nature medicine

ARTICLES

https://doi.org/10.1038/s41591-018-0177-5

Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning

Nicolas Coudray[1,2,9], Paolo Santiago Ocampo[3,9], Theodore Sakellaropoulos[4], Navneet Narula[3], Matija Snuderl[3], David Fenyö[5,6], Andre L. Moreira[3,7], Narges Razavian[8*] and Aristotelis Tsirigos[1,3*]

Visual inspection of histopathology slides is one of the main methods used by pathologists to assess the stage, type and subtype of lung tumors. Adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) are the most prevalent subtypes of lung cancer, and their distinction requires visual inspection by an experienced pathologist. In this study, we trained a deep convolutional neural network (inception v3) on whole-slide images obtained from The Cancer Genome Atlas to accurately and automatically classify them into LUAD, LUSC or normal lung tissue. The performance of our method is comparable to that of pathologists, with an average area under the curve (AUC) of 0.97. Our model was validated on independent datasets of frozen tissues, formalin-fixed paraffin-embedded tissues and biopsies. Furthermore, we trained the network to predict the ten most commonly mutated genes in LUAD. We found that six of them—STK11, EGFR, FAT1, SETBP1, KRAS and TP53—can be predicted from pathology images, with AUCs from 0.733 to 0.856 as measured on a held-out population. These findings suggest that deep-learning models can assist pathologists in the detection of cancer subtype or gene mutations. Our approach can be applied to any cancer type, and the code is available at https://github.com/ncoudray/DeepPATH.

According to the American Cancer Society and the Cancer Statistics Center (see URLs), over 150,000 patients with lung cancer succumb to the disease each year (154,050 expected for 2018), while another 200,000 new cases are diagnosed on a yearly basis (234,030 expected for 2018). It is one of the most widely spread cancers in the world because of not only smoking, but also exposure to toxic chemicals like radon, asbestos and arsenic. LUAD and LUSC are the two most prevalent types of non–small cell lung cancer[1], and each is associated with discrete treatment guidelines. In the absence of definitive histologic features, this important distinction can be challenging and time-consuming, and requires confirmatory immunohistochemical stains.

Classification of lung cancer type is a key diagnostic process because the available treatment options, including conventional chemotherapy and, more recently, targeted therapies, differ for LUAD and LUSC[2]. Also, a LUAD diagnosis will prompt the search for molecular biomarkers and sensitizing mutations and thus has a great impact on treatment options[3,4]. For example, epidermal growth factor receptor (*EGFR*) mutations, present in about 20% of LUAD, and anaplastic lymphoma receptor tyrosine kinase (*ALK*) rearrangements, present in < 5% of LUAD[5], currently have targeted therapies approved by the Food and Drug Administration (FDA)[6,7]. Mutations in other genes, such as *KRAS* and tumor protein P53 (*TP53*) are very common (about 25% and 50%, respectively) but have proven to be particularly challenging drug targets so far[5,8]. Lung biopsies are typically used to diagnose lung cancer type and stage. Virtual microscopy of stained images of tissues is typically acquired at magnifications of 20× to 40×, generating very large two-dimensional images (10,000 to > 100,000 pixels in each dimension) that are oftentimes challenging to visually inspect in an exhaustive manner. Furthermore, accurate interpretation can be difficult, and the distinction between LUAD and LUSC is not always clear, particularly in poorly differentiated tumors; in this case, ancillary studies are recommended for accurate classification[9,10]. To assist experts, automatic analysis of lung cancer whole-slide images has been recently studied to predict survival outcomes[11] and classification[12]. For the latter, Yu et al.[12] combined conventional thresholding and image processing techniques with machine-learning methods, such as random forest classifiers, support vector machines (SVM) or Naive Bayes classifiers, achieving an AUC of ~0.85 in distinguishing normal from tumor slides, and ~0.75 in distinguishing LUAD from LUSC slides. More recently, deep learning was used for the classification of breast, bladder and lung tumors, achieving an AUC of 0.83 in classification of lung tumor types on tumor slides from The Cancer Genome Atlas (TCGA)[13]. Analysis of plasma DNA values was also shown to be a good predictor of the presence of non–small cell cancer, with an AUC of ~0.94 (ref. [14]) in distinguishing LUAD from LUSC, whereas the use of immunochemical markers yields an AUC of ~0.941[15].

Here, we demonstrate how the field can further benefit from deep learning by presenting a strategy based on convolutional neural networks (CNNs) that not only outperforms methods in previously

[1]Applied Bioinformatics Laboratories, New York University School of Medicine, New York, NY, USA. [2]Skirball Institute, Department of Cell Biology, New York University School of Medicine, New York, NY, USA. [3]Department of Pathology, New York University School of Medicine, New York, NY, USA. [4]School of Mechanical Engineering, National Technical University of Athens, Zografou, Greece. [5]Institute for Systems Genetics, New York University School of Medicine, New York, NY, USA. [6]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA. [7]Center for Biospecimen Research and Development, New York University, New York, NY, USA. [8]Department of Population Health and the Center for Healthcare Innovation and Delivery Science, New York University School of Medicine, New York, NY, USA. [9]These authors contributed equally to this work: Nicolas Coudray, Paolo Santiago Ocampo. *e-mail: narges.razavian@nyumc.org; aristotelis.tsirigos@nyumc.org

published work, but also achieves accuracies that are comparable to pathologists. Most importantly, our models maintain their performance when tested on independent datasets of both frozen and formalin-fixed paraffin-embedded (FFPE) tissues as well as on images obtained from biopsies. The development of new, inexpensive and more powerful technologies (in particular, graphics processing units (GPUs)) has made possible the training of larger and more complex neural networks[16,17]. This has resulted in the design of several deep CNNs that are capable of accomplishing complex visual recognition tasks. Such algorithms have already been successfully used for segmentation[18] or classification of medical images[19] and more specifically for whole-slide image applications such as nuclei detection[20], renal tissue segmentation[21] and glomeruli localization[22], breast cancer diagnosis[23,24], colon tumor analysis[25], glioma grading in brain tumors[26], epithelial tissue identification in prostate cancer[27] and osteosarcoma diagnosis[28].

CNNs have also been studied with regard to classification of lung patterns on computerized tomography (CT) scans, achieving an f-score of ~85.5% (ref. [29]). To study the automatic classification of lung cancer whole-slide images, we used the inception v3 architecture[30] and whole-slide images of hematoxylin and eosin (H&E)-stained lung tissue from TCGA obtained by surgical excision followed by frozen section preparation. In 2014, Google won the ImageNet Large-Scale Visual Recognition Challenge by developing the GoogleNet architecture[31], which increased the robustness to translation and nonlinear learning abilities by using microarchitecture units called inception. Each inception unit includes several nonlinear convolution modules at various resolutions. Inception architecture is particularly useful for processing the data in multiple resolutions, a feature that makes this architecture suitable for pathology tasks. This network has already been successfully adapted to other specific types of classifications like skin cancers[32] and diabetic retinopathy detection[33].

Results

A deep-learning framework for the automatic analysis of histopathology images. The purpose of this study was to develop a deep-learning model for the automatic analysis of tumor slides using publicly available whole-slide images available in TCGA[34] and to subsequently test our models on independent cohorts collected at our institution. The TCGA dataset characteristics and our overall computational strategy are summarized in Fig. 1 (Methods). We used 1,634 whole-slide images from the Genomic Data Commons database: 1,176 tumor tissues and 459 normal tissues (Fig. 1a). The 1,634 whole-slide images were split into three sets: training, validation and testing (Fig. 1b). Importantly, this ensures that our model is never trained and tested on tiles (see below) obtained from the same tumor sample. Because the sizes of the whole-slide images are too large to be used as direct input to a neural network (Fig. 1c), the network was instead trained, validated and tested using 512×512 pixel tiles, obtained from nonoverlapping 'patches' of the whole-slide images. This resulted in tens to thousands of tiles per slide, depending on the original size (Fig. 1d).

Based on the computational strategy outlined in Fig. 1, we present two main results. First, we develop classification models that classify whole-slide images into normal lung, LUAD or LUSC with an accuracy that is significantly higher than previous work (AUC of 0.97 compared to 0.75 (ref. [12]) and 0.83 (ref. [13])) and comparable to results from pathologists. Unlike previous work[12,13], the performance of our classification models was tested on several independent datasets: biopsies and surgical resection specimens either prepared as frozen sections or as FFPE tissue sections. Second, starting with the LUAD regions as predicted by the LUAD versus LUSC versus normal classification model, we utilize the same computational pipeline (Fig. 1) to train a new model in order to predict the mutational status of frequently mutated genes in lung adenocarcinoma using whole-slide images as the only input. The entire workflow of our computational analysis is summarized in Supplementary Fig. 1.

