

Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning

Nicolas Coudray^{1,2}, Andre L. Moreira^{3,4}, Theodore Sakellaropoulos⁵, David Fenyö^{6,7}, Narges Razavian^{8,*}, Aristotelis Tsirigos^{1,3,*}

¹ *Applied Bioinformatics Laboratories, New York University School of Medicine, NY 10016, USA*
² *Skirball Institute, Dept. of Cell Biology, New York University School of Medicine, NY 10016, USA*
³ *USA*

⁴ *Department of Pathology, New York University School of Medicine, NY 10016, USA*
⁵ *Center for Biospecimen Research and Development, New York University, NY 10016, USA*

⁶ *School of Mechanical Engineering, National Technical University of Athens, Zografou 15780, Greece*

⁷ *Institute for Systems Genetics, New York University School of Medicine, NY 10016, USA*
⁸ *Department of Biochemistry and molecular Pharmacology, New York University School of Medicine, NY 10016, USA*

⁹ *Medicine, NY 10016, USA*
¹⁰ *Population Health and the Center for Healthcare Innovation and Delivery*
¹¹ *Science, New York University School of Medicine, NY 10016, USA*

¹² * To whom correspondence should be addressed. Tel: +1 646 501 2693; Email:
¹³ Aristotelis.Tsirigos@nyumc.org; Correspondence may also be addressed to Narges Razavian.
¹⁴ Tel: +1 212 263 2234, E-mail: Narges.Razavian@nyumc.org

¹⁵
¹⁶
¹⁷
¹⁸
¹⁹
²⁰
²¹
²²
²³
²⁴

25 **Abstract**

26 Visual analysis of histopathology slides of lung cell tissues is one of the main methods used by
27 pathologists to assess the stage, types and sub-types of lung cancers. Adenocarcinoma and
28 squamous cell carcinoma are two most prevalent sub-types of lung cancer, but their distinction
29 can be challenging and time-consuming even for the expert eye. In this study, we trained a deep
30 learning convolutional neural network (CNN) model (inception v3) on histopathology images
31 obtained from The Cancer Genome Atlas (TCGA) to accurately classify whole-slide pathology
32 images into adenocarcinoma, squamous cell carcinoma or normal lung tissue. Our method slightly
33 outperforms a human pathologist, achieving better sensitivity and specificity, with ~0.97 average
34 Area Under the Curve (AUC) on a held-out population of whole-slide scans. Furthermore, we
35 trained the neural network to predict the ten most commonly mutated genes in lung
36 adenocarcinoma. We found that six of these genes – STK11, EGFR, FAT1, SETBP1, KRAS and
37 TP53 – can be predicted from pathology images with an accuracy ranging from 0.733 to 0.856,
38 as measured by the AUC on the held-out population. These findings suggest that deep learning
39 models can offer both specialists and patients a fast, accurate and inexpensive detection of
40 cancer types or gene mutations, and thus have a significant impact on cancer treatment.

41

42 **Keywords**

43 Computational Biology; Cancer; Precision Medicine; Image Analysis; Computer Vision and
44 Pattern Recognitionr; Quantitative Methods; Deep-learning

45

46

47 Introduction

48 According to the American Cancer Society¹, over 150,000 lung cancer patients succumb to their
49 disease each year, while another 200,000 new cases are diagnosed on a yearly basis. It is one
50 of the most widely spread cancers in the world, due mostly to smoking, but also exposure to toxic
51 chemicals like radon, asbestos and arsenic. Non-small cell lung cancers represent 85% of the
52 cases and three sub-types are distinguished: Adenocarcinoma (LUAD), Squamous Cell
53 carcinoma (LUSC) and, most rarely, large-cell carcinoma. Lung biopsies are typically used to
54 diagnose lung cancer subtype and stage. Targeted therapies are applied depending on the type
55 of cancer, stage and the presence of sensitizing mutations^{1,2}. For example, EGFR (epidermal
56 growth factor receptor) mutations, present in about 20% of LUAD, and ALK mutations (anaplastic
57 lymphoma receptor tyrosine kinase), present in less than 5% of LUAD³, both have currently
58 targeted therapies approved by the Food and Drug Administration (FDA)⁴. Mutations in other
59 genes, such as KRAS and TP53 are very common (about 25% and 50% respectively), but have
60 proven particularly challenging drug-targets so far^{3,5}.

61 Virtual microscopy of stained images of tissues are typically acquired at magnifications of x20 to
62 x40, generating very large two-dimensional images (10,000 to over 100,000 pixels in each
63 dimension) that can be tricky to visually analyze in an exhaustive way. Furthermore, accurate
64 interpretation can be difficult and the distinction between LUAD and LUSC is not always clear,
65 particularly in poorly-differentiated tumors, where ancillary studies is recommended for accurate
66 classification. To assist experts, automatic analysis of lung cancer whole-slide images has been
67 recently studied for survival prognosis⁶ and classification⁷. In these studies, Yu et al. combined
68 conventional thresholding and image processing techniques with machine learning methods,
69 such as random forest classifiers, SVM or Naïve Bayes classifiers, achieving an Area Under the
70 Curve (AUC) of ~0.85 in distinguishing normal from tumor slides, and ~0.75 in distinguishing
71 LUAD from LUSC slides. Here, we demonstrate how the field can greatly benefit from deep

72 learning, by presenting a strategy based on Convolutional Neural Networks (CNNs) that not only
73 outperforms previously published work, but also achieves accuracies that are at least comparable,
74 if not superior, to human pathologists. The development of new inexpensive and more powerful
75 technologies with higher computing power (in particular Graphics Processing Units, GPUs) has
76 made possible the training of larger and more complex systems⁸⁻¹⁰. This resulted in the design of
77 several deep CNNs, capable of accomplishing complex visual recognition tasks. Such algorithms
78 have already been successfully used for segmentation¹¹ or classification of medical images¹² and
79 cancers such as breast¹³⁻¹⁵, colon cancers¹⁶ or osteosarcoma¹⁷. CNNs have also been studied
80 for classifying lung patterns on CT (Computerized Tomography) scans, achieving a f-score of
81 ~85.5%¹⁸. Here, to study the automatic classification of lung cancer tissues, we used the inception
82 v3 architecture¹⁹ and whole-slide images of hematoxylin and eosin stained histopathology images
83 from TCGA obtained by excision. In 2014, Google won the ImageNet Large-Scale Visual
84 Recognition Challenge by developing the GoogleNet architecture²⁰, derived from the work from
85 Lin et. al²¹, which increased the robustness to translation and non-linear learning abilities by using
86 multi-layer perceptrons and global averaging pooling. Inception architecture is particularly useful
87 for processing the data in multiple resolutions, a feature that makes this architecture suitable for
88 pathology tasks. This network has already been successfully adapted to other specific types of
89 classifications like skin cancers²² and diabetic retinopathy detection²³.

