Journal of Pathology
J Pathol 2024
Published online 4 March 2024 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.6263

ORIGINAL ARTICLE

AI-guided histopathology predicts brain metastasis in lung cancer patients

Haowen Zhou[1†], Mark Watson[2†], Cory T Bernadt[2], Steven (Siyu) Lin[1], Chieh-yu Lin[2], Jon H Ritter[2], Alexander Wein[2], Simon Mahler[1], Sid Rawal[2], Ramaswamy Govindan[3], Changhuei Yang[1] and Richard J Cote[2*]

[1] *Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA*
[2] *Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO, USA*
[3] *Department of Medicine, Washington University School of Medicine, Saint Louis, MO, USA*

**Correspondence to: RJ Cote, Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO, USA. E-mail: rcote@wustl.edu*

[†]These authors contributed equally to this work.

Abstract

Brain metastases can occur in nearly half of patients with early and locally advanced (stage I–III) non-small cell lung cancer (NSCLC). There are no reliable histopathologic or molecular means to identify those who are likely to develop brain metastases. We sought to determine if deep learning (DL) could be applied to routine H&E-stained primary tumor tissue sections from stage I–III NSCLC patients to predict the development of brain metastasis. Diagnostic slides from 158 patients with stage I–III NSCLC followed for at least 5 years for the development of brain metastases (Met^+, 65 patients) versus no progression (Met^-, 93 patients) were subjected to whole-slide imaging. Three separate iterations were performed by first selecting 118 cases (45 Met^+, 73 Met^-) to train and validate the DL algorithm, while 40 separate cases (20 Met^+, 20 Met^-) were used as the test set. The DL algorithm results were compared to a blinded review by four expert pathologists. The DL-based algorithm was able to distinguish the eventual development of brain metastases with an accuracy of 87% ($p < 0.0001$) compared with an average of 57.3% by the four pathologists and appears to be particularly useful in predicting brain metastases in stage I patients. The DL algorithm appears to focus on a complex set of histologic features. DL-based algorithms using routine H&E-stained slides may identify patients who are likely to develop brain metastases from those who will remain disease free over extended (>5 year) follow-up and may thus be spared systemic therapy.

Keywords: non-small cell lung cancer; deep learning; brain metastasis; digital pathology; artificial intelligence

Received 12 July 2023; Revised 30 November 2023; Accepted 16 January 2024

No conflicts of interest were declared.

Introduction

Non-small cell lung cancer (NSCLC) remains a leading cause of cancer death globally. Despite potentially curative surgery, nearly a third of early-stage (stage I–III) cases will recur with distant metastases [1]. Major advances in the treatment of primary NSCLC with therapeutic agents targeted to specific protein coding ('actionable') mutations and immune checkpoint blockade therapy targeting programmed death-1 (PD-1) or PD-1 ligands have dramatically improved primary outcomes in NSCLC. However, innate and acquired resistance to therapies and disease progression to distant metastatic sites remain a significant cause of morbidity. An increased understanding of tumor biology has suggested that the tumor microenvironment of primary NSCLC may dictate future metastatic behavior [2–4]. Brain metastases, in particular, are a common cause of morbidity and mortality in NSCLC [5]. The stage of the disease is the most commonly used predictor of outcome for NSCLC (and other cancers). However, although stage provides a general risk assessment for a population of patients with similar characteristics, staging is unable to predict which individual patients will or will not progress to metastasis. Histopathologic analysis, even when supplemented by genomic or molecular biomarkers, cannot accurately predict the metastatic potential of NSCLC, particularly in early-stage patients, where risk assessment may lead to impactful treatment decisions [6].

The growing discipline of artificial intelligence (AI), especially in the form of deep learning (DL) networks applied to image analysis, has the potential to identify subtle and complex histopathologic features that may not be appreciated by even the most experienced pathologist, and to correlate these patterns with biologic and

clinical behavior, such as tumor metastatic potential. DL algorithms have been trained to automatically and accurately identify known diagnostic histopathologic features that recapitulate the abilities of pathologists to identify these features (e.g. for the diagnosis of prostate cancer) [7–10]. However, the use of weakly/unsupervised DL to identify features that cannot be recognized by pathologists, such as progression and survival potential based on routine histologic preparations, has been less well explored [11,12]. Attention-based learning has also been utilized to identify subregions of histopathology to identify patterns of highest diagnostic value [13].

Here we demonstrate how a DL network can be effectively trained on digital images from routine H&E-stained NSCLC tumor tissue slides to predict brain metastatic progression within 5 years of the initial diagnosis and, importantly, accurately identify those cases that do not progress after 5 or more years of follow-up. Furthermore, based on the regions of interest (ROI) that most strongly contribute to the DL algorithm's ability to predict progression versus no progression, it appears that the basis of prediction relies on subtle and complex histologic features of tumor cells, non-tumor cells, and the tumor microenvironment.

Table 1. Clinical characteristics of the study population.

	Met^- ($n = 93$)	Met^+ ($n = 65$)
Gender		
Male	47	27
Female	46	38
Average age at diagnosis (years)	60 (47–78)	57 (25–73)
Histology		
Adenocarcinoma	48	44
Squamous cell	32	11
Large cell	3	0
Bronchial alveolar carcinoma	4	0
Poorly differentiated	1	5
Mixed	5	5
Grade		
I	12	4
II	48	26
III	25	27
IV	0	1
No data available	8	7
Stage		
I	85	32
II	3	12
III	0	9
IV	0	7
No data available	5	5
Median follow-up time (months)	106	12.2

Materials and methods

Ethics statement

All procedures related to this study were conducted under an Institutional Review Board-approved protocol, which allowed for the selection of tissue blocks and slides from pre-existing institutional diagnostic material, linkage to non-identifying, limited clinical datasets (where available), and de-identification of all images through an 'honest broker' mechanism.

