Active Data Enrichment by Learning What to Annotate in Digital Pathology

George Batchkala[1], Tapabrata Chakraborti[1], Mark McCole[2], Fergus Gleeson[3], Jens Rittscher[1] * **

[1] IBME/BDI, Department of Engineering Science, University of Oxford, Oxford, UK
{george.batchkala; tapabrata.chakraborty; jens.rittscher }@eng.ox.ac.uk
[2] Department of Cellular Pathology, Oxford University Hospitals NHS Trust, UK
mark.mccole@ouh.nhs.uk
[3] NCIMI/BDI, Department of Oncology, University of Oxford, Oxford, UK
fergus.gleeson@oncology.ox.ac.uk

Abstract. Our work aims to link pathology with radiology with the goal to improve the early detection of lung cancer. Rather than utilising a set of predefined radiomics features, we propose to learn a new set of features from histology. Generating a comprehensive lung histology report is the first vital step towards this goal. Deep learning has revolutionised the computational assessment of digital pathology images. Today, we have mature algorithms for assessing morphological features at the cellular and tissue levels. In addition, there are promising efforts that link morphological features with biologically relevant information. While promising, these efforts mostly focus on narrow well-defined questions. Developing a comprehensive report that is required in our setting requires an annotation strategy that captures all clinically relevant patterns specified in the WHO guidelines. Here, we propose and compare approaches aimed to balance the dataset and mitigate the biases in learning by automatically prioritising regions with clinical patterns underrepresented in the dataset. Our study demonstrates the opportunities active data enrichment can provide and results in a new lung-cancer dataset annotated to a degree that is not readily available in public domain.

Keywords: Computational Pathology · Histology Annotation Process · Unbalanced Data · Image Retrieval · Active and Continual Learning

* This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-17979-2_12. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

** GB is the corresponding author and is supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1). TC is supported by Linacre College, Oxford. The work was done as part of UKRI DART Lung Health Program.

1 Introduction

The ambition of finding new ways for linking pathology and radiology in the context of lung cancer motivates our goal of generating a comprehensive pathology report automatically. Given the current state of the art in computational pathology this is an open problem.

Lung cancer accounts for more deaths than any other type of cancer [10]. The tree main subtypes are Non-small Cell Lung Carcinoma (NSCC or NSCLC), Small Cell Carcinoma (SmCC), and Carcinoid Tumour. NSCLC accounts for more than 80% of all lung cancer cases [3,4] and is split into two main subtypes: lung adenocarcinoma (around 50% of all cases [7]) and lung squamous cell carcinoma. Based on existing clinical guidelines CT is used to detect the presence of lung cancer. Tissue samples are then taken from suspicious regions to confirm the diagnosis. The general type, sub-type, and the underlying morphological characteristics of lung cancer determines clinical prognosis [8]. Hence it is vital to identify all subtypes, including those that occur less frequently, in a robust manner. The difficulty for making an accurate diagnosis lies in the inter- and intra-tumour heterogeneity [13]. The large inter-observer variability [9] is another factor that needs to be taken into account.

Recently a number of promising approaches for automatic subtyping of specific lung cancers and identifying specific lung-cancer morphologies have been published. With the help of region-based annotation it is now possible to determine the predominant morphological pattern of lung adenocarcinoma using region-based annotations [1,11]. Other works [2,5,6] focused on subtyping NSCLC into adenocarcinoma and squamous cell carcinoma using WSI-level labels. Yang *et al.* [12] developed a model to classify lung histology images into six different types: 3 most popular lung cancer sub-types (adenocarcinoma, squamous cell carcinoma, small cell lung carcinoma), as well as pulmonary tuberculosis, organizing pneumonia, and normal lung tissue.

The drawback of all these methods is that they either support the detection of adenocarcinoma (most prominent lung cancer type) [1,11] and do not take other lung-cancer types into account or classify lung cancer types directly from the histology images [2,5,6,12] omitting the stage of explicitly finding the morphological features [8] used by the pathologists to make the diagnosis.

