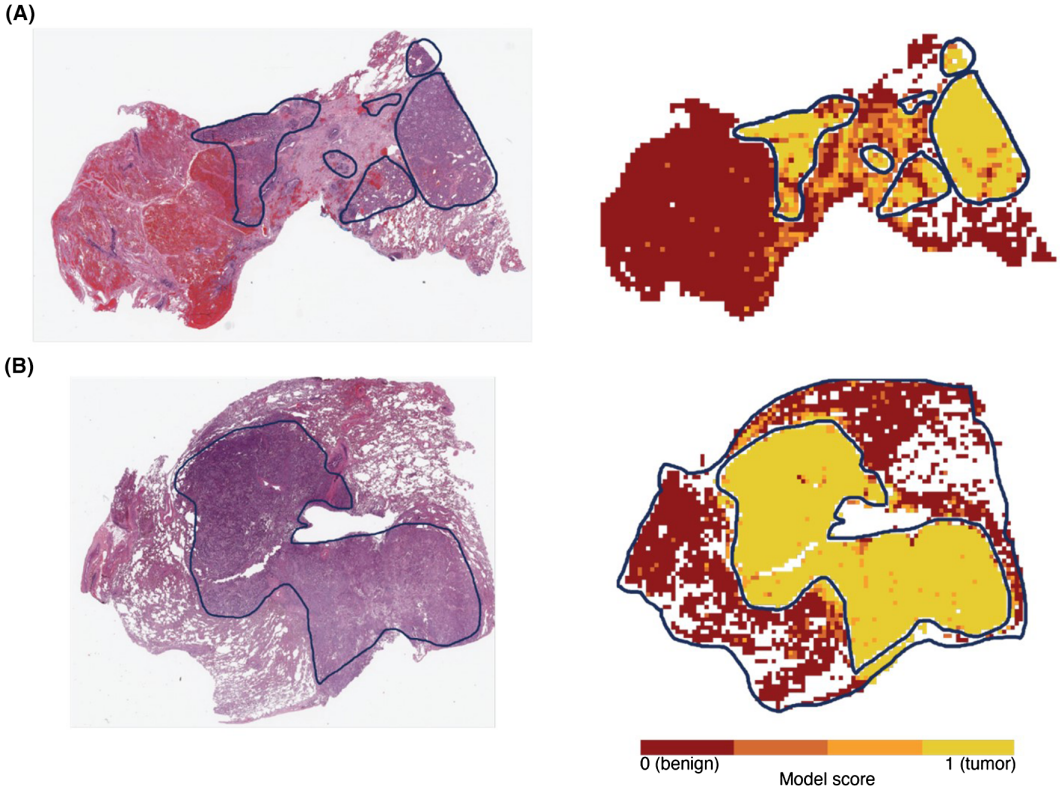

Figure 4. Prediction heatmaps for a ResNet18 trained to detect lung adenocarcinoma. (A & B) Two example whole slide images (WSIs) of lung adenocarcinoma are annotated for tumour regions in blue. The heatmaps in the right column are coloured according to the ResNet18 model score. A model score of 0 corresponds to benign tissue and 1 corresponds to tumour tissue. [Colour figure can be viewed at wileyonlinelibrary.com]

TASK 3: DNNS LEARN A ROBUST STRATEGY TO DETECT TUMOURS

Finally, we turned to a well-studied task in computational pathology: detecting lung adenocarcinoma in WSIs. This task is routine for expert pathologists, and DNNs trained to solve it have rivalled the performance of experts.[9,44–46] To test whether or not DNNs adopt deceptive visual strategies to solve tumour detection, we posed the task on the same data sets we used for molecular profiling and classifying primary tissue.

We trained and evaluated models on image patches extracted from WSIs, as was performed in prior work.[9,44–46] After training and evaluating on the Moffit data set, we found that the ShuffleNet and all three ResNets performed similarly to the state of the art,[9,44–46] reaching approximately 88% balanced accuracy and 0.95 AUROC (Supporting information, Figure S1, Table S1). We focused subsequent analyses on the best-performing model, the ResNet18.