90

91 **Results**

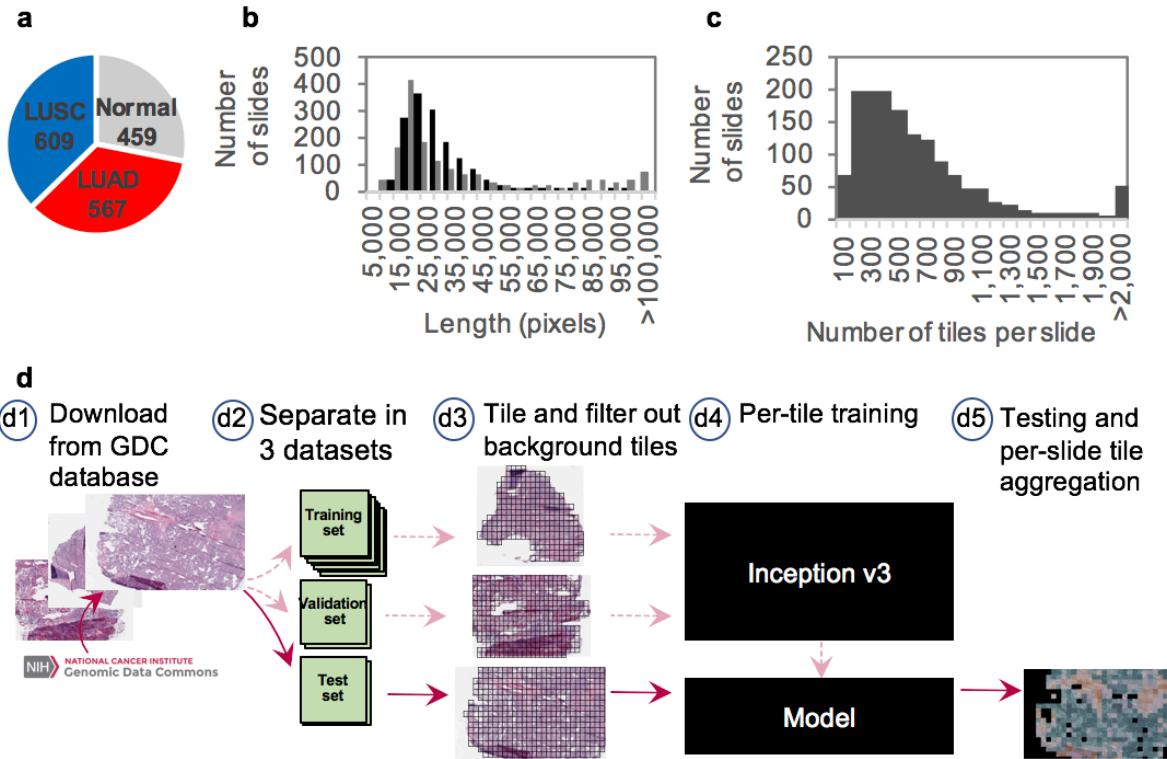
92 We are here comparing several approaches for the classification of tumor slides. First, we
93 employed a strategy similar to the one used by Yu et al.⁷, consisting of a two-step binary
94 classification of normal versus tumor slides, followed by a classification of LUAD versus LUSC
95 slides. We then explored a direct classification of the three types of whole-slide images. Finally,
96 we further analyzed LUAD slides to identify which gene mutations could be predicted from those

97 images: we modified and trained the inception v3 architecture on the 10 most common mutations
98 found in the TCGA dataset and related to lung cancer. In this study, we also compare two training
99 approaches: transfer learning versus fully training the inception architecture. In the first case, most
100 of the network keeps the weights learned after the network was trained on object recognition task
101 on the ImageNet dataset, while only the last layer (fully connected layer) of the network is trained.
102 In the second case, the weights are reinitialized randomly, and the network is trained end-to-end,
103 using exclusively lung cancer images.

104

105 ***Fully-trained inception v3 network provides accurate diagnosis (AUC=0.97) of lung***
106 ***histopathology images***

107 The TCGA dataset characteristics and our overall computational strategy are summarized in
108 **Figure 1** (see method section for details). We used 1634 whole-slide images from the Genomic
109 Data Commons database: 1176 tumor tissues and 459 normal (**Figure 1a**). These whole-slide
110 images were split into three sets: training, validation and testing (**Figure 1d**). Because the sizes
111 of the whole-slide images are too large to be used as direct input to a neural network (sometimes
112 over 100,000 pixels wide, **Figure 1b**), the network was instead trained, validated and tested using
113 512x512 pixel tiles, obtained from non-overlapping windows of the whole-slide images. This
114 resulted in tens to thousands of tiles per slide depending on the original size (**Figure 1c**). These
115 tiles were first processed individually by the network, and then, per-slide aggregation (see
116 Methods for details) of the results generated a diagnosis for each slide.