In order to support our goal of identifying new features that support the early detection of lung cancer on CT we require an approach that closely mimics the way pathologists work today. It is critical to automatically identify a broad range of WHO-defined features at different magnifications and aggregate them to make the final diagnosis. To this end, we develop a novel annotation protocol (Figure 1), which makes optimal use of the available data and expert annotation time. In order to utilise the limited time human experts can dedicate to such an annotation task, we develop an approach that actively selects specific cases to achieve a balanced training dataset. Focus of this work is to discuss and analyse novel techniques to optimize the annotation process (Figure 2, left).

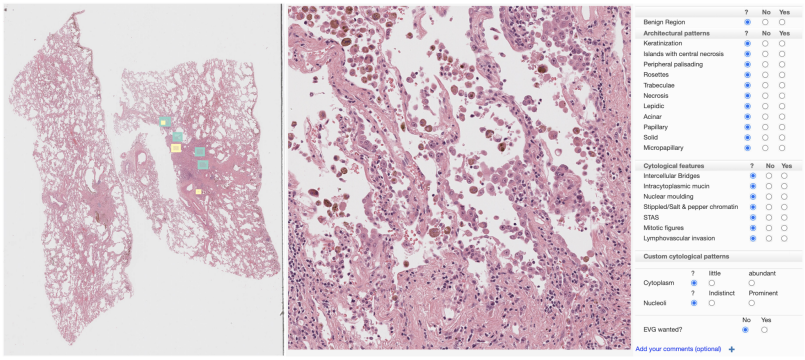

Fig. 1: Two stages of annotation process. *Left*: Pathologist is asked to choose a sufficient number of relevant regions of interest (ROIs) at different magnifications to support a diagnosis. *Right*: annotation view for one of the ROIs.

2 Methodology

Our goal is to obtain region labels at different magnifications to support automated reporting of all clinically relevant subtypes of lung cancers. Here we take the WHO guidelines as a reference. The labels should include the features and patterns used by the pathologists for making the diagnosis [8] from a WSIs.

2.1 Annotation Protocol

We propose a novel lung-cancer annotation protocol (Figure 1) that consists of two main stages: (i) selection of relevant regions on digitised histology slides; and (ii) an annotation scheme that summarizes the information from the selected regions into a number of clinically-relevant patterns.

Stage 1: Region Selection. A typical digital slide viewer set up is used present the digitised slides. The pathologist is asked to mark enough diagnostically relevant regions on the slides in order to make the diagnosis by specifying regions of Interest (ROIs) (Figure 1, left). Due to the small proportion of benign samples in our cohort, the pathologist also selects one or two regions of non-cancerous tissue on each of the slides to give explicit control of how a benign region can look and compensate for the lack of benign samples in the training data. Regions are mostly selected at two magnifications. Lower magnification allows to see *architectural patterns* (green ROIs), while higher magnification - *cytological features* (yellow ROIs).

Stage 2: Region Annotation. Aim of this step is to use the terms from the 2021 WHO Classification of Lung Tumours [8] to annotate each of the ROIs that have been selected at the previous stage. All relevant labels are shown on the right of Figure 1. The "?" label is introduced to mitigate inter-observer variability. Only unequivocal cases are given definite labels.

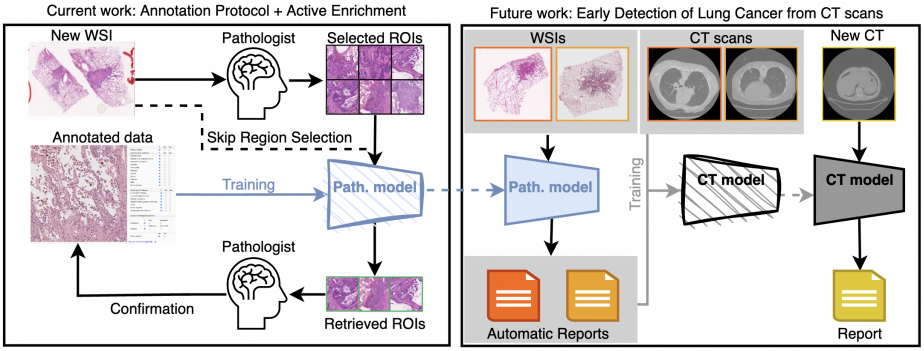

Fig. 2: Annotation protocol and active data enrichment (left) in the context of early detection of lung cancer from CT images (right). Sketch-style models represent models in the training mode, while solid-fill - trained models in the inference mode.