When evaluated out-of-distribution on lung adenocarcinoma WSIs in TCGA, the ResNet18 once again performed well, reaching nearly 82% accuracy and 0.90 AUROC (Figure S1, Table S1). This decrease in performance is relatively small but statistically significant (95% CIs do not overlap, $P < 0.05$), and probably reflects experimental measurement differences between the two data sets, such as their use of different staining protocols.[47] Notably, however, the model's out-of-distribution generalisation is still significantly better than chance ($P < 0.001$). As in the other two tasks, all models trained for tumour detection rendered similar decisions to each other on the Moffitt data set and out-of-distribution generalisation on the TCGA data set. The Pearson correlation coefficient between models' decisions on every WSI was between 0.79 and 0.89 for both data sets ($P < 0.001$ for every pair).

The ResNet18 accurately localised tumour tissue in Moffitt and TCGA WSIs (Figure 4). Image patches from the Moffitt data set, for which ResNet18 is most confident to contain tumour tissue, also have morphological features associated with lung adenocarcinoma,[47] such as irregular and enlarged nuclei, prominent nucleoli and high nuclear to

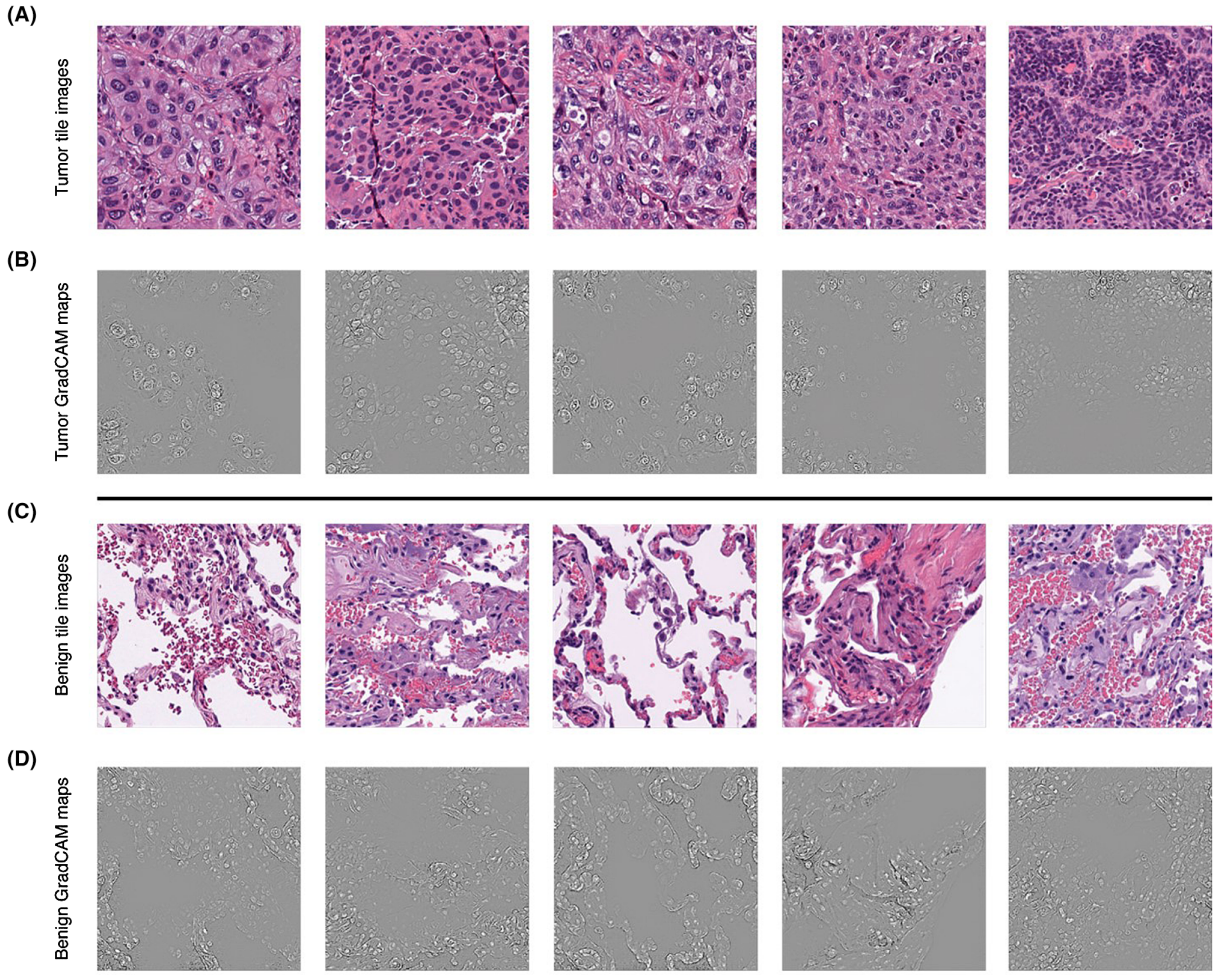

Figure 5. Deep learning models use nuclei to differentiate between tumour and benign tile images. **A,** Representative tile images inside the tumour-annotated area with high model scores (higher model score = more tumour-like). All model scores are greater than 0.999. **B,** Tumour class guided gradient classification activation maximisation (GradCAM) maps for each of the tile images in panel (**A**). The GradCAM signals often overlap with cell nuclei, and more specifically tumour cell nuclei. **C**, Representative tile images outside the tumour-annotated area with low model scores (lower model score = more benign-like). All model scores are less than $1e$-3. **D,** Benign class GradCAM maps for each of the tile images in panel (**C**). The GradCAM signals often overlap with cell nuclei and outline alveolar septal tissue. [Colour figure can be viewed at wileyonlinelibrary.com]