Patient cohort and whole-slide imaging

This study was based on a cohort of patients with stage I–III NSCLC diagnosed and treated at Washington University School of Medicine with long-term follow-up (>5 years or until metastasis). Of the patients included in the study, 113 had stage I and 41 had higher stage disease (see Table 1 for patient and tumor characteristics). One representative block of tumor tissue from a registry cohort of 198 treatment-naïve NSCLC patients was used to create a fresh H&E slide, which was then scanned at 40× magnification with an Aperio/Leica AT2 slide scanner (Leica Biosystems, Deer Park, IL, USA). All cases were initially subject to blind review to assess tumor adequacy and annotated for ROI by circling an approximate contour of the primary tumor, including the entirety of the tumor microenvironment. Forty cases were initially disqualified as being non-representative or insufficient for adequate evaluation. The clinical characteristics of the remaining 158 cases that were used for this study [65 with known CNS progression (Met^+) and 93 with no recurrence (Met^-)] are summarized in Table 1 and represented diagrammatically in Figure 2. The median time to progression or the follow-up time of these cohorts was 12.2 and 106 months, respectively. To retrieve an adequate number of cases, some heterogeneity in stage and histology features was permitted, although these were generally well represented in both Met^+ and Met^- cases. All cases were coded, and clinical parameters (stage and histology) were unknown to the DL team. Case outcomes were correlated with DL predictions only after training/validation and subsequent testing processes were complete. To compare the ability of expert pathologic assessment of the histopathology to predict progression directly against the DL model, pathologist reviewers were also blinded to outcome and stage data, although they were obviously privy to histologic subtype.

Data and image preprocessing

Figure 1 outlines the image processing algorithm employed for this study. The Otsu thresholding method [14] was implemented to exclude regions of plain glass from the annotated regions in each whole-slide image (WSI); 1,000 image tiles were randomly sampled from the ROI in each WSI, each with 256 × 256 pixels or 130 × 130 μm^2 under 20× magnification, down-sampled from a 512 × 512 pixels 40× image. On average, the sampled image tiles accounted for about 10% of the total ROI in each slide scan. Image tile colors in both the training/validation set and the testing set were normalized to the color statistics of one reference image [15]. For the training sets, data augmentation, including a random crop to a size of 224 × 224 pixels, random flips, and random rotations were performed on the