2.2 Dataset Enrichment

Due to the high cost of expert annotation, it is vital to optimize the annotation process. For us, it means, minimizing the time spent by the pathologist in order to achieve the quality of the data enough for efficient model training. When training a model to recognize multiple classes at once, it is crucial to create a balanced dataset in which all classes are well represented.

A naive sequential annotation of the available data would naturally result in an extremely unbalanced dataset in which patters of rare disease subtypes would be underrepresented. Trying to get a sufficient number of regions with underrepresented patterns in this naive sequential way would result in a sub-optimal use of limited expert annotation time. To avoid this, we propose two approaches to increase the proportion of regions with underrepresented patterns and help the models learn features distinguishing these patterns by making use of known image retrieval techniques. The approaches are illustrated on Figure 2 (left): ranking regions pre-selected by the pathologist (solid arrows) and automatically selecting regions from non-annotated WSIs in a sliding-window sweep (dashed line). We have tried supervised and unsupervised methods for the former approach. As a result we are now in a position to capture a unique reference dataset for lung histology. We plan to extend this work to the latter approach.

As our data collection is ongoing we do not know how many annotated datasets can be generated. Hence we propose a metric to measure the retrieval performance for variable number of examples: **Ranking Curve AUC**. Our metric is similar to ROC AUC used for measuring classification performance. For each n the ranking curve shows the proportion of regions with pattern of interest in top-n ranked samples from the total number of regions with the pattern of interest that are possible to pick up in n samples ($min(n, t)$), where t is the total number of relevant examples. We chose this denominator since we can not possibly find more relevant examples in any n samples. For the AUC calculation, discrete data points of the ranking curve are connected with straight lines. The

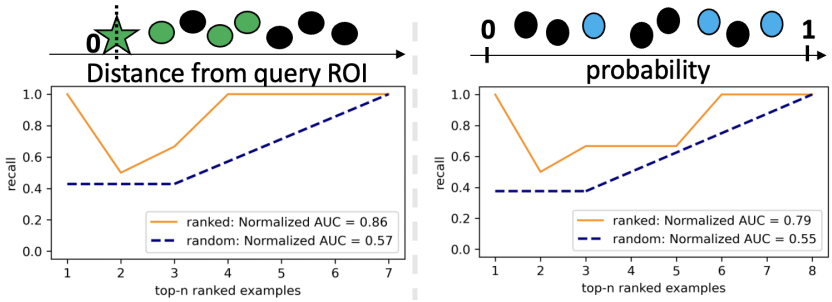

Fig. 3: Retrieval strategies. Coloured dots represent ROIs with the patterns of interest present. **Left: Unsupervised**. ROIs are ranked based on the distance of their feature vector (dots) to the query feature vector (green star). **Right: Supervised**. ROIs are ranked by their probabilities of having the pattern of interest present.

normalized version of the proposed metric has an upper bound of 1 when all relevant samples are ranked higher than irrelevant ones. See the formal definition and the derivation of properties in Supplementary Material. Figure 3 shows ranking curves for toy examples.

3 Results

We compare unsupervised, supervised, and active-learning retrieval strategies for ranking regions pre-selected by the pathologist.

3.1 Unsupervised Data Enrichment

To study the effectiveness of the enrichment methods a pathologist has annotated 2 batches with 20 and 17 WSIs that resulted in 145 and 120 selected ROIs respectively. The batches were scanned with different scanners at 20x magnification. Keratinization is the most underrepresented in the first batch with only 1 of the ROIs marked to have it by our pathologist. Having only one image of a particular class suggests unsupervised image-retrieval since we can not train a good model from one example. Furthermore, this region does not have any other patterns from our annotation list. Hence, we enrich the keratinization class as follows. We use the region as a *query region*. It is passed through a feature extractor to create a *query vector*. Not-annotated *target regions* pre-selected by the pathologist are processed by the same feature extractor. Matches with the smallest cosine similarity, a distance metric commonly used for image retrieval, are presented to the pathologist for confirmation.

