npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-01267-z

Validation of histopathology-based deep learning algorithms for detection of actionable non-small cell lung cancer biomarkers

Christian Rolfo[1,10] ✉, Efrat Ofek[2,10], Yoash Barak[2,3], Jonathan Weidenfeld[2], Razan Imad Haj[2], Yosef Molchanov[2], Alexander Loebel[2], David R. Braxton[4,5], John S. Cupp[4], Rinat Yacobi[2], Chen Mayer[2], Francesco Drago[1], Camila Avivi[2], Nurit Paz-Yaacov[6], Assaf Avinoam[6], Jonathan Zalach[6], Addie Dvir[6], Inbal Gazy[6], Nir Peled[7], Damien Urban[3,8], Jair Bar[3,8], Dov Hershkoviz[3,9] & Iris Barshack[2,3]

Non-small cell lung cancer (NSCLC) patient management relies on molecular analysis to determine eligibility for targeted therapy. Furthermore, neoadjuvant immunotherapy is primarily suitable in the absence of specific genomic alterations. However, significant challenges remain, including suboptimal molecular testing and patients being assigned to non-optimal treatment strategies. Here, we present AI classifiers for the identification of *EGFR*, *ALK*, *BRAF* and *MET* alterations directly from hematoxylin and eosin (H&E)-stained tissue using CanvOI 1.1, a digital pathology foundation model. Their performance was evaluated on an independent validation dataset of 968 NSCLC samples. The classifiers achieved AUCs of 0.87 for *EGFR*, 0.96 for *ALK*, 0.88 for *BRAF* and 0.83 for *MET*. Moreover, they demonstrated high accuracy in identifying cases lacking alterations. Our results highlight the potential of deep-learning tools for the detection of NSCLC biomarkers and specifically the identification of tumors without *EGFR* or *ALK* driver alterations, supporting more informed clinical decision-making.

The treatment of non-small cell lung cancer (NSCLC), which accounts for 80%-90% of lung cancers, relies on molecular analysis to identify the most appropriate therapeutic strategies for individual patients[1,2]. With the introduction of targeted therapies associated with a growing number of actionable genes, response rates and survival in patients have dramatically improved[3]. This progress has made biomarker testing an essential prerequisite for guiding optimal first-line systemic therapy and with the recent introduction of adjuvant *EGFR* and *ALK* targeted therapies, biomarker testing is now necessary in early-stage disease as well[4,5]. Furthermore, immunotherapy is used for the treatment of advanced NSCLC as well as neoadjuvant treatment in early-stage and locally advanced disease, primarily in cases lacking specific genomic alterations, such as in *EGFR* and *ALK*[6]. Thus, molecular diagnostics have a critical role in the management of NSCLC, identifying both the presence and absence of driver alterations, ensuring the most appropriate and effective personalized treatment.

Professional guidelines recommend molecular testing for all patients with advanced or metastatic non-squamous NSCLC, including testing for alterations in *EGFR*, *ALK*, *ROS1*, *BRAF*, *MET*, *ERBB2*, *KRAS*, *RET*, and *NTRK*1, *NTRK*2, and *NTRK*3 fusions[7]. Given the lower, but notable, occurrence of driver alterations (2-10%) in squamous cell carcinoma, it is also recommended to consider molecular testing in these cases. Comprehensive genomic testing of tumor tissue is the preferred testing approach,

[1]Division of Medical Oncology, Department of Internal Medicine, The Arthur G. James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [2]Institute of Pathology, Sheba Medical Center, Tel Aviv, Israel. [3]Faculty of Medicine and Health Sciences, Tel Aviv University, Tel Aviv, Israel. [4]Department of Pathology, Hoag Memorial Hospital Presbyterian, Newport Beach, CA, USA. [5]Hoag Family Cancer Institute, Newport Beach, CA, USA. [6]Imagene AI Ltd, Tel Aviv, Israel. [7]Helmsley Cancer Institute, Shaare Zedek Medical Center, Jerusalem, Israel. [8]Jusidman Cancer Center, Sheba Medical Center, Tel Aviv, Israel. [9]Department of Pathology, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [10]These authors contributed equally: Christian Rolfo, Efrat Ofek. This work was performed during the stated affiliation: Nurit Paz-Yaacov, Razan Imad Haj. ✉e-mail: Christian.Rolfo@osumc.edu

especially in advanced stages. However, in certain circumstances, where critical treatment decisions are necessary, consideration of gene-specific rapid testing is advised while waiting for the results of broad-based genomic testing[7]. In addition, cell-free DNA analyses can complement tissue testing for tumor genotyping[8].

Despite the advancements in diagnosing and treating NSCLC, significant challenges remain in delivering optimal care to patients. Real-world data reveal unsatisfactory molecular testing rates and that a considerable number of patients do not receive the most appropriate first-line treatment, if at all[9–12]. These gaps in practice stem from several issues, among them an incomplete or an entire lack of testing, particularly for rarer biomarkers. Studies indicate that approximately 40-50% of patients are not tested with comprehensive NGS[10,11,13], and over 10% receive no biomarker testing at all[9–11]. Furthermore, prolonged turnaround times contribute to first-line treatment initiation without comprehensive molecular test results in over 20% of patients[11,12] which results in shorter time to therapy discontinuation, lower use of matched targeted therapies, and higher use of ineffective immunotherapy-based therapies[14].

These findings underscore the need for complementary tools to ensure broader patient access to tailored cancer therapies and more appropriate patient management. Artificial intelligence (AI) technologies are emerging as transformative tools in digital pathology, offering rapid, scalable, and accessible solutions, with the potential to enhance current methods and workflows and improve clinicians' decision-making capabilities[15,16].

Leveraging AI to analyze tumor histology images, which are central to any cancer diagnosis, can facilitate the detection of genomic alterations directly from the hematoxylin and eosin (H&E)-stained tissue images, providing rapid molecular information, potentially weeks earlier than the current standard practice. Indeed, emerging technologies, such as deep learning pipelines for analyzing digitized whole slide images (WSIs), are at the forefront of this transformation[17,18]. Numerous studies have demonstrated AI's potential in detecting various biomarkers from digitized WSIs. Specifically, Saillard et al.[19] and Hu et al.[20] describe deep learning tools developed for the detection of Microsatellite Instability (MSI) in colorectal or prostate cancer WSIs, respectively. Esteva et al.[21] describe a deep learning tool for predicting long-term outcomes in prostate cancer from histopathology slides, with their multimodal AI models showing performance ranging from 0.65 to 0.83 AUC in various outcome measures, outperforming conventional risk-stratification tools. Based on this work, a Prostate test was included in the NCCN guidelines with level IB evidence for predicting therapy benefits and long-term outcomes. This incorporation into the guidelines is a promising step toward integrating AI tests into clinical workflows. In a previous study, our team demonstrated the potential of deep learning solutions for NSCLC biomarker detection using a single-site cohort of scanned H&E-stained diagnostic slides[22].

To further explore the potential of deep-learning algorithms in predicting NSCLC biomarker status directly from H&E-stained histology slides, we expanded our approach in the current study by developing deep-learning models leveraging CanvOI 1.1, our state-of-the-art Oncology Intelligence foundation model and a large training dataset of over 4000 NSCLC samples from multiple medical centers. Here, we present our work on the development and validation of AI classifiers trained to predict four NSCLC biomarkers, *EGFR*, *ALK*, *BRAF*, and *MET*. The performance of these models was validated on an independent external cohort from two international tertiary-care centers.

Results

Lung cancer biomarker-classifiers development

We developed classifier models leveraging CanvOI 1.1 (detailed in the "Methods" section and Fig. 1) to detect actionable alterations of four NSCLC biomarkers: *EGFR*, *ALK*, *BRAF* and *MET*. We focused on *EGFR* and *ALK*, given their clinical relevance as well-established actionable biomarkers. While these alterations together account for nearly half of actionable NSCLC cases in Asian populations[2], their combined prevalence in Western cohorts is lower, approximately 38% among cases with targetable alterations. Testing for these genes is recommended in both early and late cancer stages, as they guide treatment decisions of both targeted therapy and immunotherapy throughout the disease's progression. Notably, *EGFR* actionable alterations include different mutation types, such as single-nucleotide variants (SNVs) as well as insertions/deletions, while *ALK* driver alterations involve genomic rearrangements. To further explore our algorithms' limits and potential, we included *BRAF* and *MET* in this study. For *BRAF*, we developed a model to specifically identify the V600E actionable alteration (and not other oncogenic alterations found in NSCLC), and for *MET*, exon 14 skipping alterations.