cytoplasmic ratio (Figure 5A). Moffitt image patches which ResNet18 is most confident are benign, depict benign alveolar septa (the tissue that separates the alveoli or air sacs) with empty air space and maintained structure, lined by benign type I (flat cells) and type II (cuboid) epithelium (Figure 5C). Overall, the distinguishing features in the tile images with high and low model scores correspond to features that pathologists use to distinguish between malignant and benign tissue.[46] To establish whether nuclear features were driving model decisions, we again masked DNN activities corresponding to the locations of nuclei in image patches. Ablating all tumour nuclei significantly reduced balanced accuracy from 88.3 to 72.4%. We repeated this experiment by permuting the tumour nuclei masks; the accuracy dropped to 85.7%. Ablating all nuclei reduced balanced accuracy to chance-level at 49.8%, while randomly permuting the masks reduced accuracy to 59.2%.

To establish whether nuclear features were driving model decisions, we again masked DNN activities corresponding to the locations of nuclei in image patches. Ablating all tumour nuclei significantly reduced balanced accuracy from 88.3 to 72.4%. We repeated this experiment by permuting the tumour nuclei masks; the accuracy dropped to 85.7%. Ablating all nuclei reduced balanced accuracy to chance-level at 49.8%,

while randomly permuting the masks reduced accuracy to 59.2%.

Lung adenocarcinoma classification with DNNs is trustworthy. All DNNs we tested perform well both within and outside the training distribution, and render similar decisions (model-to-model correlations are all between 0.79 and 0.89, $P < 0.001$). These decisions are driven by, and depend on, nuclear morphology that is consistent with textbook and gold-standard criteria for detecting lung adenocarcinoma in histopathology. One possible explanation for the consistent and strong performance of DNNs on lung adenocarcinoma classification is that annotations are at the pixel level, rather than at the whole-slide level, as in molecular profiling or classifying the primary tissue of a tumour. This level of granularity in annotations on training data has been found to drive DNNs towards consistent and generalisable visual strategies in object classification tasks and may be necessary to avoid deceptive learning.