Deep-learning models generate accurate diagnosis of lung histopathology images. Using the computational pipeline of Fig. 1, we first trained inception v3 to recognize tumor versus normal. To assess the accuracy on the test set, the per-tile classification results were aggregated on a per-slide basis either by averaging the probabilities obtained on each tile or by counting the percentage of tiles positively classified, thus generating a per-slide classification (Methods). The former approach yielded an AUC of 0.990, and the latter yielded 0.993 (Supplementary Fig. 2a and Supplementary Table 1) for normal-versus-tumor classification, outperforming the AUC of ~0.85 achieved by the feature-based approach of Yu et al.[12], of ~0.94 achieved by plasma DNA analysis[14] and comparable or better than molecular profiling data (Supplementary Table 2).

Next, we tested the performance of our approach on the more challenging task of distinguishing LUAD and LUSC. To do this, we first tested whether convolutional neural networks can outperform the published feature-based approach, even when plain transfer learning is used. For this purpose, the values of the last layer of inception v3—previously trained on the ImageNet dataset to identify 1,000 different classes—were initialized randomly and then trained for our classification task. After aggregating the statistics on a per-slide basis (Supplementary Fig. 2b), this process resulted in an AUC of 0.847 (Supplementary Table 1); i.e., a gain of ~0.1 in AUC compared to the best results obtained by Yu et al.[12] using image features combined with random forest classifier. The performance can be further improved by fully training inception v3, leading to an AUC of 0.950 when the aggregation is done by averaging the per-tile probabilities (Supplementary Fig. 2c). These AUC values are improved by another 0.002 when the tiles previously classified as 'normal' by the first classifier are not included in the aggregation process (Supplementary Table 1).

We further evaluated the performance of the deep-learning model by training and testing the network on a direct three-way classification into the three types of images (normal, LUAD, LUSC). Such an approach resulted in the highest performance with all the AUCs improved to at least 0.968 (Supplementary Fig. 2d and Supplementary Table 1). In addition to working with tiles at 20× magnification, we investigated the impact of the magnification and field of view of the tiles on the performance of our models. As low-resolution features (nests of cells, circular patterns) may also be useful for classification of lung cancer type, we used slides showing a larger field of view to train the model by creating 512-×512-pixel tiles of images at 5× magnification. The binary and three-way networks trained on such slides led to similar results (Supplementary Fig. 2e,f and Supplementary Table 1). Supplementary Fig. 2g,h and Supplementary Table 2 summarize and compare the performance of the different approaches explored in this study and in previous work.

Comparison of deep-learning model to pathologists. We then asked three pathologists (two thoracic pathologists and one anatomic pathologist) to independently classify the whole-slide H&E images in the test set by visual inspection alone, independently of the classification provided by TCGA. Overall, the performance of our models was comparable to that of each pathologist (Supplementary Fig. 2b–f, pink cross). Supplementary Fig. 2i shows that 152 slides in our test set have a true positive probability above 0.5 (according to our model), and for 18 slides, this probability is below 0.5. Of the slides that were incorrectly classified by our model, 50% were also misclassified by at least one of the pathologists, whereas 83% of those incorrectly classified by at least one of the pathologists (45 out of 54) were correctly classified by the algorithm. We then measured the agreement between the TCGA classification

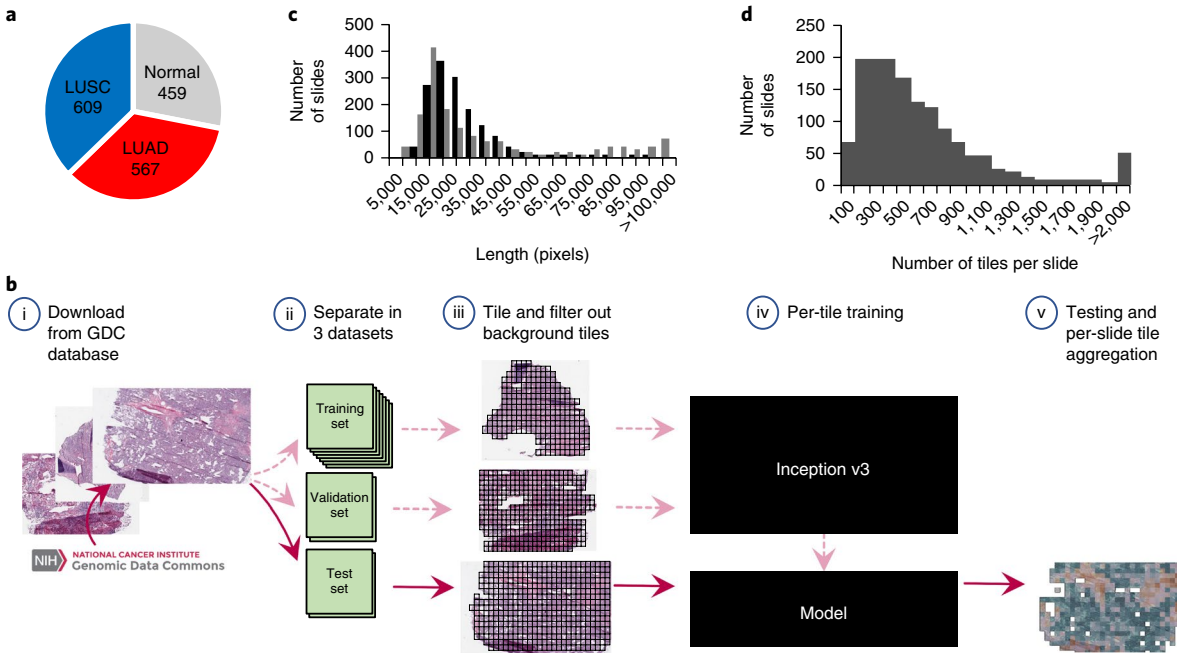

Fig. 1 | Data and strategy. a, Number of whole-slide images per class. **b**, Strategy for training. (**b**, **i**), Images of lung cancer tissues were first downloaded from the Genomic Data Commons database; (**b**, **ii**), slides were then separated into a training (70%), a validation (15%) and a test set (15%); (**b**, **iii**), slides were tiled by nonoverlapping 512- × 512-pixel windows, omitting those with over 50% background; (**b**, **iv**), the Inception v3 architecture was used and partially or fully retrained using the training and validation tiles; (**b**, **v**), classifications were performed on tiles from an independent test set, and the results were finally aggregated per slide to extract the heatmaps and the AUC statistics. **c**, Size distribution of the images widths (gray) and heights (black). **d**, Distribution of the number of tiles per slide.

and that of each pathologist, of their consensus and finally of our deep-learning model (with an optimal threshold leading to sensitivity and specificity of 89% and 93%) using Cohen's Kappa statistic (Supplementary Table 3). We observed that the agreement of the deep-learning model with TCGA was slightly higher (0.82 versus 0.67 for pathologist 1, 0.70 for pathologist 2, 0.70 for pathologist 3, and 0.78 for the consensus) but did not reach statistical significance (P values of 0.035, 0.091, 0.090 and 0.549, respectively, estimated by a two-sample two-tailed z-test score). Regarding time effort, it can take a pathologist one to several minutes to analyze a slide depending on the difficulty of distinguishing each case.

Furthermore, in the absence of definitive histologic features, confirmatory immunohistochemical stains are required and can delay diagnosis for up to 24 h. The processing time of a slide by our algorithm depends on its size; currently, it takes ~20 s to calculate per-tile classification probabilities on 500 tiles (the median number of tiles per slide is < 500) on a single Tesla K20m GPU. Considering the possibility of using multiple GPUs to process tiles in parallel, classification using our model can be executed in a few seconds. The scanning time of each slide using the Aperio scanner (Leica) is currently 2–2.5 min for a slide at 20×, but with the 2017 FDA approval of the new ultra-fast digital pathology scanner from Philipps[35], this step will probably no longer be bottleneck in the near future.

Testing on independent cohorts demonstrates generalizability of the neural network model. The model was then evaluated on independent datasets of lung cancer whole-slide images taken from frozen sections (98 slides) and FFPE sections (140 slides) as well as lung biopsies (102 slides) obtained at the New York University (NYU) Langone Medical Center (Fig. 2a–c). In this case, the pathologists' diagnosis, based on morphology and supplemented by immunohistochemical stains (TTF-1 and p40 for LUAD and LUSC respectively) when necessary, was used as the gold standard (i.e., used as a ground truth to assess the performance of our approach). Each TCGA image is almost exclusively composed of either LUAD cells, LUSC cells, or normal lung tissue. As a result, several images in the two new datasets contain features that the network has not been trained to recognize, making the classification task more challenging. We observed that features, including blood clot, blood vessels, inflammation, necrotic regions, and regions of collapsed lung are sometimes labeled as LUAD, bronchial cartilage is sometimes labeled as LUSC, and fibrotic scars can be misclassified as normal or LUAD.