Validation of the biomarker classifiers on an independent test set

The biomarker classifiers were evaluated on an independent dataset, consisting of samples collected from medical centers not included in the model development. This dataset included 993 WSIs of H&E-stained NSCLC tissue from unique cases collected from Sheba Medical Center, Israel, and Hoag Memorial Hospital Presbyterian, USA. WSI file integrity check was performed, as well as a pre-processing step to identify images unsuitable for analysis. Slides lacking sufficient analyzable cancerous cells, whether due to technical artifacts, scanning issues, or insufficient tumor tissue, were excluded from further analysis. This step ensured that only appropriate slides proceeded to analysis and were assigned a prediction score. Two (0.2%) slides did not contain the magnification in the WSI data and thus could not be analyzed. An additional 23 (2.3%) slides were excluded from analysis during the pre-processing step that evaluates whether there are sufficient analyzable cancerous cells. As a result, a total of 968 WSIs were included in the final test cohort, 510 from Sheba Medical Center and 458 from Hoag Hospital (sample characteristics are presented in Table 1). Of the 968 samples, 955 samples had a reference method status for *EGFR*, 942 for *ALK*, 948 for *BRAF* and 558 for *MET* respectively. Out of the samples with a reference method result from Sheba Medical Center, 26.4% ($n = 131$) were *EGFR*-positive, 5.7% ($n = 28$) *ALK*-positive, 0.8% ($n = 4$) *BRAF* V600E-positive, and 2.7% ($n = 10$) were *MET* exon 14 skipping positive. One case harbored both an *EGFR* exon 20 insertion mutation and *ALK*-rearrangement. Of the samples obtained from Hoag Hospital with a reference method result, 24.7% ($n = 113$) were *EGFR*-positive, 1.3% ($n = 6$) *ALK*-positive, 0.9% ($n = 4$) *BRAF* V600E-positive, and 8.8% ($n = 16$) were *MET* exon 14 skipping positive.

WSIs were analyzed by the *EGFR*, *ALK*, *BRAF*, and *MET* classifiers, and each sample in the respective biomarker cohort was assigned a label according to a 3-tier classification. For each biomarker, samples with scores above the defined upper threshold were classified as likely positive, those with scores below the defined lower threshold were classified as negative, and samples with scores between the two thresholds were assigned an intermediate status. For *EGFR*, 33.5% of the cohort was classified as either likely positive or negative, while for *ALK* 69.5%, for *BRAF* 42.3% and for *MET* 43.2% of the cases were thus classified. A confusion matrix comparing the AI classifier's predictions with the reference method results is displayed in Fig. 2A. The biomarker classifiers achieved area under receiver operating characteristic curve (AUC) values of 0.87 for *EGFR*, 0.96 for *ALK*, 0.88 for *BRAF* and 0.83 for *MET* (Fig. 2B). The *EGFR* classifier demonstrated slightly better results for common mutations (exon 19 deletions and L858R) compared to uncommon mutations (exon 20 insertions, L861Q, G719X, and S768I), with AUCs of 0.89 and 0.81, respectively (Fig. S1). Additional subgroup analyses stratified by specimen type (biopsy vs. surgical specimens) and collection site (primary *vs.* metastatic tumors) showed generally consistent model performance across. However, due to the limited number of positive cases in some subgroups, these results should be interpreted with caution (Fig. S2).

Across all genes, the AI negative group consisted almost entirely of true negative cases (Figs. 3A, 4A, 5A and 6A) with a negative predictive value (NPV) of 98.6% (95% CI: 96.0–99.7%) for the *EGFR* classifier, 99.7% (95% CI: 98.8–100.0%) for the *ALK* classifier, 100.0% (95% CI: 99.1–100.0%) for the *BRAF* classifier and 99.6% (95% CI: 97.6–100.0%) for the *MET* classifier. These rates were due to 3 false negative (FN) *EGFR* cases, 2 FN *ALK* cases

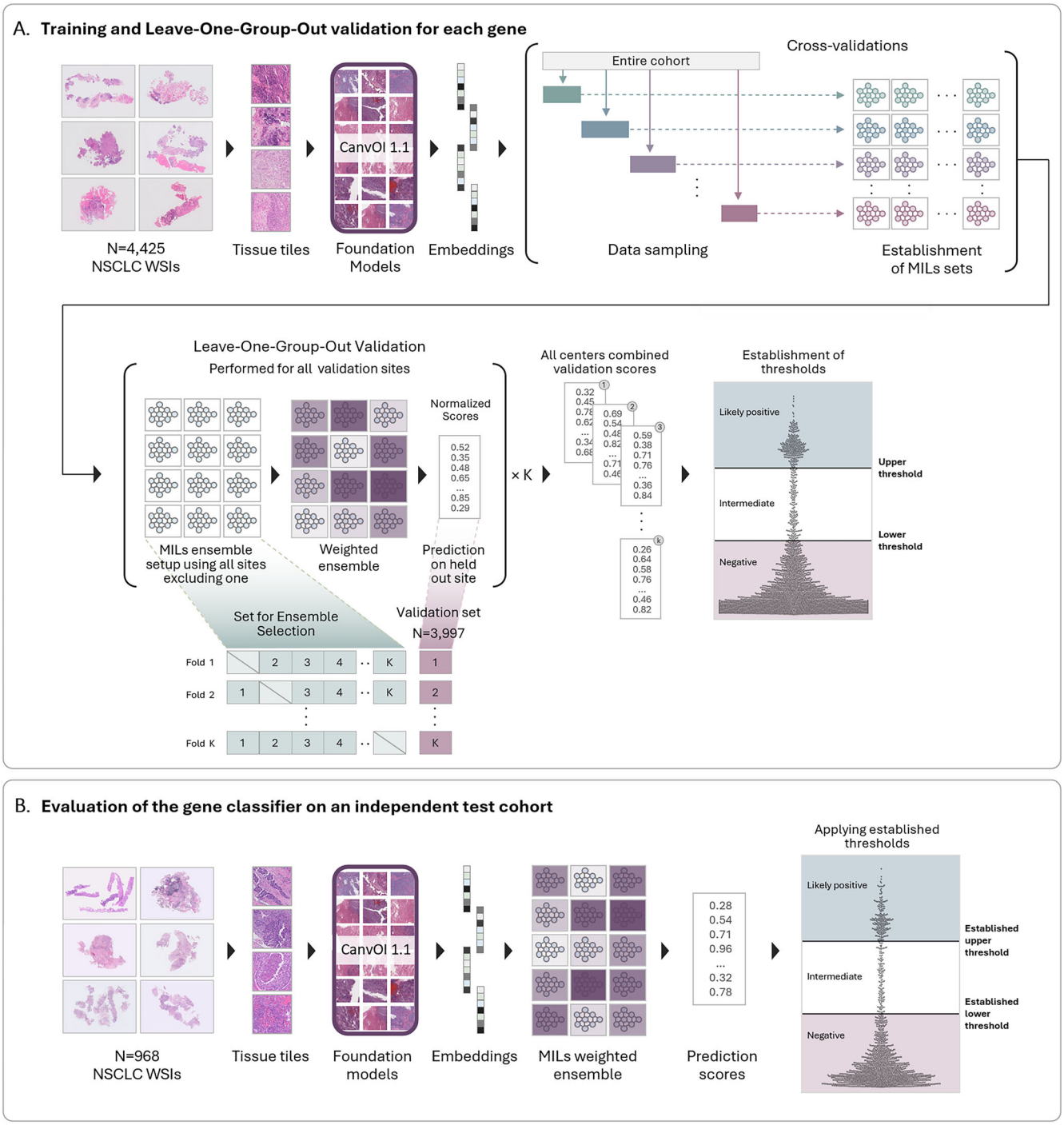

Fig. 1 | Schematic overview of the development and external validation workflow for the biomarker classifiers. The illustration represents the process applied to each of the gene's classifiers. **A** Embeddings were extracted from image tiles of 4425 NSCLC WSIs by ConvOI 1.1. These embeddings served as inputs for training using a cross-validation scheme and Multiple Instance Learning (MIL) algorithms. In each cross-validation training iteration, embeddings from a single foundation model were used, with different foundation model embeddings employed in separate iterations. This process established the set of MIL algorithms. Next, a Leave-One-group-Out validation step was conducted to evaluate the classifier's performance and define thresholds. During this step, 3997 WSIs were utilized (including only sites with sufficient positives for the performance evaluations). In each iteration of the Leave-One-group-Out step, a weighted ensemble was established and applied for biomarker prediction on the held-out site. Biomarker classifier scores from all samples were combined, and upper and lower thresholds were defined (as detailed in the "Methods" section). This process produced a three-tier classification system comprising Likely Positive, Intermediate, and Negative classes. **B** External validation was performed on an independent test set comprising 968 WSIs from Sheba Medical Center and Hoag Hospital. Thresholds defined and fixed during the development phase were applied on the test set.