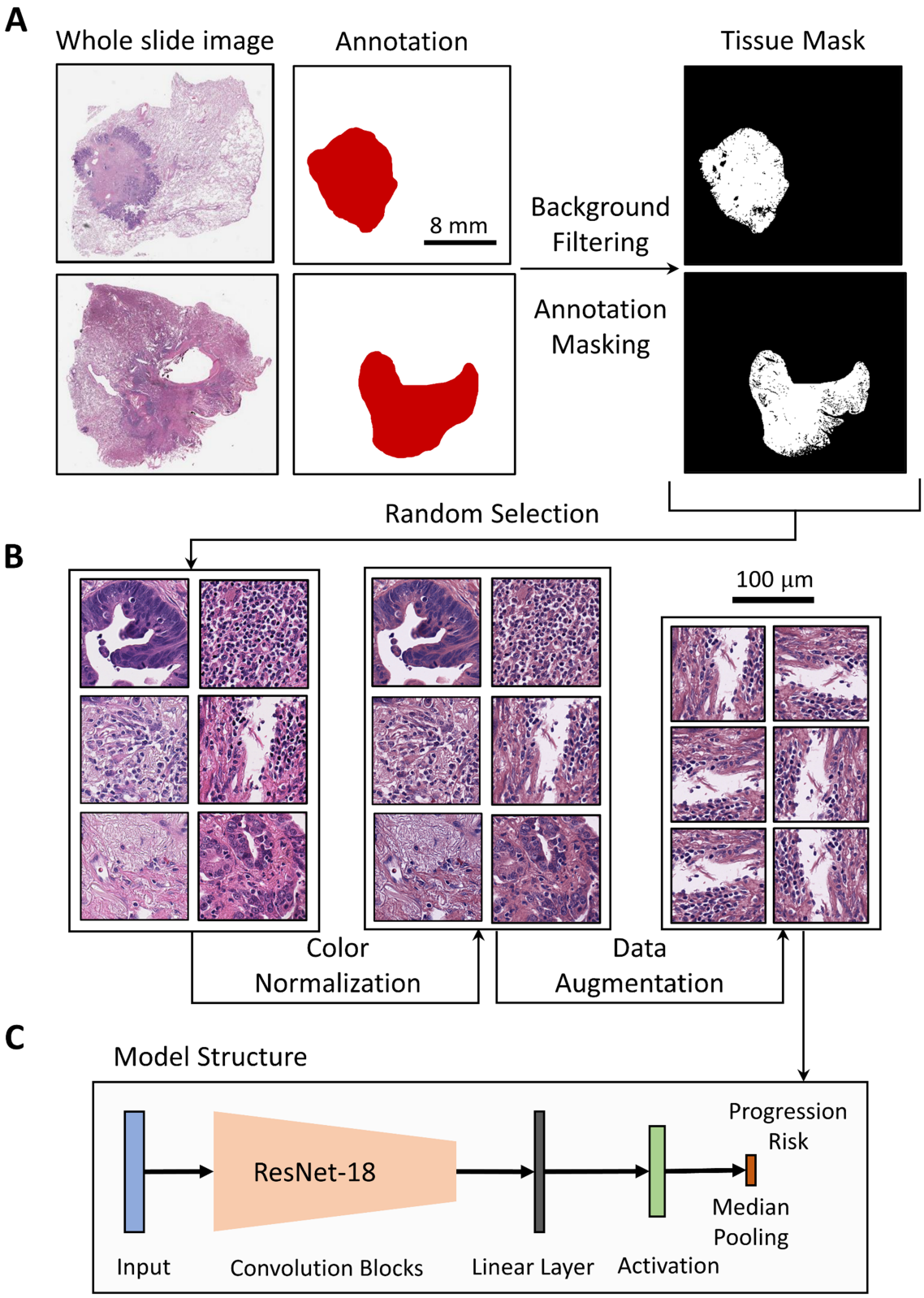

Figure 1. Data processing pipeline. (A) A single, representative H&E-stained slide of a surgically resected primary NSCLC tumor block was obtained from 158 patient cases and scanned at 40× magnification. Each scan file was coded and linked to outcome and pathology data, but blinded to both the DL team and the review pathologists until predictions were finalized. From each whole-slide scan, regions of high tumor cellularity and surrounding tumor microenvironment were annotated by one reviewing pathologist. Regions outside the tumor bed as well as areas of blank glass were masked. (B) One thousand non-overlapping image tiles from the ROI of each scan file were selected at random. Tiles were subjected to color normalization and randomized in cropping and orientation to create a data augmentation step. (C) All tiles in the training set were shuffled and fed to the convolution neural network with the ResNet-18 backbone pretrained on ImageNet, with a linear layer and sigmoid activation for model optimization. In the testing process, the weights in the model were all frozen. A median-pooling function was used to compute the final risk assessment from the collective image tiles of each patient.

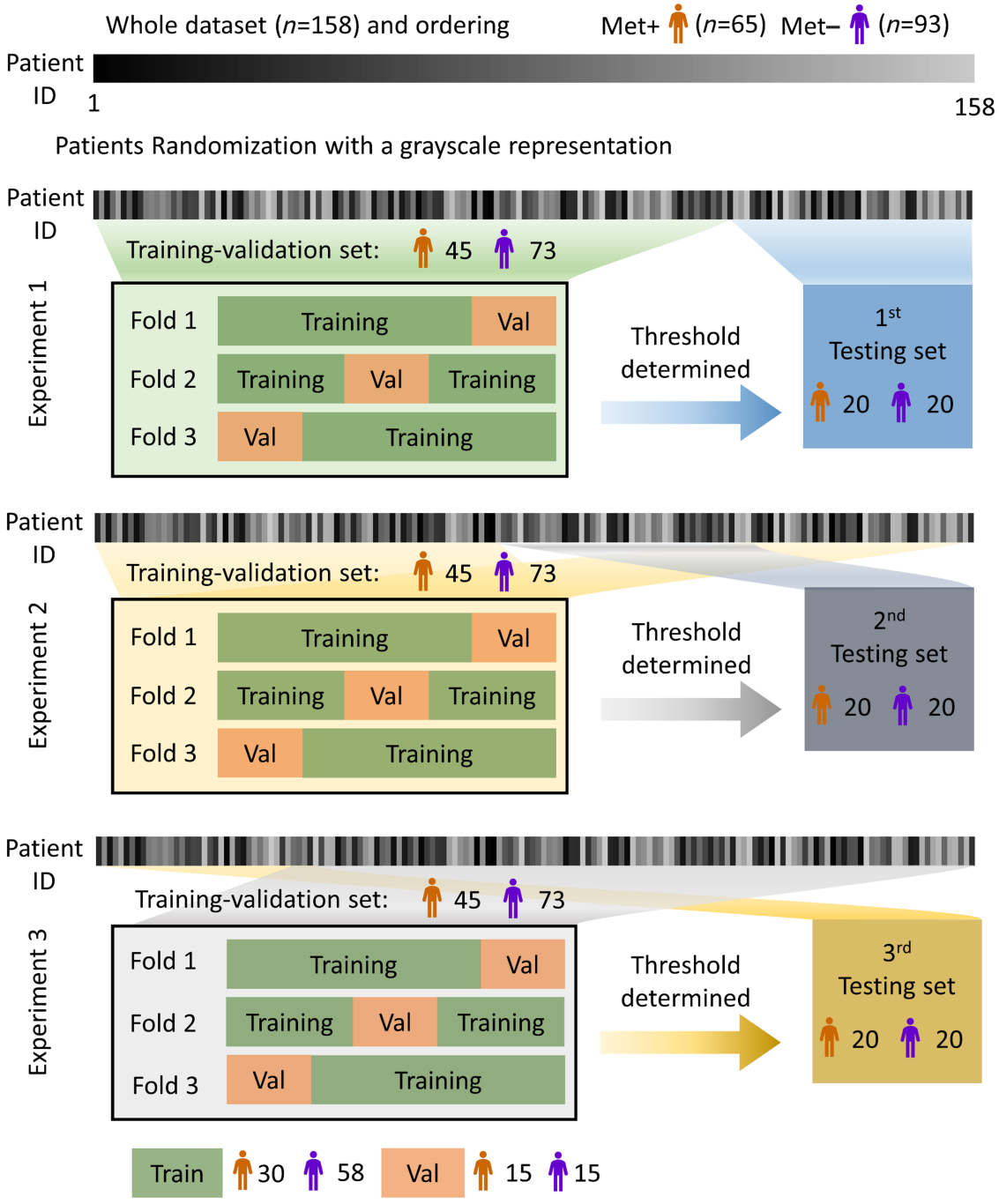

Figure 2. DL study design. The cases (slide images) were arbitrarily coded from 1 to 158 (shown in the top grayscale bar). The cases were randomized (shown in randomized grayscale bars) and split into three different partitions to create experiments 1–3. Each experiment utilized a different training set of 30 Met^+ (orange) and 58 Met^- (purple) tumor images and a validation set of 15 Met^+ and 15 Met^- images. The training/validation was performed using a three-fold cross-validation. Each subsequent set-aside testing set was composed of 20 Met^+ and 20 Met^- case images. The testing sets for experiments 1–3 in total represented ~75% of the entire 158-case cohort.

color-normalized images before using them as input to the DL model. For the validation and testing sets, the color-normalization process was mandatory, but the data augmentation was not needed.