As demonstrated in Supplementary Fig. 3a, TCGA images have significantly higher tumor content compared to the independent datasets, and tumor content correlates with the ability of the algorithm to generalize on these new unseen samples. To reduce the bias generated by some of these particular features that are found outside the tumor areas and only test the ability of our network to dissociate LUAD, LUSC and normal tissues regions, the AUCs in Fig. 2 were computed on regions of high tumor content that were manually selected by a pathologist. Considering that new types of artifacts were also observed on some older slides (dull staining, uneven staining, air bubbles under the slide cover leading to possible distortion), the results obtained on these independent cohorts are very encouraging. At 20× magnification, more tiles are fully covered by some of these 'unknown' features, whereas at 5× magnification, the field of view is larger and contains features known by the classified (tumor or normal cells) in many more tiles, allowing a more accurate per-tile classification. This, in turn, leads to a more accurate per-slide classification.

Taken together, these observations may explain why the AUC of the classifier on 5×-magnified tiles is mostly higher than the one from 20×-magnified tiles. Interestingly, even though the slides from FFPE and biopsy sections were preserved using a different technique from those in the TCGA database, the performance remains satisfactory (Fig. 2b). For the biopsies, we noticed that poor performance was not only associated with regions where

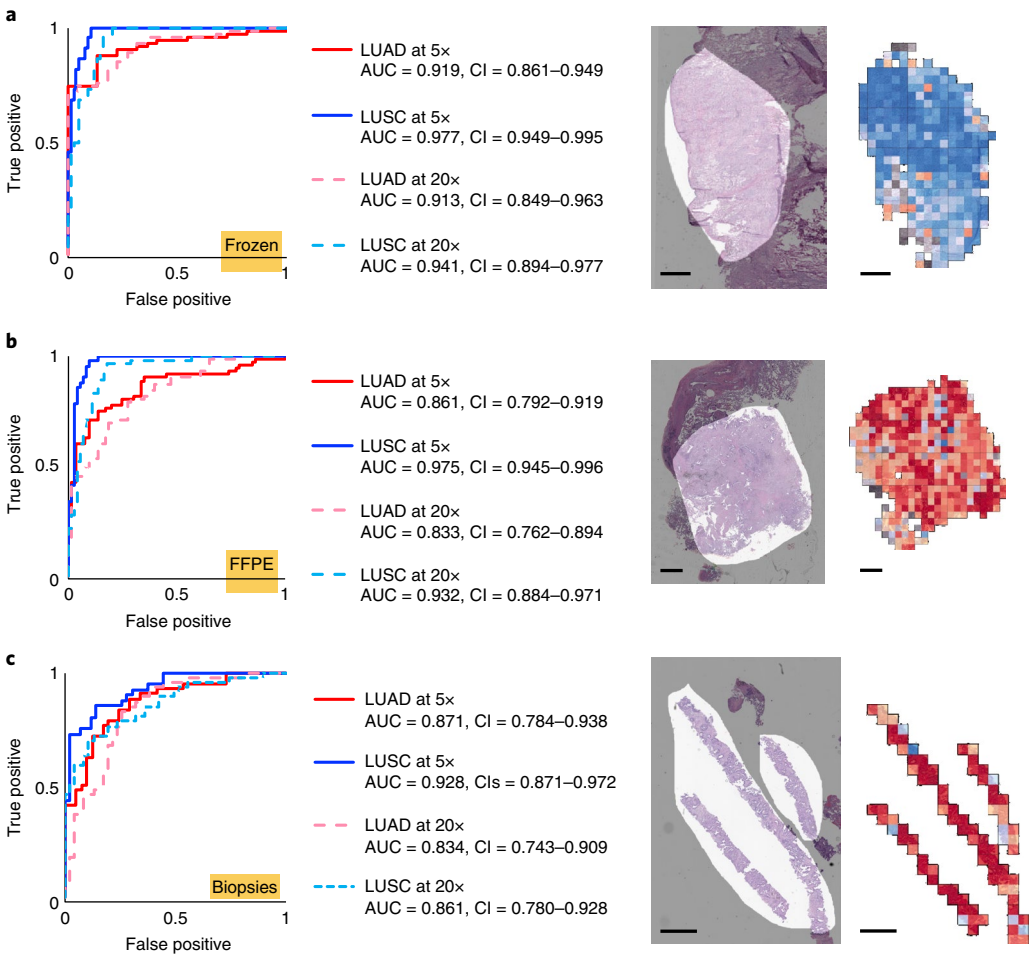

Fig. 2 | Classification of presence and type of tumor on alternative cohorts. a–c, Receiver operating characteristic (ROC) curves (left) from tests on frozen sections ($n=98$ biologically independent slides) (**a**), FFPE sections ($n=140$ biologically independent slides) (**b**) and biopsies ($n=102$ biologically independent slides) from NYU Langone Medical Center (**c**). On the right of each plot, we show examples of raw images with an overlap in light gray of the mask generated by a pathologist and the corresponding heatmaps obtained with the three-way classifier. Scale bars, 1 mm.

fibrosis, inflammation or blood was also present, but also in very poorly differentiated tumors. Sections obtained from biopsies are usually much smaller, which reduces the number of tiles per slide, but the performance of our model remains consistent for the 102 samples tested (AUC ~0.834–0.861 using 20× magnification and 0.871–0.928 using 5× magnification; Fig. 2c), and the accuracy of the classification does not correlate with the sample size or the size of the area selected by our pathologist (Supplementary Fig. 4; $r^2=9.5\times10^{-5}$). In one-third of the cases collected, the original diagnosing pathologist was not able to visually determine the tumor type; TTF-1 and p40 stains were therefore used to identify LUAD and LUSC cases, respectively. Notably, when splitting the dataset, we noticed that our model is able to classify those difficult cases as well: at 20×, the LUAD and LUSC AUCs for those difficult cases were 0.809 (confidence interval (CI), 0.639–0.940) and 0.822 (CI, 0.658–0.951), respectively, which is only slightly lower than the slides considered obvious for the pathologists (for LUAD, AUC of 0.869 (CI, 0.753–0.961) and for LUSC, 0.883 (CI, 0.777–0.962)).

Finally, we tested whether it is possible to replace the manual tumor selection process by an automatic computational selection. To this end, we trained inception v3 to recognize tumor areas using the pathologist's manual selections. Training and validation were done on two out of the three datasets, and testing was performed on the third one. For example, to test the performance of the tumor-selection model on the biopsies, we trained the model to recognize the tumor area on the frozen and FFPE samples, then applied this model to the biopsies and finally applied the TCGA-trained three-way classifier on the tumor area selected by the automatic tumor selection model. The per-tile AUC of the automatic tumor selection model (using the pathologist's tumor selection as reference) was 0.886 (CI, 0.880–0.891) for the biopsies, 0.797 (CI, 0.795–0.800) for the frozen samples, and 0.852 (CI, 0.808–0.895) for the FFPE samples. As demonstrated in Supplementary Fig. 3a (right-most bar of each graph), we observed that the automatic selection resulted in a performance that was comparable to the manual selection (slightly better AUC in frozen, no difference in FFPE, and slightly worse in biopsies; see also Supplementary Fig. 3b).