and 1 FN *MET* case with no *BRAF* FN cases. However, the proportion of negatives within the population included in the AI negative group varied, ranging from 68.4% (621 out of 908) for *ALK*, to 42.1% (396 out of 940) for *BRAF*, 42.9% (228 out of 532) for *MET*, and 29.7% (211 out of 711) for *EGFR*. To evaluate the reliability of the AI negative outcome, we assessed the sensitivity by calculating the Positive Percent Agreement (PPA) as follows. For this analysis, the AI model's three-class outputs were consolidated into two groups: samples classified as either likely positive or intermediate were

Table 1 | Test set characteristics

	All (n = 968)	Sheba (n = 510)	Hoag (n = 458)
Age (years), median (range)[a]	68 (19–>89)	71 (19–>89)	62 (21–>89)
<65 yr, *n* (%)	387 (40.0)	124 (24.3)	263 (57.4)
≥65 yr, *n* (%)	581 (60.0)	386 (75.7)	195 (42.6)
Sex, *n* (%)			
Female	507 (52.4)	239 (46.9)	268 (58.5)
Male	461 (47.6)	271 (53.1)	190 (41.5)
Biopsy site			
Primary	757 (78.2)	410 (80.4)	347 (75.8)
Metastatic	206 (21.3)	98 (19.2)	108 (23.6)
N/A	5 (0.5)	2 (0.4)	3 (0.7)
Procedure type			
Biopsy	733 (75.7)	432 (84.7)	301 (65.7)
Surgical	215 (22.2)	70 (13.7)	145 (31.7)
N/A	20 (2.1)	8 (1.6)	12 (2.6)
Tobacco use history, *n* (%)			
Current	103 (10.6)	103 (20.2)	N/A
Former	288 (29.8)	288 (56.5)	N/A
Never	95 (9.8)	95 (18.6)	N/A
N/A	482 (49.8)	24 (4.7)	N/A
NSCLC subtype, *n* (%)			
Non-squamous	824 (85.1)[b]	426 (83.5)[b]	398 (86.9)[b]
Squamous	112 (11.6)[c]	69 (13.5)[c]	43 (9.4)
N/A	32 (3.3)	15 (2.9)	17 (3.7)
Driver oncogene (%)[d]			
EGFR	244 (25.5)[e]	131 (26.4)[e]	113 (24.7)
Variant, *n* (% of mut)			
Ex 19 del, *n*	109 (44.7)	60 (45.8)	49 (43.4)
L858R	89 (36.5)	49 (37.4)	40 (35.4)
Ex 20 ins	19 (7.8)	11 (8.4)	8 (7.1)
L861Q/ G719X/S768I	27 (11.1)	11 (8.4)	16 (14.2)
ALK	34 (3.6)[e]	28 (5.7)[e]	6 (1.3)
BRAF V600E	8 (0.8)	4 (0.8)	4 (0.9)
MET ex14 skipping	26 (4.7)	10 (2.7)	16 (8.8)

[a]Age at procedure.
[b]Including adenosquamous; Sheba n = 10, Hoag n = 15.
[c]Including 1 case of Squamous cell carcinoma with mucoepidermoid.
[d]Biomarker status N/A: Sheba – *EGFR* n = 13, *ALK* n = 17, *BRAF* n = 20, *MET* n = 134; Hoag – *ALK* n = 9, *MET* n = 276. The Biomarker positive percentage was calculated from all samples with a molecular result.
[e]1 case carried both *EGFR* exon 20 insertion mutation and *ALK*-rearrangement.

combined into a single class, while the negative class remained separate. This approach allowed us to focus on the rate of FNs among all positive cases in the cohort. The calculated PPA values were 98.8% (95% CI: 96.4–99.7%) for *EGFR*, 94.1% (95% CI: 80.3–99.3%) for *ALK*, 100.0% (95% CI: 63.1–100.0%) for *BRAF*, and 96.2% (95% CI: 80.4–99.9%) for *MET*.

When examining the rates of reference method positivity across the AI groups, all genes showed a marked decrease in the proportion of positives from the AI likely-positive group to the intermediate and negative groups. For all genes, the AI negative groups included only 0–3 positive cases (i.e., FNs) (Figs. 3B, 4B, 5B and 6B). However, the proportion of positives within the AI positive groups varied widely from 79.25% for *EGFR*, 68.75% for *ALK*, and 50% for *MET* to 20% for *BRAF* with 22, 10, 6, and 4 false positives (FP) for each gene, respectively.

We next analyzed the probability distributions for *EGFR*, *ALK*, *BRAF*, and *MET*. The box plots in Figs. 3C, 4C, 5C and 6C present the probability plots for each biomarker, showing the AI classifier score percentiles across the 3-tier classification system for both reference method gene-positive and gene-negative groups. The predictiveness curves (Figs. 3D, 4D, 5D and 6D) present the risk of positivity for a biomarker as a function of the biomarker AI classifier score. These plots provide a clear visual distinction between the groups, and an increase in the risk for carrying an alteration with an increase in the AI classifiers prediction scores. Notably, intermediate cases are still assigned a probability score, which, despite lacking a definitive classification, could provide timely and meaningful information when considered alongside other relevant clinical factors.

Discussion

In this study, we present the development and validation of AI classifiers to detect actionable alterations in NSCLC biomarkers. Our study aligns with the growing body of work in digital pathology, where machine learning models are increasingly being developed to predict molecular biomarkers directly from H&E-stained slide images, and specifically NSCLC biomarkers, including *EGFR* mutations and *ALK* and *ROS1* rearrangements[23–26]. Our aim was to develop robust NSCLC biomarker classifiers that maintain consistent performance across diverse samples. Thus, we utilized CanvOI 1.1, a strong foundation model pre-trained on an extensive dataset of unlabeled histological images. This foundation model captures complex patterns and features within histological data, thereby supporting the development of more robust and adaptable downstream models[27]. Alongside leveraging the foundation model, we also trained our task-specific MIL classifiers on a large and variable cohort, incorporating NSCLC samples from multiple data sources. This diverse dataset, together with the Leave-One-Group-Out cross-validation strategy, aimed to further support the development of models that could generalize effectively across different conditions while maintaining their performance. The resulting classifiers were evaluated on an independent dataset comprising samples from two distinct medical centers, with performance assessed on the combined cohort. A single threshold was defined for each biomarker classifier during model development and was consistently applied for the Leave-One-Group-Out validation as well as across both centers in the external test set. This evaluation ensures that we select models that can operate robustly across different medical centers, accounting for their inherent variability. Classifiers for four NSCLC biomarkers were developed. We included *EGFR* and *ALK*, two key driver oncogenes, influencing treatment decisions at both early- and late-stages of the disease. *BRAF* V600E was added to assess our model's ability to differentiate between specific alterations, as V600E is the only actionable *BRAF* mutation in NSCLC, though other non-actionable alterations exist. Finally, *MET* exon 14 skipping was incorporated to cover all the actionable genomic alteration types in NSCLC, i.e., mutations, fusions, and exon skipping. To evaluate the NSCLC biomarker-classifiers performance, we used a large external retrospective test set including 968 NSCLC samples, diagnosed at two medical centers from Israel and the USA. Consistent with the known distribution of NSCLC driver mutations[7], actionable alterations in our test cohort were observed almost exclusively in adenocarcinomas, with only occasional cases annotated as squamous cell carcinomas. Of note, this study focused on evaluating the models' performance in comparison to conventional molecular testing, while prospective studies will be necessary to establish the clinical utility of these predictions.