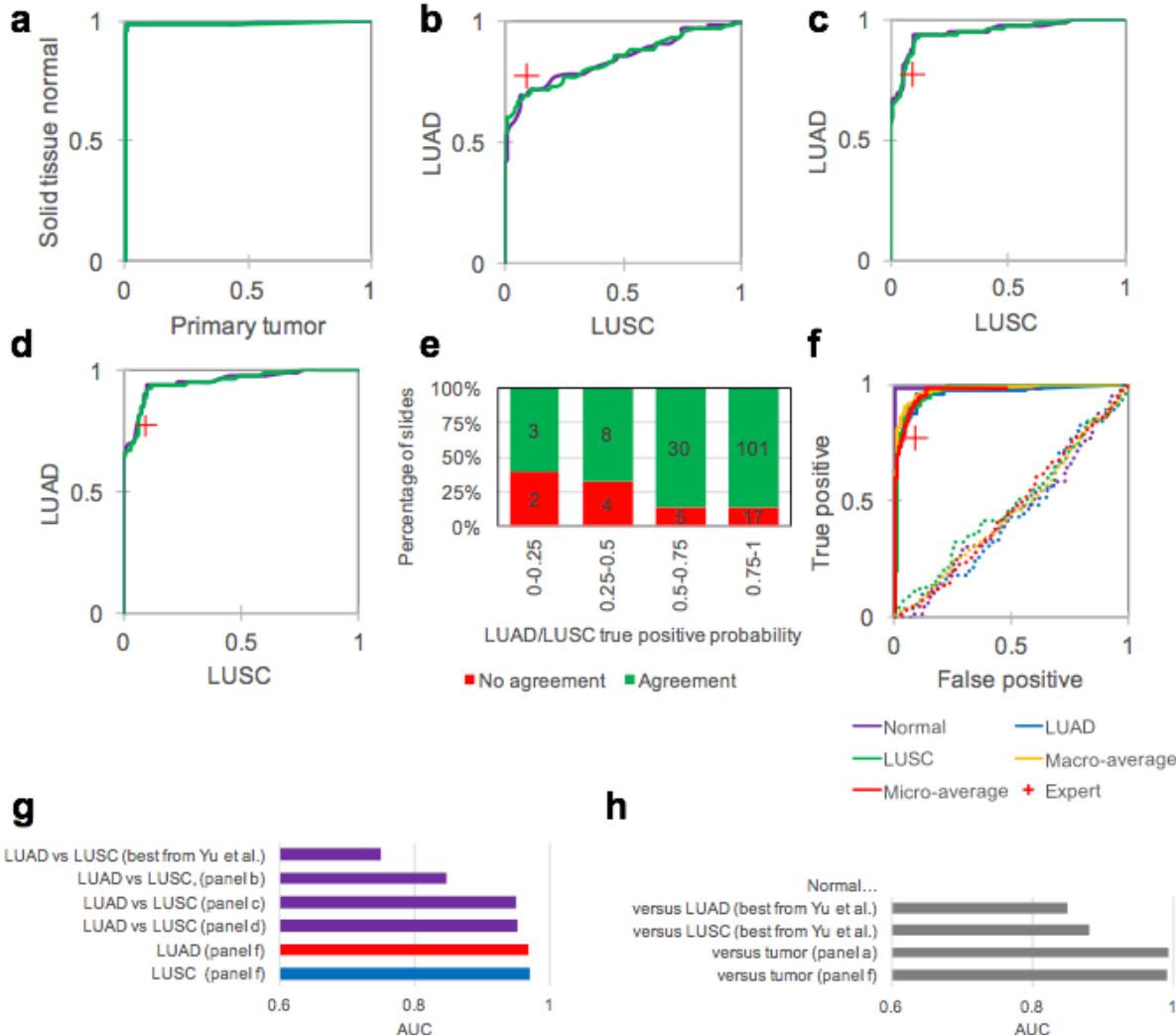
117

118 **Figure 1. Data and strategy:** (a) Number of whole-slide images per class. (b) Size distribution
119 of the images widths (gray) and heights (black). (c) Distribution of the number of tiles per slide.
120 (d) Strategy: (d1) Images of lung cancer tissues were first downloaded from the Genomic Data
121 Common database; (d2) slides were then separated into a training (70%), a validation (15%) and
122 a test set (15%); (d3) slides were tiled by non-overlapping 512x512 pixels windows, omitting those
123 with over 50% background; (d4) the Inception v3 architecture was used and partially or fully re-
124 trained using the training and validation tiles; (d5) classifications were run on tiles from an
125 independent test set and the results were finally aggregated per slide to extract the heat-maps
126 and the AUC statistics.

127

128 Our deep learning approach effectively distinguishes tumor from normal tissue, resulting in a
129 96.1% per tile accuracy. To assess the accuracy on the test set, the per-tile results were
130 aggregated on a per-slide basis either by averaging the probabilities obtained on each tile, or by
131 counting the percentage of tiles positively classified (**Figure 2a**). This process resulted in an Area
132 Under the Curve (AUC) of 0.990 and 0.993 (**Table 1**) respectively, outperforming the AUC of
133 ~0.85 achieved by the feature-based approach of Yu et al.⁷. Next, we tested the performance of
134 our approach on the more challenging task of distinguishing LUAD and LUSC. First, we tested

135 whether convolutional neural networks can outperform the published feature-based approach,
136 even when plain transfer learning is used. For this purpose, the weights of the last layer of
137 inception v3 – previously trained on the ImageNet dataset to identify 1,000 different classes –
138 were initialized randomly and then trained for our classification task. After aggregating the
139 statistics on a per slide basis (**Figure 2b**), this process resulted in an Area Under the Curve (AUC)
140 of 0.847 (**Table 1**), i.e. a gain of ~0.1 in AUC compared to the best results obtained by Yu et al⁷.
141 using image features combined with random forest classifier⁷. The performance can be further
142 improved by fully training inception v3 leading to AUC of 0.947 when aggregation is done by
143 computing the percentage of tiles positively classified, and to AUC of 0.950 when the aggregation
144 is done by averaging the per-tile probabilities (**Figure 2c**). These AUC values are improved by
145 another 0.002 when the tiles previously classified as “normal” by the first classifier are not included
146 in the aggregation process (**Table 1** and **Figure 2d**). The ROC of such a classifier shows
147 performance better than that of a specialist who was asked to classify the images in the test set,
148 independently of the classification provided in TCGA (**Figure 2d**, red cross). About a third of the
149 slides wrongly classified by the algorithm were also misclassified by the specialist, while 85% of
150 those incorrectly classified by the specialist were properly classified by the algorithm (**Figure 2e**).



151

152 **Figure 2. Accurate classification of lung cancer histopathology images:** (a) Per-slide
153 Receiver Operating Characteristic (ROC) curves after classification of normal versus tumor
154 images resulted in an almost error-free classification. Aggregation was either done by averaging
155 the probability scores (purple ROC curves in a to d) or by counting the percentage of properly
156 classified tiles (green ROC curves in a to d). (b) The ROC curves obtained after transfer learning
157 for LUAD vs LUSC images classification shows poorer results than when (c) the same network
158 has been fully trained. The red crosses correspond to the manual classification by a specialist.
159 (d) The ROC curves from (c) is only slightly improved once the tiles classified initially as "normal"
160 have been removed. (e) Proportion of slides misclassified by the specialist as a function of the
161 true positive probability assigned in (d). The number of slides are indicated on the bars. (f) Multi-
162 class ROC of the Normal vs LUAD vs LUSC classification shows the best result for overall
163 classification of cancer type. Dotted lines are negative control trained and tested after random
164 label assignments. (g) Comparison of AUCs obtained with different techniques for classification
165 of cancer type and (h) of normal slides.