Predicting gene mutational status from whole-slide images. We next focused on the LUAD slides and tested whether CNNs can be trained to predict gene mutations using images as the only input. For this purpose, gene mutation data for matched patient samples were downloaded from TCGA. To make sure the training and test sets contained enough images from the mutated genes, we only selected those which were mutated in at least 10% of the available tumors. From each LUAD slide, only tiles classified as LUAD by our classification model were used for this task in order to avoid biasing the network to learn LUAD-specific versus LUSC-specific mutations and to focus instead on distinguishing mutations relying exclusively on LUAD tiles. Inception v3 was modified to allow

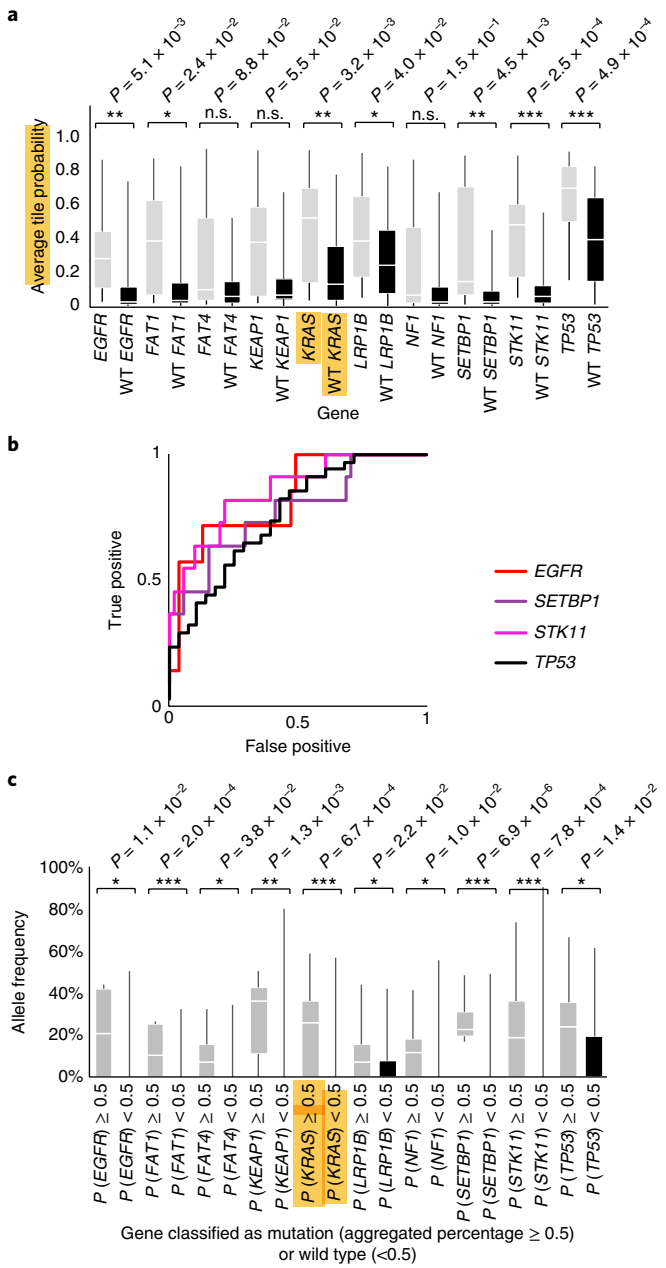

Fig. 3 | Gene mutation prediction from histopathology slides give promising results for at least six genes. a, Distribution of probability of mutation in genes from slides where each mutation is present or absent (tile aggregation by averaging output probability). **b**, ROC curves associated with the top four predictions in **a**. **c**, Allele frequency as a function of slides classified by the deep-learning network as having a certain gene mutation ($P \geq 0.5$) or the wild type ($P < 0.5$). *P* values were estimated with the two-tailed Mann–Whitney U-test and are shown as nonsignificant (n.s.; $P > 0.05$), $^*P \leq 0.05$, $^{**}P \leq 0.01$ or $^{***}P \leq 0.001$. For **a**, **b** and **c**, $n = 62$ slides from 59 patients. For the two box plots, whiskers represent the minima and maxima. The middle line within the box represents the median.

Table 1 | AUC achieved by the network trained on mutations (with 95% CIs)

Mutations	Per-tile AUC	Per-slide AUC after aggregation by…	
		… average predicted probability	… percentage of positively classified tiles
STK11	0.845 (0.838-0.852)	0.856 (0.709-0.964)	0.842 (0.683-0.967)
EGFR	0.754 (0.746-0.761)	0.826 (0.628-0.979)	0.782 (0.516-0.979)
SETBP1	0.785 (0.776-0.794)	0.775 (0.595-0.931)	0.752 (0.550-0.927)
TP53	0.674 (0.666-0.681)	0.760 (0.626-0.872)	0.754 (0.627-0.870)
FAT1	0.739 (0.732-0.746)	0.750 (0.512-0.940)	0.750 (0.491-0.946)
KRAS	0.814 (0.807-0.829)	0.733 (0.580-0.857)	0.716 (0.552-0.854)
KEAP1	0.684 (0.670-0.694)	0.675 (0.466-0.865)	0.659 (0.440-0.856)
LRP1B	0.640 (0.633-0.647)	0.656 (0.513-0.797)	0.657 (0.512-0.799)
FAT4	0.768 (0.760-0.775)	0.642 (0.470-0.799)	0.640 (0.440-0.856)
NF1	0.714 (0.704-0.723)	0.640 (0.419-0.845)	0.632 (0.405-0.845)

$n = 62$ slides from 59 patients.

multioutput classification (Methods): training and validation was conducted on ~212,000 tiles from ~320 slides, and testing was performed on ~44,000 tiles from 62 slides. Box plot and ROC curve analysis (Fig. 3a,b and Supplementary Fig. 5) show that six frequently mutated genes seem predictable using our deep-learning approach; AUC values for serine/threonine protein kinase 11 (*STK11*), *EGFR*, FAT atypical cadherin 1 (*FAT1*), SET binding protein 1 (*SETBP1*), *KRAS* and *TP53* were between 0.733 and 0.856 (Table 1). Availability of more data for training is expected to substantially improve the performance.

As mentioned earlier, EGFR already has targeted therapies. STK11, also known as liver kinase 1 (LKB1), is a tumor suppressor inactivated in 15–30% of non–small cell lung cancers[36] and is also a potential therapeutic target: it has been reported that phenformin, a mitochondrial inhibitor, increases survival in mice[37]. Also, it has been shown that *STK11* mutations in combination with *KRAS* mutations resulted in more aggressive tumors[38]. FAT1 is an ortholog of the *Drosophila* fat gene involved in many types of cancers, and its inactivation is suspected to increase cancer cell growth[39]. Mutation of the tumor suppressor gene *TP53* is thought to lead to resistance to chemotherapy, resulting in lower survival rates in small-cell lung cancers[40]. Mutation in *SETBP1*, like *KEAP1* and *STK11*, has been identified as one of the signatures of LUAD[41].

Finally, for each gene, we compared the classification achieved by our deep-learning model with the allele frequency (Fig. 3c). Among the gene mutations predicted with a high AUC, in four of them, classification probabilities (as reported by our model) were associated with allele frequency: *FAT1*, *KRAS*, *SETBP1*, and *STK11*, demonstrating that these probabilities may reflect the percentage of cells

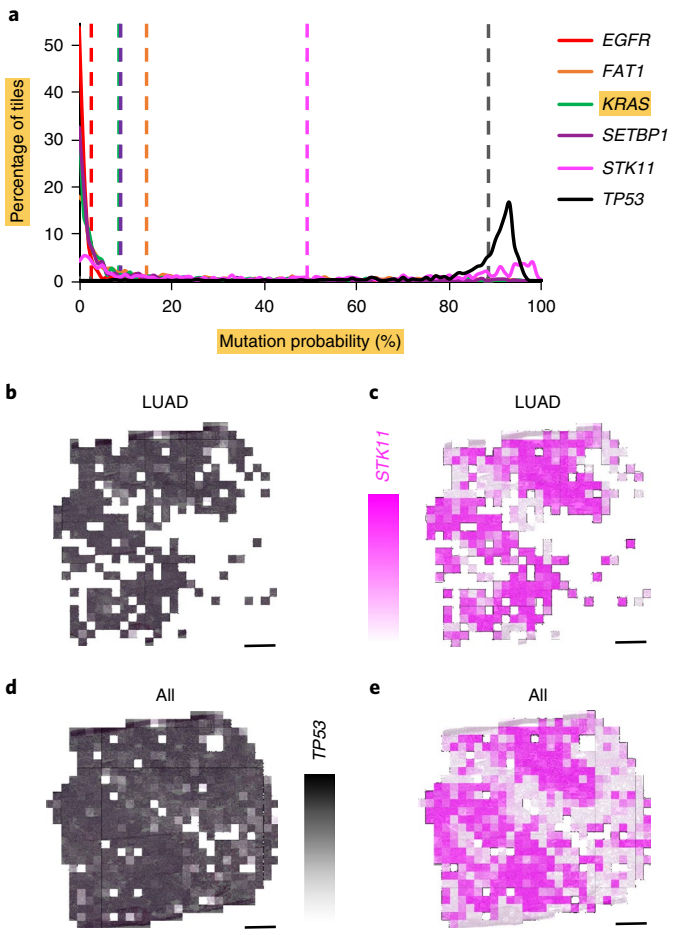

Fig. 4 | Spatial heterogeneity of predicted mutations. a, Probability distribution on LUAD tiles for the six predictable mutations, with average values in dotted lines ($n = 327$ nonoverlapping tiles). The allele frequency is 0.33 for *TP53*, 0.25 for *STK11*, and 0 for the four other mutations. **b**–**e**, Heatmaps of *TP53* (**b**,**d**) and *STK11* (**c**,**e**) when only tiles classified as LUAD are selected (**b**,**c**), and when all the tiles are considered (**d**,**e**). Scale bars, 1 mm.

effectively affected by the mutation. Looking, for example, at the predictions performed on the whole-slide image from Fig. 4a, our process successfully identified *TP53* (allele frequency of 0.33) and *STK11* (allele frequency of 0.25) as the two genes that were most likely mutated (Fig. 4a). The heatmap shows that almost all of the LUAD tiles are highly predicted to show TP53-mutant-like features (Fig. 4b), and two major regions with STK11-mutant-like features (Fig. 4c). Interestingly, when the classification is applied on all tiles, it shows that even tiles classified as LUSC present *TP53* mutations (Fig. 4d), whereas the STK11 mutant is confined to the LUAD tiles (Fig. 4e). These results are realistic considering that, as mentioned earlier, *STK11* mutation is a signature of LUAD[41], whereas *TP53* mutation is more common in all human cancers.