15 out of 120 ROIs in the second batch contain keratinization pattern. We simulate the situation of prioritising which of the 120 patterns we should annotate to maximize the number of images with keratinization in the top-20 ranked samples since it is the number of ROIs that the pathologist reported to be able to

annotate in one annotation session. Since we do not require extra annotation, we try different feature extractors: a modified version of an ImageNet pre-trained ResNet-18 without the classification layer, an ImageNet pre-trained modified ResNet-50 [6], and two ResNet-18 models pre-trained using self-supervised learning on patches extracted at 2.5x and 10x magnifications from TCGA-lung by Li *et al.*[5]. Results are shown in Table 1. The TCGA-lung pre-trained networks perform better than the ImageNet pre-trained ones. However, 20 is still an arbitrary number of ROIs to consider and this choice can seriously affect the performance evaluation. Hence, we use the Ranking Curve AUC proposed in Section 2.2. The curves for two feature extractors are shown in Figure 4 while the AUC values are presented in Table 1. ResNet-18 feature extractor pre-trained on patches of TCGA-lung extracted at 10x magnification [5] showed best retrieval results. For any 40 regions chosen at random we expect to have $40/120 \approxeq 0.33$ or 5 out of 15 regions with keratinization pattern. However, the top-40 ranked regions have 10 out of 15 regions doubling the proportion compared to random choosing ($10/15 \approxeq 0.66$). Only with a third of annotated regions from the second batch, we could get the labels for two thirds of the regions with keratinization pattern.

Feature Extractor	**Pre-training**	**Success Rate**	**Ranking AUC**
Modified ResNet-50 [6]	ImageNet	3/20	0.62
ResNet-18 w/o last layer	ImageNet	4/20	0.62
ResNet-18 w/o last layer	TCGA patches at 2.5x [5]	7/20	0.73
ResNet-18 w/o last layer	TCGA patches at 10x [5]	**8/20**	**0.79**
Random	NA	2.5/20	0.51

Table 1: Retrieval performance of different feature extractors. The proportion of regions with keratinization in the second batch is $15/120 = 1/8$ meaning that we expect $20 * 1/8 = 2.5$ examples with keratinization in any random sample of 20 examples.

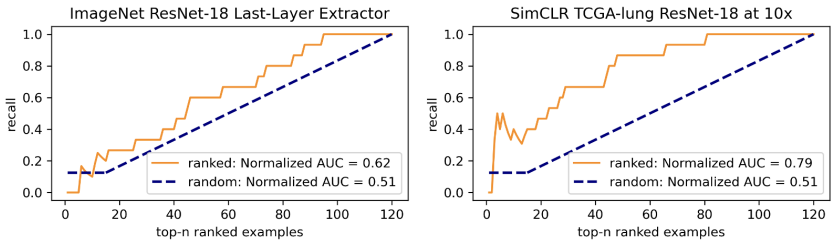

Fig. 4: Solid orange line: ranking curve (as described in Section 2.2). Dashed blue line: expected cumulative proportion if selecting samples at random. ImageNet trained extractor (left) shows worse results than TCGA-lung pre-trained extractor (right).

3.2 Supervised Active Data Enrichment

We now explore how using supervised enrichment methods can be used to prioritise the regions with acinar pattern present. We evaluate how including these regions improves the classifier performance on a previously unseen test set and mitigates the biases in learning by reducing the class imbalance.

We split the dataset into 4 subsets: Train, Validation, Pool, and Test with 20/86, 14/59, 10/60, and 10/60 examples containing acinar pattern respectively (for detailed data distribution see Supplementary Material). Pool imitates the regions next in line for annotation. Train, Validation, and Test sets serve their usual roles. Train and Validation sets come from the first batch of images, while Pool and Test sets come from the second batch. The batches were scanned with different scanners. This shows how in real life we can have training and evaluation data coming from different distributions.

To obtain a baseline, we train a single classification layer on top of a frozen ResNet-18 feature extractor pre-trained on patches from TCGA-lung at 10x magnification [5] on the Train set using Cross-Entropy loss with 3 possible labels for the acinar pattern: "yes", "no", "not sure". When calculating Cross-Entropy loss we use the same weights for all classes since the label distribution changes in different batches and we want to be able to predict all of them well in the end. We report unweighted accuracy, ROC AUC weighted by the number of support samples with each label, as well as Ranking Curve AUC described in Section 2.2, precision, and recall for the "yes" label. We save the weights of the models which showed best ROC AUC on the validation set. Weighted ROC AUC is chosen because it is is sensitive to class imbalance, gives a good understanding of the model performance, and is a popular choice in the literature [2,5,6,11,12].