166

167 **Figure 3a-r** show heatmap examples for LUAD and LUSC, comparing transfer-learning results
168 with the fully trained network. In the second case, more tiles are true positive and the distribution
169 is more homogeneous, showing for LUSC that almost all of the tiles display LUSC-like features,
170 while for the LUAD, two regions are more prominent with LUAD-like features (one horizontal at
171 the top, one vertical on the left) and some patches showing lower probabilities. Interestingly, most
172 of the LUSC tiles were previously classified as tumor by the first classifier (**Figure 3t**) while for
173 LUAD, the regions with patches having probability near 0.5 in the LUAD/LUSC classification are
174 also those classified as normal with higher probability by the first classifier (**Figure 3s**). We
175 investigated further the use of the deep-learning model by training and testing the network for a
176 direct classification of the three types of images (Normal, LUAD, LUSC in **Figure 2f**). Such an
177 approach resulted in the highest performance with all the AUCs improved to at least 0.968 (**Table**
178 **1**). **Figure 4** shows how the heat-maps are affected by such an approach: the LUSC image shows
179 most of its tiles with a strong true positive probability of LUSC while in the LUAD image, some
180 regions have strong LUAD features, with normal cells on the side (as confirmed by a specialist),
181 and some light blue tiles where LUSC probability is slightly leading.

182

183 **Table 1.** Area Under the Curve (AUC) achieved by the different classifiers

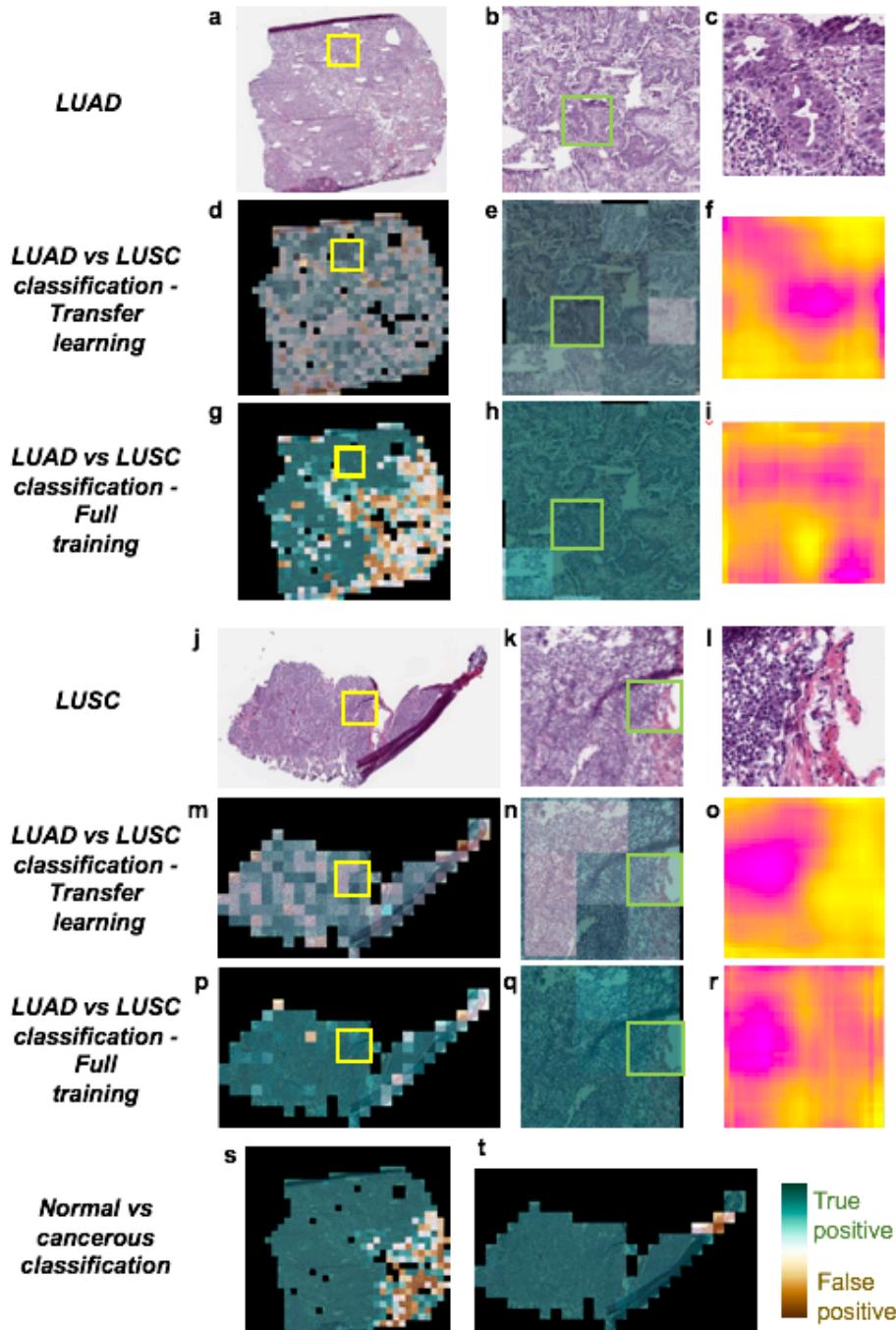
		AUC after aggregation by...	
Classification	Information	... average	... percentage
		predicted	of positively
		probability	classified tiles
Normal vs Tumor	a) Inception v3, fully-trained	0.993	0.990
LUAD vs LUSC	b) Inception v3, transfer learning	0.847	0.847
	c) Inception v3, fully-trained	0.950	0.947

	d) Same as (c) but aggregation done solely on tiles classified as “tumor” by A	0.952	0.949
3 classes.	Normal	0.991	NA
Normal vs	LUAD	0.968	NA
LUAD vs LUSC	LUSC	0.971	NA
	Micro-average	0.971	NA
	Macro-average	0.978	NA
	STK11	0.856	0.842
	EGFR	0.826	0.782
	SETBP1	0.775	0.752
	TP53	0.760	0.754
Mutations	FAT1	0.750	0.750
	KRAS	0.733	0.716
	KEAP1	0.675	0.659
	LRP1B	0.656	0.657
	FAT4	0.642	0.640
	NF1	0.640	0.632