Future work on deep-leaning model visualization tools[42] would help identify and characterize the features used by the neural network. To visualize how the mutations and tiles are organized in the multidimensional space of the network, we used as before a t-SNE representation[43] with the values of the last fully connected layer used an inputs. On the resulting plots (Supplementary Fig. 6a), each dot represents a tile, and its color is proportional to the probability of the gene to be mutated, as estimated by our model. The tile-embedded representation (Supplementary Fig. 6b) allows the visual comparison of tiles sharing similar predicted mutations. Clusters of specific mutations can be seen at the surroundings of the plot. The top left group, for example, shows tiles in which the aggressive double mutants KRAS and STK11 are both present, while the small one at the top shows tiles with KEAP1 and SETBP1 mutants and the cluster on the top right has been associated with the triple mutation of *FAT1*, *LRP1B*, and *TP53*. Future analysis with laser-capture microdissection could provide some additional spatial information and could study the limits and precision of such a method[44].

Although our current analysis does not define the specific features used by the network to identify mutations, our results suggest that such genotype–phenotype correlations are detectable. Determining mutation status from a histological image and bypassing additional testing is important in lung cancer in particular, as these mutations often carry prognostic as well as predictive information. Previous work has shown associations between clinically important mutations and specific patterns of lung adenocarcinoma[45,46] as well as with the histologic changes that correspond with the evolution of resistance[47]. More recently, Chiang et al.[48] empirically demonstrated the relationship between a defining mutation and the unique morphology of a breast cancer subtype. Some of the mutations with high AUCs highlighted in our study (like those in *STK11*, *TP53* and *EGFR*) have been shown to affect cell polarity and cell shape[49–51], two features that are not routinely assessed during the pathologic diagnosis. We note that our model was not able to detect *ALK* mutations, although such tumors have been associated with specific histologic features, such as a solid pattern with signet ring cells or a mucinous cribriform pattern[52,53]. Although the prevalence of *ALK* mutations is very low (reportedly ranging from 1.8–6.4% (ref. [54])), their presence is routinely determined via immunohistochemistry, as tumors with this mutation may respond to ALK inhibitors[6,7].

To confirm that our models can be applied to independent cohorts, we tested the prediction of the EGFR mutant using 63 whole-slide images of lung resection specimens with known EGFR mutational status: 29 EGFR mutant and 34 EGFR wild-type samples. This independent dataset has some important differences from the TCGA dataset, which may negatively impact the evaluation of the TCGA-based model: (i) the samples were not frozen but were instead preserved using FFPE, and (ii) only 22 samples had sequencing data to support the *EGFR* mutational status with high specificity and sensitivity; the rest of the samples (i.e., 65% of the test set) have been analyzed by immunohistochemical (IHC) stains[55], a technique known for its high specificity but low sensitivity[56,57] and which solely identifies the two most common mutations[55] (p.L858R and p.E746_A750del). On the other hand, data from the TCGA dataset used for training were identified with the next-generation sequencing (NGS) tools Illumina HiSeq 2000 or Genome Analyzer II. Our TCGA model has therefore been trained to detect not only p.L858R and p.E746_A75-del, but many other EGFR mutants and deletions, such as p.G719A, p.L861Q or p.E709_T710delinsD, for example.

Despite these caveats, we believed that it would still be important to demonstrate that our TCGA-derived models can at least perform significantly better than random in the independent NYU cohort. Indeed, the results showed an AUC of 0.687 (CI, 0.554–0.811), with a higher AUC (0.750; CI, 0.500–0.966) in samples validated by sequencing than in those tested by IHC (AUC, 0.659; CI, 0.485–0.826). Although the sequencing-based AUC of 0.75 is lower than the one estimated on the TCGA test set (0.83), we believe that most of this difference can be attributed to the difference in the sample preparation (frozen versus FFPE). We noticed that the discrepancy (~0.08) is similar to the difference observed in the AUCs of LUAD from the TCGA dataset (0.97) and the FFPE dataset (0.83). In the classification task, this issue was solved by lowering the magnification to 5×. However, this is not useful for the mutation prediction task, because it appears that 20× is necessary to capture predictive image features (the TCGA *EGFR* mutation prediction model at 5× has a random

performance). Still, we believe that the 0.75 AUC we obtained on the sequencing-validated subset of EGFR-mutant cases demonstrates that the model can generalize on independent datasets.

Discussion

Our study demonstrates that convolutional neural networks, such as Google's inception v3, can be used to assist in the diagnosis of lung cancer from histopathology slides: it almost unambiguously classifies normal versus tumor tissues (~0.99 AUC) and distinguishes lung cancer types with high accuracy (0.97 AUC), reaching sensitivity and specificity comparable to that of a pathologist. Interestingly, around half of the TCGA whole-slide images misclassified by the algorithms have also been misclassified by the pathologists, highlighting the intrinsic difficulty in distinguishing LUAD from LUSC in some cases. However, 45 out of 54 of the TCGA images misclassified by at least one of the pathologists were assigned to the correct cancer type by the algorithm, suggesting that our model could be beneficial in assisting the pathologists in their diagnosis. The confusions matrices in Supplementary Table 4 detail the discrepancies between the different classifications, and Supplementary Fig. 7 shows a few examples in which our model correctly classified whole-slide images misclassified by at least one of the pathologists. These images show poorly differentiated tumors that lack the classic histological features of either type (keratinization for LUSC and gland formation/recognizable histological pattern for LUAD). The high accuracy of our model was achieved despite the presence of various artifacts in the TCGA images that were related to sample preparation and preservation procedures.

However, the TCGA images used to train the deep neural network may not fully represent the diversity and heterogeneity of tissues that pathologists typically inspect, which may include additional features such as necrosis, blood vessels, and inflammation. More slides containing such features would be needed to retrain the network in order to further improve its performance. Despite this and the fact that the process was trained on frozen images, tests show very promising results on tumor classification from FFPE sections as well. Although it has been suggested that mutations could be predicted from H&E images (AUC of ~0.71 for the prediction of *SPOP* mutations from prostate cancer H&E images[58]), before this study, it was unclear whether gene mutations would affect the pattern of tumor cells on a lung cancer whole-slide image, but training the network using the presence or absence of mutated genes as a label revealed that there are certain genes whose mutational status can be predicted from image data alone: *EGFR*, *STK11*, *FAT1*, *SETBP1*, *KRAS*, and *TP53*. Notably, the presence of *STK11* mutations can be predicted with the highest accuracy (~0.85 AUC). A limiting factor in obtaining higher accuracies lies in the small number of slides that contain positive instances (i.e., the gene mutations) available for training; therefore, our models can greatly benefit from larger datasets that may become available in the near future. The ability to quickly and inexpensively predict both the type of cancer and the gene mutations from histopathology images could be beneficial to the treatment of patients with cancer given the importance and impact of these mutations[6,36–41].

Overall, this study demonstrates that deep-learning convolutional neural networks could be a very useful tool for assisting pathologists in their classification of whole-slide images of lung tissues. This information can be crucial in applying the appropriate and tailored targeted therapy to patients with lung cancer, increasing thereby the scope and performance of precision medicine that aims at developing a multiplex approach with patient-tailored therapies[59]. The diagnosis and therapy differ considerably between LUSC and LUAD and may depend on the mutational status of specific genes. In particular, when inspecting frozen section biopsies, pathologists only rely on morphology and may need immunostaining for the most difficult cases; our algorithm, which still achieves an AUC above 0.8 on biopsies that usually require immunostaining, can be used as an adjunct to telepathology to speed up diagnosis and classification during intraoperative consultation. As a result of advances in our understanding of lung cancer and a concomitant rise in the number and types of treatment options, the role of the pathologist in the diagnosis and management of this disease is substantially more complex than cancer type distinction and even determination of mutational status. Although our computational analyses may play a role in the initial diagnosis with the benefit of providing important prognostic information based on an H&E image alone, the pathologist has additional tasks, such as staging the tumor and, in an increasing number of cases, estimating response to treatment.