Having the baseline model, we vary the training data by including different portions of the Pool set into it. Given that the "yes" label is underrepresented in the first batch and our particular interest in learning to predict it better, we propose to rank the samples by sorting them in decreasing order of predicted probabilities of the "yes" label. We assess how including 10, 20, and 30 ranked examples from the Pool affects the performance metrics. For comparison, we repeat the experiments with 10, 20, and 30 randomly-chosen examples from the Pool. To account for randomness when choosing a subset of examples, we repeat the experiments 10 times for each subset size. Finally, we include all 60 Pool set examples to get the largest-training data baseline (results in Table 2.).

We observe that including more ($0 \rightarrow 10 \rightarrow 20 \rightarrow 30$) ranked or non-ranked samples into the training data increases both the performance of the classifier (weighted ROC AUC) and the ability of the classifier to rank the regions with acinar pattern higher than the ones without it (Ranking Curve AUC). Furthermore, adding 10 ranked examples improves ROC AUC, Ranking Curve AUC, and Precision more than adding 10, 20, or even 30 random examples. Finally, including all 60 Pool examples improves accuracy and recall, but results in lower weighted ROC AUC, Ranking AUC, and Precision. We believe that this happens because we optimize the parameters of the network using a non-weighted Cross-Entropy loss which optimizes the weights better for more prevalent classes. The

	Accuracy	**ROC AUC**	**Rank AUC**	**Pre (yes)**	**Re (yes)**
Train	0.65	0.82	0.752	0.333	0.2
Train + 10 Rand	0.75 ± 0.028	0.83 ± 0.026	0.82 ± 0.036	0.5 ± 0.191	0.39 ± 0.158
Train + 10 Rank	**0.8**	**0.864**	**0.874**	**0.667**	**0.6**
Train + 20 Rand	0.75 ± 0.04	0.83 ± 0.018	0.83 ± 0.042	0.59 ± 0.164	**0.46 ± 0.143**
Train + 20 Rank	**0.783**	**0.904**	**0.923**	**1.0**	0.2
Train + 30 Rand	0.76 ± 0.031	0.84 ± 0.017	0.85 ± 0.011	0.55 ± 0.097	0.57 ± 0.1
Train + 30 Rank	**0.817**	**0.924**	**0.954**	**0.833**	0.5
Train + Pool	**0.833**	0.865	0.894	0.636	**0.7**

Table 2: Experiments were conducted by training the model on different training sets. For each $N \in \{10, 20, 30\}$, adding N ranked examples results in better models then adding N random examples. For random selection, 10 sets of N examples were taken from the Pool set with mean ± one standard deviation reported for each metric.

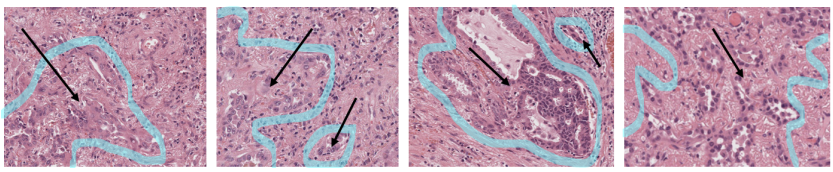

Fig. 5: Regions containing acinar pattern from the top-10 ranked Pool set examples returned by our method. Solid arrows point at areas confirmed and delineated by the pathologist to contain acinar patterns, thus validating the results.

proportion of the "yes" class in Train + Pool is 0.205, while it is 0.25. 0.245, 0.241, for Train + 10, +20, +30 Ranked images respectively (see data distribution in Supplementary Material). This change in proportion results in the improved unweighted accuracy, but removes the precise focus from the "yes" class, which hurts ROC AUC (sensitive to class imbalance), Ranking AUC and Precision which are both measured for the "yes" class.

Figure 5 shows examples from the top-10 ranked samples from the Pool set. The samples were ranked using predicted probabilities of the acinar pattern presence. These four regions were later confirmed and delineated by the pathologist to contain acinar pattern, thus adding further validation to our method.