184

185

186



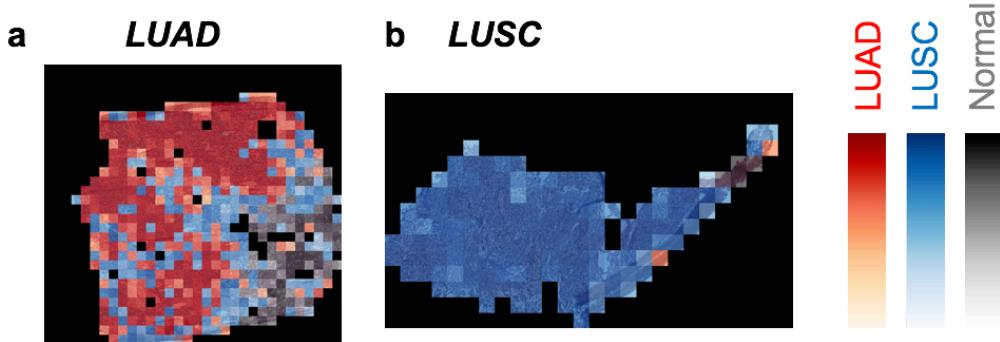
187

188 **Figure 3. Examples of heatmaps for different binary classifications strategies:** (a) Typical
189 slide of Lung Adenocarcinoma (LUAD) tissue. (b) Zoom region corresponding to the yellow box
190 in (a). (c) Tile corresponding to the green box in (b). (d) and (e) are the heat-maps corresponding
191 to images (a) and (b), with probability assigned to each tile from brown (false positive) to green
192 (true positive). (f) Per-tile heat-map generated after having applied a rolling mask on part of the
193 tile. Yellow pixels show regions not affected by masking while the pink pixels show regions where
194 features were important for proper classification. Images (d) to (f) were obtained after transfer

195 learning while images **(g)** to **(i)** were obtained after fully training inception v3. Images **(j)** to **(r)**
196 show similar examples for a Lung Squamous Cell (LUSC) tissues. **(s)** and **(t)** show the results of
197 the “normal vs tumor” tiles classifier.

198

199



200

201 **Figure 4. Heatmaps for classification of Normal vs LUAD vs LUSC:** (a) and (b) Heatmaps
202 corresponding to images of **(Figure 3a)** and **(Figure 3b)** with probabilities of the winning class
203 assigned to each tile such as: red for tiles classified as LUAD, blue for LUSC and grey for Normal.

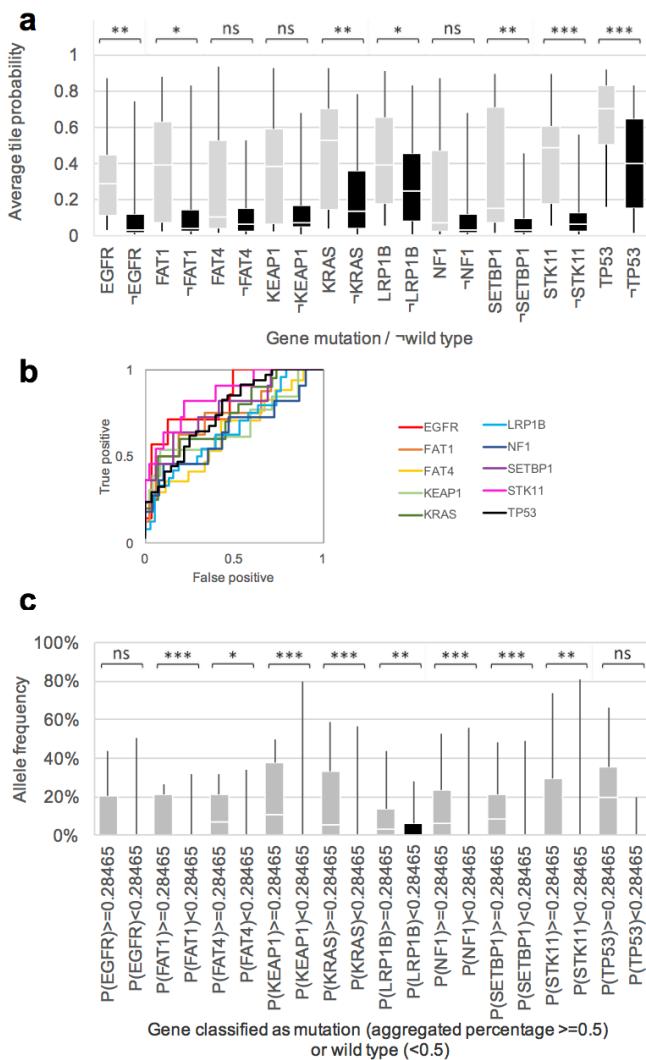
204

205

206 **Whole-slide images can predict 6 mutations with an AUC above 0.74**

207 The LUAD whole-slide images were further trained to predict gene mutations. Inception v3 was
208 modified to allow multi-output classification and tests were conducted using around 44,000 tiles
209 from 62 slides. Box plot and ROC curves analysis (**Figure 5a-b** and **Figure supp 1**) show that at
210 least six frequently mutated genes seem predictable using our deep learning approach: AUC
211 values for STK11, EGFR, FAT1, SETBP1, KRAS and TP53 were found between 0.733 and 0.856
212 (**Table 1**). As mentioned earlier, EGFR already has targeted therapies. **STK11** (Serine/Threonine
213 protein Kinase 11), also known as Liver Kinase 1 (**LKB1**), is a tumor suppressor inactivated in 15-
214 30% of non-small cell lung cancers²⁴ and is also a potential therapeutic target: it has been shown
215 on mice that phenformin, a mitochondrial inhibitor, increases survival²⁵. Also, it has been shown
216 that STK11 mutations may play a role in KRAS mutations which, combined, result in more
217 aggressive tumors²⁶. FAT1 is an ortholog of the Drosophila fat gene involved in many types of

218 cancers and its inactivation is suspected to increase cancer cell growth²⁷. Mutation of the tumor
219 suppressor gene TP53 is thought to be more resistant to chemotherapy leading to lower survival
220 rates in small-cell lung cancers²⁸. As for SETBP1 (SET 1 binding protein), like KEAP1 and STK11,
221 has been identified as one of the signature mutations of LUAD²⁹. Finally, for each gene, we
222 compare the classification achieved by the deep learning approach with the allele frequency
223 (Figure 5c). Among the gene mutations predicted with a high AUC, four of them seem to show
224 probabilities related to the allele frequency: FAT1, KRAS, SETBP1 and STK11.