Discussion

There is growing consensus that DNNs can automate tasks on biomedical imaging data, achieving performance that matches or exceeds human experts.[39,48] It is also becoming clear that these DNN achievements are due to visual strategies that are not necessarily aligned with those used by human experts.[35,37,39] When humans and machines use different strategies to solve tasks it can be a positive development, with the chance to reveal new insights into biology and generate testable hypotheses for understanding the development of disease.[31] However, there is no guarantee that the visual features that machines, but not humans, use to solve tasks are meaningful. There has been extensive work demonstrating that DNNs are in fact vulnerable to learning shortcuts in standard computer vision tasks posed on natural images, such as object recognition.[15,49] These shortcuts represent visual strategies that achieve high performance by focusing on biases that are unique to the training and testing data—from low-level cues such as lightness, contrast or colour to object-centric cues, such as their size in pixels or common appearance in a specific context. However, little is known about DNN shortcut learning in the context of biomedical imaging, and specifically histopathological analysis. Through our experiments, we have begun to address this critical question of shortcut learning in histopathological analysis, and demonstrated the steps needed to avoid it. In contrast to proposals from computer vision that shortcut learning can be avoided by simply testing model generalisation out-of-distribution,[15] we find that it is also essential to analyse the morphological features used by models to solve tasks to ensure that their performance is not exploiting a biologically trivial and ultimately deceptive visual strategy.

There is now evidence that DNNs can solve histopathological tasks that expert pathologists cannot. One foundational example of this trend are recent demonstrations that DNNs can learn molecular profiling from morphology.[9,11,12] These models could potentially revolutionise oncology, providing rapid accurate prognoses that replace the expensive and time-consuming molecular panels that are standard in the field, and opening up new vistas for studying cell–cell interactions in the development of cancers. However, on further examination of these findings, we see evidence that DNNs learn molecular profiling by focusing on a shortcut rather than a previously unknown morphological pattern. DNNs render molecular decisions by categorising the morphological subtype of tumours, which has a weak association with genomic mutations that has been known for at least a decade.[21] Strikingly, all DNNs that we tested here produce highly similar patterns of decisions, suggesting that the problem we observed is widespread in DNNs trained on histopathological tasks using TCGA-scale data sets.

DNNs learn to rely on this shortcut because they are trained to associate all of the potentially genetically heterogeneous image patches from a single WSI with a single sequence taken from the entire WSI. When mutation predictions from these models are tested at a more granular level than the WSI, such as the LCM image patches we introduce in this work, the correlation between subtype and mutation is weakened, which causes model performance to drop to chance. DNNs learning to focus on subtype rather than a stronger morphological signal for molecular profiling is not necessarily a shortcut—subtype is not completely spurious—but understanding this visual strategy puts a low ceiling on the utility of DNNs for molecular subtyping using existing data sets and training routines. For this reason, we refer to this DNN strategy as deceptive, and significant work is needed to develop trustworthy DNNs for extracting genomic insights from morphology. We release all the images, labels and LCMs from our Moffitt Data set to support this goal.

Not all tasks that are difficult or impossible for expert pathologists are vulnerable to shortcut learning in DNNs. When training DNNs to identify the originating tissue for cancer of unknown primary[13,14]

(CUP) on the same WSIs we used for molecular profiling, we found that models learned a robust visual strategy that did not exploit shortcuts. The models learned a novel combination of nuclear and non-nuclear features to solve the task, which generalised effectively, and which will need additional experimental study to understand their relationship to the development of lung adenocarcinoma. Overall, these findings indicate that recent findings on the ability of DNNs to learn to identify the originating tissue of CUP[13,14] are trustworthy and an important line of future computational research.

Our success on tumour detection, while not novel, points to a general strategy for ensuring that DNNs are trustworthy in histopathological analysis. In that task, unlike molecular profiling or CUP, annotations are provided at the level of pixels in WSIs. Such a 'per-pixel' labelling strategy has proved successful in training DNNs for segmentation tasks on natural images and can promote visual strategies that align with human perception.[37,50] When this level of extensive model supervision is not available, because such data are expensive or difficult to gather, the explainability framework we laid out in this paper is necessary for distinguishing between trustworthy and deceptive DNN visual strategies.

Our explainability framework is especially important for continued progress on training DNNs to learn molecular profiling from H&E. It is possible that we simply need more data to build models that can do this. Indeed, data and compute scale have been essential for advancing the capabilities of DNNs in natural language processing and computer vision tasks over recent years to rival and exceed humans. However, the scale of data needed to solve the issues we face today in histopathology are unclear, and we advise researchers to adopt our explainability framework to rigorously measure progress as they 'scale up' their DNNs.