DL study design

The DL model was based on the ResNet-18 convolutional neural network pretrained on the ImageNet dataset [16]. The pretrained weights were taken as the initialization for our model training and all model weights were unfrozen during the training process. A linear layer, attached with a sigmoid activation layer, was used as the classifier and the output was a 'prediction score' for an individual image tile; that is, tiles that the DL assessed as associated with Met^+ versus tiles that the DL assessed as associated with Met^-. The prediction scores of all individual image tiles from each WSI were subjected to a median-pooling layer to produce the final overall progression risk assessment for each slide (case).

To avoid potential bias in the testing set selection from a single experiment, we designed three individual experiments with different training–testing splits (Figure 2). The entire cohort of patients was arbitrarily numbered from 1 to 158 and then randomized. The randomized patient sequence was used to divide the cohort into a training/validation set ($n = 118$; Met^+ $n = 45$, Met^- $n = 73$) and a testing set ($n = 40$; Met^+ $n = 20$, Met^- $n = 20$) in each experiment, where different (overlapping) subsets of patients were selected for the training/validation set and completely different (non-overlapping) patients were selected for the testing set in each experiment (Figure 2).

To perform three rounds of cross-validation training for each experiment, the entire training/validation set (118 patients) was divided into an 88-patient training set (Met^+ $n = 30$, $Met^- = 58$) and a 30-patient validation set (Met^+ $n = 15$, Met^- $n = 15$). In each round, the model was trained on 88,000 image tiles derived from the 88 training cases and validated on 30,000 image tiles derived from the 30 validation cases, with progression outcomes as training and validation labels. The training/validation process was iterated three times, using a different set of 88 cases for training and 30 cases for validation. During the process, validation accuracy was optimized by altering the model hyperparameters (learning rate, batch size, weight decay, number of epochs, and learning scheduler) and the model was then retrained on the entire set using the optimized parameters, before testing the model for progression risk in the independent 40 case test set. The thresholds for converting prediction risk scores to binary outcomes were determined in the validation process (see supplementary material, Figure S1). This same procedure was repeated for the three independent experiments. The preprocessed dataset is publicly available at CaltechData (https://doi.org/10.22002/dw66e-mbs82). The codes are publicly available at GitHub (https://github.com/hwzhou2020/NSCLC_ResNet).

Statistical analysis

To assess the effectiveness of our DL-based classifier in predicting progression risk, we plotted the receiver operating characteristic (ROC) curve; the area under the ROC curve (AUC) was calculated to provide a measure of the overall performance of the model. To compare the model outputs with clinical progression outcomes, we binarized the model prediction scores and reported the accuracy metric. p values were calculated to show the performances of the model compared with those of the pathologists and a random classifier assuming the null-hypothesis of the random classifier.

Results

Pathologist versus DL-based risk prediction

As brain metastasis is a frequent progression event in NSCLC and no clinically useful histologic characteristics or biomarkers exist to predict this behavior, particularly in earlier stage disease, our analyses were focused on a relatively homogenous patient cohort with well-defined and clinically relevant endpoints (Met^+ versus Met^-). The predictive performance of a DL-based classifier was evaluated in three rounds using non-overlapping/separate/distinct patients in the testing set for each round, as described in Materials and methods. The resulting AUC after each training and validation session that utilized different cases was 0.96, 0.98, and 0.95, respectively (Figure 3A) with optimal sensitivities of 80%, 90%, and 95% and specificities of 90%, 95%, and 70%, respectively (see supplementary material, Figure S2 and Table S1). In contrast, a similar training and validation session that employed a random assignment of progression phenotype labels yielded an AUC of 0.51, as might be expected. When the optimal cut-off was chosen from each training and validation session and applied to the separate set of test cases, an average accuracy of 87% was achieved ($p < 0.00001$, compared with accuracy based on random classifier training). Although there are no reliable histopathologic features that can be routinely used to predict metastatic risk progression in NSCLC, we tested the ability of four independent expert lung pathologists to provide a similar binary diagnosis from the identical, blinded set of H&E-stained slide images. Pathologists were provided with WSIs, divided into the same three testing sets as implemented for the DL testing. Compared with the trained DL model, the average accuracy across the three test sets among four pathologists was 55.0%, 60.8%, 54.2%, and 59.2%, respectively, with an average accuracy of 57.3%, not significantly different compared with prediction accuracy based on random classifier training ($p > 0.05$). The individual sensitivity and specificity of prediction from the four pathologists were 57%, 93%, 75%, 75% and 53%, 28%, 33%, 43%, respectively (see supplementary material, Figure S2 and Table S1), suggesting that the DL model identified features that were not readily discernable by a trained pathologist (such as tumor grade, necrosis, lymphocytic infiltration, spread to airway spaces [17]) and that the DL model outperformed careful histologic review by experienced pathologists ($p < 0.001$).