In the future, we would ideally extend the classification to other types of less common lung cancers (large-cell carcinoma, small-cell lung cancer) and histological subtypes of LUAD (acinar, lepidic, papillary, micropapillary, and solid) as well as to non-neoplastic features including necrosis, fibrosis, and other reactive changes in the tumor microenvironment, though the amount of data currently available is insufficient. We hope that by extending our algorithm to recognize a wider range of histologic features, followed by providing a quantitative and spatial assessment as in our heatmaps, we will be able to aid aspects of the pathologists' evaluation that are well-suited to automated analyses. We hope that this computational approach could play a role in both routine tasks and difficult cases (for example, distinguishing intrapulmonary metastases from multiple synchronous primary lung cancers) in order to allow the pathologist to concentrate on higher-level decisions, such as integrating histologic, molecular, and clinical information in order to guide treatment decisions for individual patients.

URLs. American Cancer Society, http://www.cancer.org/; Cancer Statistics Center http://cancerstatisticscenter.cancer.org/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0177-5.

Received: 22 November 2017; Accepted: 6 July 2018;
Published online: 17 September 2018

References

1. Travis, W. D. et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* **6**, 244–285 (2011).
2. Hanna, N. et al. Systemic therapy for stage IV non–small-cell lung cancer: American Society of Clinical Oncology clinical practice guideline update. *J. Clin. Oncol.* **35**, 3484–3515 (2017).
3. Chan, B. A. & Hughes, B. G. Targeted therapy for non–small cell lung cancer: current standards and the promise of the future. *Transl. Lung Cancer Res.* **4**, 36–54 (2015).
4. Parums, D. V. Current status of targeted therapy in non–small cell lung cancer. *Drugs Today (Barc)* **50**, 503–525 (2014).
5. Terra, S. B. et al. Molecular characterization of pulmonary sarcomatoid carcinoma: analysis of 33 cases. *Mod. Pathol.* **29**, 824–831 (2016).
6. Blumenthal, G. M. et al. Oncology drug approvals: evaluating endpoints and evidence in an era of breakthrough therapies. *Oncologist* **22**, 762–767 (2017).
7. Pérez-Soler, R. et al. Determinants of tumor response and survival with erlotinib in patients with non–small-cell lung cancer. *J. Clin. Oncol.* **22**, 3238–3247 (2004).
8. Jänne, P. A. et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. *Lancet Oncol.* **14**, 38–47 (2013).
9. Thunnissen, E., van der Oord, K. & den Bakker, M. Prognostic and predictive biomarkers in lung cancer. A review. *Virchows Arch.* **464**, 347–358 (2014).
10. Zachara-Szczakowski, S., Verdun, T. & Churg, A. Accuracy of classifying poorly differentiated non–small cell lung carcinoma biopsies with commonly used lung carcinoma markers. *Hum. Pathol.* **46**, 776–782 (2015).

11. Luo, X. et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J. Thorac. Oncol.* **12**, 501–509 (2017).
12. Yu, K.-H. et al. Predicting non–small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
13. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* **27**, 317–328 (2018).
14. Sozzi, G. et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. *J. Clin. Oncol.* **21**, 3902–3908 (2003).
15. Terry, J. et al. Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples. *Am. J. Surg. Pathol.* **34**, 1805–1811 (2010).
16. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
17. Greenspan, H., Ginneken, Bv & Summers, R. M. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* **35**, 1153–1159 (2016).
18. Qaiser, T., Tsang, Y.-W., Epstein, D. & RajpootEma, N. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In *Medical Image Understanding and Analysis: 21st Annual Conference on Medical Image Understanding and Analysis*. (Eds. Valdes Hernandez, M. & González-Castro, V.) 320–329 (Springer International Publishing, New York, 2018).
19. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
20. Xing, F., Xie, Y. & Yang, L. An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans. Med. Imaging* **35**, 550–566 (2016).
21. de Bel, T. et al. Automatic segmentation of histopathological slides of renal tissue using deep learning. In *Medical Imaging 2018: Digital Pathology* Vol. 10581 (Eds. Tomaszewski, J. E. & Gurcan, M. N.) 1058112 (International Society for Optics and Photonics, Bellingham, WA, USA, 2018).
22. Simon, O., Yacoub, R., Jain, S., Tomaszewski, J. E. & Sarder, P. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. *Sci. Rep.* **8**, 2032 (2018).
23. Cheng, J.-Z. et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* **6**, 24454 (2016).
24. Cruz-Roa, A. et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).
25. Sirinukunwattana, K. et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **35**, 1196–1206 (2016).
26. Ertosun, M. G. & Rubin., D. L. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. In *AMIA Annual Symposium Proceedings*. 1899–1908 (American Medical Informatics Association, Bethesda, MD, USA).
27. Bulten, W., Kaa, C.A.H.-d., Laak, J.d. & Litjens, G.J. Automated segmentation of epithelial tissue in prostatectomy slides using deep learning. In *Medical Imaging 2018: Digital Pathology*. Vol. 10581 (Eds. Tomaszewski, J. E. & Gurcan, M. N.) 105810S (International Society for Optics and Photonics, Bellingham, WA, USA, 2018).
28. Mishra, R., Daescu, O., Leavey, P., Rakheja, D. & Sengupta, A. Histopathological Diagnosis for Viable and Non-viable Tumor Prediction for Osteosarcoma Using Convolutional Neural Network. In *International Symposium on Bioinformatics Research and Applications* Vol. 10330 (Eds. Cai, Z., D. Ovidiu, & Li, M.) 12–23 (Springer International Publishing, New York, 2018).
29. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. Lung Pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans. Med. Imaging* **35**, 1207–1216 (2016).
30. 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 (Boston, MA, USA, 2015).
31. Szegedy, C. et al. Going Deeper With Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition*. 1–9 (Boston, 2015).
32. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
33. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
34. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
35. Abels, E. & Pantanowitz, L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. *J. Pathol. Inform.* **8**, 23 (2017).
36. Sanchez-Cespedes, M. et al. Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer Res.* **62**, 3659–3662 (2002).
37. Shackelford, D. B. et al. LKB1 inactivation dictates therapeutic response of non–small cell lung cancer to the metabolism drug phenformin. *Cancer Cell* **23**, 143–158 (2013).
38. Makowski, L. & Hayes, D. N. Role of LKB1 in lung cancer development. *Br. J. Cancer* **99**, 683–688 (2008).
39. Morris, L. G. et al. Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.* **45**, 253–261 (2013).
40. Mogi, A. & Kuwano, H. *TP53* mutations in nonsmall cell lung cancer. *J. Biomed. Biotechnol.* **2011**, 583929 (2011).
41. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
42. Zeiler, M.D. & Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. 818–833 (Springer International Publishing, New York, 2015).
43. Maaten, L. J. Pd Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
44. Bonner, R. F. et al. Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997). 1483.
45. Ninomiya, H. et al. Correlation between morphology and *EGFR* mutations in lung adenocarcinomas significance of the micropapillary pattern and the hobnail cell type. *Lung Cancer* **63**, 235–240 (2009).
46. Warth, A. et al. *EGFR*, *KRAS*, *BRAF* and *ALK* gene alterations in lung adenocarcinomas: patient outcome, interplay with morphology and immunophenotype. *Eur. Respir. J.* **43**, 872–883 (2014).
47. Sequist, L. V. et al. Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Sci. Transl. Med.* **3**, 75ra26 (2011).
48. Chiang, S. et al. *IDH2* mutations define a unique subtype of breast cancer with altered nuclear polarity. *Cancer Res.* **76**, 7118–7129 (2016).
49. Baas, A. F., Smit, L. & Clevers, H. LKB1 tumor suppressor protein: partaker in cell polarity. *Trends Cell Biol.* **14**, 312–319 (2004).
50. Gloushankova, N., Ossovskaya, V., Vasiliev, J., Chumakov, P. & Kopnin, B. Changes in p53 expression can modify cell shape of ras-transformed fibroblasts and epitheliocytes. *Oncogene* **15**, 2985–2989 (1997).
51. Yatabe, Y. et al. EGFR mutation is specific for terminal respiratory unit type adenocarcinoma. *Am. J. Surg. Pathol.* **29**, 633–639 (2005).
52. Yoshida, A. et al. Comprehensive histologic analysis of ALK-rearranged lung carcinomas. *Am. J. Surg. Pathol.* **35**, 1226–1234 (2011).
53. Rodig, S. J. et al. Unique clinicopathologic features characterize ALK-rearranged lung adenocarcinoma in the western population. *Clin. Cancer Res.* **15**, 5216–5223 (2009).
54. Dearden, S., Stevens, J., Wu, Y.-L. & Blowers, D. Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). *Ann. Oncol* **24**, 2371–2376 (2013).
55. Yu, J. et al. Mutation-specific antibodies for the detection of *EGFR* mutations in non-small-cell lung cancer. *Clin. Cancer Res.* **15**, 3023–3028 (2009).
56. Houang, M. et al. *EGFR* mutation specific immunohistochemistry is a useful adjunct which helps to identify false negative mutation testing in lung cancer. *Pathology* **46**, 501–508 (2014).
57. Dimou, A. et al. Standardization of epidermal growth factor receptor (EGFR) measurement by quantitative immunofluorescence and impact on antibody-based mutation detection in non–small cell lung cancer. *Am. J. Pathol.* **179**, 580–589 (2011).
58. Schaumberg, A. J., Rubin, M. A. & Fuchs, T. J. H&E-stainedwhole slide deep learning predicts spop mutation state in prostate cancer. Preprint at https://doi.org/10.1101/064279 (2016).
59. Donovan, M. J. et al. A systems pathology model for predicting overall survival in patients with refractory, advanced non-small-cell lung cancer treated with gefitinib. *Eur. J. Cancer* **45**, 1518–1526 (2009).