225
226 **Figure 5. Gene mutation prediction from histopathology slides give promising results for**
227 **at least 6 genes: (a)** Mutation probability distribution for slides where each mutation is present
228 and absent after tile aggregation done by averaging output probability. **(b)** ROC curves associated

229 with (a). **(c)** Allele frequency as a function of slides classified by the deep-learning network as
230 having a certain gene mutation ($P \geq 0.5$), or the wild-type ($P < 0.5$). p-values estimated with Mann-
231 Whitney U-test are shown as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$) or *** ($p \leq 0.001$).
232

233 Discussion

234 The analysis of lung cancer slide images using the inception v3 convolutional neural network
235 shows a clear improvement over classification techniques combining random forest classifiers,
236 SVM or Naïve Bayes classifiers with conventional image processing tools⁷ (**Figure 2g-h**). For
237 LUAD vs LUSC classification, while transfer learning outperforms this previous approach by about
238 10% and another ~10% is gained by fully training the network, at the expense of a much longer
239 training period. Finally, another ~2.8% is gained on cancer type classification when “normal”
240 tissues are immediately considered and binary classification is replaced by a direct three-class
241 analysis. This latest approach results in performances slightly better than those achieved by a
242 specialist. It is interesting to notice that around a third of the slides misclassified by the algorithms
243 have also been misclassified by the specialist, showing the intrinsic difficulty to distinguish LUAD
244 from LUSC in some cases. However, 22 out of 26 of the slides misclassified by the specialist were
245 assigned to the correct cancer type by the algorithm showing that it could be beneficial in assisting
246 the specialist in their prognosis. As for classification of normal versus tumor cells, the classification
247 is nearly unambiguous with CNN. Per-slide heat-maps (**Figure 3**) show that true positive tiles
248 appear with a stronger probability when the network has been fully trained. For the LUAD sample,
249 it also shows more consistency with tiles in the same adjacent regions being assigned similar
250 probabilities while the bottom right side of the slide seems to contain less LUAD-like tumor cells
251 according to the classification (**Figure 3g**) and is consistent with visual inspection of that region
252 by a specialist. An example of the important features used for classification of individual is shown
253 for LUAD (**Figure 3f,i**) and LUSC (**Figure 3o,r**). In both cases, the per-tile heat-map of the fully
254 trained network shows a more varied gradient of colors while the tests done after transfer-learning
255 shows more of a binary-like heat-map with regions either very yellow or very pink. The

256 development of appropriate tools for visualizing deep learning models will help in the future to
257 better understand the features used by the classifier³⁰. In the current strategy, the only selection
258 used for early tile removal is to make sure that there are enough information and the portion of
259 background present is low. Afterwards, all the remaining tiles belonging to a given slide are used
260 for training and all are associated with the label associated with it. This assumption gives good
261 result since AUCs of 0.95 to 0.97 performance is achieved for LUSC vs LUAD, but it is unlikely
262 that 100% of the tiles are indeed representative of the labelled cancer type. Oftentimes, the tumor
263 is only local and some regions of the slides are not affected by the tumor. Performing an initial
264 classification of “normal” vs “tumor” partially addresses this issue removing normal-like tiles. The
265 fact that these are excisions of lung cancer, the tumor cells spread over the whole slide images
266 available and not a small portion of which has clearly been beneficial for this classification. Finally,
267 it is surprising to note the high AUCs achieved considering that several slides present artifacts
268 inherent to freezing techniques used to prepare those samples. However, it should be noted that
269 the available images may not fully represent the diversity that specialists have to deal with and it
270 may be interesting in the future to assess how the network performs under the less than ideal
271 circumstances that can occur (poor staining quality, focus not optimal or autofocus failure, lack of
272 homogeneity in the illumination, etc.). Before this study, it was a priori unclear if and how a given
273 gene mutation would affect the pattern of tumor cells on a whole-slide image but the training of
274 the network using mutated genes as a label lead to very promising prediction results for 6 genes:
275 EGFR, STK11, FAT1, SETBP1, KRAS and TP53. STK11 mutation leads to the highest prediction
276 rate with AUC above 0.85 using aggregation by averaging tile probabilities. Though the number
277 of cases is low (44,000 tiles from 62 test slides), it is interesting to see that our training protocol
278 gives non-random values for several genes, showing that mutation of these particular genes could
279 be predicted from whole-slide images. Hopefully, these predictions will be confirmed once more
280 data are made available. It means that those mutations somehow affect the way the tumor cells
281 look like. Future work on deep-learning models visualization tools³⁰ would help identifying those

282 features. These probabilities could be reflecting the percentage of cells effectively affected by the
283 mutation, the allele frequency being significantly higher for at least 4 genes when they were
284 predicted as mutated by the neural network (**Figure 5c**). Looking, for example, at the predictions
285 done on the whole-slide image from **Figure 3a**, our process successfully identifies TP53 (allele
286 frequency of 0.33) and STK11 (allele frequency of 0.33) are two gene most likely mutated (**Figure**
287 **supp 2a**). The heatmap shows that almost all the LUAD tiles are highly predicted as showing
288 TP53-mutant-like features (**Figure supp 2b**), and two major regions with STK11-mutant-like
289 features (**Figure supp 2c**). Interestingly, when the classification is applied on all tiles, it shows
290 that even tiles classified as LUSC present TP53 mutations (**Figure supp 2d**) while the STK11
291 mutant is confined to the LUAD tiles (**Figure supp 2e**). These results are realistic since, as
292 mentioned earlier, STK11 is a signature mutations of LUAD ²⁹ while TP53 is more common in
293 human cancers. Being able to predict gene mutations at this stage could be beneficial regarding
294 the importance and impact of those mutations^{4,24-29}. This study shows that using deep-learning
295 convolutional neural networks for cancer analysis greatly improve the state-of-the-art automatic
296 classification and could be a very promising tool to assist the specialist in their classification of
297 whole-slide images of lung tissues. Histopathology slides are very large, they usually contain
298 artifacts and be noisy with features of cancer type ambiguous, and making a prognosis manually
299 based on every single region of it can be challenging. Those new techniques can efficiently
300 highlight regions associated with a certain cancer type. Finally, we have shown for the first time
301 the potential to use deep-learning on histopathology images to predict some gene mutations at
302 an early stage.