The majority of the patients included in this study had stage I NSCLC (Table 1) and it is in these patients that AI/DL may have the most impact. An analysis of the stage I cases revealed that the specificity of the DL model in predicting Met^- was 95.7% and the negative predictive value was 92.9% (Table 2). That is, of the 47 cases the model predicted Met^-, only two cases resulted in metastasis. The AI was less accurate in predicting Met^+ patients, with a sensitivity of 74.3%, but a positive predictive value of 83.3% (Table 2).

Model understanding

The complexity of DL models can make model interpretation difficult, but the framework employed in this study allowed investigation of the model's attention at tile-level resolution over the WSI. Although the model

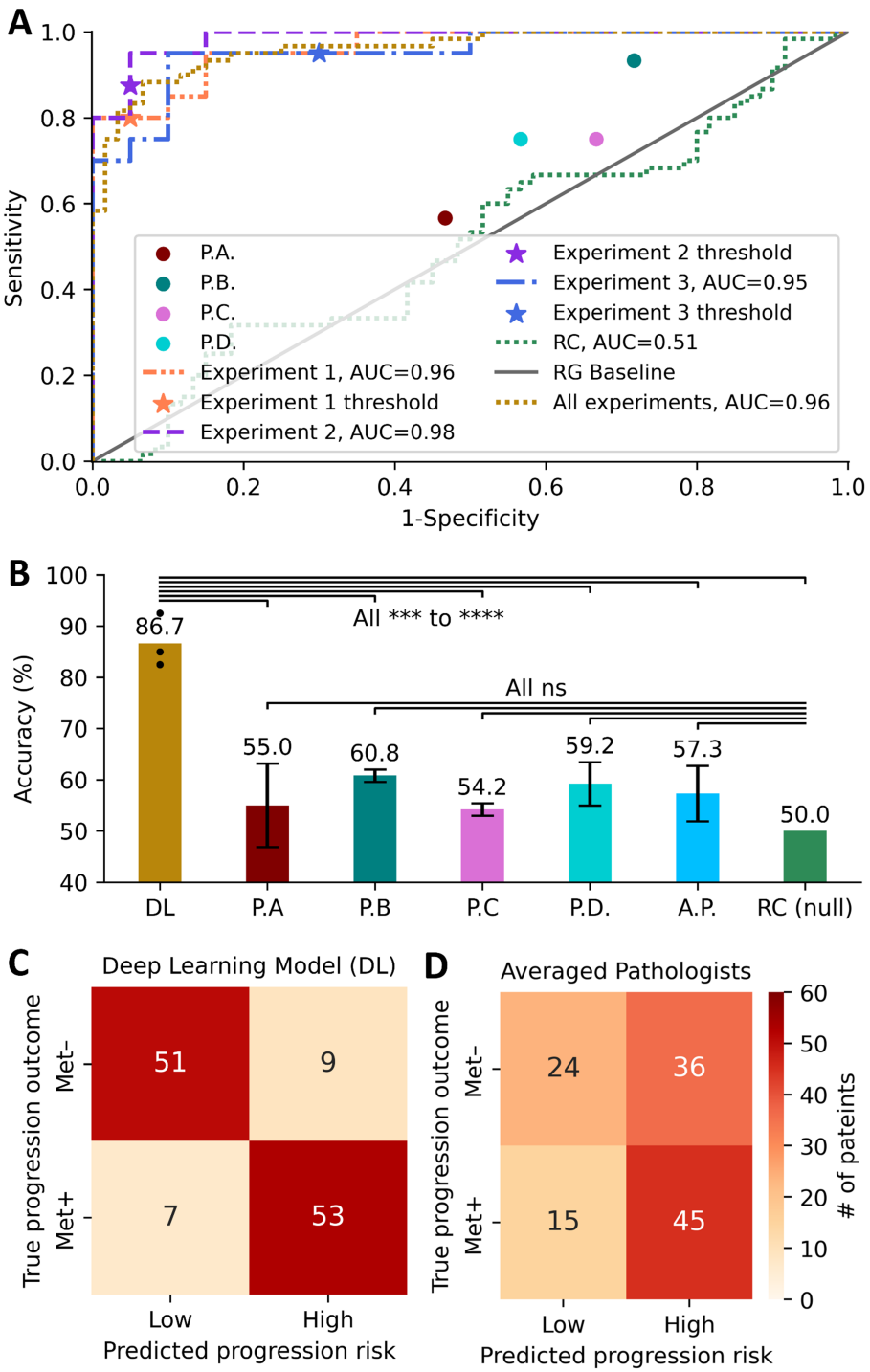

Figure 3. DL versus pathologist prediction of progression. (A) ROC curves generated from three experimental rounds of training and validation using three-fold cross-validation. For comparison, the ROC curve generated from the identical process using random phenotype assignment is shown by the dotted green line. The threshold for calculating testing set prediction accuracies in each experimental session is indicated with a star. For comparison, the sensitivity and specificity for prediction by four independent pathologists are shown as points. (B) The bar plot shows testing accuracies for the DL model, four independent pathologist reviews (PA, PB, PC, PD, and average-AP), and the random classifier (RC) across three experiments. For each pathologist, the predictions were performed using the same slides in three testing sets implemented in the DL model, resulting in three accuracy scores. The error bars represent one standard deviation from the mean value. ns, not significant; ***, $p < 0.0001$; ****, $p < 0.00001$. (C) Confusion matrix of DL model performance combined over three experiments. (D) Confusion matrix of averaged pathologists' evaluations on the three testing sets.

was trained and tested on 1,000 image tiles in annotated regions of primary tumor and the immediately surrounding tumor microenvironment in each case to determine the prediction score for each tile, the prediction scores were aggregated over the entire image to make a slide-level prediction. Figure 4 demonstrates several examples of such 'prediction/attention maps' showing the areas that DL determined as significant for determining outcome for

Table 2. AI prediction of the development of progression (Met^+) versus no progression (Met^-) for stage I patients: AI was able to accurately predict which patients do not progress to brain metastasis, with a specificity of 95.7% and a negative predictive value of 92.9%. The AI was less accurate at predicting those patients that develop metastases, with a sensitivity of 74.3% and a positive predictive value of 83.3%.