Integrating AI biomarker classifiers early in the diagnostic process has the potential to provide timely and relevant insights into the likelihood of the presence or absence of targetable alterations. These tools could prioritize subsequent testing, streamline time management, optimize tissue use and reduce the need for additional procedures. Moreover, a rapid and accessible tool for identifying actionable oncogenic driver mutations in NSCLC could significantly impact therapy selection, leading to better clinical outcomes by

A

		Reference method							
		EGFR		*ALK*		*BRAF*		*MET*	
		Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
AI	Negative	211	3	621	2	396	0	228	1
	Intermediate	478	157	277	10	540	7	298	19
	Likely positive	22	84	10	22	4	1	6	6
	Total	711	244	908	34	940	8	532	26

B

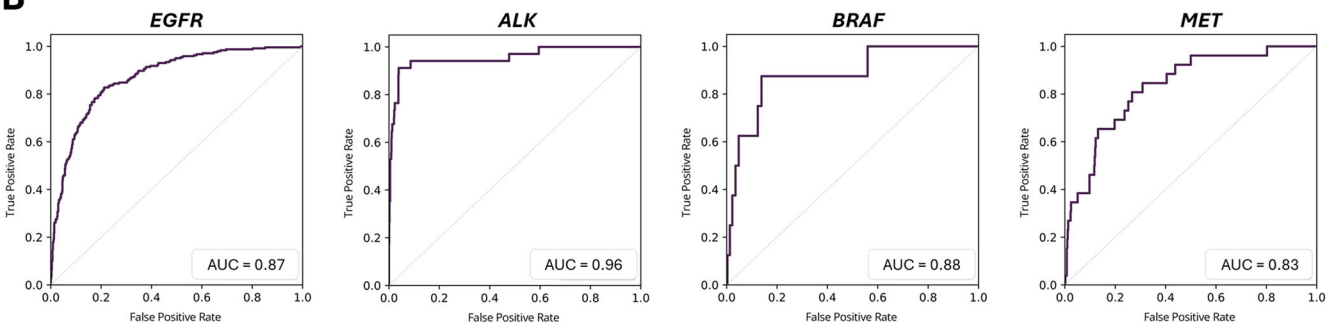

Fig. 2 | NSCLC biomarkers performance. A Summary of concordance for biomarkers classification between the AI classifiers and the conventional reference methods. **B** Receiver operating characteristic (ROC) curves, and the calculated Area under the curve (AUC) for *EGFR*, *ALK*, *BRAF*, and *MET* biomarker classifiers.

enabling treatments tailored to specific molecular profiles. Importantly, once H&E slides are digitized, the inference time of the model is measured in minutes rather than the several days or even weeks often required for molecular assays. Although this approach is not yet positioned to replace molecular testing in clinical practice, it has the potential to support patient management particularly in light of challenges such as undertesting and initiation of treatment without comprehensive molecular test results[10–13]. Moreover, as the analysis is performed directly on routine digitized H&E slides without additional tissue consumption, the method provides a clear advantage in contexts of limited material, such as small biopsies, where conventional molecular testing may deplete the specimen, and the AI tool can help guide tissue utilization decisions. A prospective clinical study will be required to evaluate the feasibility and clinical utility of this approach in real-world practice.

Despite professional guidelines recommending biomarker testing, implementation in real-world clinical settings remains suboptimal[9–11,13]. This gap often results in patients receiving therapies without all the recommended biomarker test results, leading to suboptimal treatment decisions. A data-driven AI tool, when combined with other patient variables such as clinical and risk factors and additional test results, can enhance the clinician's ability to guide patient care and ultimately support more informed decision-making.

The classifiers for *EGFR*, *ALK*, *BRAF*, and *MET* presented in this study demonstrated high accuracy in identifying negative cases that lack actionable alterations in these genes. However, the percentage of positives within the likely positive samples varied, ranging from 20% of *BRAF* likely positives confirmed to carry a V600E mutation by conventional methods, to 50% for *MET* likely positives, 68.8% for *ALK* and 79.3% for *EGFR*.

Healthcare providers managing NSCLC patients could benefit significantly from integrating such an AI solution, especially when making decisions related to immunotherapy. In early-stage disease, quickly identifying patients with tumors lacking *EGFR* or *ALK* driver alterations is crucial for guiding neoadjuvant immunotherapy, which has been shown to improve event-free survival and potentially downstage tumors previously considered unresectable, making surgery a viable option[28]. These patients require quick characterization of these oncogenes and often have small quantities of biopsied tissue. Thus, a highly accurate tool that can provide rapid results and does not require any dedicated tissue can enhance the clinical workflow by providing information on the appropriateness of neoadjuvant immunotherapy.

In advanced NSCLC, first-line immunotherapy is the standard of care for tumors without driver mutations, while targeted therapies are preferred for tumors with actionable alterations in genes such as *EGFR*, *ALK*, *BRAF*, and *MET*, given their higher response rates. Rapid identification of patients eligible for immunotherapy or those with actionable driver mutations is crucial for ensuring timely and appropriate treatment decisions. An AI tool capable of flagging potential cases with a higher likelihood of driver alterations, especially when the prevalence of these mutations is low, would be highly beneficial. This tool would enable clinicians to identify patients who may benefit from a gene-specific rapid test or highlight cases where comprehensive testing results could be especially informative, in line with professional guideline recommendations.

Our study has several limitations. First, the performance evaluation of the *BRAF* and *MET* classifiers was limited by the small number of positive samples available. Our research did not include KRAS mutations or less frequent fusions such as RET and ROS1, mainly due to the limited number of available cases, which precludes robust algorithm training. While KRAS G12C is one of the most prevalent oncogenic drivers in NSCLC, the development of an AI-based prediction model

A

AI negative	Ref. method negative	NPV	Adjusted NPV*
214	211	98.6% (95% CI: 96.0-99.7%)	99.3% (95% CI: 97.8-99.8%)

* NPV adjusted based on an assumed prevalence of 15% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

B

		AI	Ref. method positive (P)	% P / AI group	PPA**
AI classification	Likely positive	106	84	79.25%	98.8% (95% CI: 96.4-99.7%)
	Intermediate	635	157	24.72%	
	Negative	214	3	1.40%	

** PPA calculated using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

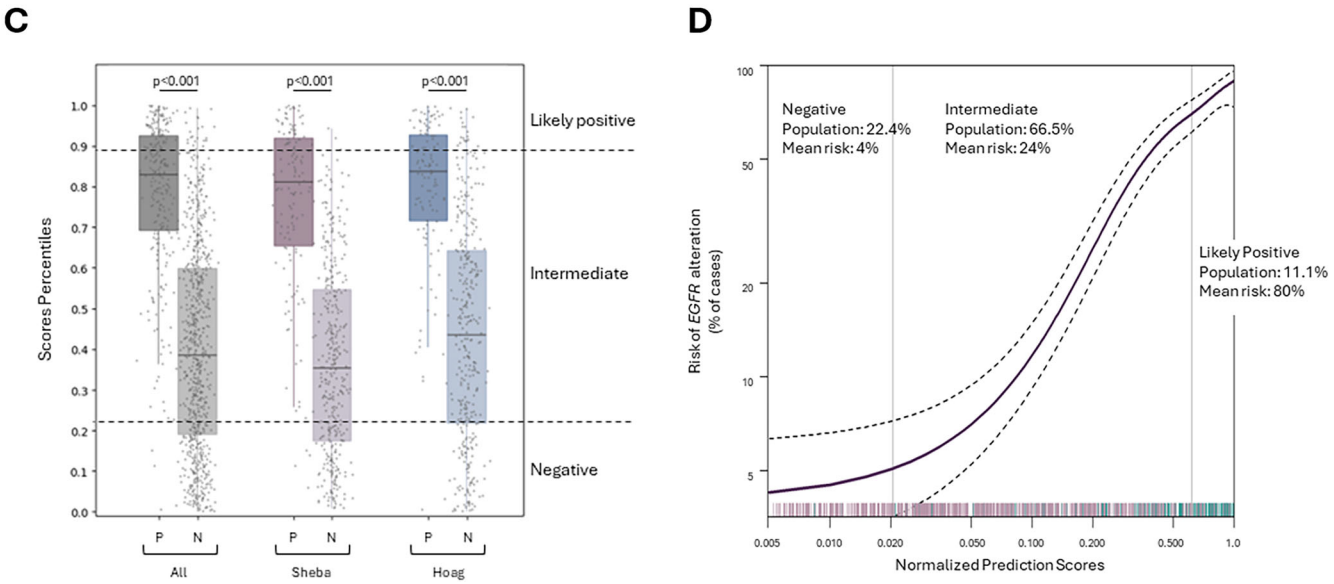

Fig. 3 | Performance of the *EGFR*-classifier. A Rate of reference method negatives in the *EGFR* AI negative group. Negative predictive value (NPV) adjusted based on an assumed prevalence of 15% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples. **B** Occurrences of reference method positivity across the *EGFR* AI groups. The proportion of reference method positive cases was calculated for each AI group, demonstrating a significant decrease in reference method positives, from a relatively high percentage in the AI likely-positive group to a very low percentage in the AI negative group. Sensitivity was assessed by calculating the Positive Percent Agreement (PPA) using a 2-tier classification, including a negative group and a likely positive group combined with intermediate samples. **C** Distribution of the test set samples according to their AI scores percentile ranks. Samples were plotted according to the reference method, positive and negative groups for all test set samples, and divided according to the source sites. Samples were plotted according to their AI scores and percentile distribution (see "Methods" for details). Box plots display the median, interquartile range (IQR), and whiskers (extending 1.5 times the IQR). Dashed lines indicate the defined thresholds, dividing the samples into the three-tier classification system. *P* value was calculated using the Mann–Whitney U test. Sheba test set: negatives (N) = 366, positives (P) = 131; Hoag test set: negatives (N) = 345, positives (P) = 113. **D** Risk for positivity as a function of the prediction score. Dashed curves represent the 95% confidence interval, while the rug plot above the *x*-axis depicts individual cases, with pink indicating reference method-negative and green indicating reference method-positive for the alteration. Risk and scores are shown on a logarithmic scale. Vertical lines represent the lower and upper thresholds, dividing the curves into negative, intermediate, and likely positive score categories. For each class, the percentage of the population included in the group and the mean risk of carrying the gene alteration are indicated.