Our work shows the utility and dangers of applying deep learning models to histopathology tasks. When training DNNs for histopathological analysis, practitioners should have access to high-quality labelled data as well as explainable artificial intelligence methods to build robust and clinically useful models. Our mixed findings on the trustworthiness of recent efforts in histopathological analysis underscores this message, and indicates that the field needs a far greater emphasis on model interpretability to create automated systems that can aid biomedical research and increase the efficiency of expert pathologists in the clinic.

Acknowledgements

The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding information

National Institute of General Medical Sciences (grant numbers: U54GM115677, R35GM139628); National Institutes of Health (grant numbers: S10OD025181); Brown University; Moffitt Cancer Center.

Conflicts of interest

None of the authors have any conflicts of interest to disclose.

Data availability statement

The Moffit data set is available from the corresponding author on reasonable request. The TCGA data set is available online: https://www.cancer.gov/tcga.

References

1. Rubin R, Strayer DS, Rubin E *et al*. Rubin's pathology: clinicopathologic foundations of medicine.2008.
2. Esteva A, Kuprel B, Novoa RA *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**; 115–118.
3. Tschandl P, Rinner C, Apalla Z *et al*. Human-computer collaboration for skin cancer recognition. *Nat. Med.* 2020; **26**; 1229–1234.
4. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit. Med.* 2019; **2**; 92.
5. Gulshan V, Peng L, Coram M *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316**; 2402–2410.
6. Ting DSW, Cheung CY-L, Lim G *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; **318**; 2211–2223.
7. Lotter W, Diab AR, Haslam B *et al*. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Med.* 2021; **27**; 244–249.
8. McKinney SM, Sieniek M, Godbole V *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577**; 89–94.
9. Coudray N, Ocampo PS, Sakellaropoulos T *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 2018; **24**; 1559–1567.
10. Coudray N, Tsirigos A. Deep learning links histology, molecular signatures and prognosis in cancer. *Nat. Cancer* 2020; **1**; 755–757.