Acknowledgements

We would like to thank the Applied Bioinformatics Laboratories (ABL) at the NYU School of Medicine for providing bioinformatics support and helping with the analysis and interpretation of the data. The Applied Bioinformatics Laboratories are a Shared Resource, partially supported by the Cancer Center Support Grant, P30CA016087 (A.T.), at the Laura and Isaac Perlmutter Cancer Center (A.T.). For this work, we used computing resources at the High-Performance Computing Facility (HPC) at NYU Langone Medical Center. The slide images and the corresponding cancer information were uploaded from the Genomic Data Commons portal (https://gdc-portal.nci.nih.gov) and are in whole or in part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/). These data were publicly available without restriction, authentication or authorization necessary. We thank the GDC help desk for providing assistance and information regarding the TCGA dataset. For the independent cohorts, we only used

whole-slide images; the NYU dataset we used consists of slide images without identifiable information and therefore does not require approval according to both federal regulations and the NYU School of Medicine Institutional Review Board. For this same reason, written informed consent was not necessary. We thank C. Dickerson, from the Center for Biospecimen Research and Development (CBRD), for scanning the whole-slide images from the NYU Langone Medical Center. We also thank T. Papagiannakopoulos, H. Pass and K.-K. Wong or their valuable and constructive suggestions.

Author contributions

N.C. performed the experiments; N.C., A.T. and N.R. designed the experiments; N.C. and T.S. wrote the code to achieve different tasks; T.S. gathered the mutation information and contributed to their analysis; M.S. helped identify cases validated by next-generation sequencing; A.L.M. and P.S.O. collected and labeled the independent cohorts. A.L.M, P.S.O and N.N. manually labeled the TCGA dataset; N.C., A.L.M., P.S.O., N.R. and A.T. contributed to the analysis of the data; D.F., N.R. and A.T. conceived and directed the project; N.C., A.T., N.R., A.L.M. and P.S.O. wrote the manuscript with the assistance and feedback of all the other co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0177-5.

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

Correspondence and requests for materials should be addressed to N.R. or A.T.

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

Methods

TCGA lung cancer whole-slide image dataset. Our dataset comes from the NCI Genomic Data Commons[34], which provides the research community with an online platform for uploading, searching, viewing and downloading cancer-related data. All freely available slide images of Lung cancer were uploaded from this source. We studied the automatic classification of 'solid tissue normal' and 'primary tumor' slides using a set of 459 and 1,175, respectively, H&E- stained histopathology whole-slide images[60,61]. Then, the 'primary tumor' images were classified between LUAD and LUSC types using a set of 567 and 608, respectively, of those whole-slide images. The labels provided by the TCGA database were used as our gold standard. Those labels were the result of a consensus as explained by the GDC data curator (personal communication): first, the submitting institutions were asked to review each sample prior sending it to confirm the diagnosis. Then, a slide from the sample was reviewed by a TCGA contracted expert thoracic pathologist. In the event of a disagreement, the slide would be reviewed by one or more other expert thoracic pathologists. Out of the 170 slide images in our test set, only 1 image was tagged as leading to inconsistent labels (and about 30 images had no information about it).

Image preprocessing generating 987,931 tiles. The slides were tiled in non-overlapping 512×512 pixel windows at a magnification of 20× using the openslide library67 (533 of the 2,167 slides initially uploaded were removed because of compatibility and readability issues at this stage). The slides with a low amount of information were removed; i.e., all the tiles where > 50% of the surface was covered by background (for which all the values are below 220 in the RGB color space). This process generated nearly 1,000,000 tiles. For information regarding the number of tiles and slides as well as LUAD and LUSC classification, see Supplementary Tables 5 and 6.

Deep learning with convolutional neural networks. We used 70% of those tiles for training, 15% for validation, and 15% for final testing (Supplementary Tables 5 and 6). The tiles associated with a given slide were not separated but associated as a whole to one of these sets to prevent overlaps between the three sets. Typical convolutional neural networks (CNNs) consist of several levels of convolution filters, pooling layers and fully connected layers. We based our model on inception v3 architecture36. This architecture makes use of inception modules which are made of a variety of convolutions having different kernel sizes and a max pooling layer. The initial 5 convolution nodes are combined with 2 max pooling operations and followed by 11 stacks of inception modules. The architecture ends with a fully connected and then a softmax output layer. For normal versus tumor tile classification, we fully trained the entire network. For the classification of type of cancer, we followed and compared different approaches to achieve the classification: transfer learning, which includes training only the last fully connected layer, and training the whole network. Tests were implemented using the Tensorflow library (http://tensorflow.org).

Transfer learning on inception v3. We initialized our network parameters to the best parameter set that was achieved on ImageNet competition. We then fine-tuned the parameters of the last layer of the network on our data via back propagation. The loss function was defined as the cross entropy between predicted probability and the true class labels, and we used RMSProp69 optimization, with learning rate of 0.1, weight decay of 0.9, momentum of 0.9, and epsilon of 1.0 method for training the weights. This strategy was tested for the binary classification of LUAD versus LUSC.

Training the entire inception v3 network. The inception v3 architecture was fully trained using our training datasets, following the procedure previously described[30]. Similar to transfer learning, we used back-propagation, cross entropy loss, and the RMSProp optimization method as well as the same hyperparameters as the transfer-learning case for the training. In this approach, instead of only optimizing the weights of the fully connected layer, we also optimized the parameters of previous layers, including all of the convolution filters of all layers. This strategy was tested on three classifications: normal versus tumor, LUAD versus LUSC and normal versus LUAD versus LUSC. The training jobs were run for 500,000 iterations. We computed the cross-entropy loss function on the train and validation datasets, and we used the model with best validation score as our final model. We did not tune the number of layers or hyperparameters of the inception network, such as size of filters. As this training gave the best results, we also investigated the importance of training the network on a larger field of view at the expense of a lower resolution. Whole-slide images were tiled at a magnification of 5×(keeping the tile size at 512×512 pixels) and the network was again fully trained.

Identification of gene mutations. To study the prediction of gene mutations from histopathology images, we modified the inception v3 to perform multitask classification rather than a single-task classification. Each mutation classification was treated as a binary classification, and our formulation allowed multiple mutations to be assigned to a single tile. We optimized the average of the cross entropy of each individual classifier. To implement this method, we replaced the final softmax layer of the network with a sigmoid layer, to allow each sample to be associated with several binary labels[62]. We used the RMSProp algorithm for the optimization, and fully trained this network for 500,000 iterations using only LUAD whole-slide images, each one associated with a 10-cell vector, and each cell associated to a mutation and set to 1 or 0 depending on the presence or absence of the mutation. Only the most commonly mutated genes were used (Supplementary Table 7), leading to a training set of 223,185 tiles. Training and validation were done over 500,000 iterations (Supplementary Fig. 8). The test was then achieved on the tiles, and aggregation on the $n=62$ test slides for which at least one of these mutations is present was done only if the tile was previously classified as 'LUAD' by the normal, LUAD and LUSC three-classes classifier.