Stage I	Low risk	High risk	Row totals
Met^-	45	9	54
Met^+	2	26	28
Column totals	47	35	82

each case, created from cases correctly identified as Met^+ or Met^- by the DL model. Surprisingly, the tumor regions identified by the model as high or low risk had little in the way of discernable histologic differences, and both showed tumor cells, tumor microenvironment, and acellular stromal components. This suggests that the DL training considered a broad and perhaps unappreciated set of histologic features and not (at least not only) the characteristic nuclear or cellular characteristics that are used in traditional tumor grading methods. Moreover, cases correctly identified by the DL model as Met^+ and

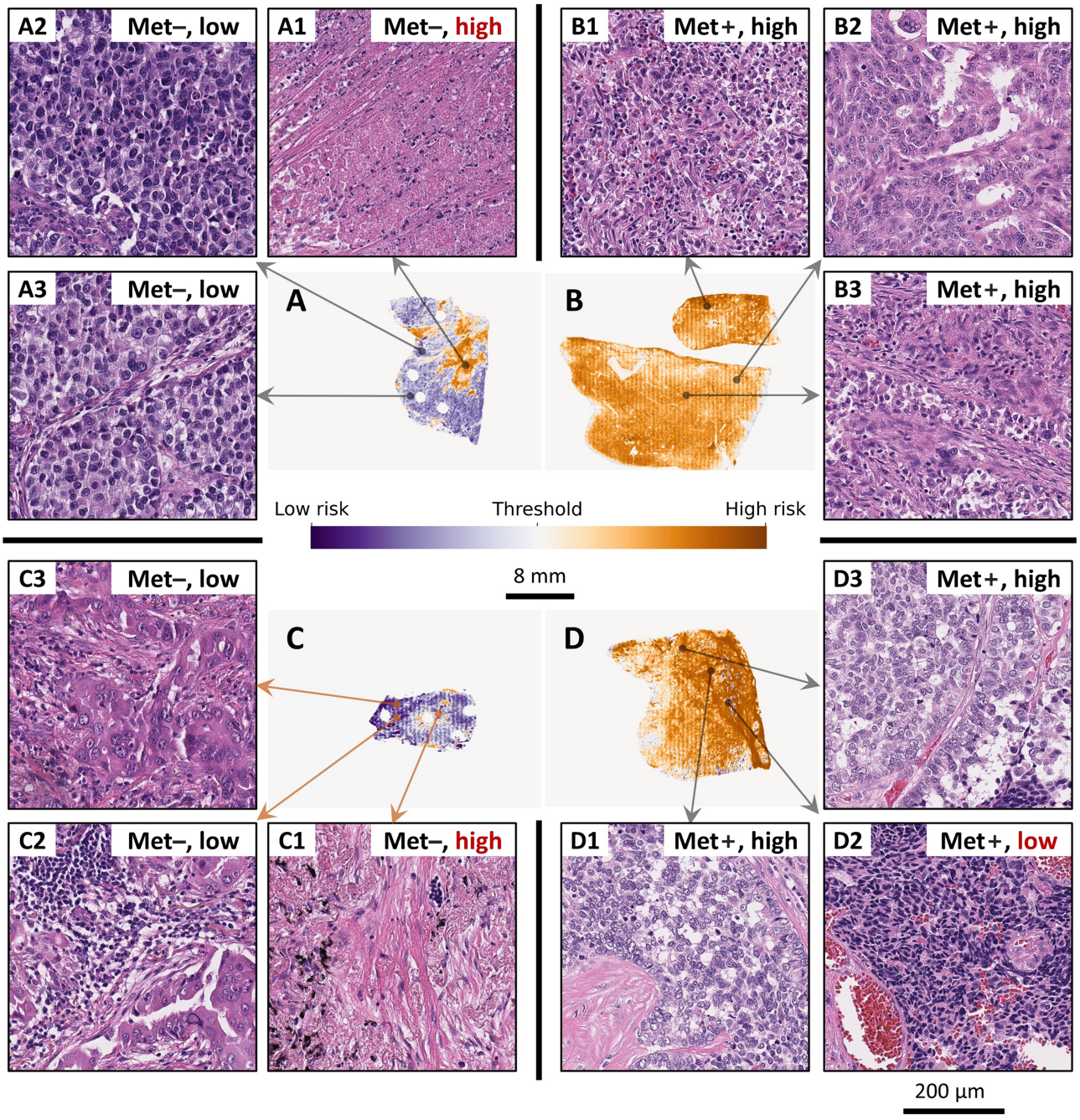

Figure 4. Model prediction maps. (A and C) Prediction maps of non-progression cases (Met^-); (B and D) prediction maps of progression cases (Met^+). Dark orange pixels in the prediction maps indicate image features scored with a high progression risk; deep purple pixels represent features scored as a low progression risk. (A1–A3), (B1–B3), (C1–C3), and (D1–D3) are corresponding high-power images of the H&E-stained regions indicated within each prediction map. (A1), (C1), and (D2) show high-power H&E-stained images of regions with an incorrectly predicted risk.

Met$^-$ did not necessarily demonstrate uniform scores across the individual tiles in the entire tissue section. That is, many cases correctly called as Met$^-$ nonetheless contained elements (tiles) predicted as high risk, and vice versa. It is possible that such incorrect calls were a deficit of the DL model or perhaps, more intriguingly, reflect known tumor heterogeneity with respect to molecular phenotype and metastatic potential.