11. Kather JN, Heij LR, Grabsch HI *et al*. Pan-cancer image based detection of clinically actionable genetic alterations. *Nat. Cancer* 2020; **1**; 789–799.
12. Fu Y, Jung AW, Torne RV *et al*. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* 2020; **1**; 800–810.
13. Lu MY, Chen TY, Williamson DFK *et al*. AI-based pathology predicts origins for cancers of unknown primary. *Nature* 2021; **594**; 106–110.
14. Jiao W, Atwal G, Polak P *et al*. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* 2020; **11**; 728.
15. Geirhos R, Jacobsen J-H, Michaelis C *et al*. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2020; **2**; 665–673.
16. Lapuschkin S, W¨aldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* 2019; **10**; 1096.
17. Pfungst O, Rahn L. *Clever hans (the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. New York: Holt, Rinehart and Winston, 1911.
18. Diao JA, Wang JK, Chui WF *et al*. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.* 2021; **12**; 1–15.
19. Mobadersany P, Yousefi S, Amgad M *et al*. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA* 2018; **115**; 2970–2979.
20. Saltz J, Gupta R, Hou L *et al*. Spatial organization and molecular correlation of tumorInfiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* 2018; **23**; 181–1937.
21. Rekhtman N, Ang DC, Riely GJ, Ladanyi M, Moreira AL. KRAS mutations are associated with solid growth pattern and tumor-infiltrating leukocytes in lung adenocarcinoma. *Mod. Pathol.* 2013; **26**; 1307–1319.
22. Kather JN, Pearson AT, Halama N *et al*. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 2019; **25**; 1054–1056.
23. Espina V, Wulfkuhle JD, Calvert VS *et al*. Laser-capture microdissection. *Nat. Protoc.* 2006; **1**; 586–603.
24. Knepper TC, Bell GC, Hicks JK *et al*. Key lessons learned from moffitt's molecular tumor board: the clinical genomics action committee experience. *Oncologist* 2017; **22**; 144–151.
25. Ellrott K, Bailey MH, Saksena G *et al*. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 2018; **6**; 271–2817.
26. Macenko M, Niethammer M, Marron JS *et al*. *A method for normalizing histology slides for quantitative analysis*. Boston, MA: IEEE, 2009; 1107–1110.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015. Available at: https://arxiv.org/abs/1512.03385 [cs.CV].
28. Ma N, Zhang X, Zheng H-T, Sun J. ShuffleNet v2: practical guidelines for efficient CNN architecture design. 2018. Available at: https://arxiv.org/abs/1807.11164 [cs.CV].
29. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. *ImageNet: a large-scale hierarchical image database*. Miami, FL: IEEE, 2009; 248–255.
30. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* preprint arXiv:1412.6980. 2014.
31. Selvaraju RR, Cogswell M, Das A *et al*. *Grad-CAM: visual explanations from deep networks via Gradient-Based localization*. Venice: IEEE, 2017; 618–626.
32. Gamper J, Koohbanani NA, Benes K *et al*. PanNuke dataset extension, insights and baselines. 2020. Available at: https://arxiv.org/abs/2003.10778 [eess.IV].
33. Lin T-Y, Maire M, Belongie S *et al*. Microsoft COCO: common objects in context. 2014. Available at: https://arxiv.org/abs/1405.0312 [cs.CV].
34. Boyle TA, Mondal AK, Saeed-Vafa D *et al*. Guideline-adherent clinical validation of a comprehensive 170-gene DNA/RNA panel for determination of small variants, copy number variations, splice variants, and fusions on a next-generation sequencing platform in the CLIA setting. *Front. Genet.* 2021; **12**; 503830.
35. Linsley D, Eberhardt S, Sharma T, Gupta P, Serre T. What are the visual features underlying human versus machine vision? In *2017 IEEE International Conference on Computer Vision Workshops*. Venice: IEEE, 2017; 2706–2714.
36. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. 2013. Available at: https://arxiv.org/abs/1311.2901 [cs.CV].
37. Linsley D, Shiebler D, Eberhardt S, Serre T. Learning what and where to attend. International Conference on Learning Representations. 2020.
38. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. 2013. Available at: https://arxiv.org/abs/1312.6034 [cs.CV].
39. Linsley JW, Linsley DA, Lamstein J *et al*. Superhuman cell death detection with biomarkeroptimized neural networks. *Sci. Adv.* 2021; **7**; 8142.
40. Graham S, Vu QD, Raza SEA *et al*. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* 2019; **58**; 101563.
41. Kumar V, Abbas AK, Aster JC. *Robbins basic pathology E-book*. New York: Elsevier Health Sciences, 2017.
42. Noorbakhsh J, Farahmand S, Foroughi Pour A *et al*. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat. Commun.* 2020; **11**; 6367.
43. Edgington ES. Randomization tests. *J. Psychol.* 1964; **57**; 445–449.
44. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* 2019; **9**; 3358.
45. Gertych A, Swiderska-Chadaj Z, Ma Z *et al*. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci. Rep.* 2019; **9**; 1483.
46. Travis WD, Brambilla E, Noguchi M *et al*. Diagnosis of lung cancer in small biopsies and cytology: implications of the 2011 International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society Classification. *Arch. Pathol. Lab Med.* 2013; **137**; 668–684.
47. Campanella G, Hanna MG, Geneslaw L *et al*. Clinical grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* 2019; **25**; 1301–1309.
48. Lee K, Zung J, Li P *et al*. Superhuman accuracy on the SNEMI3D connectomics challenge. 2017. Available at: https://arxiv.org/abs/1706.00120 [cs.CV].
49. DeGrave AJ, Janizek J, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* 2021; **3**; 610–619.
50. Linsley D, Kim J, Ashok A, Serre T. Recurrent neural circuits for contour detection. International Conference on Learning Representations. 2020.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Summary of performance at the slide and patch level for detecting KRAS mutations in lung adenocarcinoma.

Table S1. Model decisions are correlated for the KRAS detection task.