303

304 **Methods**

305 The overall steps described in this section are summarized in **Figure 1** and described in the
306 following sections.

307

308 ***Dataset of 1,634 whole-slide images***

309 Our dataset comes from the NCI Genomic Data Commons³¹ which provides the research
310 community with an online platform for uploading, searching, viewing and downloading cancer-
311 related data. All freely available slide images of Lung cancer were uploaded from this source. We
312 studied the automatic classification of “solid tissue normal” and “primary tumor” slides using a set
313 of respectively 459 and 1175 eosin stained histopathology whole-slide images. Then, the “primary
314 tumor” were classified between LUAD and LUSC types using a set of respectively 567 and 608
315 of those whole-slide images.

316

317 ***Image pre-processing generates 987,931 tiles***

318 The slides were tiled in non-overlapping 512x512 pixel windows at a magnification of x20 using
319 the openslide library³² (533 of the 2167 slides initially uploaded were removed because of
320 compatibility and readability issues at this stage). The slides with a low amount of information
321 were removed, that is all the tiles where more than 50% of the surface is covered by background
322 (where all the values are below 220 in the RGB color space). This process generated nearly
323 1,000,000 tiles.

324

325 **Table 2.** Dataset information for normal vs tumor classification: number of tiles / slides in each
326 category.

	Training	Validation	Testing
Normal	132,185 / 332	28,403 / 53	28,741 / 74
Primary tumor	556,449 / 825	121,094 / 181	121,059 / 180

327

328

329 **Table 3.** Dataset information for LUAD vs LUSC classification: number of tiles / slides in each
330 category.

	Training	Validation	Testing
LUAD	255,975 / 403	55,721 / 85	55,210 / 79
LUSC	300,474 / 422	65,373 / 96	65,849 / 91

331

332

333 ***Deep learning with Convolutional Neural Network***

334 We used 70% of those tiles for training, 15% for validation, and 15% for final testing (**Table 2** and
335 **Table 3**). The tiles associated with a given slide were not separated but associated as a whole to
336 one of these sets to prevent overlaps between the three sets. Typical CNN consist of several
337 levels of convolution filters, pooling layers and fully connected layers. We based our model on
338 inception v3 architecture¹⁹. This architecture makes use of inception modules which are made of
339 a variety of convolutions having different kernel sizes and a max pooling layer. The initial 5
340 convolution nodes are combined with 2 max pooling operations and followed by 11 stacks of
341 inception modules. The architecture ends with a fully connected and then a softmax output layer.
342 For “normal” vs “tumor” tiles classification, we fully trained the entire network. For the classification
343 of type of cancer, we followed and compared different approaches to achieve the classification:
344 transfer learning, which includes training only the last fully-connected layer, and training the whole
345 network. Tests were implemented using the Tensorflow library (tensorflow.org).

346

347 ***Transfer learning on inception v3***

348 We initialized our network parameters to the best parameter set that was achieved on ImageNet
349 competition. We then fine-tuned the parameters of the last layer of the network on our data via
350 back propagation. The loss function was defined as the cross entropy between predicted

351 probability and the true class labels, and we used RMSProp³³ optimization, with learning rate of
352 0.1, weight decay of 0.9, momentum of 0.9, and epsilon of 1.0 method for training the weights.
353 This strategy was tested for the binary classification of LUAD vs LUSC.

354

355 ***Training the entire inception v3 network***

356 The inception v3 architecture was fully trained using our training datasets and following the
357 procedure described in ³⁴. Similar to transfer learning, we used back-propagation, cross entropy
358 loss, and RMSProp optimization method, and we used the same hyperparameters as the transfer
359 learning case, for the training. In this approach, instead of only optimizing the weights of the fully
360 connected layer, we also optimized the parameters of previous layers, including all the
361 convolution filters of all layers. This strategy was tested on three classifications: normal vs tumor,
362 LUAD vs LUSC and Normal vs LUAD vs LUSC. The training jobs were run for 500,000 iterations.
363 We computed the cross-entropy loss function on train and validation dataset, and used the model
364 with best validation score as our final model. We did not tune the number of layers or hyper-
365 parameters of the inception network such as size of filters.

366

367 ***Identification of gene mutations***

368 To study the prediction of gene mutations from histopathology images, we modified the inception
369 v3 to perform multi-task classification rather than a single task classification. Each mutation
370 classification was treated as a binary classification, and our formulation allowed multiple
371 mutations to be assigned to a single tile. We optimized the average of the cross entropy of each
372 individual classifier. To implement this method, we replaced the final softmax layer of the network
373 with a sigmoid layer, to allow each sample to be associated with several binary labels ³⁵. We
374 used RMSProp algorithm for the optimization, and fully trained this network for 500k iterations
375 using only LUAD whole-slide images, each one associated with a 10-cell vector, each cell

376 associated to a mutation and set to 1 or 0 depending on the presence or absence of the mutation.
377 Only the most common mutations were used (**Table 4**), leading to a training set of 223,185 tiles.
378 Training and validation were done over 500,000 iterations (**Figure supp 3**). The test was then
379 achieved on the tiles, and aggregation on the 62 test-slides where at least one of these mutations
380 is present was done only if the tile was previously classified as “LUAD” by the Normal/LUAD/LUSC
381 3-classes classifier.

382
383 **Table 4.** Gene included in the multi-output classification and the percentage of patients with LUAD
384 in the database where the genes are mutated.

Gene mutated	TP53	LRP1B	KRAS	KEAP1	FAT4	STK11	EGFR	FAT1	NF1	SETBP1
%Patients	50	34	28	18	16	15	12	11	11	11

385
386
387 **Results analysis**
388 Once the training phase was finished, the performance was evaluated using the testing dataset
389 which is composed of tiles from slides not used during the training. We then aggregated the
390 probabilities for each slide using two methods: either average of the probabilities of the
391 corresponding tiles, or percentage of tiles positively classified. The ROC (Receiver Operating
392 Characteristic) curves and the corresponding AUC (Area Under the Curve) were computed in
393 each case. Tumor slides could contain a certain amount of “normal” tiles. Therefore, we also

394 checked how the ROC & AUC were affected when tiles classified as “normal” were removed from
395 the aggregation. Heat-maps were also generated for some tested slide to visualize the differences
396 between the two approaches and identify the regions associated with a certain cancer type. To
397 visualize the regions of a given tile which were important for the algorithm to take a decision, a
398 rolling mask was applied to the tile. The masked tile was then fed to the network to analyze how

399 the classification is affected. 128x128 pixel overlapping masks were generated over the whole tile
400 with 87.5% overlapping between adjacent masks.