Statistical analysis. Once the training phase was finished, the performance was evaluated using the testing dataset, which is composed of tiles from slides not used during the training. We then aggregated the probabilities for each slide using two methods: either average of the probabilities of the corresponding tiles, or percentage of tiles positively classified. For the binary LUAD and CLUSC classifiers, $n=170$ slides from 137 patients; and for the normal and tumor and for the three-way classifiers, $n=244$ slides from 137 patients. The ROC curves and the corresponding AUC were computed in each case[63] using the python library sklearn[64]. CIs at 95% were estimated by 1,000 iterations of the bootstrap method[65]. Tumor slides could contain a certain amount of normal tiles. Therefore, we also checked how the ROC and AUC were affected when tiles classified as normal were removed from the aggregation. We asked three pathologists to manually label the TCGA test LUAD and LUSC images and compared the agreements between the ratings using the Cohen's Kappa statistic[66,67], comparing it to the binary LUAD and LUSC deep-learning classifier using the optimal threshold of 0.4 and 0.6 (optimal threshold is here defined as the point of the ROC curve which is closest to the perfect (1,0) coordinate). Heatmaps were also generated for some tested slides to visualize the differences between the two approaches and to identify the regions associated with a certain cancer type. To analyze more thoroughly the network trained on gene mutations, we used the Barnes-Hit implementation of the t-SNE technique[43] to reduce the dimensionality and facilitate the visualization of the classes. The values associated with the last fully connected layer were used as an input, and theta was set to 0.5, perplexity to 50, and 10,000 iterations. For the LUAD and LUSC classifier, the t-SNE plot was generated using $n=149{,}790$ tiles of 244 slides from 137 patients. For the gene mutation prediction task, the t-SNE plot was generated using $n=24{,}144$ tiles of 62 slides from 59 patients. Mutation probability distributions and relationship to allele frequency were analyzed with the two-tailed Mann–Whitney U-tests and computed using the same dataset (62 slides from 59 patients).

Visualization of features identified by the three-way classifier in high-confidence tiles. In Supplementary Fig. 9, we present examples of LUSC and LUAD slides, together with heatmaps generated by our algorithm, in which the color of each tile corresponds to the class assigned by our algorithm (LUAD, LUSC or normal), and the color shade is proportional to the classification probability. The LUSC image shows most of its tiles with a strong true positive probability for LUSC classification, while in the LUAD image, the largest regions indeed have strong LUAD features, with normal cells on the side (as confirmed by our pathologist), and some light blue tiles indicating the existence of LUSC-like features in this tumor. In Supplementary Fig. 10, the values of the last fully connected layer are visualized using a t-SNE representation, which generates 2-D scatterplots of high-dimensional features[43]. For tiles associated with LUSC, we note a predominance of areas of keratinization and dyskeratotic cells as well as rare foci of cells with prominent intracellular bridging. Among the tiles denoted LUAD, the predominant feature noted is the presence of distinct gland forming histological patterns, such as lepidic and acinar (well differentiated) and micropapillary (poorly differentiated). These include well-differentiated patterns (lepidic and acinar) as well as poorly differentiated types (micropapillary). At the center of the t-SNE, regions that cannot be clearly associated with either LUAD or LUSC are composed of tiles with conspicuous preservation artifact, minute foci of tumor, or areas of interstitial/septal fibrosis. Then, the area designated as normal is composed of tiles showing benign lung parenchyma, focal fibrosis or inflammation, as well as rare LUAD with preservation artifacts. Interestingly, the area with tiles which could not be designated normal, LUAD or LUSC with high confidence, shows both benign and malignant lung tissue in a background of dense fibrosis and/or inflammation.

Tests on independent cohorts. To challenge the trained algorithm and identify its limitations, we tested the three-way classifier with different cohorts. Images of lung cancers were obtained from the New York University Langone Medical Center from both frozen (75 of LUAD and 23 of LUSC), FFPE sections (74 LUAD, 66 LUSC) and biopsies (51 LUAD and 51 LUSC). The diagnosis used as true positive for these cases are based on morphology (gland formation for adenocarcinoma and keratinization and intracellular bridges for squamous cells), with the cases classified according to the World Health Organization; for the more challenging cases, immunostaining was performed. Because biopsies can be much narrower, during the tiling process at 5× magnification, a tile was kept if at least 20% was covered by the tissue instead of 50%. As those external slides also contained a lot of elements the

network was not trained to identify (blood clot, cartilage), we ran the final tests on regions of interests (ROI) selected by a pathologist. Those regions were selected manually using Aperio ImageScope (Leica Biosystems), and tiles were kept only if it was covered by at least 50% of the ROI for 20×-magnified tiles, and 10% for the 5×-magnified tiles. Additionally, we trained several networks to automatically select those ROIs for the NYU dataset (tumor or nontumor): the first network was trained with the FFPE + biopsies slides and tested on the frozen ones, the second trained with the FFPE + frozen slides and tested on the biopsy ones, and the third trained with the frozen + biopsy slides and tested on the FFPE ones. For each test, we therefore applied this automatic ROI selection followed by the three-way classifier trained on the TCGA dataset, allowing us to compare the performance of the independent cohorts at different levels: using the whole slide image, using ROIs selected by a pathologist, and using ROIs selected by a trained deep-learning architecture. For the mutations, we identified 63 FFPE sections which were tested for *EGFR* mutations; 34 were identified as wild-type and 29 as mutant. Most of them (41) were analyzed using markers used as immunochemical stains to detect the mutations L858R and E746_A750del. 17 others were analyzed by PCR, and 5 others were analyzed with NGS. The tests were run using tumor regions manually selected by a pathologist.

Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code Availability. The source code can be accessed at https://github.com/ncoudray/DeepPATH.

Data availability

All relevant data used for training during the current study are available through the Genomic Data Commons portal (https://gdc-portal.nci.nih.gov). These datasets were generated by TCGA Research Network (http://cancergenome.nih.gov/), and they have made them publicly available. Other datasets analyzed during the current study are available from the corresponding author on reasonable request.

References

60. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
61. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
62. Hershey, S. et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135 (New Orleans, LA, USA, 2017).
63. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
64. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
65. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* 56 (CRC Press, Boca Raton, FL, USA, 1994).
66. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
67. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)* **22**, 276–282 (2012).

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

Images were downloaded from the open TCGA database,
NYU-images were scanned using the Aperio scanner and labelled using Aperio ImageScope (Leica Biosystems)

EGFR mutation status was determined via these tests and software:
1. FoundationOne comprehensive genomic profile for solid tumors, Foundation Medicine:
NGS sequence data processing
Sequence data were mapped to the human genome (hg19) using BWA aligner v0.5.9. PCR duplicate read removal and sequence metric collection was done using Picard 1.47 (http://picard.sourceforge.net/) and Samtools 0.1.12a33. Local alignment optimization was performed using GATK 1.0.4705. Variant calling was done only in genomic regions targeted by the test.
2. EGFR mutation assay by Genzyme, performed by LabCorp:
Analysis was performed on their proprietary software, cobas 4800 SR2 System Software version 2.0 or higher configured with the EGFR Analysis Package
3. NGS50, NYU
For NGS50 (IonTorrent), the pipeline is used is accessible here: https://github.com/stevekm/reportIT

The code used for image pre-processing, training and testing is freely available at https://github.com/ncoudray/DeepPATH and is using tensorflow 1.0, Python-language and libraries. It uses other open-source codes (inception v3) also available from that link and openslide 3.4.1 (https://github.com/openslide/openslide-python).

Data analysis	Data analysis code is also available on https://github.com/ncoudray/DeepPATH (ROC curve, heatmap) and has also used python language and libraries as well as inception v3 architecture coded with tensorflow 1.0 (https://github.com/tensorflow/).

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 upon request. 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

All relevant data used for training during the current study are available through the Genomic Data Commons portal (https://gdc-portal.nci.nih.gov). These datasets were generated by TCGA Research Network (http://cancergenome.nih.gov/) and they have made them publicly available. Other datasets analyzed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-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 predetermine sample size. Sample size was determined by the number of cases available in the databases mined. The TCGA database images are composed of more than 1,500 slides leading to hundreds of thousands of tiles for training. For the independent cohorts, we increased the number of slides to around 100 for each of them (98 frozen sections, 140 FFPE and 102 biopsies) as requested by the reviewers.
Data exclusions	No exclusion
Replication	The results were further tested using independent cohorts. Training, validation and testing was done once for each task with no attempt to replicate.
Randomization	Datasets were randomly assigned to the different sets (training, test and validation).
Blinding	The 170 images from the TCGA used as a test set were classified by 3 pathologists independently, and using only the visual information from each slide.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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