Discussion

An increasing number of targeted therapies are extending the survival of NSCLC patients. However, the cost and potential adverse effects of such therapies are creating an ever-increasing need for strategies that can stratify the risk of disease progression for optimal patient management, particularly in early-stage patients with no other defined risk factors, the majority of whom are likely cured by surgery alone and may not need to suffer the toxicity and expense of systemic therapy. Routine histopathologic analysis of primary NSCLC cannot provide an accurate assessment of progression risk (as also demonstrated in this study), while mutation analysis and other biomarkers have proven equally unreliable in risk assessment [18].

There have been very few efforts to predict the future development of brain metastases outside of standard clinical and pathologic stage and grade prognosis. A recent example studied 395 patients with all stages of NSCLC who were followed for the subsequent development of brain metastases. A predictive model that included radiomics, histology, lymph node involvement, and tumor grade was shown to have a maximal sensitivity of ~70% and specificity of ~60% for the prediction of subsequent brain metastases, far below our results [19]. Most studies have focused on detecting brain metastases at the earliest possible time through sensitive blood or imaging tests [20–23]. However, these studies do not address the prediction of the development of subsequent brain metastases at the time of initial diagnosis. Ours is the first study to our knowledge that has explored the prediction of future brain metastases based on the histology of the primary tumor in early-stage NSCLC.

In this study, we evaluated the ability of a DL-based analysis of high-quality digital images of routinely stained histological slides of primary NSCLC to predict clinically significant events – the development of brain metastasis versus remaining disease free over a long follow-up (Met$^+$: median = 12.2 months, Met$^-$: median = 106 months). Although the cohort in this study was relatively small, we consistently achieved an accuracy to predict progression of 87% using separate validation and testing sets. Importantly, this was improved over routine assessment by surgical pathologists (87% versus 57% average, $p < 0.001$). Although focused mainly on stage I disease, our cohort had an unavoidable heterogeneity of stages and histologic types. Nonetheless, the algorithm was trained and accurately predicted progression risk regardless of histologic subtype and stage. High performance among mixed histologic tumor types is concordant with the observation that the prediction 'attention map' is not specific to tumor cell elements themselves, but rather is focused on multiple features of the tumor landscape.

In the management of early-stage (I, II, and III) NSCLC, the most difficult decisions are for stage I patients, as these patients have the highest chance of remaining disease free with no or minimal systemic therapy, with 5-year survival rates of 90% to 73% depending on substage [24]. Nevertheless, contemporary management of stage I patients recommends systemic treatment, as studies have shown benefit overall for this group [24–26], which has been enhanced with the advent of targeted and immunotherapies [27–29]. The patients included in the present study were of mixed early-stage NSCLC, but the majority were stage I. Analysis of this group showed that the DL model was highly specific in identifying the patients who do not develop brain (or any other) metastases. On the other hand, the DL model was less effective in identifying patients who develop brain metastases; 25% of such patients predicted to develop brain metastases by the model did not develop metastases of any kind during the course of the follow-up. From a practical point of view, it will be much more important to correctly specify which patients with stage I NSCLC will not develop metastases with high accuracy, as this is the group that might be spared harmful and expensive systemic therapy. A study utilizing a larger cohort restricted to only stage I disease will further establish the importance of this clinically relevant observation. As a retrospective, standard of care patient cohort, it is also recognized that some degree of adjuvant therapeutic diversity existed in this study, which makes the accuracy of metastatic progression prediction in the absence of knowledge of treatment effect even more impressive, and actually may explain why model performance was limited to 87% accuracy. It is likely that, when combined with other relevant clinical data (such as treatment data or other molecular biomarkers) in a multimodal model, our DL algorithm will achieve even higher accuracy.

An important aspect of this study was to explore the potential of DL to elucidate features in histologic images that predict progression; that is, to 'see' features that a human pathologist cannot see or interpret based on conventional training. In this regard, we are beginning to leverage the attention maps generated by our algorithm to understand better the molecular and cellular neighborhood features associated with metastatic progression. We expect that the existing DL model will require iterative refinement with larger, independent cohorts from our own institution, and then from patients treated at multiple sites; we are currently working toward that goal. In addition, we plan to expand this study to represent a wider range of clinically relevant outcomes (e.g. metastasis to other distant organs). Nonetheless, these results demonstrate the potential for AI-guided histopathologic image analysis to

augment and/or exceed traditional histopathology review for risk assessment and improved medical management of early-stage NSCLC patients.

Acknowledgements

This study was supported by U01CA233363 and by the Washington University in St. Louis School of Medicine Personalized Medicine Initiative (RJC). HZ, SL, SM and CY are supported by Sensing to Intelligence (S2I) (grant no. 13520296) and Heritage Research Institute for the Advancement of Medicine and Science at Caltech (grant no. HMRI-15-09-01). MW and RG were supported by the National Cancer Institute (grant no. 5R01CA182746).

Author contributions statement

HZ and MW wrote the first draft of the paper. RJC, CY, HZ and MW conceived the experimental design. HZ, SL, SM and CY performed the DL and data analysis in this study. MW, CTB, SR and RJC designed the clinical and pathologic section of the experiments. CTB, CL, JHR and AW provided the clinical evaluation. MW, RG and RJC provided essential data resources. HZ, MW, CTB, SL, SM, SR, RG, CY and RJC contributed to the writing of this paper.

Data availability statement

The data that support the findings of this study are openly available in CaltechData at https://doi.org/10.22002/dw66e-mbs82. The code for processing the data is publicly available on GitHub at https://github.com/hwzhou2020/NSCLC_ResNet.