401

402 Acknowledgments

403 This work has utilized computing resources at the High-Performance Computing Facility at NYU
404 Langone Medical Center. The slide images and the corresponding cancer information were
405 uploaded from the Genomic Data Commons portal (<https://gdc-portal.nci.nih.gov>) and are in
406 whole or part based upon data generated by the TCGA Research Network
407 (<http://cancergenome.nih.gov/>). The data used were publicly available without restriction,
408 authentication or authorization necessary.

409

410 References

- 411
- 412 1 American Cancer Society, <<https://www.cancer.org/>> (2017).
- 413 2 Chan, B. A. & Hughes, B. G. Targeted therapy for non-small cell lung cancer: current
414 standards and the promise of the future. *Translational Lung Cancer Research* **4**, 36-54
415 (2015).
- 416 3 Terra, S. B. *et al.* Molecular characterization of pulmonary sarcomatoid carcinoma:
417 analysis of 33 cases. *Modern Pathology* **29**, 824-831 (2016).
- 418 4 Blumenthal, G. M. *et al.* Oncology Drug Approvals: Evaluating Endpoints and Evidence in
419 an Era of Breakthrough Therapies. *The Oncologist* **22**, 762-767 (2017).
- 420 5 Jänne, P. A. *et al.* Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell
421 lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. *The Lancet
422 Oncology* **14**, 38-47 (2013).
- 423 6 Luo, X. *et al.* Comprehensive Computational Pathological Image Analysis Predicts Lung
424 Cancer Prognosis. *Journal of Thoracic Oncology* **12**, 501-509 (2017).
- 425 7 Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated
426 microscopic pathology image features. *Nature Communications* **7** (2016).
- 427 8 Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**,
428 85-117 (2015).
- 429 9 Hinton, G. E., Osindero, S. & Teh, Y. W. A fast learning algorithm for deep belief nets.
430 *Neural Computation* **18**, 1527-1554 (2006).
- 431 10 Greenspan, H., Ginneken, B. v. & Summers, R. M. Guest Editorial Deep Learning in
432 Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE
433 TRANSACTIONS ON MEDICAL IMAGING* **35**, 1153-1159 (2016).

- 434 11 Qaiser, T., Tsang, Y.-W., Epstein, D. & RajpootEma, N. in *Medical Image Understanding*
435 and Analysis: 21st Annual Conference on Medical Image Understanding and Analysis. (ed
436 Springer International Publishing).
- 437 12 Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annual Review of*
438 *Biomedical Engineering* **19**, 221-248 (2017).
- 439 13 Wang, D., Khosla, A., Gargava, R., Irshad, H. & Beck, A. H. Deep learning for identifying
440 metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).
- 441 14 Cheng, J.-Z. et al. Computer-Aided Diagnosis with Deep Learning Architecture:
442 Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans.
443 *Scientific Reports* **6** (2016).
- 444 15 Cruz-Roa, A. et al. Accurate and reproducible invasive breast cancer detection in whole-
445 slide images: A Deep Learning approach for quantifying tumor extent. *Scientific Reports*
446 **7** (2017).
- 447 16 Sirinukunwattana, K. et al. Locality sensitive deep learning for detection and
448 classification of nuclei in routine colon cancer histology images. *IEEE TRANSACTIONS ON*
449 *MEDICAL IMAGING* **35**, 1196-1206 (2016).
- 450 17 Mishra, R., Daescu, O., Leavey, P., Rakheja, D. & Sengupta, A. in *International*
451 *Symposium on Bioinformatics Research and Applications*. (ed Springer) 12-23.
- 452 18 Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. Lung
453 Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural
454 Network. *IEEE TRANSACTIONS ON MEDICAL IMAGING* **35**, 1207-1216 (2016).
- 455 19 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception
456 Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer*
457 *Vision and Pattern Recognition*, 2818-2826 (2015).
- 458 20 Szegedy, C. et al. Going Deeper With Convolutions. *The IEEE Conference on Computer*
459 *Vision and Pattern Recognition*, 1-9 (2015).
- 460 21 Lin, M., Chen, Q. & Yan, S. Network In Network. *ArXiv arXiv:1312.4400*, 1-10 (2013).
- 461 22 Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural
462 networks. *Nature* **542**, 115-118 (2017).
- 463 23 Gulshan, V. et al. Development and Validation of a Deep Learning Algorithm for
464 Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410
465 (2016).
- 466 24 Sanchez-Cespedes, M. et al. Inactivation of LKB1/STK11 is a common event in
467 adenocarcinomas of the lung. *Cancer Research* **62**, 3659-3662 (2002).
- 468 25 Shackelford, D. B. et al. LKB1 Inactivation Dictates Therapeutic Response of Non-Small
469 Cell Lung Cancer to the Metabolism Drug Phenformin. *Cancer Cell* **23**, 143-158 (2013).
- 470 26 Makowski, L. & Hayes, D. N. Role of LKB1 in lung cancer development. *British Journal of*
471 *Cancer* **99**, 683-688 (2008).
- 472 27 Morris, L. G. et al. Recurrent somatic mutation of FAT1 in multiple human cancers leads
473 to aberrant Wnt activation. *Nature genetics* **45**, 253-261 (2013).
- 474 28 Mogi, A. & Kuwano, H. TP53 Mutations in Nonsmall Cell Lung Cancer. *Journal of*
475 *Biomedicine and Biotechnology* **2011**, 9 (2011).
- 476 29 Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types.
477 *Nature* **502**, 333-339 (2013).

- 478 30 Zeiler, M. D. & Fergus, R. in *European Conference on Computer Vision*. 818-833.
- 479 31 Grossman, R. L. et al. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* **375**, 1109-1112 (2016).
- 480 32 Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A Vendor-Neutral Software Foundation for Digital Pathology. *Journal of Pathology Informatics* **4**, 27 (2013).
- 481 33 Hinton, G., Srivastava, N. & Swersky, K. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning. (2012).
- 482 34 *Inception in TensorFlow*, <<https://github.com/tensorflow/models/tree/master/inception>> (
- 483 35 Hershey, S. et al. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- 490