requires a sufficiently large cohort. Future work will therefore focus on expanding the framework to additional alterations, including KRAS, RET, and ROS1, as soon as adequate datasets become available. Moreover, further refinement of our classifiers could potentially reduce the size of the intermediate groups and minimize the proportion of samples without a definitive classification. An additional limitation is that, similar to molecular testing, analysis with the AI tool can only be performed on cases with available tissue samples. Lastly, the nature of

A

AI negative	Ref. method negative	NPV	Adjusted NPV*
623	621	99.7% (95% CI: 98.8-100.0%)	99.5% (95% CI: 98.3-99.9%)

* NPV adjusted based on an assumed prevalence of 5% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

B

		AI	Ref. method positive (P)	% P / AI group	PPA**
AI classification	Likely positive	32	22	68.75%	94.1% (95% CI: 80.3-99.3%)
	Intermediate	287	10	3.48%	
	Negative	623	2	0.32%	

** PPA calculated using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

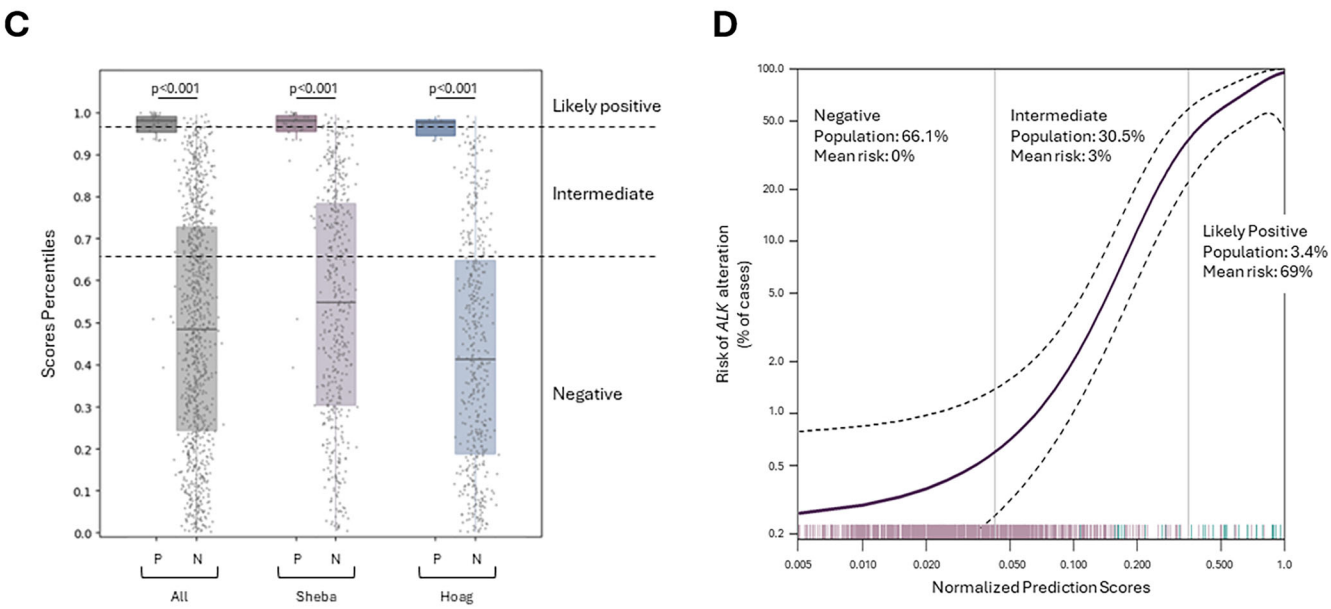

Fig. 4 | Performance of the *ALK*-classifier. A Rate of reference method negatives in the *ALK* AI negative group. Negative predictive value (NPV) adjusted based on an assumed prevalence of 5% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples. **B** Occurrences of reference method positivity across the *ALK* AI groups. The proportion of reference method positive cases was calculated for each AI group, demonstrating a significant decrease in reference method positives, from a relatively high percentage in the AI likely-positive group to a very low percentage in the AI negative group. Sensitivity was assessed by calculating the Positive Percent Agreement (PPA) using a 2-tier classification, including a negative group and a likely positive group combined with intermediate samples. **C** Distribution of the test set samples according to their AI scores percentile ranks. Samples were plotted according to the reference method positive and negative groups for all test set samples and divided according to the source sites. Samples were plotted according to their AI scores and percentile distribution (see "Methods" for details). Box plots display the median, interquartile range (IQR), and whiskers (extending 1.5 times the IQR). Dashed lines indicate the defined thresholds, dividing the samples into the three-tier classification system. *P* value was calculated using the Mann–Whitney U test. Sheba test set: negatives (N) = 465, positives (P) = 28; Hoag test set: negatives (N) = 443, positives (P) = 6. **D** Risk for positivity as a function of the prediction score. Dashed curves represent the 95% confidence interval, while the rug plot above the *x*-axis depicts individual cases, with pink indicating reference method-negative and green indicating reference method-positive for the alteration. Risk and scores are shown in a logarithmic scale. Vertical lines represent the lower and upper thresholds, dividing the curves into negative, intermediate, and likely positive score categories. For each class, the percentage of the population included in the group and the mean risk of carrying the gene alteration are indicated.

A

AI negative	Ref. method negative	NPV	Adjusted NPV*
396	396	100.0% (95% CI: 99.1-100.0%)	100.0% (95% CI: 100-100.0%)

* NPV adjusted based on an assumed prevalence of 2% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

B

		AI	Ref. method positive (P)	% P / AI group	PPA**
AI classification	Likely positive	5	1	20.00%	100.0% (95% CI: 63.1-100.0%)
	Intermediate	547	7	1.28%	
	Negative	396	0	0.00%	