References

1. Ganti AK, Klein AB, Cotarla I, *et al*. Update of incidence, prevalence, survival, and initial treatment in patients with non-small cell lung cancer in the US. *JAMA Oncol* 2021; **7:** 1824.
2. Binnewies M, Roberts EW, Kersten K, *et al*. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med* 2018; **24:** 541–550.
3. Shintani Y, Kimura T, Funaki S, *et al*. Therapeutic targeting of cancer-associated fibroblasts in the non-small cell lung cancer tumor microenvironment. *Cancers (Basel)* 2023; **15:** 335.
4. Wood SL, Pernemalm M, Crosbie PA, *et al*. The role of the tumor-microenvironment in lung cancer-metastasis and its relationship to potential therapeutic targets. *Cancer Treat Rev* 2014; **40:** 558–566.
5. Waqar SN, Morgensztern D, Govindan R. Systemic treatment of brain metastases. *Hematol Oncol Clin North Am* 2017; **31:** 157–176.
6. Tsui DCC, Camidge DR, Rusthoven CG. Managing central nervous system spread of lung cancer: the state of the art. *J Clin Oncol* 2022; **40:** 642–660.
7. Tsai P-C, Lee T-H, Kuo K-C, *et al*. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun* 2023; **14:** 2102.
8. Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24:** 1559–1567.
9. Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–1309.
10. Yu K-H, Zhang C, Berry GJ, *et al*. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* 2016; **7:** 12474.
11. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021; **124:** 686–696.
12. Bychkov D, Linder N, Turkki R, *et al*. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep* 2018; **8:** 3395.
13. Chen RJ, Chen C, Li Y, *et al*. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vol. **2022**), 2022; 16123–16134.
14. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979; **9:** 62–66.
15. Vahadane A, Peng T, Sethi A, *et al*. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging* 2016; **35:** 1962–1971.
16. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognoition (CVPR)*, 2016; 770–778.
17. Stögbauer F, Lautizi M, Kriegsmann M, *et al*. Tumour cell budding and spread through air spaces in squamous cell carcinoma of the lung – determination and validation of optimal prognostic cut-offs. *Lung Cancer* 2022; **169:** 1–12.
18. An N, Jing W, Wang H, *et al*. Risk factors for brain metastases in patients with non-small-cell lung cancer. *Cancer Med* 2018; **7:** 6357–6364.
19. Visonà G, Spiller LM, Hahn S, *et al*. Machine-learning-aided prediction of brain metastases development in non-small-cell lung cancers. *Clin Lung Cancer* 2023; **24:** e311–e322.
20. Souza VGP, de Araújo RP, Santesso MR, *et al*. Advances in the molecular landscape of lung cancer brain metastasis. *Cancers (Basel)* 2023; **15:** 722.
21. Yokoi K, Kamiya N, Matsuguma H, *et al*. Detection of brain metastasis in potentially operable non-small cell lung cancer: a comparison of CT and MRI. *Chest* 1999; **115:** 714–719.
22. Choi H, Puvenna V, Brennan C, *et al*. S100B and S100B autoantibody as biomarkers for early detection of brain metastases in lung cancer. *Transl Lung Cancer Res* 2016; **5:** 413–419.
23. Carolan H, Sun AY, Bezjak A, *et al*. Does the incidence and outcome of brain metastases in locally advanced non-small cell lung cancer justify prophylactic cranial irradiation or early detection? *Lung Cancer* 2005; **49:** 109–115.
24. Godoy LA, Chen J, Ma W, *et al*. Emerging precision neoadjuvant systemic therapy for patients with resectable non-small cell lung cancer: current status and perspectives. *Biomark Res* 2023; **11:** 7.
25. Bunn PA. Early-stage NSCLC: the role of radiotherapy and systemic therapy. *J Natl Compr Canc Netw* 2004; **2:** S31–S40.
26. Johnson BE, Rabin MS. Patient subsets benefiting from adjuvant therapy following surgical resection of non-small cell lung cancer. *Clin Cancer Res* 2005; **11:** 5022s–5026s.
27. Wakelee HA, Altorki NK, Zhou C, *et al*. IMpower010: primary results of a phase III global study of atezolizumab versus best supportive care after adjuvant chemotherapy in resected stage IB-IIIA non-small cell lung cancer (NSCLC). *J Clin Oncol* 2021; **39:** 8500.
28. FDA. Approves KEYTRUDA® (pembrolizumab) as adjuvant treatment following surgical resection and platinum-based chemotherapy for patients with Stage IB (T2a ≥4 centimeters), II, or IIIA non-small cell lung cancer (NSCLC). 2023. [Accessed 23 May 2023]. Available from: https://www.businesswire.com/news/home/2023012

7005078/en/FDA-Approves-KEYTRUDA%C2%AE-pembrolizumab-as-Adjuvant-Treatment-Following-Surgical-Resection-and-Platinum-Based-Chemotherapy-for-Patients-With-Stage-IB-T2a-%E2%89%A54-Centimeters-II-or-IIIA-Non-Small-Cell-Lung-Cancer-NSCLC.

29. Weaver CH. Treatment of stage I–IIIA non-small cell lung cancer. CancerConnect 2023. [Accessed 23 May 2023]. Available from: https://news.cancerconnect.com/lung-cancer/treatment-of-stage-i-iiia-non-small-cell-lung-cancer.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Plots of validation accuracy versus threshold cut-offs for each of three rounds (fold) of cross-validation, in each of three experiments

Figure S2. Confusion matrices for pathologist evaluations and DL three-train-test splits

Table S1. Sensitivities and specificities for pathologist evaluations and DL analysis