4 Conclusion

We present a new comprehensive annotation protocol for lung histopathology. In order to increase the value of expert annotations, we propose a simple yet effective method for prioritizing the annotation of regions extracted from whole-slide images that are likely to contain underrepresented patterns. The method achieves this by utilising a region-retrieval model created using the annotated data. The proposed method is evaluated on a new in-house lung-pathology dataset. We conclude that even with little supervision we can enrich the dataset for a pattern of interest. This method is now being used to actively balance the distribution of

labels in the annotated portion of our dataset. This, in turn, will result in better classification and retrieval models. We plan to use better retrieval models to bypass the region selection stage (Figure 2, left, dashed line). Finally, we will use the classification models for automatic histology report generation. The generated reports will be used to learn an new set of radiology features from histology in order to improve the early detection of lung cancer (Figure 2, right).

References

1. Alsubaie, N., Shaban, M., Snead, D.R.J., Khurram, S.A., Rajpoot, N.M.: A Multi-resolution Deep Learning Framework for Lung Adenocarcinoma Growth Pattern Classification. In: Nixon, M.S., Mahmoodi, S., Zwiggelaar, R. (eds.) Medical Image Understanding and Analysis - 22nd Conference, MIUA 2018, Southampton, UK, July 9-11, 2018, Proceedings. vol. 894, p. 311. Springer (2018). https://doi.org/10.1007/978-3-319-95921-4_1
2. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine **24**(10), 1559–1567 (Oct 2018). https://doi.org/10.1038/s41591-018-0177-5
3. Davidson, M.R., Gazdar, A.F., Clarke, B.E.: The pivotal role of pathology in the management of lung cancer. Journal of Thoracic Disease **5 Suppl 5**, S463–478 (Oct 2013). https://doi.org/10.3978/j.issn.2072-1439.2013.08.43
4. Huang, T., Li, J., Zhang, C., Hong, Q., Jiang, D., Ye, M., Duan, S.: Distinguishing Lung Adenocarcinoma from Lung Squamous Cell Carcinoma by Two Hypomethylated and Three Hypermethylated Genes: A Meta-Analysis. PLOS ONE **11**(2), e0149088 (Feb 2016). https://doi.org/10.1371/journal.pone.0149088
5. Li, B., Li, Y., Eliceiri, K.W.: Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14318–14328 (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01409
6. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 2021 5:6 pp. 555–570 (Mar 2021). https://doi.org/10.1038/s41551-020-00682-w
7. Meza, R., Meernik, C., Jeon, J., Cote, M.L.: Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973–2010. PLOS ONE **10**(3), 1–14 (Jan 2015). https://doi.org/10.1371/journal.pone.0121323
8. Nicholson, A.G., Tsao, M.S., Beasley, M.B., Borczuk, A.C., Brambilla, E., Cooper, W.A., Dacic, S., Jain, D., Kerr, K.M., Lantuejoul, S., Noguchi, M., Papotti, M., Rekhtman, N., Scagliotti, G., van Schil, P., Sholl, L., Yatabe, Y., Yoshida, A., Travis, W.D.: The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer **17**(3), 362–387 (Mar 2022). https://doi.org/10.1016/j.jtho.2021.11.003
9. Stang, A., Pohlabeln, H., Müller, K.M., Jahn, I., Giersiepen, K., Jöckel, K.H.: Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. Lung Cancer **52**(1), 29–36 (Apr 2006)

10. Torre, L.A., Siegel, R.L., Jemal, A.: Lung Cancer Statistics. In: Lung Cancer and Personalized Medicine: Current Knowledge and Therapies, pp. 1–19. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-24223-1_1
11. Wei, J.W., Tafe, L.J., Linnik, Y.A., Vaickus, L.J., Tomita, N., Hassanpour, S.: Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific Reports **9**(1), 3358 (Mar 2019). https://doi.org/10.1038/s41598-019-40041-7
12. Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., Wang, Y., Huang, L., Chen, Y., Peng, S., Ke, Z., Li, W.: Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. BMC Medicine **19**(1), 80 (Mar 2021). https://doi.org/10.1186/s12916-021-01953-2
13. Zhao, B., Tan, Y., Tsai, W.Y., Qi, J., Xie, C., Lu, L., Schwartz, L.H.: Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific Reports **6**(1), 23428 (2016). https://doi.org/10.1038/srep23428