** PPA calculated using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

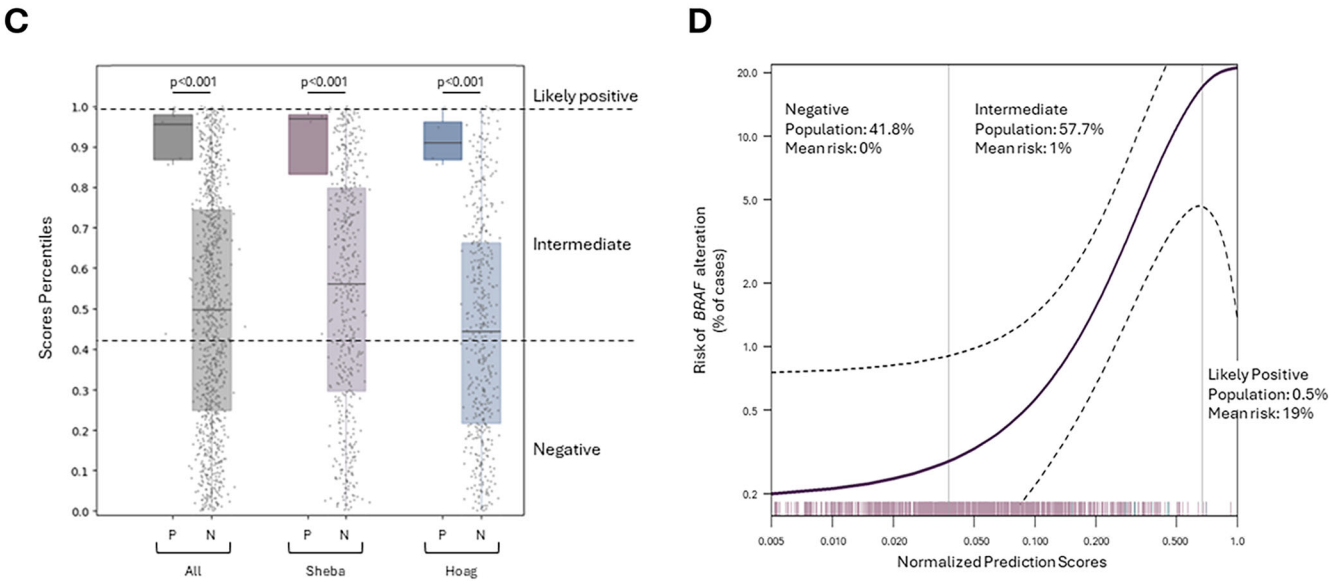

Fig. 5 | Performance of the *BRAF*-classifier. A Rate of reference method negatives in the *BRAF* AI negative group. Negative predictive value (NPV) adjusted based on an assumed prevalence of 2% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples. **B** Occurrences of reference method positivity across the *BRAF* AI groups. The proportion of reference method positive cases was calculated for each AI group, demonstrating a significant decrease in reference method positives, from a relatively high percentage in the AI likely-positive group to a very low percentage in the AI negative group. Sensitivity was assessed by calculating the Positive Percent Agreement (PPA) using a 2-tier classification, including a negative group and a likely positive group combined with intermediate samples. **C** Distribution of the test set samples according to their AI scores percentile ranks. Samples were plotted according to the reference method positive and negative groups for all test set samples and divided according to the source sites. Samples were plotted according to their AI scores and percentile distribution (see methods for details). Box plots display the median, interquartile range (IQR), and whiskers (extending 1.5 times the IQR). Dashed lines indicate the defined thresholds, dividing the samples into the three-tier classification system. *P* value was calculated using the Mann–Whitney U test. Sheba test set: negatives (N) = 485, positives (P) = 4; Hoag test set: negatives (N) = 454, positives (P) = 4. ns not significant. **D** Risk for positivity as a function of the prediction score. Dashed curves represent the 95% confidence interval, while the rug plot above the *x*-axis depicts individual cases, with pink indicating reference method-negative and green indicating reference method-positive for the alteration. Risk and scores are shown in a logarithmic scale. Vertical lines represent the lower and upper thresholds, dividing the curves into negative, intermediate, and likely positive score categories. For each class, the percentage of the population included in the group and the mean risk of carrying the gene alteration are indicated.

this study and the inherent constraints in model validation may not fully capture the capabilities of our models, as only samples with reference test results could be included in the cohort for performance evaluation. Consequently, we were unable to assess the classifiers' ability to generate predictions for a broader range of samples, including those not suitable for comprehensive testing.

Integrating an AI image-based method for the detection of NSCLC biomarkers into the standard pathological toolkit offers significant

A

AI negative	Ref. method negative	NPV	Adjusted NPV*
229	228	99.6% (95% CI: 97.6-100.0%)	99.6% (95% CI: 97.5-99.9%)

* NPV adjusted based on an assumed prevalence of 4% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

B

		AI	Ref. method positive (P)	% P / AI group	PPA**
AI classification	Likely positive	12	6	50.00%	96.2% (95% CI: 80.4-99.9%)
	Intermediate	317	19	5.99%	
	Negative	229	1	0.44%	

** PPA calculated using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples.

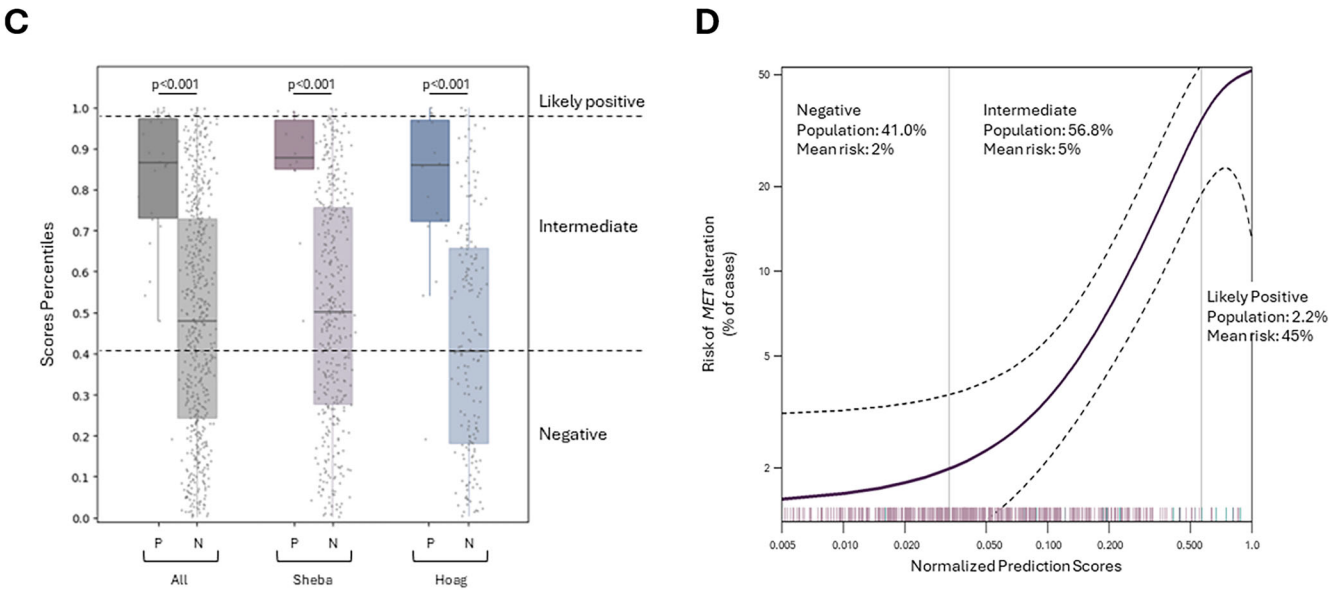

Fig. 6 | Performance of the *MET*-classifier. A Rate of reference method negatives in the *MET* AI negative group. Negative predictive value (NPV) adjusted based on an assumed prevalence of 4% in the population, using a 2-tier classification of a negative group and a likely positive group combined with intermediate samples. **B** Occurrences of reference method positivity across the *MET* AI groups. The proportion of reference method positive cases was calculated for each AI group, demonstrating a significant decrease in reference method positives, from a relatively high percentage in the AI likely-positive group to a very low percentage in the AI negative group. Sensitivity was assessed by calculating the Positive Percent Agreement (PPA) using a 2-tier classification, including a negative group and a likely positive group combined with intermediate samples. **C** Distribution of the test set samples according to their AI scores percentile ranks. Samples were plotted according to the reference method positive and negative groups for all test set samples and divided according to the source sites. Samples were plotted according to their AI scores and percentile distribution (see "Methods" for details). Box plots display the median, interquartile range (IQR), and whiskers (extending 1.5 times the IQR). Dashed lines indicate the defined thresholds, dividing the samples into the three-tier classification system. *P* value was calculated using the Mann–Whitney U test. Sheba test set: negatives (N) = 366, positives (P) = 10; Hoag test set: negatives (N) = 166, positives (P) = 16. **D** Risk for positivity as a function of the prediction score. Dashed curves represent the 95% confidence interval, while the rug plot above the *x*-axis depicts individual cases, with pink indicating reference method-negative and green indicating reference method-positive for the alteration. Risk and scores are shown in a logarithmic scale. Vertical lines represent the lower and upper thresholds, dividing the curves into negative, intermediate, and likely positive score categories. For each class, the percentage of the population included in the group and the mean risk of carrying the gene alteration are indicated.

advantages, facilitating accessible, rapid and accurate biomarker profiling within routine pathological diagnosis for NSCLC. Our retrospective findings provide valuable insights for guiding future analyses and predictions, supporting ongoing and future research efforts.

Methods

Study samples

This multicenter retrospective study has received ethical approvals with informed consent waiver due to the retrospective nature of the study from

the Institutional Review Boards (IRBs) at Sheba Medical Center, Israel (IRB 7451-20-SMC) and Hoag Memorial Hospital Presbyterian, USA (IRB 20233193). The study included WSIs of Formalin-Fixed Paraffin-Embedded (FFPE) H&E-stained histological specimens from cases with a primary NSCLC diagnosis collected between 2012 and 2023. Research involving human participants and their biological materials was conducted in accordance with the Declaration of Helsinki.

Test cohort

NSCLC samples were collected from two independent medical centers, Sheba Medical Center, Israel and Hoag Memorial Hospital Presbyterian, USA.

The performance of the AI models was evaluated against the diagnostic test results for each biomarker. The test cohort consisting of H&E-stained FFPE NSCLC tissue samples from both Sheba Medical Center and Hoag Hospital archives included unique cases collected either by biopsy or by surgical procedures from either the primary lung site, lymph nodes, or distant metastatic sites. Slides were scanned using the Philips IntelliSite Ultra Fast (Philips Digital Pathology Solutions, Best, Netherlands). Cytology biospecimens, including cell blocks and freshly frozen samples, were excluded from the cohort.

The alterations included in the *EGFR* positive ground truth were exon 19 deletions, exon 20 insertions, L858R, L861Q, G719X, and S768I mutations. The *ALK* model was trained to detect *ALK* rearrangements, the *MET* model exon 14 skipping alteration, while the *BRAF* model focused on identifying the *BRAF* V600E mutation (with all other alterations labeled as negatives). At Sheba Medical Center, *EGFR*, *BRAF* and *MET* alterations were assessed using the NGS assays Oncomine™ Dx Target Test or Oncomine™ comprehensive assay plus kit (Thermo Fisher Scientific, Waltham, MA, USA). Oncomine™ Solid Tumor DNA Kit (Thermo Fisher Scientific, Waltham, MA, USA) was also used to detect *EGFR* and *BRAF* mutations. In some samples, the *EGFR* mutations were assessed by the real-time PCR Idylla™ EGFR Mutation Test (Biocartis, Mechelen, Belgium). *ALK* rearrangements were assessed either by NGS assays Oncomine™ Dx Target Test or Oncomine™ comprehensive assay plus kit (Thermo Fisher Scientific, Waltham, MA, USA); by FISH using the Vysis ALK Break Apart FISH Probe Kit (Abbott Molecular, Abbott Park, IL, USA); or by IHC using the clone D5F3 antibody, (Ventana Medical Systems Inc., Oro Valley, AZ, USA). For Hoag hospital samples, genomic large panel testing was done through a single commercial vendor laboratory accredited by the College of American Pathologists (CAP) as previously described[29–31]. Briefly, Hoag Hospital tumor testing was completed using NGS of the entire exons of over 600 cancer-relevant genes with the issue of a curated report. Later iteration of the test included whole-exome and transcriptome sequencing to better detect fusions. RNA Whole Transcriptome Sequencing (WTS) uses a hybrid-capture method to pull down the full transcriptome from FFPE tumor samples using the Agilent SureSelect Human All Exon V7 bait panel (Agilent Technologies, Santa Clara, CA, USA) and the Illumina NovaSeq platform (Illumina, Inc., San Diego, CA, USA).

Only samples with a definitive result for a specific biomarker were included in the corresponding biomarker cohort. Samples showing discrepancies in biomarker status between two different assays were excluded from the respective biomarker cohort.

Algorithm development

The classifier for each biomarker was developed using the following general approach, described in detail below. Embeddings generated by CanvOI 1.1 foundation models were used as input for task-specific multiple instance learning (MIL) models. A Leave-One-Group-Out cross-validation step was conducted using 3997 WSIs from seven sites as a validation set. During this step, scores for all samples were combined and used to determine the thresholds for the final classifiers (Fig. 1).

For the training of the NSCLC biomarker classifier models, CanvOI 1.1, Imagene's proprietary foundation model, was utilized. CanvOI 1.1 is an upgraded version of our previous model, CanvOI[27], with an expanded pre-training dataset derived exclusively from Imagene's internal data, excluding public datasets, with more than 833k tissue samples of almost exclusively FFPE H&E-stained tissues. Non-overlapping tissue-containing 380×380 pixels tiles cropped from the WSI and converted into 1×1536 vector embeddings by two following training checkpoints of CanvOI 1.1 served as input for downstream MIL models. The two checkpoints represent models saved at different stages of the self-supervised training process. Leveraging both allows for a consistent embedding space while introducing moderate, beneficial variation enhancing the ensemble with subtly diverse feature representations. The MIL step included 4425 NSCLC labeled data from 8 source sites (obtained from internal datasets including Tel-Aviv Sourasky Medical Center as well as the public datasets TCGA and CPTAC) scanned using various Leica scanners, including the Leica Aperio AT Turbo, Leica Aperio AT2 and Leica Aperio GT 450 (Leica Biosystems, Wetzlar, Germany), as well as the Philips IntelliSite Ultra Fast Scanner (Philips Digital Pathology Solutions, Best, Netherlands).

Prior to the extraction of embedding, a WSI file integrity check was performed, as well as a pre-processing step to ensure the exclusion of unsuitable tiles. This step includes Canny edge detection and color averaging techniques to exclude background tiles and those that do not contain tissue, and a segmentation model, which excludes slides with insufficient analyzable cancerous cells, due to technical artifacts, scanning issues, or insufficient tumor tissue cells.

For the development of the biomarker classifiers, we employed a cross-validation Leave-One-Group-Out scheme. This training approach was implemented to develop models with strong generalization capabilities, ensuring the robustness of the algorithms. Several MIL models were trained using a 5-fold cross-validation approach, with different subsets of the training dataset's source sites. An ensemble of MIL models, created during this step, was then used to assess their performance on an out-of-domain (OOD) site, not included in the cross-validation process of the ensembled MILs. This procedure was repeated for all source sites, except TCGA, which was only utilized during the cross-validation phase, as it contained low amounts of positive cases with a relatively low prevalence of some of the alterations. In total, the validation cohort utilized in the cross-validation Leave-One-Group-Out step included 3997. For the final classifier, an ensemble of the MIL models evaluated across all source sites, excluding TCGA, was selected. A schematic of the methodology is shown in Fig. 1.

An upper and lower thresholds were established to create a three-tier classification system. Images scoring above the upper threshold were labeled as "likely-positive", those below the lower threshold as "negative," and those in between were classified as "intermediate". While setting these thresholds, we considered the balance between accuracy and the proportion of the population classified as either negative or likely-positive, minimizing the intermediate group. For the lower threshold, the goal was to achieve that among the cases classified as negatives, no more than 1% were incorrectly classified, to align with clinical tests, with adjustments made for population prevalence to reflect real-world distributions. For the upper threshold, the goal was to aim for no more than 50% misclassification among the positives, with adjustments to the prevalence. When necessary, thresholds were further adjusted to encompass approximately 30% of negative or positive cases for the lower and upper thresholds, respectively. During the development of each of the biomarker classifiers, the lower- and upper-thresholds were selected independently, fixed, and applied in the external validation.

Algorithm external validation

The models' performance was evaluated on an independent test set (described above). 519 and 474 samples of unique cases were collected from Sheba Medical Center and Hoag Hospital, respectively. These two sites were not represented in either the CanvOI 1.1 pre-training set or the NSCLC biomarker classifiers training set. Of these, 510 WSIs from Sheba and 458

WSIs from Hoag passed the preprocessing step and were included in the final test set.

For each biomarker, the AI-classifier assigned a "likely-positive", "intermediate" or "negative" status to each analyzable sample. The algorithm's predictions were then compared to the reference method results for all samples with a definitive reference method outcome, i.e., "likely-positive" and "negative". Concordance between the reference method(s) results and the AI-classifier results was assessed.

Statistical analysis

For the score distribution plots, the AI prediction scores of the external validation test cohort were evenly distributed into percentiles. For example, in the percentile above 90% all scores in this top decile were higher than 90% of all samples in the combined cohorts. Samples were divided into a reference method positive and negative subgroup for each biomarker and plotted according to the percentiles of their AI scores. Box plots were generated to display the median, interquartile range (IQR), and whiskers (extending 1.5 times the IQR) for each subgroup and jittered scatter points represent individual samples. The medians and IQRs were compared with Mann–Whitney U test's t-test. A p-value of less than 0.05 (two-sided) in the likelihood-ratio test was considered to indicate a significant result. The relationship between model scores and the probabilities of genetic alterations was analyzed and visualized using a predictiveness curve. Curves were generated with a Generalized Additive Model (GAM) in R, utilizing smooth spline functions.

Data availability

Data used in the research is subject to controlled access for privacy and proprietary reasons. Derived data are within the paper and its supplementary.

Code availability

The code used for this study is not publicly available and is Imagene's intellectual property.

Received: 15 May 2025; Accepted: 24 December 2025;
Published online: 20 January 2026

References

1. Meyer, M. L. et al. Lung cancer research and treatment: global perspectives and strategic calls to action. *Ann. Oncol.* **35**, 1088–1104 (2024).
2. Tan, A. C. & Tan, D. S. W. Targeted therapies for lung cancer patients with oncogenic driver molecular alterations. *J. Clin. Oncol.* **40**, 611–625 (2022).
3. Araghi, M. et al. Recent advances in non-small cell lung cancer targeted therapy; an update review. *Cancer Cell Int.* **23**, 162 (2023).
4. Wu, Y. L. et al. Osimertinib in resected EGFR-mutated non-small-cell lung cancer. *N. Engl. J. Med.* **383**, 1711–1723 (2020).
5. Wu, Y. L. et al. Alectinib in resected ALK-positive non-small-cell lung cancer. *N. Engl. J. Med.* **390**, 1265–1276 (2024).
6. Gainor, J. F. et al. EGFR mutations and ALK rearrangements are associated with low response rates to PD-1 pathway blockade in non-small cell lung cancer: a retrospective analysis. *Clin. Cancer Res.* **22**, 4585–4593 (2016).
7. National Comprehensive Cancer Network (NCCN), *Clinical Practice Guidelines in Oncology, Non-Small Cell Lung Cancer (Version 11.2024)*, https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf.
8. Rolfo, C. et al. Liquid biopsy for advanced NSCLC: a consensus statement from the International Association for the Study of Lung Cancer. *J. Thorac. Oncol.* **16**, 1647–1662 (2021).
9. Sadik, H. et al. Impact of clinical practice gaps on the implementation of personalized medicine in advanced non-small-cell lung cancer. *JCO Precis. Oncol.* **6**, e2200246 (2022).
10. Schwartzberg, L. et al. Improving biomarker testing in advanced non-small-cell lung cancer and metastatic colorectal cancer: experience from a large community oncology network in the USA. *Future Oncol.* **19**, 1397–1414 (2023).
11. Bruno, D. S., Hess, L. M., Li, X., Su, E. W. & Patel, M. Disparities in biomarker testing and clinical trial enrollment among patients with lung, breast, or colorectal cancers in the United States. *JCO Precis. Oncol.* **6**, e2100427 (2022).
12. Scott, J. A. et al. Compromised outcomes in stage IV non-small-cell lung cancer with actionable mutations initially treated without tyrosine kinase inhibitors: a retrospective analysis of real-world data. *JCO Oncol. Pract.* **20**, 145–153 (2024).
13. Robert, N. J. et al. Biomarker testing and tissue journey among patients with metastatic non-small cell lung cancer receiving first-line therapy in the US Oncology Network. *Lung Cancer* **166**, 197–204 (2022).
14. Yorio, J. et al. Association of timely comprehensive genomic profiling with precision oncology treatment use and patient outcomes in advanced non-small-cell lung cancer. *JCO Precis. Oncol.* **8**, e2300292 (2024).
15. Echle, A. et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* **124**, 686–696 (2021).
16. Zhang, C. et al. Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *J. Hematol. Oncol.* **16**, 114 (2023).
17. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
18. Baxi, V., Edwards, R., Montalto, M. & Saha, S. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod. Pathol.* **35**, 23–32 (2022).
19. Saillard, C. et al. Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. *Nat. Commun.* **14**, 6695 (2023).
20. Hu, Q. et al. Development and validation of a deep learning-based microsatellite instability predictor from prostate cancer whole-slide images. *npj Precis. Oncol.* **8**, 88 (2024).
21. Esteva, A. et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *npj Digit. Med.* **5**, 71 (2022).
22. Mayer, C. et al. Direct identification of ALK and ROS1 fusions in non-small cell lung cancer from hematoxylin and eosin-stained slides using deep learning algorithms. *Mod. Pathol.* **35**, 1882–1887 (2022).
23. Campanella, G. et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nat. Med.* **31**, 3002–3010 (2025).
24. Amidi, E. et al. Predicting ROS1 and ALK fusions in NSCLC from H&E slides with a two-step vision transformer approach. *npj Precis. Oncol.* **9**, 266 (2025).
25. Pao, J. J. et al. Predicting EGFR mutational status from pathology images using a real-world dataset. *Sci. Rep.* **13**, 4404 (2023).
26. Kosaraju, S. C. et al. Evidential deep learning-based ALK-expression screening using H&E-stained histopathological images. *npj Digit. Med.* **8**, 610 (2025).
27. Zalach, J. et al. CanvOI, an oncology intelligence foundation model: scaling flops differently. Preprint at https://doi.org/10.48550/arXiv.2409.02885 (2024).
28. Sorin, M. et al. Neoadjuvant chemoimmunotherapy for NSCLC: a systematic review and meta-analysis. *JAMA Oncol.* **10**, 621–633 (2024).
29. Darabi, S. et al. Programmatic efforts increase adoption of genomic precision medicine in cancer care in a community cancer center. *JCO Precis. Oncol.* **6**, e2200090 (2022).
30. Darabi, S. et al. Transcriptional profiling of malignant melanoma reveals novel and potentially targetable gene fusions. *Cancers* **14** https://doi.org/10.3390/cancers14061505 (2022).

31. Darabi, S., Zuazo, C. E., Braxton, D. R., Eisenberg, B. L. & Demeure, M. J. Precision medicine in a community cancer center: pan-cancer DNA/RNA sequencing of tumors reveals clinically relevant gene fusions. *Biologics* **3**, 198–208 (2023).

Acknowledgements

This work was funded by Imagene. The results here are partly based upon data generated by the TCGA Research Network: (https://www.cancer.gov/tcga). We would like to thank Drew Watson for support with statistical analysis.

Author contributions

J.Z., N.P.Y., and I.G. conceptualized the study. N.P. and C.R. provided clinical guidance. E.O., N.P.Y., J.Z., I.G., and I.B. designed the study. E.O., Y.B., J.W., R.I.H., F.D., Y.M., A.L., D.R.B., J.S.C., R.Y., C.M., C.A., D.U., J.B., and D.H. contributed to data collection. A.A., J.Z., A.D., and I.G. performed data analysis and interpretation. A.A. and J.Z. developed the models. C.R., E.O., A.D., and I.G. were the main contributors to writing the manuscript. All authors read and approved the final paper.

Competing interests

C.R. reports personal fees for advisory board membership from: Novocure; institutional fees for advisory board membership from AstraZeneca, Abbvie, Imagene, MedStar, Amgem, Boeringer-Ingelheim, Hoffmann-La Roche Ltd, Janssen Pharmaceutical, NeoGenomics, Pfizer, Inc., Regeneron, Roche, IDEOlogy, Medical Educators Consortium (MEC), Eli Lily, EXACT Sciences Corp, ThermoFisher, Foundation Medicine. Research collaboration non-remunerated: Guardant, Foundation One. Institutional fees as an invited speaker from COR2ED, HPM education, IDEOlogy, Merck, Roche, OneCell Dx, MJH Life Sciences, TACTICS, AIOM, FONICAP, MD Education, Precis CA, Suzhou Liangyihui Network Technology Co, Ltd., Emirates Oncology Conference. Podcasts: Health EOR, Total Health Conferencing. Non-renumerated leadership roles as a scientific board member for the European School of Oncology (ESO), Past Chair on the educational committee for the International Association for Study of Lung Cancer (IASLC), President for the International Society of Liquid Biopsy (ISLB) and Educational Chair for the Oncology Latin American Association (OLA); a renumerated role as Editor in Chief for Critical Reviews in Oncology Hematology (CROH); AIRC Board member. Non-renumerated role as ESMO Faculty Group/Speciality and Faculty Coordinator for metastatic non-small cell lung cancer for European Society for Medical Oncology (ESMO); non-renumerated roles as Scientific Board Member at ESO (European School of Oncology), External advisor Board member of Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO), External Advisor of the School of Public Health, University of Granada, Spain, and non-renumerated analysis of liquid biopsies in a lung cancer trial for Guardant Health. At the time of study was running C.R.'s affiliation was Center for Thoracic Oncology, The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. D.U. declares consulting services for Medison Pharma, Amgen GmBH, Merck Sharp & Dohme, and declares honoraria for Merck Sharp & Dohme, AstraZeneca, Pfizer, Megapharm, BMS, Roche, Takeda, and J-C Healthcare. J.B. declared receiving research funding from Immunai, OncoHost, MSD, AstraZeneca, Roche, and AbbVie, consulting services for AbbVie, Amgen, AstraZeneca, Bayer, Beigene, Janssen, MSD, Merck-Serono, Medison, Roche and Takeda, honoraria from BMS, Medison and Pfizer and holding roles in Committee IASLC and Lung Ambition Israel leadership. D.R.B received research support from Imagene, C.M. was reimbursed by Imagene AI for attending a conference. D.H. received payment to institution for lectures from AstraZeneca, MSD, Merck, Roche, Medison, BMS and Sanofi. J.Z. is a cofounder and a stockholder of Imagene, A.A., N.P.Y., A.D., I.G. are currently, or were formerly, employees of Imagene and have stock options. E.O., Y.B., J.W., R.I.H., F.D., Y.M., A.L., J.S.C., R.Y., and C.A. declare no conflict of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-01267-z.

Correspondence and requests for materials should be addressed to Christian Rolfo.